 ISSN 1759-586X

Curated Databases in the Life Sciences: The Edinburgh Mouse Atlas Project

SCARP Case Study No. 4

Elizabeth Fairley, EFB Services
for the Digital Curation Centre

Edited by
Sarah Higgins and Angus Whyte
Digital Curation Centre

DCC SCARP INTERIM CASE STUDY REPORT
Deliverable B4.8.5.1
Version No. 1.1
Status FINAL
Date 13 July 2009

Copyright

Catalogue Entry

Title Curated Databases in the Life Sciences: The Edinburgh Mouse Atlas Project
Creator Elizabeth Fairley (author) Sarah Higgins and Angus Whyte (editors)
Subject Data curation; formats, processes and issues; system development; standards; legal factors; methodology, and problems overcome; human factors
Description This case study scopes and assesses the data curation aspects of the Edinburgh Mouse Atlas Project (EMAP), a programme funded by the Medical Research Council (MRC). The principal goal for EMAP is to develop an expression summary for each gene in the mouse embryo, which collectively has been named the Edinburgh Mouse Atlas Gene-Expression Database (EMAGE).
Publisher University of Edinburgh; UKOLN, University of Bath; HATII, University of Glasgow; Science and Technology Facilities Council.
Date 29 June 2009 (creation)
Type Text
Format Adobe Portable Document Format v.1.3
Resource Identifier ISSN 1759-586X
Language English
Rights © 2009 DCC, University of Edinburgh

Citation Guidelines

Fairley, E. (2009), " Curated Databases in the Life Sciences: The Edinburgh Mouse Atlas Project. SCARP Case Study No. 4", Digital Curation Centre, Retrieved <date>, from http://www.dcc.ac.uk/scarp

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

1.1 Introduction

This report has been produced for the Digital Curation Centre (DCC) [1] as a SCARP Life Sciences case study. The DCC SCARP project, funded by the Joint Information Systems Committee (JISC), investigates disciplinary attitudes and approaches to data deposit. The study concerned the data curation aspects of the Edinburgh Mouse Atlas Project (EMAP), a programme funded by the Medical Research Council (MRC). The principal goal for EMAP is to develop an expression summary for each gene in the mouse embryo, which collectively has been named the Edinburgh Mouse Atlas Gene-Expression Database (EMAGE).

1.2 Scope of the Study

The purpose of this case study is to profile and scope the work of the Edinburgh Mouse Atlas Project in relation to digital curation processes and activities undertaken by the researchers working on the project, and the users and stakeholders for the services and products provided by the project. One aspect of digital curation is the process of establishing and developing infrastructure to provide for current and future reference materials including the curation, preservation, maintenance, collection and archiving of the digital assets.

The approach taken in SCARP is to reflect the researchers' views and understanding of what they are doing, follow the sequence of stages or phases in which information (data) is produced, manipulated and used as a scientific product, and involve the researchers in considering any changes in curation approach that might be relevant, using the DCC Curation Lifecycle Model as an 'ideal type' to summarise the results. The aim was to identify factors that might provide for curation appropriate to the disciplinary setting; life sciences and specifically model organism research, the production and use of curated databases, image based studies of development (wild-type) linked to gene expression, interdisciplinary work and international collaboration.

The case study encompasses:

- Characterisation of the field in terms of the research questions (or class of question) addressed
- Organisational form of the group and its work with other research groups
- The research group's drivers for curation and what the group undertakes in terms of digital curation

[1] The Digital Curation Centre at www.dcc.ac.uk/

- The stakeholders for the mouse atlas and profile of its users
- Mapping of the research group's curation processes against the DCC Curation Lifecycle Model.

1.3 Methodology

This case study report was produced by Elizabeth Fairley of EFB Services, acting as consultant to the DCC and edited by Sarah Higgins and Angus Whyte of the DCC.

As a short study, the methods adopted aimed for a broad profile of the curation practices employed in support of the research being undertaken. Principally, the study was based upon a series of site visits to identify individual team members' roles and activities, with informal interviews and the demonstration of operational processes by key staff, supplemented by attendance as observer at research group meetings. In addition, lab-based observation over a series of half days enabled the acquisition of more detailed context, for the analysis.

A brief review (Appendix 3) was completed with the help of the Edinburgh Mouse Atlas Project of documentation (primarily EMAGE documentation) used by the research group to describe their processes and product (e.g. Mouse Atlas) both for their own internal work and also any produced or published for an external audience.

The mapping of the EMAGE curation processes against the DCC Curation Lifecycle Model enables the team to view the curation requirements and challenges through their individual role(s) in the project. The aim of the DCC Curation Lifecycle Model is to provide a graphical high-level overview of the stages required for successful curation and preservation of data. "The model can be used to plan activities within an organisation or consortium to ensure that all necessary stages are undertaken, each in the correct sequence. The model enables granular functionality to be mapped against it; to define roles and responsibilities, and build a framework of standards and technologies to implement. It can help with the process of identifying additional steps which may be required, or actions which are not required by certain situations or disciplines, and ensuring that processes and policies are adequately documented."[2]

1.4 Key Findings / Outcomes

EMAP is on course to produce a digital atlas of mouse development that can be used effectively to facilitate further research. The scope for the development of the EMAGE database has been well defined and there is evidence that the EMAGE team is effectively meeting all set objectives. Furthermore, market research has shown

[2] The Digital Curation Centre at www.dcc.ac.uk/

that the EMAGE database is the first and only UK scientific product to provide a spatially mapped gene-expression repository with associated tools for data mapping, submission and analysis.

Notwithstanding the perceived quality and effectiveness of the EMAGE database, when considering its specific data curation aspects the study identified a number of issues requiring further monitoring and resolution. Foremost among these were: third party **copyright**, which continues significantly to inhibit the display of images; the need to address the **standardisation** of experimental details and an incidence of variability between data sources that leads to a number of 'unspecified' entries; the practice of manual **data entry** from an Excel spreadsheet into the EMAGE database, which limits the tracking and error checking of data into the EMAGE database; a **quality assurance** process in which the high quality of data displayed appears to depend upon human intervention, where curated data is being checked and corrected by the senior editor, and where there is no formal process for the correction of errors.

Nonetheless, with high **data throughput** crucial to increasing the opportunities for discovering novel genes, the team has recognised that greater use of improved curation tools and methods represents a means of potentially increasing effectiveness; and on an associated theme, developments in **tools and methods** are accepted as essential for the continued maintenance and sustainability of the EMAGE infrastructure and interface. That said, for a project of this scale, the importance of the **human infrastructure** remains significant, particularly with respect to the sharing of expertise in data management.

Recommendations from the study focus on optimising data management and the rate of data entry. In particular, attention to a revision of the EMAGE data management (administration) tool is expected to pay early dividends, particularly through a consequent increase in the rate of curation, and automation of a number of steps in the process is also to be encouraged. In support of this objective, documenting the curation policies and activities applied should include the production of practical step-by-step guidelines for the practice of curation. Given the relationship established through this study, working with the DCC to achieve these goals is likely to prove mutually beneficial.

2 THE LIFE SCIENCES DATA SHARING LANDSCAPE

Researchers and research-based organisations in the life sciences are not simple consumers of information services provided by publishers, libraries and others because they actively produce and maintain their own information sources and services. This is confirmed by the growing number of curated databases in many life sciences research fields (Galperin, 2008).

The majority of life sciences databases and tools are publicly funded and their broad aim is to enable life sciences researchers to access, analyse and contribute to accurately represented sources of information[3]. Biocuration is accomplished through the convergent work of biocurators (highly educated, experienced scientists who catalogue, annotate and analyse data), software developers, researchers and journal publishers. Recently, it has been recognised that there is a requirement for a formal organisation, currently known as the International Society for Biocuration (ISB), to build relationships and facilitate communication, the sharing of information (data and documentation), training and future funding[4].

Digital atlases are being recognised as a useful data sharing resource by acting as a scaffold in which data from multiple resources, can be shared, visualised, analysed and mined. Thus, the semantic and spatial infrastructure of an atlas adds a dimension to data that increases its potential use and reusability (Boline *et al*, 2008). Recent life sciences studies, such as the Joint Data Standards Study[5] (2005), have demonstrated the value of sharing and re-using data and address the importance of; standards, planning and management, the existence of a framework that supports researchers in their data collection and submission activities, software tools, good communication, incentives for individuals to submit data, the availability and quality of data, the importance of consent for data use and confidentiality of data, funding and legislation.

2.1 Market Size and Drivers

The European bioinformatics industry is predicted to be worth $720 million by 2011 (Feick, 2005). This is mainly due to the support of national governments, promoting the benefits of bioinformatics and increasing their overall research and development investments (Chu, 2005). However, despite projects of strong growth there are challenges in the market which mainly focus on:

[3] Case Studies in Life Sciences. Understanding Researchers' Information needs and Uses at http://www.rin.ac.uk/case-studies

[4] The International Society for Biocuration at http://www.biocurator.org/

[5] Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models (The "Joint Data Standards Study") by Digital Archiving Consultancy, e-Science Centre, Bioinformatics Research Centre, University of Glasgow, 2005 at http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC002552

- The mainstream acceptance of bioinformatics solutions
- The perception that bioinformatics tools are restricted for use to only specialised end user groups
- The continual consolidation of bioinformatics and life sciences companies.

The bioinformatics market is being driven by the exponential growth of novel biological discoveries as it has been estimated that approximately 1 terabyte of biological information is generated per week. Thus, there is a need for information technology to aid in the development and maintenance of curated databases. For biology, the use and effectiveness of text data mining and natural language techniques is challenging. This is mainly due to the nomenclature and ontologies used. However, the development and use of information extraction techniques for entities and relations between entities from the literature and the application of semi-automation for curation will assist data collection from the literature and be of great value to the life sciences community.

2.2 History and Development of Curated Databases

In the last three decades, biology has yielded an immense amount of data. This has been mainly due to researchers being able to explore the functional significance of genome sequencing data leading to more data about gene expression, gene positioning and phenotypic analysis (genotype-phenotype associations) being generated. One of the main objectives in bioinformatics is to exploit new technologies to construct databases that are and easily available for consultation (Leonelli, 2008).

The term 'curated database' describes a database, or repository, whose content, often about a specialised subject, has been obtained by extensive human effort through consultation, verification, aggregation of existing sources and the interpretation of new data (often experimental). Thus, curated databases tend to represent the efforts of a dedicated group of people that wish to produce a definitive description of a specific subject area. As scientific research data has become more available by being published electronically there has been a significant increase in the number of databases and the value of such curated databases is dependant on its organisation and the quality of the data (Buneman *et al*, 2008).

There are a number of challenges in developing and maintaining a curated database. These include;

- Obtaining the source, quality and reliability of annotation data
- Curation of relevant data consistently and accurately

- Provenance and citation of data through the use of unique identifiers; this is of particularly relevance to the cross-referencing of data from other databases as much of the work of a curator is to annotate existing data

- Updating of curated data; to periodically publish versions of the database to enable users to cite and retrieve particular a version of the database

- Evolution of the database schema and structure; to accommodate research and development of the database, new scientific discoveries and highlight relations from the data collected

- Finding a vocabulary and format that allows data to be accessible and retrievable to all research groups

- The economic and social factors that effect the long-term usefulness of curated databases

- Copyright and intellectual property issues that are especially relevant to open access curated databases.

2.3 The Value of Curated Databases

The value of curated databases for life sciences research is ultimately the re-use of data[6]. Thus, curated databases benefit data creators, researchers, funders and users by:

- Improving the quality of research data

- Providing access to reliable data

- Allowing researchers to form new hypotheses, analyse results, validate conclusions and guide future research

- Encouraging good record-keeping standards for discovered research data and consistency in working practices to enable data to be analysed and researched further

- Addressing the relationships between the different dynamic, evolving datasets

- Facilitating linkage between related research

- Ensuring that valuable, non-reproducible knowledge and data is preserved

- Allowing data sets to be combined in new and innovative ways

[6] Curating e-Science Data at www.dcc.ac.uk/

- Enabling the provenance of data to be verified.

2.4 Existing Bioinformatics Databases

There are several hundred public-domain databases in the field of biology (Galperin, 2006). Few contain raw experimental data as the majority focus on mapping curated data by organising, interpreting, annotated data from other sources.

Examples of biology databases include:

- The National Center for Biotechnology Information[7] (NCBI) which was established in 1988 as a national resource for molecular biology information. Databases include:
- The genetic sequence database, GenBank
    - Molecular databases such as; Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, Expression and Chemical databases
    - Literature databases such as; PubMed, Medline, OMIM (Online Mendelian Inheritance in Man), OMIA (Online Mendelian Inheritance in Animals) and the medicine's controlled vocabulary, MeSH (Medical Subject Headings)

- A curated database which is used extensively and regarded as being of a high standard within the life sciences domain is 'UniProt' (Bairoch and Apweiler, 1997). UniProt (formally known as SwissProt) is currently the standard reference for protein sequence data and currently consists of over 300,000 entries (Buneman *et al*, 2008).

- A small curated biological database is the 'IUPHAR receptor database' which describes the molecules that transmit information across cell membranes[8]. Unlike UniProt most of the curation is completed by volunteers and very few people are involved in its direct maintenance.

A number of curated databases are now being referred to as 'ontologies' rather than databases. This is mainly due to the hierarchical classification of information and the ability to perform queries on structured data. For example, 'Gene Ontology' has a number of hierarchies constructed over an underlying database of entries[9]. The Open Biomedical Ontologies[10] (OBO), sanctioned by a consortium of specialists,

[7] National Center for Biotechnology Information at www.ncbi.nlm.nih.gov/

[8] IUPHAR receptor database at http://www.iuphar-db.org.

[9] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25-29, 2000.

[10] The Open Biomedical Ontologies at www.obofoundry.org/

provides information on many of the other reliable, highly standardised, freely available, well-structured controlled vocabularies.

2.5 The Cost and Scalability of Curated Databases

The economic model for the distribution of research papers has shifted from academics paying to get their research into print and disseminated, to papers becoming accessible through the publishing of articles electronically. The idea of open access is that the initial costs should be paid for by the person (or institution or grant) responsible for a publication and thereafter the research article should become freely available. A key question is whether this economic model is suitable for curated databases which are open access, because unlike research papers curated databases are constantly updated and it is often difficult to obtain funding for future maintenance and sustainability (Houghton *et al*, 2009). Whether users should be charged to view the information saved within curated databases is complicated due to the fact that some payment should potentially go to the data source (Buneman *et al*, 2008).

The cost of curated databases is extensive; for example, there are over 150 people working full-time on the proteomic database, UniProt (Buneman *et al*, 2008). A breakdown of the Edinburgh Mouse Atlas Project costs is discussed in Section 12.1; however, determining the process and feasibility of scaling up the project, particularly the EMAGE team, would be interesting to review in more detail.

2.6 The Standards of Curated Databases

Standardising and structuring forms of data, such as life sciences activities and outputs, is becoming increasingly important in the progress and development of life sciences research. Researchers' involvement in developing and conforming to international standards is vital. However the challenges which lead to inefficiencies and lack of coordination include the variation in publication formats of existing databases and accessibility to certain scientific information sources.

The vast majority of researchers disclose their results through publication in a refereed journal. However, even with new biological database and curation journals becoming available[11], for instance using open access author pay models, researchers may find it difficult to select the most relevant journal for the publication of their scientific findings and due to the data publication policy of journals, or strict selection criteria, a large amount of data produced in the course of experiments is discarded without being circulated to the wider community. These restrictions do not apply to curated databases; however it is crucial that the quality of data collected by curated databases is high, and that the researchers' results are reported

[11] Database: The Journal of Biological Databases and Curation at http://www.oxfordjournals.org/our_journals/databa/about.html

in sufficient detail that the methods of data collection and analysis can be performed independently.

The Microarray Gene Expression Data (MGED) Society is an international organisation of biologists, computer scientists and data analysts that aim to facilitate biological and biomedical discovery through data integration. In 2001, the European Bioinformatics Institute (EBI) published a standard for presenting and exchanging microarray data known as Minimum Information About a Microarray Experiment (MIAME) (Brazma *et al* , 2001) for MGED and recently, the society has set up a working group, known as the Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE), to develop and promote standardisation as a community effort[12,] (Deutsch *et al*, 2008).

There are also a number of other organisations that are working together to increase the consistency of biological information and to maintain and raise the standards of data integration, exchange and comparison. For example, in 2007 the European Molecular Biology Laboratory (EMBL) published guidelines to aid proteomic data integration and comparison (Wilkinson, 2007) and in 2008, researchers described a new bioinformatics tool ('MisPred') that can identify and correct abnormal, incorrect and mis-predicted protein annotations in public databases (Wilkinson, 2008).

[12] The MISFISHIE working group at mged.sourceforge.net/misfishie/

3 THE EDINBURGH MOUSE ATLAS PROJECT (EMAP)

3.1 Background

The EMAP was initially a collaborative effort between the MRC Human Genetics Unit, Edinburgh and the Section of Biomedical Sciences, University of Edinburgh. Currently, EMAP is solely funded by the MRC.

Dr. Duncan Davidson and Prof. Richard Baldock received funding to commence the project in 1994. Initially much time and effort was spent on building the infrastructure, establishing and developing novel software for the reconstruction and mapping of curated data. In 2001, the editorial office became established and since then Dr. Jeff Christiansen has been key to driving the developments of the Edinburgh Mouse Atlas Gene-Expression (EMAGE) database forward. An important aspect of the success of the project is the knowledge and expertise of the team and the successful working relationship that the scientists and software developers have established and maintained.

3.2 The Aim of EMAP

The overall aim of EMAP is to produce a digital atlas of mouse development and accompanying databases to be a community resource for spatially mapped data during mouse embryonic development. Ultimately, to allow users to view complex expression spatially, develop hypotheses and reduce the misinterpretation of published findings.

The Edinburgh Mouse Atlas Project (EMAP) is described as a time-series of mouse-embryo volumetric models that provide a context-free spatial framework onto which structural interpretations and experimental data can be mapped. This enables users to compare and query complex spatial patterns to each other and other known or hypothesised structure. The atlas also includes a time-dependent anatomical ontology to enable mapping between the ontology and the spatial models in the form of delineated, anatomical regions or tissues. Thus, the models provide a natural, graphical context for browsing and visualising complex data (Baldock *et al*, 2007).

3.3 The EMAP Team

The principal investigators of EMAP (Table 1, Appendix 4) are Prof. Richard Baldock (Project leader, computing) and Dr. Duncan Davidson (Project leader, biological). The cross-section of scientists and software developers (approximately ratio 1:3) is impressive and vital for the success of their work. Time was spent with both Prof. Richard Baldock and Dr. Duncan Davidson, and the other members of the EMAP team and EMAGE editorial staff listed below (Table 1). All were approachable and took time to address questions asked. The EMAGE editors and database service administrator also completed a short questionnaire to obtain their individual perspective on the project (Appendix 5).

Both principal investigators have their own office, the editorial staff work together in close proximity in one room and the software developers are situated together in a separate room. This allows for the necessary direct communication between editors without disrupting others from their work.

The expertise of the EMAGE editorial team (Table 1) would be difficult to replicate, as they have invaluable experience, appear to work together well and efficiently. A number of internal meetings were also attended in which open issues and forward planning was addressed in an informal, interactive manor. No external meetings were attended and unfortunately it was not possible to attend the Scientific Advisory Board (SAB) meeting in December 2008.

Name	Position (additional responsibilities)
Dr. Duncan Davidson	Project leader: biology
Prof. Richard Baldock	Project leader: computing
Dr. Jeff Christiansen*	Senior Editor EMAGE database (curator, completes all quality assurance of all curated data)
Ms. Lorna Richardson*	Editor, EMAGE database (full time curator and responsible for data flow management of external submissions and partial management of the anatomy ontology)
Dr. Shanmugasundaram Venkataraman*	Editor, EMAGE database (full time curator and involved in developing methods for incorporating 3D gene expression data into EMAGE)
Mr. Peter Stevenson*	Database Service Administrator (EMAGE Computer Support and Database Systems Manager, involved in the ETL of data from many different data sources)
Dr. Yiya Yang	Database Architect for EMAGE and other EMAP projects
Mr. Nicholas Burton	Interface developer for EMAGE and other EMAP projects
Dr. Jianguo Rao	Image processing, 3D warping and matching, HRP collaboration (devises methods for text and image curation to move towards semi-automation)
Liz Graham	EMAP 3D embryo model development (digitisation of data)
Mr. Bill Hill	EMAP imaging research

Note: No time was spent with Dr. Yang, Mr. Burton or Liz Graham.

Table 1: Core MRC funded staff involved in EMAP (which includes the EMAGE editorial team*)

Interestingly, Prof. Richard Baldock commented that EMAP did not necessarily need to be located in Edinburgh, as many of the e-science and bioinformatic techniques work well in networked and virtual environments, although if EMAP were relocated it could be difficult to retain or replace the current expertise and experience of the EMAP team.

3.4 Stakeholders of EMAGE

The stakeholders for the EMAGE database are the MRC, the board of advisors, collaborators and users of the database.

The Board of Advisors meet annually to provide direction and assess the progress of the EMAGE project represent the fields of Developmental Biology, Mouse Genetics, Databases and Commercialisation. Current members include[13]:

- Dr. David Wilkinson (Chair): Head of Developmental Neurobiology Division and Head of Mammalian Development Division. MRC National Institute for Medical Research, London, UK.
- Dr. Alvis Brazma: Head of ArrayExpress. European BioInformatics Institute, European Molecular Biology Laboratory, Hinxton, UK.
- Prof. Steve Brown: Director, MRC Mammalian Genetics Unit, Harwell, UK.
- Dr. Janan Eppig: Senior Staff Scientist, Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, USA.
- Dr. Graham Kemp: Associate Professor, Bioinformatics Research Group, Chalmers University of Technology, Goteborg, Sweden.
- Dr. Suzanna Lewis: Informatics Group Leader, Berkeley Drosophila Genome Project, Berkeley, USA.
- Dr. Martin Ringwald: Associate Staff Scientist and Head of GXD Database, Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, USA.
- Prof. Claudio Stern: Head of the Department of Anatomy and Developmental Biology and JZ Young Professor of Anatomy, University College London, UK. (2005-)
- Dr. Sarah Wedden: Medical Research Council Technology Scotland, Edinburgh, UK.

3.5 Other Known Genome Databases and Ontologies

Prof. Richard Baldock believes that the EMAP provides a global centre for mammalian spatial mapping and text annotation. Currently, there are no mouse embryo databases that are in direct competition with EMAGE, and few databases that focus on spatial mapping. The closest competitor is the genome-wide image database of gene expression in the mouse brain released by the Allen Institute for Brain Science[14]. The International Neuroinformatics Coordinating Facility (INCF) program on Digital Brain Atlasing was launched following the recommendations of the 1st INCF Workshop on Mouse and Rat Brain Digital Atlasing Systems[15]. The workshop report gives an introduction to digital atlasing research and the need for open standards and protocols.

[13] The EMAGE Board of Advisors at www.emouseatlas.org/testemage/about/about_EMAGE.html#Ad_Board

[14] The Allen Institute for Brain Science at www.brain-map.org/

[15] Digital Brain Atlasing at http://www.incf.org/about/programs/atlasing/digital-brain-atlasing

There are a number of databases (in addition to those described in Section 8) that provide a platform to query and compare microarray and gene expression data such as (Galperin , 2006):

- 4DXpress: The EMBL database for cross species expression pattern comparisons
- ABA (Ascidian Body Atlas): The 3D atlas of ascidian embryo development and gene expression patterns
- ArrayExpress: A new public repository for microarray based gene expression data
- Axeldb: A database storing and integrating gene expression patterns and DNA sequences identified *Xenopus laevis* embryos
- BGED (Brain Gene Expression Database): A database that contains gene expression data for various physiological and pathological processes in the mouse brain
- BodyMap: A human and mouse gene expression database
- CGED (Cancer Gene Expression Database): A database of gene expression and clinical information
- FLIGHT: A database that enables integration of *Drosophila* phenotypes, gene expression and protein interactions
- Gene Expression in Tooth: A database of gene expression in detail tissue
- GEISHA (Gallus Expression *In Situ* Hybridization Analysis): A centralized and comprehensive repository of precise spatial and temporal information on chicken embryonic gene expression created through in situ hybridization
- GenePaint: A digital atlas of gene expression patterns in the mouse
- GENSAT (Gene Expression Nervous System ATlas): A database that captures information on gene expression in mouse brain at several developmental ages

As part of the Coordination and Sustainability or International Mouse Informatics Resources (CASIMIR) initiatives information on ontologies and resources for mouse biology, genetics and functional genomics was co-ordinated and is kept up to date with the latest information[16].

[16] Informatics resources for mouse functional genomics at www.i-mouse.org/

4 EMAP DIGITAL CURATION PROCESSES AND ACTIVITIES

4.1 The EMAP Data

The DCC Curation Lifecycle Model (Figure 1) (Higgins, 2008) has data as its focus, with the data defined as "any information in binary form". The focus of the EMAP's digital curation effort is on the dataset held in the web-accessible EMAGE Database, which may also be queried on the web using the EMAP Atlas. Both the Atlas and Database are described below, and the study will focus on the curation activities the EMAP team undertakes to develop and maintain the EMAGE Database.

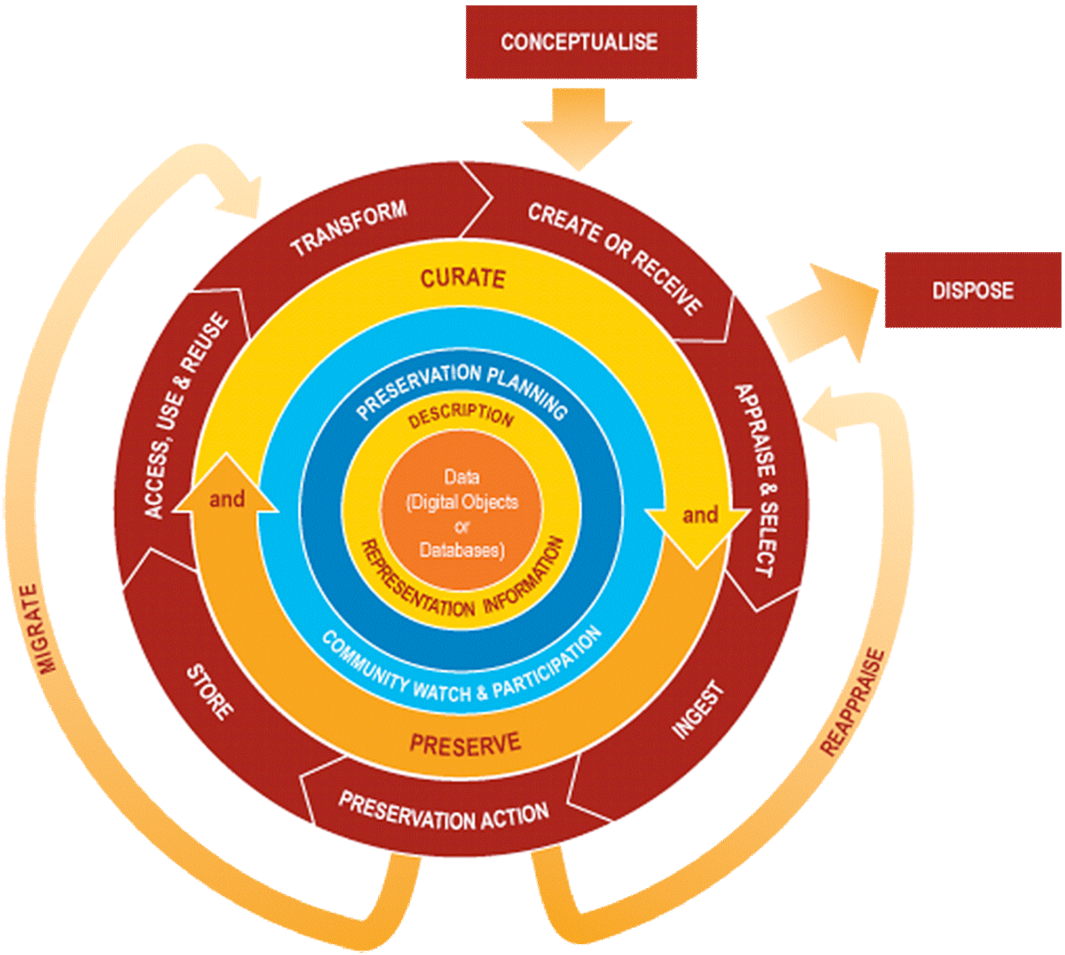

Figure 1: DCC Curation Lifecycle Model.

4.1.1 The EMAP Atlas

The EMAP Atlas is a digital atlas of mouse embryonic development and is based on the definitive publications of mouse embryonic development by Theiler (Theiler, 1989) and Kaufman (Kaufman, 1992). From these studies a series of interactive three-dimensional (3D) computer models of mouse embryos at successive stages of

development with defined anatomical domains were mapped to a stage-by-stage ontology of anatomical names. The atlas represents a 3D model with a comprehensive list (EMAGE's anatomy ontology[17]) of anatomical structures for every Theiler Stage (TS) of the mouse embryo.

4.1.2 The EMAGE Database

The Edinburgh Mouse Atlas Gene-Expression Database (EMAGE) is one of the first applications of the EMAP framework and provides a spatially mapped gene-expression database with associated tools for data mapping, submission, and query (Baldock *et al*, 2007).

The aim of EMAGE is to[18]:

- Provide a focal point for biomedical and clinical researchers to access mouse embryo *in situ* gene expression data sourced from the community
- Offer high-quality annotation and curation of gene expression data in the spatio-temporal and anatomical framework of the EMAP Digital Atlas
- Generate and offer methods for analysis of gene expression data
- Be used in the broader context with other bioinformatics resources to generate a tool for understanding the genetic control of mouse development.

EMAGE data comprises the original raw data, processed data (mapping, image size compression) and the website interface of descriptive and image files with their related EMAGE identifier and metadata information. The database structure comprises the gene expression and anatomy ontology. There is both a private local copy and the publicly available databases.

4.2 Conceptualisation of the Data to be Curated

The EMAP Project uses a visual data model for the Edinburgh Mouse Atlas Gene-Expression Database (EMAGE). This illustrates their conceptualisation of the data to be curated, the first of the sequential actions in the DCC Curation Lifecycle Model.

The EMAGE data model is schematically represented below (Figure 2) and illustrates the external data sources, EMAGE internal data flow and curation processes and the publicly visible EMAGE database. The model also represents the Project's planning of objectives, which includes increasing collaborations and international projects,

[17] The nomenclature database at genex.hgu.mrc.ac.uk/Databases/Anatomy/new/

[18] EMAGE at www.emouseatlas.org/testemage/about/about_EMAGE.html

understanding the requirements of stakeholders and users of the EMAGE database and strengthening community relationships.

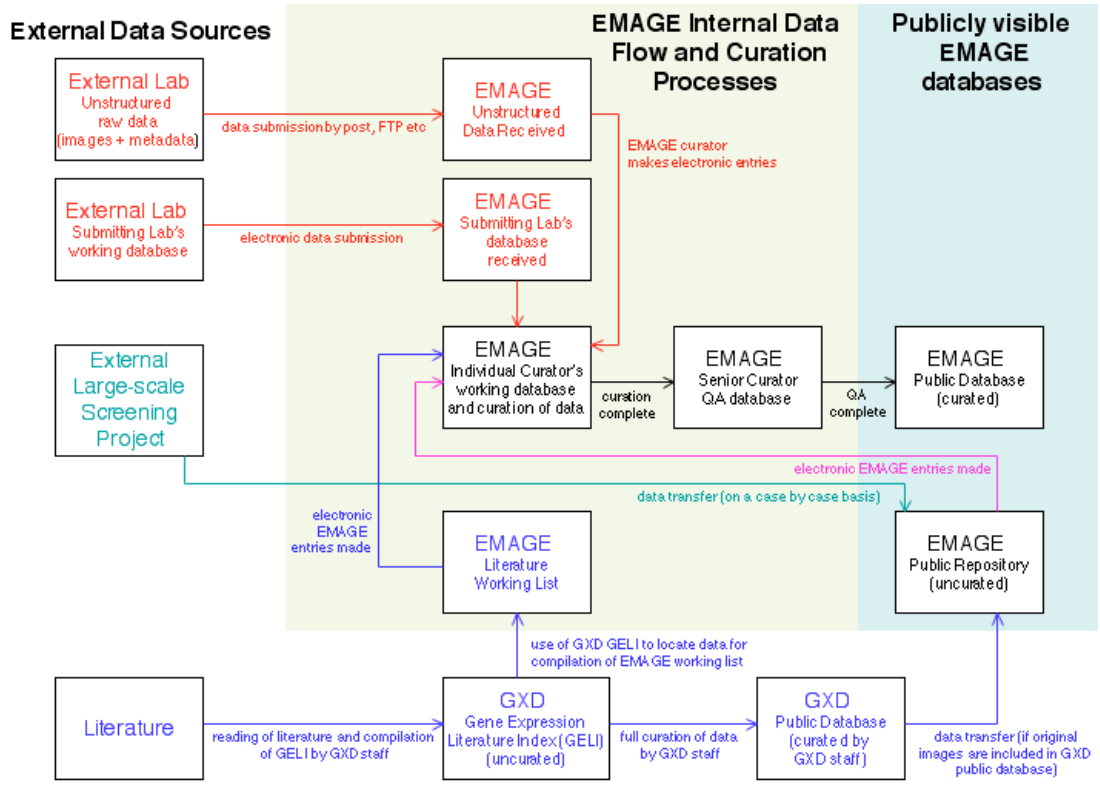

Figure 2: The data model for EMAGE.

4.3 Data Creation and Receipt

The EMAP Project sources data for inclusion in the EMAGE dataset from a variety of places detailed below. Data is received in a number of different formats, then selected and transformed to a structured format. The data is then described, using the MISFISHIE metadata standard, to enable the data to be discovered through their web interfaces, and additional annotation added. Sourcing and describing data corresponds to the "Create or Receive" action of the DCC Curation Lifecycle Model.

4.3.1 Data Sources

Data for the EMAGE database is sourced and shared[19]:

From the literature, published data from journals such as; Development, Developmental Biology, Mechanisms of Development and Gene Expression Patterns.

- In collaboration with the Gene Expression Database (GXD, Section 8.1); the Gene Expression Literature Index (GELI) is an index complied by the GXD of

[19] EMAGE at www.emouseatlas.org/testemage/about/about_EMAGE.html

scientific publications from over 150 journals that contain mouse *in situ* expression data. This includes information on the authors, gene/protein assay, whether the samples were whole-mount or sectioned and the age of the specimens involved. To date, information for 3986 images for 1188 genes has enabled at least one whole-mount image per gene to be annotated.

- From large-scale projects and screens, such as EURExpress, Mahoney Transcription Factor data and FaceBase (Section 8). To date, 3D images of approximately 300 embryo samples have been incorporated into the public EMAGE database as part of the FaceBase pilot study.

- Directly from numerous laboratories as mouse embryologists and geneticists are actively encouraged to deposit their *in situ* gene expression data in the EMAGE database (ideally using the Java EMAGE data submission interface). For example, data has been received from Dr. Paula Murphy at Trinity College Dublin[20] and Dr. Janet Rossant at Toronto medical Discovery Tower, Canada [21]. To date, approximately 2,500 images (of all different formats) at all stages of mouse development have been obtained.

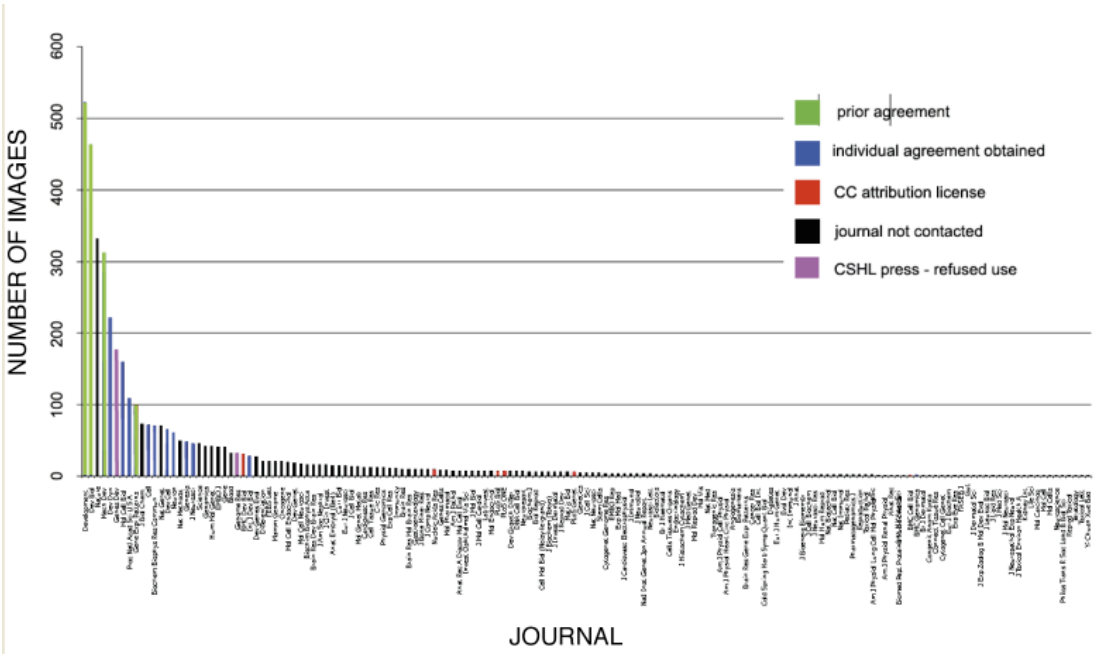

Figure 3: Survey of relevant image content of over 100 journals and status with respect to image reproduction by EMAGE. A subset of information within the GXD Gene Expression Literature Index (GELI) was surveyed (whole-mount data for TS15-19 embryos). Appendix 6 lists the x-axis in full, together with the figures on the y-axis representing the number of images for this dataset in each journal in the GELI. The

[20] Paula Murphy's Laboratory at http://www.tcd.ie/Zoology/research/WntPathway/

[21] The Rossant laboratory at www.sickkids.ca/rossant/

colours of the bars represent the status of image reproduction rights for EMAGE for each journal (e.g. whether agreements have been reached between EMAGE and each journal publisher or if a journal publishes under a Creative Commons Attribution License).

4.3.2 User Submission Options for Data

The EMAGE data model (Figure 2) enables users to search the central EMAGE database, make their own private local database for in-lab data management or submit gene expression data for curation and inclusion in the publicly available EMAGE database. The data submission options, that the EMAGE team actively encourage, are to[22]:

- Primarily, follow the EMAGE electronic data submission instructions to make one or more local (private) databases in which selected entries can be submitted to the EMAGE editorial office for curation and subsequent inclusion in the public EMAGE database.

- Submit data directly to the EMAGE Editorial Office to enable electronic entries of the information to be created for curation and subsequent inclusion in the public EMAGE database. Information can be sent by post (e.g. on compact disc), FTP (File Transfer Protocol) transfer, email attachment etc. and all common file formats for text-based information (plain text, Excel, Word) and images (jpeg, gif, tiff, png etc.) are acceptable.

4.3.3 Identifying New Data

The EMAGE team have begun to negotiate access to data for entry into the database from a number of different data sources. These include:

- EMBRYS ISH data, Hiroshi Asahara *et al*, National Research Institute for Child Health and Development, Japan
- European Conditional Mouse Mutagenesis (EUCOMM) / Knockout Mouse Project (KOMP), Wellcome Trust Sanger Institute, UK
- VISTA23, Lawrence Berkeley National Laboratory, USA

4.3.4 Data Creation Statistics

The current number of data entries in the public EMAGE database is approximately 5,500 spatial and text annotations and the number of genes/proteins represented is approximately 2,400 (Figure 4). Initially, there was a steady increase in EMAGE database growth, however after May 2006 to November 2007 the number of data

[22] Informatics resources for mouse functional genomics at www.i-mouse.org/

[23] VISTA Enhancer Browser at enhancer.lbl.gov/

entries and genes/proteins slowed due to the primary source at this time being data that had been published in the literature. The main reasons for the annotation rate decrease were the time required to assess and find images (the exact section plane) that were suitable for data mapping and the length of time to check probe/antibody details.

The focus over the last year has been to increase gene coverage by obtaining whole-mount data at a specific range of stages of mouse embryo development (TS15-19) and the current target is 1,500 data entries per annum (Figure 3). Thus, there is greater value in entering data for more genes at fewer stages than for fewer genes at more stages. Data has been obtained primarily from the literature, large-scale projects and screens (Section 8). A recent publication on EMAGE stated that 8% of data had been obtained from direct submissions, 52% from data previously published in the literature and 40% from screening consortia (Venkataraman *et al*, 2008).

Currently, the data entry rate for an EMAGE curator is 20 entries per week. Thus, with two curators the team currently curate 40 entries per week. On average 40 curated entries per week are reviewed and entered into the EMAGE database, although the quality assurance is the limiting step as one senior editor reviews all curations. Once data has been checked by the senior editor and submitted to the central database the data appears online, for public use, in less than an hour (initially a thumbnail represents the curated image until the image processing is complete).

Data entry rate is dependant on the data source meaning that the amount of data entered into the EMAGE database is project specific. For example, due to the fact that the EURExpress project data has been annotated to an acceptable standard it is likely that more than 40 entries will be made per week prior to review by EMAGE editors.

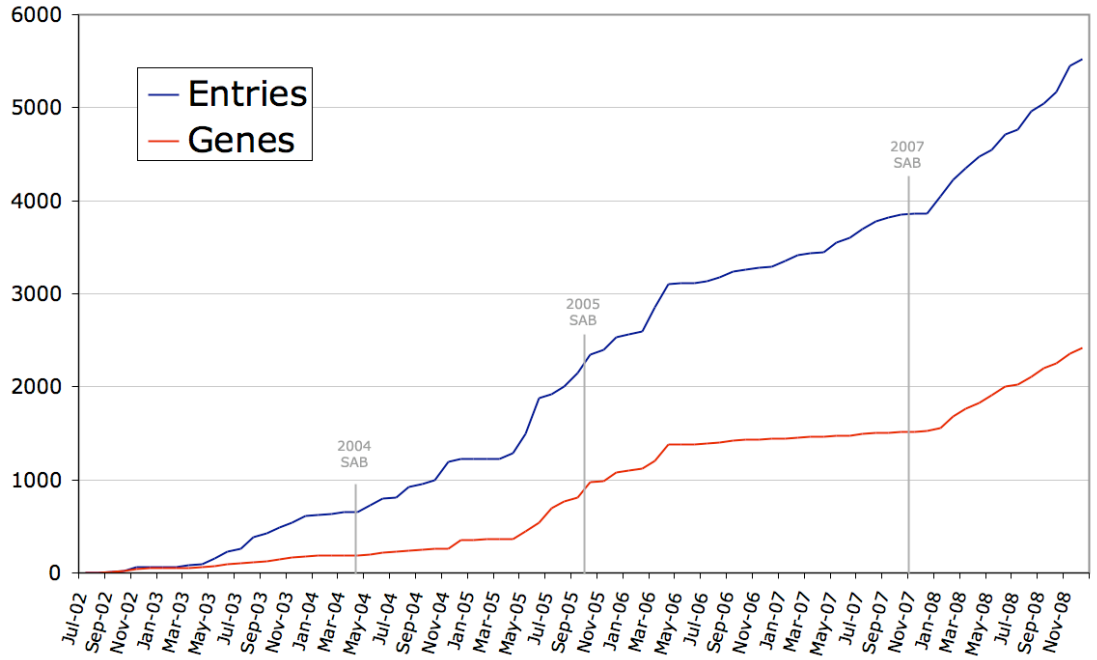

Figure 4: EMAGE database growth, the rate of spatial/text annotation and data entry into the public EMAGE database. The number of individual entries and genes represented in the public EMAGE database over time; one entry = one annotated representation of the sites of expression for one gene from one or more original assay images from one specimen. SAB is the times of previous Scientific Advisory Board Meetings.

4.3.5 Metadata for Discovery and Administration

When compiling the information, the EMAGE editorial team strongly encourage those entering data to follow the proposed MISFISHIE guidelines[24], to ensure the information given is sufficient so that the experiment can be interpreted and/or repeated by others.

The seven basic parts required to make an EMAGE entry, preferably with a list of any pertinent references and any other relevant information, are:

1) Name and contact details.
2) The detection reagent used (the probe or antibody used, specified as accurately and as unambiguously as possible e.g. full nucleotide sequences of probes or catalogue numbers of antibodies are preferred).
3) The gene or protein whose expression is being detected (use an identifier where possible e.g. MGI, Entrez or Ensembl gene identifier).
4) Information about the specimen (e.g. stage of development, strain, mutations).
5) The full method used (*in situ* hybridisation, immunohistochemistry or *in situ* reporter).
6) Original data image(s), at least one is required.
7) A text-based description of the sites where expression is detected.

EMAGE database entries also include:

- The source; whether from a journal, screen or direct submission with the submitter's contact details
- The detection regent; detailing whether the method used to detect expression was either probe or antibody
- The experimental conditions including assay information
- Associated references and relevant links to data in other databases.

4.3.6 Annotation Methods

The sites of gene expression, detected (strong, moderate, weak) and not detected, are described by[25]:

[24] The MISFISHIE working group at mged.sourceforge.net/misfishie/

[25] EMAGE at www.emouseatlas.org/testemage/about/about_EMAGE.html

- The process of denoting appropriate regions in the EMAP virtual embryos to capture spatial-based data, known as 2D spatial annotation;
- Text annotation, which can be performed manually by using the original information provided by the author or automatically inferred from a 2D spatial annotation to refer to the appropriate terms in the anatomy ontology to write text-based descriptions;
- Full 3D spatial annotation which is currently being developed by the EMAGE team.

4.4 Appraise and Select

Appraisal and selection of the available data to curate in EMAGE is undertaken as part of the data sourcing. This means that the "Appraise and Select" action of the DCC Curation Lifecycle Model, to evaluate data and select for long-term curation and preservation, is largely undertaken in parallel with the "Create or Receive" action. Additionally selection and appraisal activities concentrate on the quality of the data, and their associated metadata and image files, which are subject to rigorous quality assurance checks. Reappraisal of data that fails integrity and quality checks is undertaken in accordance with the occasional "Reappraise" action of the DCC Curation Lifecycle Model.

4.4.1 Data Quality

All data curated is checked and corrected by EMAGE's senior editor. Primarily, this is to ensure accuracy of text and spatial description of sites of expression, detection reagent information and experimental conditions. The quality assurance process steps are to:

- Open the curated information with the paper from which data has been extracted.
- Check the external data identifiers (for example, gene or sequence identifier) and links to external sources are correct.
- Check probe/antibody details.
- Check spatial annotation.
- Check confidence assignments.
- Once complete the curated data can be submitted to the central database.

Information, for example probe details, can be saved by the senior editor for future use. The senior editor also notes errors that have been identified by the editorial team to be corrected in the future. For example, for the correction of MGI/GELI data (GXD collaboration) the following information would be recorded:

> EMAGE identifier|MGI panel label|MGI assay identifier|EMAGE Editor comments|Corrections for MGI/GELI

A potential limiting step is that the senior editor then writes down the EMAGE identifier information on paper and this is passed back to the curator to update themselves (as work is divided between curators).

The current quality assurance process that is being followed by the EMAGE editor could potentially be improved, as there is no formal process for the correction of errors. Currently, errors that have been identified by the senior editor are also being corrected by the senior editor and feedback is provided to the curator informally. For the training and development of the curators it may be best for feedback to be provided in a more structured manner so that the curators can learn from the mistakes that are being made and all could work together on data that is difficult to curate. It is likely that a more formal approach will not only improve efficiency but also reduce the number of curation errors that remain uncorrected. However, those using the EMAGE database have highlighted few errors, which suggests that the quality of curated data within the EMAGE database is of a high standard.

There are no formal procedures (inter-annotation/curation agreement scores) for consistency tests between curators and the senior editor. Although at the start of a project there are a number of discussion groups to resolve annotation and curation issues, the EMAGE editorial team would extract text and spatial annotations from the same papers to determine the level of consistency between them.

4.4.2 Rating Data

Recently, EMAGE curators have begun to score the quality of incoming data images by assigning a confidence score (good, moderate or poor) on how closely each spatial annotation reflects the data observed in the data image (Venkataraman *et al*, 2008). These scores have also been retrospectively assigned to all previous spatial annotations in the database. Two factors that contribute to the overall confidence of an annotation are:

- The clarity and ease of interpretation of the staining pattern
- The degree of morphological similarity between the data embryo and the EMAP embryo template that the data is spatially mapped onto.

The scores can be used to gauge the potential quality of each spatial annotation and for filtering data sets for spatial analyses (such that only the highest quality annotations are used, for example).

This simple approach of using a confidence score rather than percentages has been designed to enable the usage and subjective nature of the rating process to be reviewed over time and potentially adapted in the future.

4.5 Ingest and Preservation Action

After data is sourced and appraised for inclusion in the EMAGE dataset the "Ingest" action of the DCC Curation Lifecycle Model is undertaken. EMAGE ingest procedures involve checking and correcting the conversion of the non-standard data, from the different sources, to the standard, structured format which allows for subsequent data interrogation and exchange. Thus, data can be stored, accessed and discovered by ensuring they are: in an acceptable file format for inclusion in the dataset; described consistently; and annotated appropriately. Data that is not in an acceptable file format may require to be migrated to ensure long-term preservation. This "Preservation Action" corresponds to the occasionally used "Migrate" action of the DCC Curation Lifecycle Model.

4.5.1 File Formats

The EMAGE database contains gene expression data in the mouse embryo from the following methods; *in situ* hybridisation (directed against RNA), immunohistochemistry (directed against proteins) and *in situ* reporter (data generated by genetic methods such as transgenics, animal modifications).

These methods can be performed on whole tissues, the entire embryo, which the EMAGE team also referred to as whole-mount (WM), or on tissue sections of the embryo and the raw data images (saved as jpeg). These are shown in EMAGE as conventional photographs, movies (saved as QuickTime or MPEG1) or 3D images (woolz format[26]) derived from techniques such as Optical Projection Tomography (OPT)[27].

4.5.2 Preservation Action

Unique identifiers are assigned to the data. This helps maintain provenance information and ensures that the data can be cited. Previously versioned curated data are updated and where necessary the format is migrated. Metadata can be received in many different formats and is saved as XML in the EMAGE database. The potential loss of experimental context information has not been explored here.

4.6 Access, Use and Reuse

EMAGE data is made freely available to the community through a Web interface that requires no authentication procedure. The EMAP team ensure data accessibility to both themselves and users of the EMAGE database, although there are some copyright issues. This "Access, Use and Reuse" step of the DCC Curation Lifecycle

[26] The woolz image processing software developed by the MRC Human Genetics Unit at genex.hgu.mrc.ac.uk/Software/woolz/

[27] EMAGE at www.emouseatlas.org/testemage/about/about_EMAGE.html

Model is supported through: published papers, attendance at international meetings, and an outreach programme that trains users and contributors alike.

4.6.1 Database Access - Searching and Browsing

Data stored within the EMAGE database can be analysed by either text- or spatial-based methods. For example, it is possible to perform Boolean operations between two sets of EMAGE entries (gene, stage of development, expression pattern, anatomical structure) using the three most basic Boolean logic operators 'and', 'or' and 'not'. Alternatively, genes can be hierarchically clustered into potential synexpression groups that contain highly similar expression profiles based on the spatial expression patterns themselves rather than intermediate text description[28]. From a digital curation perspective the database can itself be regarded as a digital analytical object supporting re-analysis and the production of new knowledge.

In addition, information about the EMAGE data can be obtained by running SQL queries and scripts.

4.6.2 Data Usage

Initially usage of the EMAGE database, mainly by researchers in the USA and Europe, was tracked regularly by obtaining information on requests for mouse atlas data (on a compact disc) and online EMAGE software, and from publication references and website usage. Since 2007, exact use of the EMAGE interface has not been logged, however, approximately 2 million requests per year for static EMAGE web-pages are made public per year and there has been a significant increase (from 105,000 to 170,000 requests) in the Repository Browse / Quick Search function since 2007.

4.6.3 The Outreach Programme

The EMAGE editorial team publish papers outlining the functionalities of the EMAGE database and continue to promote EMAGE to the community and educate users (termed the 'outreach programme') by attending and presenting at conferences and developmental biology meetings. Information is also obtainable directly from the EMAP/EMAGE resources (on-line demonstrations, tutorials (on-line and course lead) and technical documentation), which cover many aspects of the EMAP digital atlas and EMAGE. This work not only ensures that users and potential users understand how to use the EMAGE database for their research but also enables users to understand what information is of key importance to ensure that their data can be displayed in an accurate and informative manner in the database.

[28] EMAGE analysis options at www.emouseatlas.org/testemage/analysis/all_analysis_tools.html

4.6.4 Copyright of Data Sources

Agreements originally organised by Dr. Martin Ringwald (Associate Staff Scientist and Head of GXD Database) enabled the reproduction of original data images on the EMAGE website from 4 journals (Development, Developmental Biology, Mechanisms of Development and Gene Expression Patterns).

To date, the EMAGE team have arranged individual legal agreements with the publishers of 24 journals (that collectively house over 80% of published *in situ* gene expression images in the mouse) that do not license its material under a suitable Creative Commons Attribution License (Figure 2, Appendix 6). This allows the EMAGE team to reproduce copyrighted images from these journals on the EMAGE website. If the EMAGE team do not have permission to reproduce the original data image, it is their policy to use a generic image showing the copyright symbol on the EMAGE website that also includes a relevant link to the original data at either PubMed entry or a digital object identifier (DOI) link direct to the data at the journal website.

4.6.5 Intellectual Property (IP)

The Optical Projection Tomography (OPT) imaging software is patented and sold commercially however many of the aspects of the EMAP are not patent protected due to the following reasons:

- The software developed is open source, as a business decision was taken not to industrialise the research code

- The selected gene expression sources from which data is obtained.

4.6.6 Commercialisation Position

The EMAP database is not currently commercialised and there are no plans to commercialise the database in the future. Thus:

- The annotation and curation of all gene expression data is freely available through EMAGE.

- The mouse models developed by the EMAP are freely available although payment is taken for sending the mouse models on compact disc to cover the cost of providing this service.

There is potential for the EMAGE atlas model to be used across organisms and in a commercial medical setting, although the market and IP landscape would need to be researched in detail. For example, the EMAGE atlas paradigm could enable a database to be developed that records and maps patient medical information to a human anatomy framework. Ongoing and future collaborations may enable this path to be explored in more detail and to assess what EMAGE developments (changes to the database schema and EMAGE interface) would be required.

4.7 Transform

Transformation of existing data to create newly derived results from original datasets, by selection or query, is an important action in the DCC Curation Lifecycle Model. EMAP is making continual technological advances to enable EMAGE database users to perform complex analysis and obtain statistical information on data stored on either their own private local database or the publicly available EMAGE database.

4.8 Description and Representation Information, and Community Watch and Participation

The standardisation of EMAGE data is of key importance to the success of the EMAGE database (Figure 5). The use of metadata standards enables current and future interoperability with other relevant projects, and ensures that the data created can be accessed and administered over the long-term. The DCC Curation Lifecycle Model recommends that standards are used for description throughout the curation lifecycle, and that the representation information required to both understand and render digital materials, and their metadata, are collected.

The EMAGE Team are actively participating in the development of appropriate descriptive and interoperability standards, as recommended by the "Community Watch and Participation" full lifecycle action of the DCC Curation Lifecycle Model. Close collaboration with other related projects ensures that the data can be created in an interoperable manner for sharing across projects.

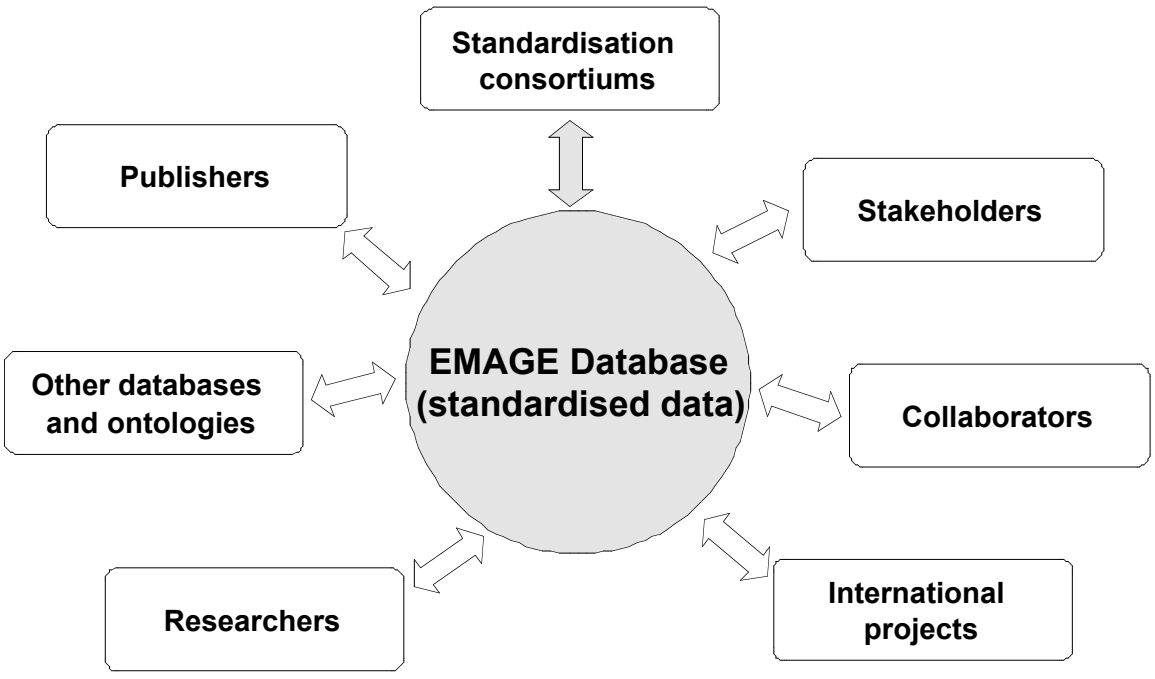

Figure 5: The importance of the standardisation of EMAGE data.

4.8.1 Data Standardisation

The EMAGE team actively promote standardisation within the life sciences community by:

- Being part of an international consortium that is developing a minimum specification for *in situ* hybridisation and immunohistochemistry experiments (known as 'MISFISHIE'[29]), and has developed a schema that can be used to record all aspects of an *in situ* experiment. In addition, a separate *in situ* detection reagent database of all probes and antisera used in EMAGE will be developed.

- Promoting database interoperability and integration by sharing EMAGE information on the Mouse Resource Browser (MRB) developed by the BioIT Unit at Alexander Fleming Biomedical Sciences Research Center as an electronic aid for searching and retrieving information about online mouse resources[30].

- Working with key networking bodies such as The Coordination and Sustainability of International Mouse Informatics Resources (CASIMIR) and The European Life-Sciences Infrastructure for Biological Information (ELIXIR).

4.8.2 Collaborations and International Projects

4.8.2.1 The Mouse Gene Expression Information Resource

EMAGE is part of the Mouse Gene Expression Information Resource (MGEIR) project, which is the collaboration between EMAP and the Gene Expression Database[31] (GXD) project at the Jackson laboratory (Dr. Martin Ringwald), USA. GXD collects and integrates the gene expression information in the Mouse Genome Informatics[32] (MGI) databases to enable the scientific community to view gene expression information about the mouse in the context of genetic, sequence, functional and phenotypic information (Smith *et al*, 2007). The ultimate aim of the MGEIR is to provide a unified resource that combines text-based and spatial-based methods to store, display, and analyse mouse developmental gene expression information. Key points to note are that:

[29] The MISFISHIE working group at mged.sourceforge.net/misfishie/

[30] EMAGE Mouse Resource Browser information at bioit.fleming.gr/mrb/Controller?workflow=ViewModel&eid=18

[31] Gene Expression Database at www.informatics.jax.org/expression.shtml

[32] Mouse Genome informatics at www.informatics.jax.org/

- GXD and EMAGE obtain gene expression data from the literature and by direct submission, both incorporate data from *in situ* techniques however GXD also incorporates data from 'non-spatial' expression profiling techniques such as RT-PCR, Northern blots etc.

- A key difference between GXD and EMAGE is that GXD incorporates text data only whereas EMAGE incorporate text and spatial data

- GXD and EMAGE follow common guidelines that result in consistent descriptions that can be shared between the two databases

- The EMAP mouse nomenclature database has been incorporated in the GXD gene-expression database to enable the unstructured, text descriptions of gene expression patterns to be converted into a standardised description that is available for database storage and query.

4.8.2.2 EURExpress

EURExpress[33] is a transcriptome atlas database for mouse embryo development and the EURExpress consortium are currently working on a 4 year project to generate mRNA *in situ* hybridisation data for approximately 20,000 mouse genes on sagittal sections at embryonic day 14.5 (~24 evenly spaced sections for each gene), and performing a text-based annotation of the sites of expression seen in all 480,000 images. The aim of the EMAGE team is to assess the text annotation data, initially by using automated methods, before visually assessing each image to enable the EMAP anatomy ontology to be used to describe the sites of expression spatially. EMAGE plans for the EURExpress data to be imported into EMAGE in 2009 and in addition to the information already compiled by the EURExpress consortium, EMAGE has developed automated signal extraction and alignment methods to allow spatial-based annotation and analyses to be applied to this dataset.

4.8.2.3 EuReGene

The goal of the European Renal Genome (EuReGene) is to discover genes responsible for renal development and disease to enable their proteins and actions to be researched further. This is achieved by a consortium of leading scientists, clinicians and industry partners (particularly small and medium-sized enterprises) working together to develop novel technologies and discovery tools that could be applied to kidney research[34]. In March 2008, the EuReGene Kidney Atlas and Expression databases with movies of kidney development, the ontology database and EuReGene's mutant phenotype data were made publically available.[35]

[33] EURExpress at www.eurexpress.org/

[34] EuReGene at www.euregene.org/

[35] EuReGene Kidney Atlas Data Portal at www.euregene.org/euregene/pages/kidney_atlas.htm

4.8.2.4 Mahoney Transcription Factor Data

This dataset (Gray *et al*, 2004) is published by the Dana-Farber Cancer Institute and contains WM mRNA *in situ* hybridisation data for approximately 1350 transcription and other nuclear factors.

4.8.2.5 GUDMAP

The Genito Urinary Development Molecular Anatomy Project (GUDMAP) is a consortium of laboratories that work together to provide the scientific and medical community with tools to facilitate research[36]. The aim of the 5 year project (funded by the NIH) with the EMAGE team is to build the GUDMAP morphological atlas and the GUDMAP *in situ* and micro-array gene expression database to facilitate genitourinary development and disease research. The linking of GUDMAP gene expression data to the EMAGE database will provide spatially mapping (curation) data for approximately 34,000 GUDMAP annotated images.

4.8.2.6 DGEMap

The Developmental Gene Expression Map (DGEMap), an EU project that Newcastle University is coordinating, is the first "Design Study" for a pan-European research infrastructure dedicated to the analysis of gene expression patterns in early human development[37]. The project is arranged into four complementary and multidisciplinary activities which include; laboratory-based technologies, computer-based informatics technologies, ethical framework study and feasibility study to determine the organisational and collaborative structure necessary for a new research infrastructure designed by and dedicated to the scientific community.

4.8.2.7 FaceBase

FaceBase is a 2 year pilot study that aims to produce image data (2D and 3D images) depicting mRNA *in situ* hybridisation patterns for approximately 500 genes involved in craniofacial development, at several stages of mouse embryo development. This work is currently being produced in the laboratories of Dr. David FitzPatrick (MRC Human Genetics Unit, Edinburgh) and Dr. Mike Dixon (School of Dentistry, Manchester University) and is funded as part of the National Institute of Dental and Craniofacial Research (NIDCR) Center[38].

[36] GenitoUrinary Development Molecular Anatomy Project at www.gudmap.org/

[37] The Developmental Gene Expression Map at www.dgemap.org/

[38] The FaceBase Project at www.nidcr.nih.gov/GrantsAndFunding/See_Funding_Opportunities_Sorted_By/ConceptClearance/CurrentCC/FaceBase.htm

4.8.2.8 e-CHICKATLAS

The aim of the e-CHICKATLAS project (researchers from the University of Bath, the Roslin Institute (University of Edinburgh), the MRC Human Genetics Unit (Edinburgh), University College London and Trinity College Dublin) funded by the BBSRC is to develop a three-dimensional atlas and gene expression database for chick development with cross comparisons to the mouse via the EMAGE database[39]. Expression data will focus on approximately 1,000 genes (identified as having expression in several organiser regions) at two stages of development.

4.9 Planning Curation and Preservation

The EMAGE Team have a number of processes, tools and resources which are currently used, throughout the curation lifecycle, to help plan and undertake management and administrative tasks. Using these help to ensure that the "Curate and Preserve" actions of the DCC Curation Lifecycle Model, pertinent throughout the curation lifecycle, are considered.

EMAGE will continue to source and develop processes to "Curate and Preserve" spatial data in the developing mouse embryo. It is of key importance to develop tools for curation and analysis, while planning for future preservation needs. Obtaining good reviews on EMAGE will increase the profile of the database, ensuring that it is more widely used among researchers.

Future access is dependant on both the curation methodologies employed, and the continuation of funding to ensure that curation can continue. Planning future funding is an important part of the "Preservation Planning" action in the DCC Curation Lifecycle.

4.9.1 Standard Operating Procedures

Standard Operating Procedures (SOPs) are written and made available for others to view (internally only) on the EMAGE Wiki (Appendix 3). Some of the SOPs are project specific, for example there are specific SOPs for the EURExpress project, and the amount of detail described is procedure dependant, for example detailed SOPs describe clustering and specimen preparation.

4.9.2 Curation Tools and Methods

Tools (Table 2) that are used by the EMAGE team are either developed internally or obtained externally. EMAGE database users can directly access the software systems that are externally sourced, however, from the tools that are developed in-house only the submission interface Java Client and MAPaint tool can be obtained directly from the EMAGE team.

[39] Chick Atlas, University of Bath at www.bath.ac.uk/news/2008/11/14/chick-atlas.html

The development process for the tools that are produced in-house is driven by the scientists proposing a new tool to aid curation. If the tool is deemed to be important for curation, time and resources are allocated for the design, development, testing, bug fixing and release of the tool. Tools that are only used internally tend to be tested as the scientists use the tool whereas those tools that are made publicly available are thoroughly tested across multiple platforms. The maintenance and longevity of tools developed internally is a potential issue as time is not always taken to document and describe how code is written. This means that previously written code is rewritten rather than being reviewed and modified unless the developer who initially wrote the code is still working for the EMAP.

Tool name	Description (externally sourced or developed in-house)
Mantis Bug Tracker	A popular free web-based, project management, bug tracking system (externally sourced)
Wiki	A project management software that captures SOPs, communication, R&D discussions and action points, between the team members (developed in-house)
Submission Interface Java Client	A tool to create, submit and edit entries (developed in-house)
EMAGE AdminTool	A tool to track and record curation and quality assurance actions (developed in-house)
Axiope Catalyzer	To view data received externally and to categorise the embryo stage of data (externally sourced)
MAPaint	A tool to map curation to the correct embryo model (developed in-house)
AMIRA	A tool that is currently being used for 3D visualisation and warping (externally sourced)
Visualisation ToolKit (VTK)	An open-source, freely available software system for 3D computer graphics, image processing and visualisation (externally sourced)

Table 2: **Tools used for EMAGE curation**

4.9.3 Curation Resources

There is not a handbook or any guidelines for EMAGE curation. Resources used for data curation include the mouse development atlas information that provides definitions for all stages and individual Theiler Stages and several external sources (information on gene/protein symbol and name, mouse strains, mouse alleles, nucleic acid sequences, amino acid sequences, probes and antisera and mouse embryo anatomy descriptions)[40].

[40] EMAGE at www.emouseatlas.org/testemage/about/about_EMAGE.html

4.9.4 Current Developments

Current developments in curation methodology include (Venkataraman *et al*, 2008):

- The development of automated methods (scripts currently in development) for signal extraction and tissue section registration to allow a partial automated approach to spatial annotation

- Complete redesign of the EMAGE website (incorporated drop-down menus, quick search functions and more extensive user help information) and associated User Query Interfaces (search by gene/protein name symbol, anatomical structure name, spatial region)

- Continual database development from an object-oriented to relational database structure, which includes new SQL access for text data in EMAGE (the EMAGE database software will continues to have a client-server architecture, however, separate databases will store the gene expression and anatomy ontology information)

- Continual development of the Image Internet Protocol image delivery system to include a 3D object sectioning and 2D section delivery component to allow sectioning of the 3D EMAP models in a web browser application

- A slide scanner for the collection and processing of slides received externally to enable data to be efficiently processed, recorded and entered into the EMAGE database

- The rewriting of the administration tool to incorporate; curation and quality assurance status, versioning, links back to the original data, user privileges and enable data to be captured inline with changes to the database schema and structure

- Database development to enable data entered from a specific source to be managed, extracted and analysed.

4.9.5 Proposed Developments

Work proposed for 2009 includes increasing data entry and analysis functionality, interface refinements and new search and analysis methods.

New concepts for the EMAGE database include community annotation for data in EMAGE. The notion of community annotation was recently been highlighted by the fact that some annotation were missed by the EURExpress dataset in which sites of gene expression were described by a consortium using a text annotation approach (Section 8.2). This raises the question of whether methods should be developed to allow community annotation for data held in EMAGE and similar databases.

4.9.6 Breakdown of Project Costs

The diversity, source and level of funding required for the research and development of a gene expression database, such as EMAGE, are shown in Table 3

Project	Funding source	Level of funding
EMAP	MRC	Less than £500,000 per annum
EMAGE (excluding EMAP)	MRC	Less than £250,000 per annum
EURExpress	EU	€1.5-2 million over 4 years
GUDMAP	NIH	$2.5-3 million over 5 years*
EuReGene	EU	€0.5-1 million over 4 years

* The funding of the GUDMAP project is greater than the value stated as the project is being extended and the final costs are being negotiated (the total is likely to be greater than $3 million)

Table 3: **The cost of funding a gene expression database**

The actually cost of funding a gene expression database, for example, EMAGE and GUDMAP, is more likely to be double the values stated (Table 3) once overhead and manpower costs are included. From information provided by Dr. Duncan Davidson it could be estimated that the EMAGE database has cost over £1 million (2001-2008 financial years).

4.9.7 Future Funding Opportunities

Currently, the core funding for the work completed by the EMAP is from the MRC. However, there are a number of additional external funding bodies which include; BBSRC, NIH and the EU. In the past funding has been obtained from the pharmaceutical company, GlaxoSmithKline, for the reconstruction of a new model (there were no intellectual property restrictions which enables EMAP to incorporate the new model into their atlas).

Obtaining funding for the development of database and ontology resources is possible; however the maintenance of ongoing databases is currently under-funded and obtaining funding for that is more difficult.

There are a number of key organisations that Dr. Duncan Davidson and Prof. Richard Baldock are working with, establishing and strengthening relations, in the hope that these organisations will be the gateway to funding opportunities. These include:

- The Coordination and Sustainability of International Mouse Informatics Resources (CASIMIR) which focuses on the co-ordination and integration of databases that contain experimental data relevant to the use of the mouse as a model organism for human disease. The aim is to set standards and benchmarks to allow data sharing and integration between European and International databases.

- The European Life-Sciences Infrastructure for Biological Information (ELIXIR), a consortium (32 research organisations, universities and companies from 13 countries), led by European Bioinformatics Institute (EBI) Director Prof. Janet Thornton[41], is working together to determine how to transform European biological databases into a bioinformatic network for life sciences (Marx, 2008).

- The UK e-Science Programme[42] that supports the generic facilities (National Grid Service, Open Middleware infrastructure Institute, e-Science Centres) for users and potential users of e-Science tools and techniques to further their research.

[41] The Thornton group at www.ebi.ac.uk/Thornton/

[42] The e-Science Core Programme structure and key activities at www.rcuk.ac.uk/escience/coreprog/

5 FINDINGS AND ANALYSIS OF THE CASE STUDY

5.1 Key Findings

It was found that the aim of the Edinburgh Mouse Atlas Project (EMAP) to produce a digital atlas of mouse development that can be used by the community to facilitate research was being accomplished. The scope for the development of the publicly funded Edinburgh mouse Atlas Gene-Expression (EMAGE) database was found to be well defined and the EMAGE team work efficiently to ensure that all set objectives are being achieved. Interestingly, market research showed that the EMAGE database is one of the first (and only UK) scientific products that provide a spatially mapped gene-expression repository with associated tools for data mapping, submission, and analysis.

No business model documentation was reviewed for either EMAP or EMAGE although EMAGE's mission is stated on their website[43] and their objectives are primarily driven by the Scientific Advisory Board (SAB) which consists of researchers who are experts within their fields and have an excellent depth of understanding of biocuration. There appeared to be good communication within the EMAP team and project collaborators although the EMAGE team may benefit from obtaining a clearer understanding of the future plans for EMAP and how this will direct development, ongoing and future collaborations, and funding opportunities for EMAGE.

The process of curation is highly skilled, based on expert judgement, and there may always be a manual component which is subjective. This is especially the case for spatial curation, which by nature is more complex than text curation because of the mapping of data (for example, gene expression) to a structured framework (for example, digital atlas of mouse embryonic development). However by exploring computational methods there is potential for the process to be partially automated. Through evaluation and benchmarking studies it would be interesting to determine whether changes planned by the EMAGE team lead to an increase in efficiency and accuracy of data curation.

The initial analysis focused on the inputs (stakeholders, funding) and outputs (data, stakeholders) of EMAP (Appendix 7) prior to mapping of the EMAGE curation processes against the DCC Curation Lifecycle Model (Appendix 8). This ensured that the case study summary report covered many aspects of the curation practices without exploring any in specific detail. A key finding of the case study is that, to optimise data curation a number of lifecycle management issues need to be continually assessed and further steps taken.

[43] EMAGE at www.emouseatlas.org/testemage/about/about_EMAGE.html

From conducting this short case study of EMAP, primarily focusing on the curation of text and spatial data that can be viewed using the EMAGE database, there are a number of next steps that can be recommended. The majority of these were formed from suggestions made by the EMAGE team.

5.2 Lifecycle Management Issues and Next Steps

5.2.1 Full Lifecycle Actions

DATA
SCOPE Data, any information in binary digital form, is at the centre of the Curation Lifecycle. This includes: Digital Objects: simple digital objects (discrete digital items such as text files, image files or sound files, along with their related identifiers and metadata) or complex digital objects (discrete digital objects made by combining a number of other digital objects, such as websites) Databases: structured collections of records or data stored in a computer system
LIFECYCLE MANAGEMENT ISSUES None

DESCRIPTION AND REPRESENTATION INFORMATION
SCOPE Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand and render both the digital material and the associated metadata.
LIFECYCLE MANAGEMENT ISSUES **Data standardisation:** The standardisation of experimental details and variability between the data sources is an issue for the EMAGE editorial team. Even though the team strive to ensure that data is correctly curated there are still a number of 'unspecified' entries in the database. Due to time and resource constraints it is not possible for the team to work with researchers individually however the team realise the importance of standardisation and actively address and promote standardisation methods within the life sciences community.
NEXT STEPS As the number of collaborations and externally funded projects increase the standardisation of curated data and the transfer of the mouse atlas knowledge will continue to be of key importance. It is important that the EMAGE team continue to drive forward the use of standardisation and transfer their experiences, lessons learnt to others that are working on similar projects. For example, Dr. Jeff Christiansen is currently working with Mr. Michael Wicks to ensure that data is correctly curated for the e-CHICKATLAS Project.

COMMUNITY WATCH AND PARTICIPATION
SCOPE Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software.
LIFECYCLE MANAGEMENT ISSUES None
NEXT STEPS Continually review the progress of the EMAGE collaborations and international projects and where possible share resources and tools to improve communication and efficiency.

CURATE AND PRESERVE AND PRESERVATION PLANNING
SCOPE Be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle. Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions.
LIFECYCLE MANAGEMENT ISSUES **Development of tools and methods:** There are various developments ongoing in tools and methods and the team recognise that for the continual maintenance and sustainability of EMAGE there is a requirement to support the database infrastructure as well as the interface. **Teamwork:** It was found that the EMAGE editorial team were strong in sharing their expertise internally and externally. The fact that the team is small and work in close proximity with each other may contribute to the successful working relationship between individuals. **Limited time and resources:** One or the main challenges that has been highlighted by the EMAGE editorial team is that there is a lot of work, mainly defined by the Scientific Advisory Board, which requires to be completed in a short time frame with few people.

5.2.2 Sequential Actions

CONCEPTUALISE
SCOPE Conceive and plan creation of data, including capture method, storage options.
LIFECYCLE MANAGEMENT ISSUES None

CREATE OR RECEIVE

SCOPE

Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation. Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata.

LIFECYCLE MANAGEMENT ISSUES

Data entry: All data received is entered into the EMAGE database using the administration tool directly or saved in an Excel spreadsheet before being manually entered. Within a small team the use of Excel spreadsheets does not appear to cause the EMAGE editorial team any issues however the tracking and error checking of data is limited.

Data throughput: High throughput is of key importance for the success of the EMAGE database. It is hoped that the focus on entering data for more genes at fewer stages will increase the likelihood of discovering novel genes that potentially overlap functionally and regulatory with known genes. The continual advance to increase curation efficiency through the use and development of curation tools and methods was evident and is an area of value that could potentially be investigated in more detail.

Data integration: The EMAGE team continues to research and develop data integration. For example, EMAGE and GXD are currently working together to produce a Mouse Gene Expression Information Resource (MGEIR) that will unify annotated data.

NEXT STEPS

Optimise data management and entry rate by rewriting the EMAGE data management tool (administration tool) by increasing the rate of curation and steps that could be automated.

Research and develop the use of computational methods to partially automate spatial integration, annotation and curation. Evaluate efficiency and data quality in relation to cost savings.

Optimise the cataloguing of data received externally.

Continue to educate users and potential users to make sure that their data can be imported and viewed using the EMAGE database.

APPRAISE AND SELECT

SCOPE

Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.

LIFECYCLE MANAGEMENT ISSUES

Data quality: The quality of data that is saved in the EMAGE database is dependent on a number of factors which include the probe or antibody description, morphology (colour and quality) of the raw data images and whether the exact age (Theiler Stage) of the mouse embryo can be correctly determined. If the information is published the EMAGE team curates the presented annotated data stating that the curation is their interpretation of the findings. However, if the information is not published the EMAGE editorial team either work with those that submitted the annotated data to ensure that data is correctly curated or state 'unspecified' which is not so informative.

Data quality assurance: The quality assurance process for EMAGE is defined and ensures that the data displayed in the database is of a high quality. This could be due to the fact that all curated data is checked and corrected by the senior editor. Although there is no formal process for the correction of errors significant errors made by the editors are discussed to avoid error repetition.

NEXT STEPS

Review the quality assurance process for curated data prior to being saved in the EMAGE database. Potentially, the Mantis Bug Tracker system could be used initially to log curation errors that require to be corrected by the curators. Efficiency could also be improved by the use of quality assurance reports so that corrections that were made to curated data could be recorded and saved. The advantages of recording and saving curated data, curation corrections and issues would be to monitor efficiency, aid in the optimisation of the process and support curation training.

INGEST AND PRESERVATION ACTION
SCOPE Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements. Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.
LIFECYCLE MANAGEMENT ISSUES None

STORE
SCOPE Store the data in a secure manner adhering to relevant standards.
LIFECYCLE MANAGEMENT ISSUES **Database costs:** For EMAGE, digital storage costs need to be considered however as the costs of computer storage are constantly reducing the primary costs for the EMAP staff is the long-term presentation and preservation of data.
NEXT STEPS Review database storage and capacity so that no issues will occur as a result of the planned increase in data entry. Optimise the storage of data received externally.

ACCESS, USE AND REUSE

SCOPE

Ensure that data is accessible to both designated users and reusers, on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable.

LIFECYCLE MANAGEMENT ISSUES

Data access: To resolve any copyright issues the EMAGE team have sought advice from Digital Curation Centre's Legal Services Associate. Although the majority of EMAGE's activities are covered by copyright agreements with the relevant publishers, there is a small portion of their work where the status remains unclear. It has been determined that 'fair dealing' (or 'fair use' in the USA) may have limited applicability for these activities. Discussions are ongoing as to the most appropriate way forward and include further investigation of copyright case law, as well as the possibility of joining with other bio-curators to make collective approaches to journals for permissions.

Data analysis: It was found that no structured documentation of queries was made or scripts documented in detail; however, current advances that are planned to be available in 2009 begin to address how the analysis of EMAGE data can be preformed and recorded by the EMAGE team and users of the database.

Data presentation: The presentation of whole-mount, sectioned and 3D OPT data presents a number of new challenges for the EMAGE team. The data file size, whether to show images statically or in rotation and how users will be able to download, view and correct annotations are all areas that require to be investigated further.

Ethical issues: The EMAGE database publishes information rather than performing research experiments which requires ethical approval. Potentially, ethical issues could be reduced by decreasing the number of duplicated experiments if researchers used the EMAGE database to review experiments that had been previously performed.

NEXT STEPS

Although much work has been completed on the EMAGE interface it is also important that time and resources are allocated to developing and improving the EMAGE processing tools.

Overcome copyright issues to enable images to be added to the database without explicit permission:

Continue to explore legal agreements with the publishers and to work with legal services of the DCC and the University of Edinburgh to resolve any copyright issues.

Explore the possibilities of working with publishers to enable published data to be

extracted and imported into the EMAGE database more efficiently, for example, researchers could submit their results for publication in a document and database format. Track EMAGE usage more extensively and determine an increase of usage of the database by: Developing a process for interpreting and presenting statistics of logged EMAGE data usage. Obtaining feedback from users on database requirements, improvement suggestions, data errors and demand for additional links to external data sources. Use local (contacts at the University of Edinburgh, MRC-HGU) and externally funded projects to obtain feedback on EMAGE. Continue to raise the profile of the EMAGE database by publishing information on the EMAGE database and ensuring that users of the database publish results on how the resource was used for their research.

TRANSFORM
Scope Create new data from the original, for exampleby migration into a different format, or by creating a subset, by selection or query, to create newly derived results, perhaps for publication.
LIFECYCLE MANAGEMENT ISSUES None

5.2.3 Occasional Actions

DISPOSE
SCOPE Dispose of data, which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian. In some instances data is destroyed. The data's nature may, for legal reasons, necessitate secure destruction.
LIFECYCLE MANAGEMENT ISSUES None
NEXT STEPS Understand the data deletion process, how data is disposed of and whether there are any legal obligations that require to be adhered to.

REAPPRAISE
SCOPE Return data which fails validation procedures for further appraisal and re-selection.
LIFECYCLE MANAGEMENT ISSUES None
NEXT STEPS Use the quality assurance process to check integrity and quality of data.

MIGRATE
SCOPE Migrate data to a different format. This may be done to accord with the storage environment or to ensure the data's immunity from hardware or software obsolescence.
LIFECYCLE MANAGEMENT ISSUES None
NEXT STEPS Understand the migration policy for moving datasets to new storage formats in more detail.

6 CONCLUSIONS AND RECOMMENDATIONS

This SCARP life sciences case study scoping report begins to explore the scientific product produced by the Edinburgh Mouse Atlas Project (EMAP). The digital curation activities undertaken by the researchers working on the construction of the publicly available Edinburgh Mouse Gene-Expression (EMAGE) database are documented. The lifecycle management issues and the next steps identified resulted from analysis of information on how the project has been scoped and driven, who the stakeholders are, and the numerous collaborations and international projects enabled the case study to be analysed and key findings reported. An initial mapping of the EMAGE curation processes against the DCC Curation Lifecycle Model is provided however a more detailed review of this case study will result in a clearer understanding of the project's risks which will enable more definitive recommendations to be made. Thus, a more in depth review of the EMAP would be beneficial.

The DCC and appropriate funders, including JISC, should support further investigation with the following recommended scope:

- Detailed discussions of the key areas of EMAGE, such as data curation and entry rate, identification of data for future entry, analysis and the development of software capabilities and computational methods.

- Further mapping of the EMAGE curation process with the DCC Lifecycle Model.

- Obtain a better understanding of feature detection and how EMAGE processes image information.

- A detailed analysis of the optimal time for data curation and what potential advances could be made to increase scientific product efficiency (for example, an increase in curation resources versus curation tool development).

- Explore the process and feasibility of scaling up the EMAGE team by working with the DCC to run a focus group, to capture the strengths and weaknesses of how the team work together, and their ideas on recruitment, management (ratio of senior editor to editors to database service administrators, working onsite versus offsite), training, estimating curation rates (the increase in volume of curated data) and assessing the quality of curated data (checking a percentage of curated data rather than all).

- Spend more time with the EMAP software developers; understand their roles and activities in the project in greater detail, how data is stored in a secure manner, what and how data is disposed, whether there is a migration policy for moving datasets to new storage formats and how the maintenance and sustainability of curation tools and methods could be improved.

- To map the relationships between data quality, scalability and users of the EMAGE database.

- Explore current and future opportunities of sharing EMAGE data.

- Complete a detailed review of the costs involved in developing a gene expression database, understanding the effort required and the value of the resulting product.

- Explore the funding opportunities for the EMAGE database and how the DCC can aid in the cost of development, digital storage, presentation and long-term preservation of the EMAGE data.

- Explore whether the EMAGE atlas model could be used commercially in a clinical setting.

7 REFERENCES

Bairoch, A and Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement trEMBL. *Nucleic Acids Research*, 25(1):31-36, 1997.

Baldock, R.A. *et al*. EMAP and EMAGE A Framework for Understanding Spatially Organized Data. *Neuroinformatics* 1(4): 1559-0089, 2007.

Boline, J., Lee, E., Toga, A.W. Digital atlases as a framework for data sharing. *Front. Neurosci*, 2(1): 100-106, 2008.

Brazma, A. *et al.*, Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics* 29: 365-371, 2001.

Buneman, P., Cheney, J., Tan, W-C., Vansummeren, S. Curated databases. *Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (Vancouver, Canada, June 09 – 12), 1-12, 2008.

Chu, W.L. Report identifies bioinformatics as growth area. *DrugResearcher*, August, 2005.

Deutsch, E.W. *et al.,* Minimum information specification for in situ hybridization and immunohistrochemistry experiments (MISFISHIE). *Nature Biotechnology* 26: 305-312, 2008.

Feick, K. Bioinformatics Market Surges Ahead. *Frost and Sullivan*, 2005.

Galperin, M.Y. The Molecular Biology Database Collection: 2006 update. *Nucleic Acids Research*, 34, Database issue D3-D5, 2006.

Galperin, M.Y, The Molecular Biology Database Collection: 2008 update. *Nucleic Acids Research*, 36, Database issue D2-D4, 2008.

Gray *et al.*, Mouse Brain Organization Revealed Through Direct Genome-scale TF Expression Analysis. *Science* 306(5705):2255-7, 2004.

Higgins, S. The DCC Curation Lifecycle Model. *The International Journal of Digital Curation*, Issue 1, Volume 3, 2008.

Houghton, J*. et al.*, Economic Implications of Alternative Scholarly Publishing Models: Exploring the cost and benefits. *JISC Publication*, 2009 at http://www.jisc.ac.uk/publications/publications/economicpublishingmodelsfinalreport.aspx

Kaufman, M.H. The Atlas of Mouse. *Development.Academic Press*, 1992.

Leonelli, S. Circulating Evidence Across Research Contexts: The Locality of Data and Claims in Model Organism Research. Working Paper 25/08, Nature of Evidence: How Well Do 'Facts' Travel?, Department of Economic History, London School of Economics and Political Science, 2008.

Marx, V. EBI-Led Consortium to Study How to Turn EU Bio-Databases Into Bioinformatic Network. *GenomeWeb*, June 06, 2008.

Smith, C.M., Finger, J.H., Hayamizu, T.F., McCright, I.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., Ringwald, M. The mouse Gene Expression Database (GXD): 2007 update. *Nucleic Acids Res.,* 35, Database issue D618-23, 2007.

Theiler, K. The House Mouse Atlas of Embryonic Development. *Springer-Verlag* New York Inc, 1989.

Venkataraman, S., Stevenson, P., Yang, Y., Richardson, L., Burton, N., Perry, T., Smith, P., Baldock, R.A., Davidson, D.R., Christiansen, J.H. EMAGE - Edinburgh Mouse Atlas of Gene Expression: 2008 update. *Nucleic Acids Res.,* 36:D860-5, 2008.

Wilkinson, M. Controlling the quality of protein datasets. *DrugResearcher*, August 2008.

Wilkinson, M. Raising the standard of proteomics data. *DrugResearcher*, August 2007.

Appendices

7.1 APPENDIX 1. Glossary of Terms

Term	Definition
2D	Two dimensional
3D	Three dimensional
ABA	Ascidian Body Atlas
Annotation	The manual, partial automation or automation of text and image/spatial data from a published source
Antibody	Proteins that are found in the blood and are used by the immune system to identify and neutralise foreign objects
Antiserum	Blood serum containing polyclonal antibodies
BBSRC	Biotechnology and Biological Sciences Research Council
BGED	Brain Gene Expression Database
CASIMIR	Coordination and Sustainability of International Mouse Informatics Resources
CGED	Cancer Gene Expression Database
Curation	The manual processing of structuring, formatting and correcting annotated data
DCC	Digital Curation Centre
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
ELIXIR	European Life Sciences Infrastructure for Biological Information
EMAGE	Edinburgh Mouse Atlas Gene-Expression Database
EMAP	Edinburgh Mouse Atlas Project
EMBL	European Molecular Biology Laboratory
ETL	Extraction, transform, load
EU	European Union
EUCOMM	European Conditional Mouse Mutagenesis
EuReGene	European Renal Genome Project
FTP	File Transfer Protocol
GEISHA	*Gallus* Expression In Situ Hybridization Analysis
GELI	Gene Expression Literature Index
GENSAT	Gene Expression Nervous System ATlas
GUDMAP	GenitoUrinary Development Molecular Anatomy Project
GXD	MGI-Mouse Gene Expression Database
HGU	Human Genetics Unit
IP	Intellectual Property
ISB	International Society for Biocuration
ISH	*In situ* hybridisation
JISC	Joint Information Systems Committee
KOMP	Knockout Mouse Project
Metadata	Definitional data that provides information on structure, context and meaning of raw data managed within an application
MeSH	Medical Subject Headings
MGI	Mouse Genome informatics
MISFISHIE	Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments
MRB	Mouse Resource Browser
MRC	Medical Research Council

Term	Definition
MGED	Microarray Gene Expression Data
MGEIR	Mouse Gene Expression Information Resource
mRNA	Messenger RNA
NCBI	National Center for Biotechnology information
NIDCR	National Institute of Dental and Craniofacial Research
NIH	National Institutes of Health
OBO	Open Biomedical Ontologies
OMIA	Online Mendelian Inheritance in Animals
OMIM	Online Mendelian in heritance in Man
OPT	Optical Projection Tomography
Probe	A labelled or tagged segment of DNA or RNA that can be used to identify a corresponding gene or sequence of interest
RNA	Ribonucleic acid
RT-PCR	Reverse transcription polymerase chain reaction
SCARP	Sharing, Curation, Re-use and Preservation
SQL	Structured Query language
TS	Theiler Stage
WM	Whole-mount
Unique identifier	A unique label given to a data item so that the origin source of the item can be traced and there can be no confusion between items
VTK	Visualization Toolkit
XML	eXtensible Markup Language

7.2 APPENDIX 2. Websites Viewed

Website name	URL
Alexander Fleming Biomedical Sciences Research Center	www.fleming.gr/
Allen Institute for Brain Science	www.brain-map.org/
AMIRA	www.amiravis.com/
Axiope	www.axiope.com/
BSRC Fleming's BioIT Unit	bioit.fleming.gr/
Coordination and Sustainability or International Mouse Informatics Resources	www.casimir.org.uk/
Dana-Farber Cancer Institute	www.dana-farber.org/
Developmental Gene Expression Map	www.dgemap.org/
Digital Curation Centre	www.dcc.ac.uk/
Edinburgh Mouse Atlas Project Homepage	genex.hgu.mrc.ac.uk/intro.html
Edinburgh Mouse Atlas Gene-Expression Database	genex.hgu.mrc.ac.uk/Emage/database/emageIntro.html
EuReGene	www.euregene.org/
EURExpress	www.eurexpress.org/
European Life-science Infrastructure for Biological Information	www.elixir-europe.org
GenitoUrinary Development Molecular Anatomy Project	www.gudmap.org/
IUPHAR receptor database	www.iuphar-db.org/
MGED Society	www.mged.org/
MGI-Mouse Gene Expression Database	www.informatics.jax.org/
Mouse Gene Expression at the BC Cancer Agency	www.mouseatlas.org/
Mouse Resource Browser	bioit.fleming.gr/mrb/welcome.jsp
National Center for Biotechnology Information	www.ncbi.nlm.nih.gov/
National Research Institute for Child Health and Development, Japan	www.nch.go.jp/TOP/indexE.htm
National Institute of Dental and Craniofacial Research	www.nidcr.nih.gov/
The European Molecular Biology Laboratory	www.embl.de/
The Open Biomedical Ontologies	www.obofoundry.org/
The Jackson Laboratory	research.jax.org/
UniProt	www.uniprot.org/
Visualization Toolkit	www.vtk.org/
Wellcome Trust Sanger Institute	www.sanger.ac.uk/

7.3 APPENDIX 3. EMAGE Documentation

Note: Documentation (internal and external) used by the EMAGE research group to describe their processes and product. A list of the publications from the Edinburgh Mouse Atlas Project (EMAP) and collaborative projects can be found at http://genex.hgu.mrc.ac.uk/Papers/intro.html

Document Name	Reference
EMAGE website (in development)	http://www.emouseatlas.org/testemage/
Most recent EMAGE publications	http://www.emouseatlas.org/testemage/about/publications.html
A link to an abstract and zipped Microsoft PowerPoint presentation about EMAGE curation from the 2007 Biocurator Meeting	http://genome-www.stanford.edu/biocurator/IBCM2007/abshtml/23.html
A Microsoft PowerPoint presentation about EMAGE curation from a 2008 DCC Curation Workshop	http://www.emouseatlas.org/testemage/temp/EMAGE_eSci_DCC_short.ppt
Previous Scientific Advisory board reports	http://www.emouseatlas.org/testemage/about/about_EMAGE.html#Ad_Board
Information on MISFISHIE	http://www.emouseatlas.org/testemage/info/misfishie.html
Edinburgh Mouse Atlas Wiki (requires authorisation)	http://aberlour.hgu.mrc.ac.uk/Twiki/bin/view/TWiki/WelcomeGuest
SOPs (requires authorisation)	Internal website or Wiki
Mouse Resource Browser, lists technical information related to the maturity of the EMAGE database	http://bioit.fleming.gr/mrb/Controller?workflow=ViewModel&eid=18

7.4 APPENDIX 4. Additional EMAP Staff

Staff funded by MRC-HGU Core Scientific Services involved in EMAP 3D embryo model development include:	
Name	**Position**
Allyson Ross	EMAP 3D embryo model development (histology)
Julie Moss	EMAP 3D embryo model development (specimen collection, OPT imaging)
Staff in Dr. Duncan Davidson and Prof. Richard Baldock's research groups funded by external grants include:	
Name	**Position (funding source)**
Derek Houghton	Database development GUDMAP (NIH)
Xingjun Pi	Database development GUDMAP (NIH)
Mehran Sharghi	Database development GUDMAP (NIH)
Ying Cheng	Database development GUDMAP (NIH)
Zsolt Husz	Imaging research (NIH)
Mike Wicks	Database development eCHICKATLAS (BBSRC)
Lalit Kumar	Database development EURExpress (EU)
Mei Sze Lam	Database development EURExpress (EU)
Staff in Dr. David FitzPatrick's research group funded by external grants include:	
Name	**Position (funding source)**
Dr. Malcolm Fisher	FaceBase curator, analyst (NIH), based in the EMAGE Editorial Office and line managed by Dr. Jeff Christiansen

7.5 APPENDIX 5. EMAGE Editorial Team Questionnaire

Note: None of the responses were included in the report unless permission was obtained.

	Question	Response
1	What are your roles (or role) in the project?	
2	Is your role (or roles) in the project well defined?	
3	What are your main challenges?	
4	Who do you believe the key stakeholders for the EMAGE database are?	
5	Who do you believe the main users of the EMAGE database are?	
6	How would you improve the EMAGE database curation process?	
7	How would you improve communication with collaborators?	
8	How would you rate the management of the project?	(Good or Poor and please describe some of the factors that support your decision)
9	How would you rate communication within the team?	(Good or Poor and please describe some of the factors that support your decision)
10a	What is your assessment of the quality of EMAGE data for text curation?	(High or Low and please describe some of the factors that support your decision)
10b	What is your assessment of the quality of EMAGE data for spatial curation?	(High or Low and please describe some of the factors that support your decision)

7.6 APPENDIX 6. Journals Listed on the X-Axis of Figure 2

Note: The *bracketed letters* after journal names indicate the status of image reproduction rights for EMAGE for each journal at the time this case study was conducted. *(a)* = prior agreement, *(b)* = individual agreement obtained, *(c)*= Creative Commons Attribution License, *(x)* = CSHL press – refused use , no letter = journal not contacted.

Journal name	Image count	Journal name	Image count
Development *(a)*	521	Nat Neurosci	5
Dev Biol *(a)*	435	Stem Cells	5
Nature	330	Cytogenet Genome Res	4
Mech Dev *(a)*	309	EMBO Rep	4
Dev Dyn *(b)*	220	Exp Mol Med	4
Genes Dev *(x)*	176	J Cardiovasc Electrophysiol	4
Mol Cell Biol *(b)*	138	J Immunol	4
Proc Natl Acad Sci U S A *(b)*	107	J Neuropathol Exp Neurol	4
Gene Expr Patterns *(a)*	97	Natl Inst Genet Jpn Annual Report	4
Biochem Biophys Res Commun *(b)*	87	Neurosci Lett	4
J Biol Chem *(b)*	73	Biofactors	3
Cell *(b)*	72	Br J Haematol	3
Nat Genet	70	Cells Tissues Organs	3
Dev Cell *(b)*	65	Endocrinology	3
Neuron *(b)*	60	J Histochem Cytochem	3
Nat Methods	49	J Med Genet	3
Genesis *(b)*	48	Mol Reprod Dev	3
J Neurosci *(b)*	46	Mol Vis	3
Science	45	Nat Med	3
Hum Mol Genet	41	Transgenic Res	3
EMBO J	40	Am J Physiol Cell Physiol	2
Gene	40	Am J Physiol Heart Circ Physiol	2
Genomics	32	Angiogenesis	2
Blood	31	Biomarkers	2
Genome Res	31	Brain Res Gene Expr Patterns	2
BMC Dev Biol *(c)*	30	Cancer Res	2
Int J Dev Biol *(b)*	28	Cell Biol Int	2
Dev Genes Evol	27	Cold Spring Harb Symp Quant Biol *(x)*	2
Oncogene	21	Diabetes	2
Differentiation	20	Eur J Hum Genet	2
FEBS Lett	20	Evol Dev	2
Mamm Genome	20	Int Immunol	2
Mol Cell Endocrinol	20	J Anat	2
Dev Genet	18	J Bioenerg Biomembr	2
Mol Cell Neurosci	17	J Cell Biochem	2
Biochim Biophys Acta	16	J Neurosci Res	2
Brain Res Dev Brain Res	16	Nat Biotechnol	2
J Am Soc Nephrol	16	Nat Cell Biol	2
Anat Embryol (Berl)	15	Nat Immunol	2
Eur J Neurosci	15	Pediatr Res	2

Journal name	Image count	Journal name	Image count
J Clin Invest	15	Pharmacogenomics J	2
Curr Biol	14	Reproduction	2
J Cell Biol	13	Toxicol Pathol	2
Mol Genet Metab	13	Traffic	2
Mol Hum Reprod	13	J Neurobiol	1
Cardiovasc Res	12	Am J Hum Genet	1
Physiol Genomics	12	Am J Physiol Lung Cell Mol Physiol	1
Cell Tissue Res	11	Am J Physiol Renal Physiol	1
Exp Cell Res	11	Anat Rec	1
Immunity	11	Ann N Y Acad Sci	1
Brain Res	10	Biomed Pept Proteins Nucleic Acids	1
Brain Res Mol Brain Res	10	BMC Cell Biol *(c)*	1
Gastroenterology	10	BMC Genomics *(c)*	1
J Bone Miner Res	10	Br J Dermatol	1
J Comp Neurol	10	Congenit Anom Kyoto	1
Mol Pharmacol	10	Connect Tissue Res	1
Nucleic Acids Res *(c)*	10	Cytogenet Cell Genet	1
Circ Res	9	Eur J Biochem	1
Genes Cells	9	Exp Eye Res	1
PLoS ONE *(c)*	9	FASEB J	1
Anat Rec A Discov Mol Cell Evol Biol	8	Gut	1
Invest Ophthalmol Vis Sci	8	J Dermatol Sci	1
J Mol Biol	8	J Exp Zoolog B Mol Dev Evol	1
J Mol Cell Cardiol	8	J Leukoc Biol	1
Lab Invest	8	J Lipid Res	1
Mol Endocrinol	8	J Med Sci	1
PLoS Biol	8	J Mol Neurosci	1
Am J Pathol	7	J Neuropathol Exp Neurol	1
Dev Growth Differ	7	J Toxicol Environ Health A	1
DNA Cell Biol	7	Kidney Int	1
Neuroreport	7	Life Sci	1
PLoS Genet *(c)*	7	Mol Carcinog	1
Biochem J	6	Mol Cell	1
Biol Reprod	6	Mol Cells	1
Cell Mol Biol (Noisy-le-grand)	6	Neuroscience	1
J Biochem (Tokyo)	6	Philos Trans R Soc Lond B Biol Sci	1
J Invest Dermatol	6	Reprod Toxicol	1
J Reprod Dev	6	Teratology	1
Matrix Biol	6	Toxicol Lett	1
Genetics	5	Yi Chuan Xue Bao	1
J Cell Sci	5		

7.7 APPENDIX 7. Initial Scope for the Analysis of the Case Study

Stakeholders (input)	Funding (input)
• The researcher and their expertise – data creators, data curators/producers, data-holders, data re-users • Data sources used • Publishers used • Annotation – data, training (internally and externally sourced) • Curation of data • DCC's strategic leadership in digital curation and preservation	• Funding opportunities • Commercialisation opportunities • Software techniques/tools • Curation (digital aspects) • Workflow • Data storage options • Courses for users, training • Develop curation services • Pilot development for recording and monitoring file formats • Evaluation of process/tools • Preservation, planning tools • To assist well established archives • Develop curation services
Data (output)	**Stakeholders (output)**
• Sharing of data • User interface • Documentation (internal and external) – manual, partial automation, automation guidelines for digital curation, promotional materials • Normalisation of imaging • Who's viewing the data globally • Who's downloading the data • What are people doing with the data • External reports • Expert advice and guidance • Collaborative networks – relationships with key players, gain recognition, greater exposure in the broader community, links to other sciences, non-science projects • Support services • Audit and certification – standards and practice	• Who • Where • Global collaborations • Cost • Quality – accountability and efficiency • Analysis, assessment of data • Preservation – maintaining value, research lifecycle • Value of resources – expertise, process development • Testing of techniques/tools developed • Raise awareness of curation issues • Review other e-sciences strategies that will impact the digital curation process