



# AUDIT AND REQUIREMENTS ANALYSIS FINDINGS

---

## EMBEDDING INSTITUTIONAL DATA CURATION SERVICES IN RESEARCH (EIDCSR)

[eidcsr.oucs.ox.ac.uk](http://eidcsr.oucs.ox.ac.uk)

**Author**

Luis Martinez-Uribe ([luis.martinez-uribe@oerc.ox.ac.uk](mailto:luis.martinez-uribe@oerc.ox.ac.uk))  
Project Manager and Analyst

**Affiliation**

Oxford e-Research Centre and Computing Services

A collaborative project between

OFFICE OF THE DIRECTOR OF IT  
Enabling Oxford University to make optimal use of IT



Oxford University Library Services

**Research Services** Computational Biology Group The Cardiac Mechano-Electric Feedback Group

Funded by





Project Information			
<b>Project Acronym</b>	EIDCSR		
<b>Project Title</b>	Embedding Institutional Data Curation Services in Research		
<b>Start Date</b>	01/04/09	<b>End Date</b>	30/09/10
<b>Lead Institution</b>	University of Oxford		
<b>Project Director</b>	Dr. Mike Fraser		
<b>Project Manager &amp; contact details</b>	Luis Martinez-Urbe Oxford e-Research Centre 7 Keble Road, Oxford OX1 3QG		
<b>Partner Institutions</b>			
<b>Project Web URL</b>	<a href="http://eidcsr.oucs.ox.ac.uk">http://eidcsr.oucs.ox.ac.uk</a>		
<b>Programme Name (and number)</b>	JISC Information Environment 2009-11 12/08		
<b>Programme Manager</b>	Neil Grindley		

Document Name			
<b>Document Title</b>	EIDCSR Audit and Requirements Analysis Findings		
<b>Reporting Period</b>			
<b>Author(s) &amp; project role</b>	Luis Martinez-Urbe, Project Manager and Analyst		
<b>Date</b>	15/09/2009	<b>Filename</b>	EIDCSR-AnalysisFindings v2.1.doc
<b>URL</b>			
<b>Access</b>	<input checked="" type="checkbox"/> Project and JISC internal	<input checked="" type="checkbox"/> General dissemination	

Document History		
Version	Date	Comments
1.0	17/7/09	Sections circulated to interviewees
1.1	28/7/09	EIDCSR Analyst first draft presented to EIDCSR Team
2.0	10/9/09	After EIDCSR Team feedback
2.1	14/9/09	Revised by Mike Fraser

## Table of contents

<b>1. Introduction.....</b>	<b>1</b>
<b>2. Methodology.....</b>	<b>1</b>
Scope.....	1
Identification and organization of interviewees.....	1
The interview process.....	1
Data analysis, reporting and quality assessment.....	2
<b>3. Audit and Requirements Analysis Findings.....</b>	<b>2</b>
The research project .....	2
Funding agency policies and requirements.....	3
The research data lifecycle process.....	3
Wet lab.....	4
Magnetic Resonance Imaging .....	5
Segmentation, registration, mesh generation and simulation.....	7
Requirements and needs.....	11
Secure storage.....	11
Data transfer.....	11
Metadata .....	11
<b>Appendix 1. Questionnaire.....</b>	<b>12</b>
<b>Appendix 2. Research Workflow Process.....</b>	<b>14</b>
<b>Appendix 3. Sample methodology descriptions.....</b>	<b>15</b>
<b>Appendix 4. Tools .....</b>	<b>19</b>
<b>Appendix 5. Storage resources.....</b>	<b>21</b>

## 1. Introduction

The Embedding Institutional Data Curation Services in Research (EIDCSR) project is a JISC funded activity aiming to scope and address the data management and curation requirements of two collaborating research groups in Oxford. The project is an institutional collaboration bringing together two research groups – the Cardio Mechano-Electric Feedback Group and the Computational Biology Group – with a range of institutional service providers: the Library and Computing services, the Research Services Office, the Oxford e-Research Centre, and the Office of the Director of IT. In addition, IBM contributes consultancy time to the project.

The initial stage of the EIDCSR project involved undertaking an audit of data assets and data management practices as well as a requirements analysis of the two research groups based on the Data Audit Framework (DAF) Methodology and building on previous DAF work conducted in Oxford. This report presents the results of this analysis exercise and it is intended to feed in to other project activities including identification of adequate metadata standards, the development of a workflow module and cost models.

The report is organized as follows: the methodology used for this exercise is presented describing the scope, the organization of the interviews and desk research and how the data were analysed. After this, the findings of the audit and requirements analysis are reported by describing the research project, the requirements from the funding agency to then describe their processes and research data generated at each stage as well as the tools and storage resources used. Finally, researchers' requirements for services and tools to support their data management activities are presented.

## 2. Methodology

### Scope

The interviews aimed to audit data assets as well as to capture and document data management practices in order to undertake an analysis of researchers requirements for access and preservation of their data. The focus was on researchers of the two research groups in Oxford, the Cardiac Mechano-Electric Feedback Group (CMEFG) and the Computational Biology Group (CBG), and the data they produce during their research activities as part of a BBSRC funded project named "*Technologies for 3D histologically-detailed reconstruction of individual whole hearts*".

The information gathered from the interviews was complemented with other documentation such as the BBSRC project proposal, research articles and presentations produced by researchers participating in the project.

### Identification and organization of interviewees

A choice of candidates was originally guided by suggestions from the EIDCSR Co-PIs from within the research departments. For each researcher identified, the *friend of a friend* approach or snowball sampling was used to identify further candidates to be interviewed. The interviewees identified were emailed individually providing them with the project brief and asking them to take part on the study. If they agreed, the interview questions in Appendix 1 were circulated and time and place were arranged for the interview to take place.

### The interview process

The EIDSCR Analyst conducted eleven interviews, these took no longer than one hour although due to the agile approach adopted for requirements elicitation further iterations may be required. The interview itself started with a

brief introduction to the project and a reminder of the iterative nature of the interview process. Researchers were also informed of the intention to record the interviews, with permission, and take notes. They were also asked to sign a consent form. During the interview, the Project Analyst attempted to understand the data that is produced during the research process, how these are used, the approach taken to manage these and the requirements for managing their data more effectively. A de-brief served to ask the interviewee about how they liked the interview framework, the benefits of participating and if they know someone who might be a good interview candidate.

### **Data analysis, reporting and quality assessment**

After each interview, the Analyst produced a short summary and completed a data register form based on the DAF Methodology with specific information about the datasets. The recorded interviews were transcribed using the services of a professional audio transcription company. The interview transcripts and other literature including research papers were colour coded to organize and group the information relevant to the following themes and categories to facilitate searching, making comparisons and identify patterns in the interviews:

- Types of data including information about formats, sizes, ownership, etc.
- Storage resources used at different stages.
- Metadata compiled about the data resources created
- Tools for data creation, manipulation and visualization
- Requirements for services and tools

Once the findings report was completed it was circulated to the researchers participating in the interviews, the EIDCSR Project Working Group, Executive Board and JISC Programme Manager for feedback and sign-off.

## **3. Audit and Requirements Analysis Findings**

This section presents the findings from the interviews and desk research. It starts by describing the BBSRC research project with the policy and requirements from this funding agency. After this, the research process involving different research groups producing, sharing and re-purposing data is presented with special focus on the types of data and metadata, storage resources, tools used. Finally, the top requirements for services and tools to help researchers manage their data are presented.

This multidisciplinary research collaboration between distinct and diverse research groups in Oxford in the areas of Life Sciences, Medical Physics, Image Analysis and Computing represents an extraordinary exemplar of data generation, sharing and reuse where a range of complimentary skills and knowledge are brought together to the benefit of all groups involved.

The research groups' data management practices are evolving to deal with the emerging challenges they face due to the creation of large amounts of valuable and heterogeneous data that need to be securely stored and shared, readily accessible and preserved over the long-term for future reuse.

### **The research project**

The Computational Biology Group (CBG) and the Cardiac Mechano-Electric Feedback Group (CMEFG) together with the Department of Cardiovascular Medicine (DCM) gained funds from the Biotechnology and Biological Science Research Council (BBSRC) for a three-year project starting in January 2007 named "*Technologies for 3D histologically-detailed reconstruction of individual whole hearts*". The case for support explained how after decades of research into ventricular tissue architecture there is still controversy around basic issues. This, it is argued, may

be caused by the fact that traditional histological techniques, including tissue preparation and sectioning, are not only time consuming but 'destructive' by nature. Thus the study proposed to use novel imaging techniques like Magnetic Resonance Imaging (MRI) and Diffusion Tensor MRI (DTMRI) considered to be "non-destructive" and combine them with traditional histological techniques as well as with image processing with data registration and computational models for bio-mathematical simulation.

### **Funding agency policies and requirements**

As explained in a recent Digital Curation Centre (DCC) report<sup>1</sup>, since 2007 BBSRC has had a statement on access to published research outputs and a Data Sharing Policy where they recognize and support the international efforts in data sharing. Main requirements include:

- Applicants must submit a statement on data sharing to be assessed by reviewers. This should include concise plans for data management and sharing or provide explicit reasons why data sharing is not possible or appropriate.
- Researchers should make use of current best practice and generate data and documentation using widely accepted formats, methodologies and standards.
- Data should be accompanied by contextual information (documentation / metadata) to provide a secondary user with any necessary details on the origin or manipulation of the data in order to prevent any misuse, misinterpretation or confusion.
- Data should be made available through existing community resources or databases where possible.
- BBSRC expects research data to be made available for subsequent research with as few restrictions as possible in a timely and responsible manner. Timely release would generally be no later than publication of the main findings and should be in-line with established best practice or within three years if no best practice exists.
- Data must be kept securely in paper or electronic form for a period of ten years after the completion of a research project.
- Researchers are expected to ensure appropriate data management strategies are in place throughout the research project.
- Institutions receiving BBSRC funding must have guidelines setting out responsibilities and procedures for keeping data.

### **The research data lifecycle process**

The research process of these research groups has been described in Gemot et. al. in 2009<sup>2</sup>, see Appendix 2 and figure 1 and it has been mapped to a spiral lifecycle inspired by the DCC Lifecycle Model<sup>3</sup>. It starts with the generation of complementary images stacks that are then processed in different ways to generate meshes that can be used for computational modelling of the heart.

---

<sup>1</sup> Jones, Sarah (2009) "A report on the range of policies required for and related to digital curation"  
[www.dcc.ac.uk/docs/reports/DCC\\_Curation\\_Policies\\_Report.pdf](http://www.dcc.ac.uk/docs/reports/DCC_Curation_Policies_Report.pdf)

<sup>2</sup> Gemot Plank et al. (2009) "Generation of histo-anatomically representative models of the individual heart: tools and application"  
Phil Trans R Soc A 2009 367: 2257-2292.

<sup>3</sup> Higgins, Sarah (2008) "The DCC Lifecycle Model" International Journal of Digital Curation, Vol 3, No 1  
<http://www.ijdc.net/index.php/ijdc/article/view/69>

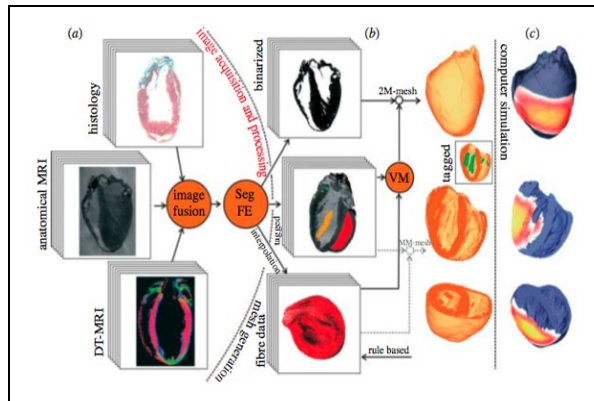


Figure 1. The research process by Gernot et. al 2009

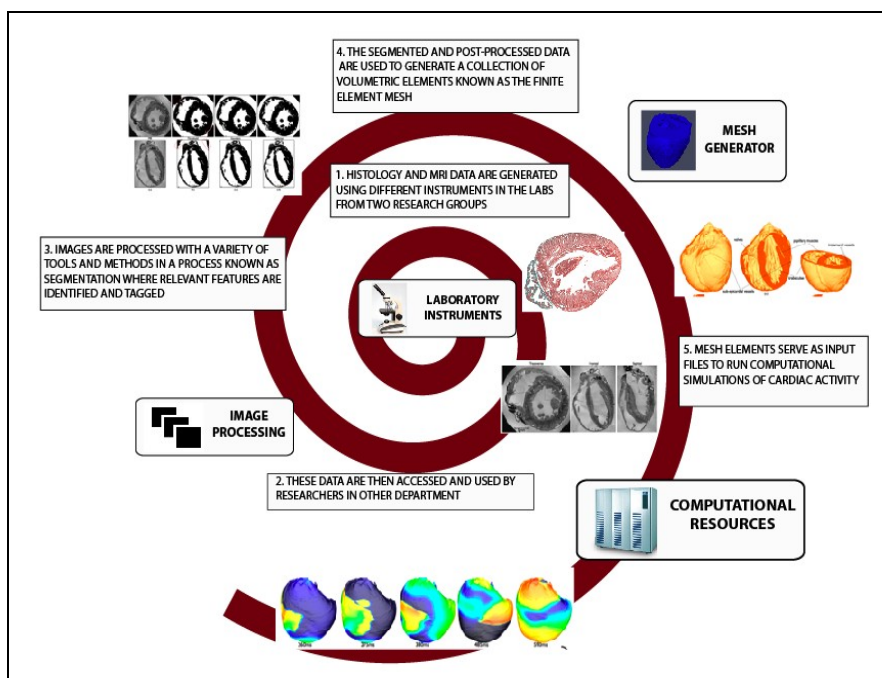
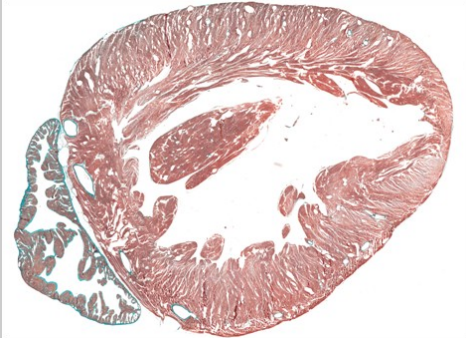


Figure 2. The research process based on the DCC Lifecycle Model

### Wet lab

The research process starts with the isolation of rat hearts which undergo fixation in different physiological states (see detailed process explained in Appendix 3). The whole process beginning from isolation, MRI (DTMRI and anatomical MRI) to histology requires about 6 weeks (per heart). The imaging of these sections takes a further 4 weeks. The hearts, of 2x3 cm in dimension, are embedded in black wax. Prior to cutting each section (at 10  $\mu\text{m}$  thickness), the surface of the wax block is imaged using a stereo-microscope fitted with polarizers to obtain low-resolution pre-section images. The whole heart is then stained and imaged using a technique by which a microscope moves along and takes 300-350 pictures and sticks them together to get a montage image.

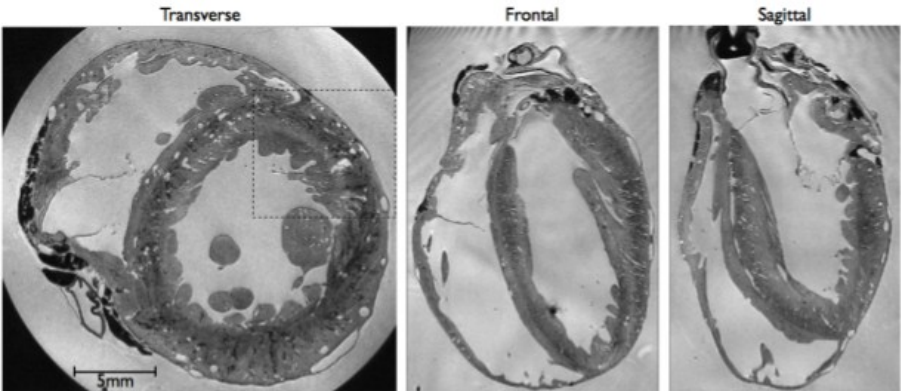
## Histology data

<b>Description</b>	Images generated by microscopes representing a whole heart or two-dimensional sections of a heart
<b>Type</b>	Experimental
<b>Creator</b>	Rebecca Burton, Cardio Mechano-Electric Feedback Group
<b>Format</b>	BMP format
<b>Size</b>	1TB per heart
<b>Current amounts</b>	4TB
<b>Forecast amounts</b>	10TB
<b>Source of data</b>	Treated guinea pig hearts
<b>Storage</b>	The data is initially stored on the lab computer and moved onto a NAS system at DPAG whenever the lab computer's hard disk is full. The transfer is done directly between the lab computer and a NAS system via an Ethernet cable See appendix 5
<b>Back up policy</b>	The NAS systems are configured as RAID 5 systems. Some of the data is also stored on DVDs and at one of our collaborators' institution (Auckland, New Zealand).
<b>Metadata</b>	The way hearts are treated, sectioned and imaged is recorded on printed lab notebooks. Some of this information is also stored on the NAS systems as ReadMe files. See appendix 3
<b>Responsible of data management</b>	Alan Garry looks after the NAS system
<b>Tools and Instruments</b>	Leica MZ 95 stereomicroscope to image low-resolution "in the wax" images of the heart; Leica SM2400 heavy duty sledge-type microtome to section block; Leica AutoStainer XL, ST50-10 to stain sections; Microscopes Leica QWin and Leica QGO software for imaging and BMP Viewer developed internally to visualize the images See appendix 4
<b>Used by</b>	Researchers at the Computational Biology Group
<b>Lifespan of the data</b>	10 years or more
<b>To be made publicly available</b>	Yes
<b>Requirements</b>	Secure storage Provenance metadata with information about the experiment Fast data transfer for access, sharing and publication
<b>Sample</b>	 <p style="text-align: center;">Histology data representing a stained heart section.</p>

## Magnetic Resonance Imaging

After the hearts have been treated in the Department of Physiology, Anatomy and Genetics they are sent to the Department of Cardiovascular Medicine to be scanned. Three types of data are then produced; the raw data generated by the magnets and two forms of derived data. The raw data, held in plain-text files, provides information

about the voltage decays. These are then transformed by processing the data to produce standard anatomical MRIs (tiff images) and using mathematical routines for extraction of useful information to generate a second type of MRI known as diffusion tensor MRI (DTMRI) that contains more information specially about fibre orientation. The 2D MRIs are collated to produce a 3D MRI-based dataset that can then be sectioned at any angle.

<b>Anatomical and Diffusion Tensor MRI data</b>	
<b>Description</b>	Two types of Magnetic Resonance Imaging (MRI): Anatomical and Diffusion Tensor MRIs generated from raw data produced by the magnet.
<b>Type</b>	Experimental
<b>Creator</b>	Dr. Patrick Hales, Department of Cardiovascular Medicine
<b>Format</b>	Raw data are plain text files with numbers; Derived data are a stack of tiff images
<b>Size</b>	Up to 2Gb per heart
<b>Current amounts</b>	
<b>Forecast amounts</b>	
<b>Source of data</b>	Treated guinea pig hearts
<b>Storage</b>	Initially stored on lab computers, then creator's computer for production of derived data. Raw stored on creator's computer and derived upload to NAS system. See appendix 5
<b>Back up policy</b>	Raw data back-up on DVDs and derived data on NAS system
<b>Metadata</b>	A file with the parameters set when running the scan is produced which contains the values of the parameter, date of scan, ID of sample, etc... If data are published some more information would need to be made available like the published methods (already known in the field) used for processing the data.
<b>Responsible of data management</b>	Creator for raw data, the derived data is on NAS system
<b>Tools and Instruments</b>	Varian Magnet A 11.7 T (500 MHz) MR system, consisting of : - a vertical magnet (bore size 123 mm; Magnex Scientific, Oxon UK), - a Bruker Avance console (Bruker Medical, Ettlingen, Germany), and - a shielded gradient system (548 mT/m, rise time 160 $\mu$ s; Magnex Scientific, Oxon, UK). Quadrature driven birdcage coils with an inner diameter of 28 mm and 40 mm (Rapid Biomedical, Wurzburg, Germany) used to transmit/receive the MR signals. See appendix 4.
<b>Used by</b>	Researchers in Computational Biology Group
<b>Lifespan of the data</b>	5 years or more
<b>To be made publicly available</b>	Yes
<b>Requirements</b>	Secure storage for both the raw and the derived data Provenance metadata with information about the experiment
<b>Sample</b>	<div style="display: flex; justify-content: space-around; text-align: center;"> <div>Transverse</div> <div>Frontal</div> <div>Sagittal</div> </div> 

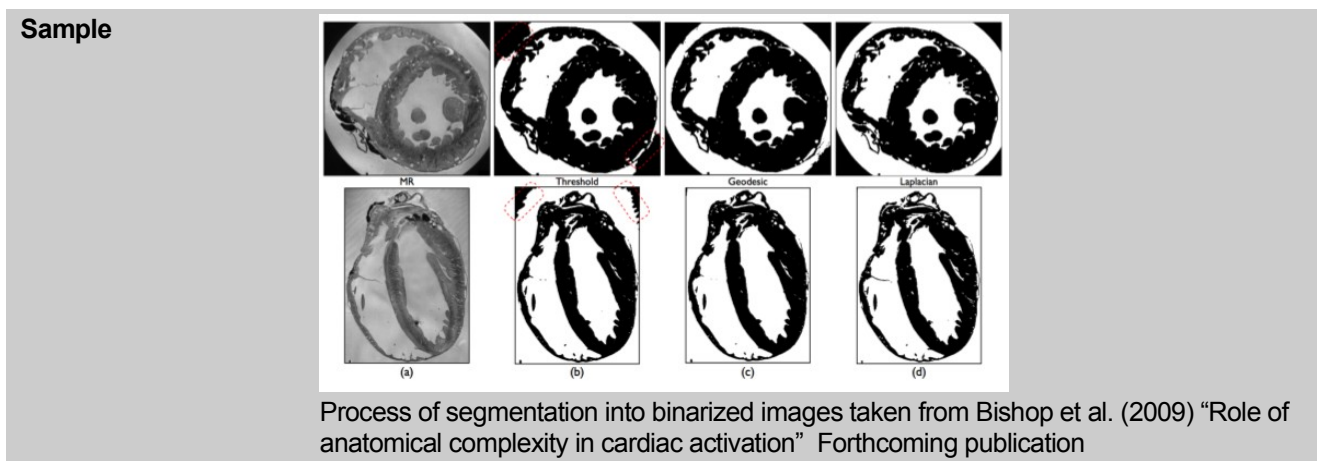
MRI images taken from Bishop et al. (2009) "Role of anatomical complexity in cardiac activation" Forthcoming publication.

**Segmentation, registration, mesh generation and simulation**

Once the histology and MRI data have been created and uploaded to the NAS system, researchers from the Computational Biology Group access them for a next stage of the data on the pipeline. These data go through the process of segmentation and post-processing to identify and classify the information found in the data.

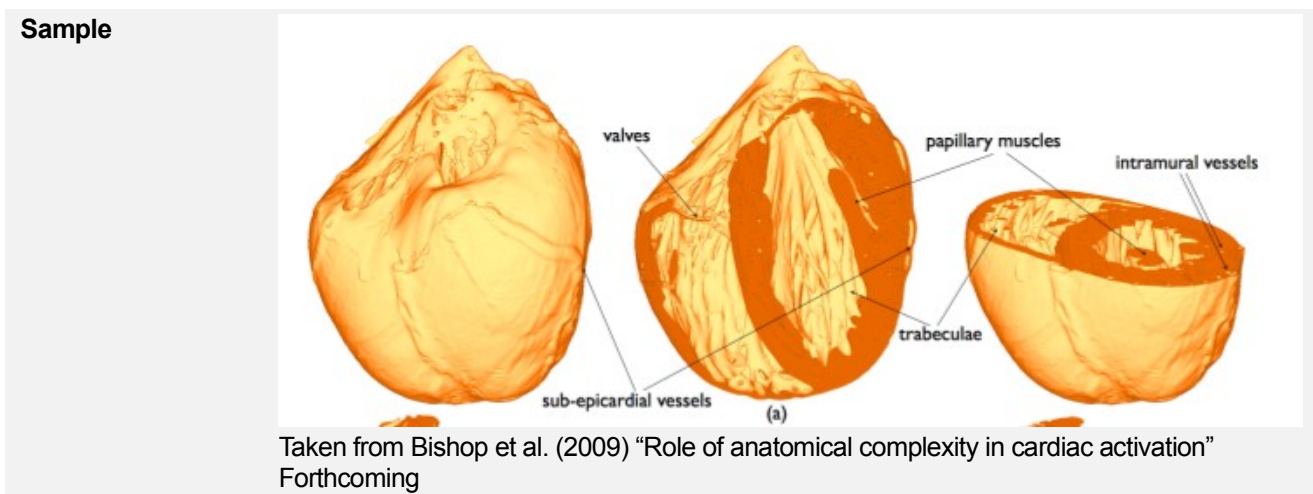
One way in which this happens involves segmenting the stained histology data using colour thresholding and the anatomical MRI data applying grey level thresholds to discriminate different tissues types in each pixel, in some case some image processing is required. Then these datasets are tagged to identify different relevant features and this in some cases requires certain manual interaction.

<b>Segmentation data</b>	
<b>Description</b>	Data resulting from applying segmentation to the histology and MRI data sets.
<b>Type</b>	Derived from experimental data
<b>Creator</b>	Various researchers at the Computational Biology Group
<b>Format</b>	MetaImage, DICOM, NRRD, VFF
<b>Size</b>	
<b>Current amounts</b>	
<b>Forecast amounts</b>	
<b>Source of data</b>	Histology and MRI data
<b>Storage</b>	Initially stored on the creator's computer, then the successful models are uploaded on the NAS system. See appendix 5
<b>Back up policy</b>	Author uses HFS service for own desktop and HFS own back up strategy.
<b>Metadata</b>	Different segmentation algorithms are applied with different filters and there is also a manual process involved. This information including the parameters values is all recorded as input files used. See appendix 3
<b>Responsible of data management</b>	Creator while work in progress and then uploaded onto NAS system
<b>Tools and Instruments</b>	Insight Toolkit (ITK) for the 3D image segmentation Seg 3D with ITK functions for manual interaction including landmark selection, labelling or air bubble removal Matlab scripts and functions for a variety of tasks including image down-sampling See appendix 4
<b>Used by</b>	Researcher that creates them to generate mesh
<b>Lifespan of the data</b>	Around 5 years, at which point it is likely a better model has been developed from more complete and accurate raw data
<b>To be made publicly available</b>	Yes with mesh resulting from it and MRI and histology sources
<b>Requirements</b>	Secure storage Linkage to MRI and histology data



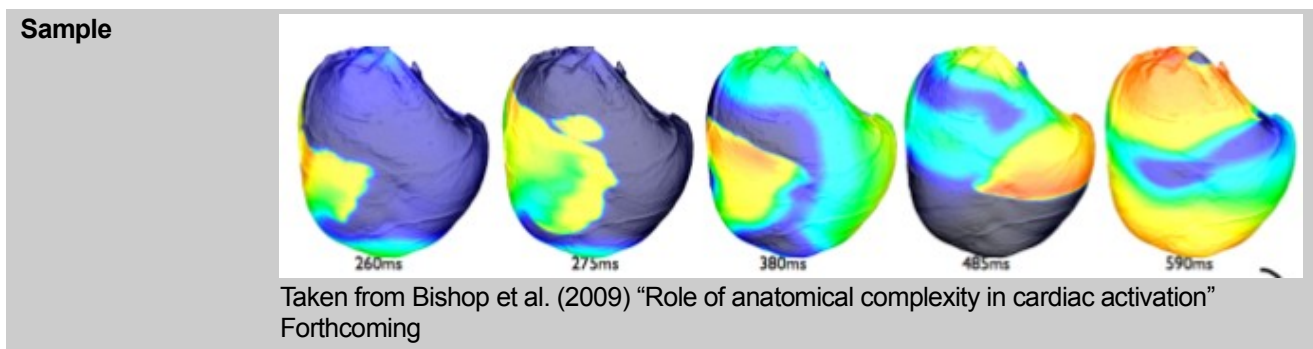
The final segmented and post-processed data are then used to generate a tetrahedral finite element mesh, a collection of volumetric individual elements, that are treated as myocardial element or non-myocardial element, which contain fibre orientation information and allow for electrical conduction in the myocardium to be described.

<b>Mesh data</b>	
<b>Description</b>	Data generated from segmented data using a mesh generator
<b>Type</b>	Derived from segmented research data a tetrahedral finite element mesh is generated
<b>Author</b>	Dr. Martin Bishop and others in the Computational Biology Group
<b>Format</b>	Three types of ASCII files (elements, notes and fibre)
<b>Size</b>	In the order of few MBs per mesh
<b>Current amounts</b>	In the order of few GBs
<b>Forecast amounts</b>	In the order of few GBs
<b>Source of data</b>	Segmented histology and MRI data
<b>Storage</b>	Initially stored on the creators' computer, then the successful experiments are uploaded on the NAS system. See appendix 5
<b>Back up policy</b>	Author uses HFS service for own desktop and HFS own back up strategy.
<b>Metadata</b>	The information about the mesh generation i.e. different parameter values is recorded as the input files. See appendix 3
<b>Responsible of data management</b>	Creator while work in progress and then uploaded onto NAS system
<b>Tools and Instruments</b>	<b>Tarantula</b> to generate the mesh <b>TetGen</b> to generate element meshes of enclosed regions tagged with numerical values (a feature not available in Tarantula) <b>Meshalzyer</b> to visualize the mesh See appendix 4
<b>Used by</b>	Some in Computational Biology with an interest on cardiac modelling. Potentially many others will use them outside of Oxford once they are made available
<b>Lifespan of the data</b>	Around 5 years, at which point it is likely a better model has been developed from more complete and accurate raw data
<b>To be made publicly available</b>	Yes with sources of and MRI , histology and segmentation data
<b>Requirements</b>	Secure storage Linkage to the segmentation data



Finally the mesh elements serve as input files to run computational simulations of cardiac activity.

<b>Simulation data</b>	
<b>Description</b>	Electrophysiological simulations
<b>Type</b>	Simulation
<b>Creators</b>	Many in the Computational Biology Group.
<b>Format</b>	Output files of the simulation
<b>Size</b>	
<b>Current amounts</b>	
<b>Forecast amounts</b>	
<b>Source of data</b>	Meshes are input data
<b>Storage</b>	Creators' desktops Comlab Heart server See appendix 5
<b>Back up policy</b>	
<b>Metadata</b>	The simulation is the result of using the meshes and other input files that define the models (using CellML) and the parameters used for the simulation. All this information is needed to record provenance.
<b>Responsible of data management</b>	Creators
<b>Tools and Instruments</b>	CHASTE or CARP for running the electrophysiological simulations Oxford Supercomputing Centre, National Grid Service and other such resources in Europe for computational resources to run the simulations See appendix 4
<b>Used by</b>	The analysed simulation may be useful to others but not the simulation output files
<b>Lifespan of the data</b>	
<b>To be made publicly available</b>	Not part of the BBSRC funded project
<b>Requirements</b>	Secure storage Provenance metadata with information about the simulation



The segmentation process can also be used as a step in generating probabilistic three-dimensional heart atlases. These atlases combine the imaged hearts anatomical MRI data to find an average representation of the rat cardiac anatomy, mainly the ventricles, to then study variability between subjects. In order to generate the atlas, the anatomical MRI data are converted to 3D volumes that are cropped to cut off the empty space around the cube where the heart is contained. The volumes are segmented and then go through a process of registration where images are aligned to compare the anatomical landmarks in the heart. The outputs of the registration process are the 3D atlas.

<b>Probabilistic 3D Heart Atlas</b>	
<b>Description</b>	Three dimensional cardiac atlases being an average representation of rat heart ventricles
<b>Type</b>	Derived from anatomical MRI data
<b>Creator</b>	Ramon Casero
<b>Format</b>	Atlas will be a set of 3D images or in a mesh format
<b>Size</b>	
<b>Current amounts</b>	
<b>Forecast amounts</b>	
<b>Source of data</b>	Anatomical MRI data from the NAS system
<b>Storage</b>	Download the MRI data to desktop, stored on personal computer for processing and then upload onto NAS system
<b>Back up policy</b>	
<b>Metadata</b>	The process includes using specific functions in ITK that need to be included as part of the methodology when publishing papers based on data.
<b>Responsible of data management</b>	Creator
<b>Tools and Instruments</b>	<i>Insight Toolkit (ITK)</i> for registration and atlas building <i>Seg3D</i> <i>Slicer 3</i> <i>Tarantula</i>
<b>Used by</b>	The data resulting from generating the 3D volumes and cropping them may be useful to researchers in CMEFG. The atlas can be used to model anatomy through simulations or in clinical settings to compare subjects with the model to identify deviations from healthy models.
<b>Lifespan of the data</b>	
<b>To be made publicly available</b>	Yes
<b>Requirements</b>	Secure storage

## Requirements and needs

The researchers interviewed as part of this audit and requirements analysis exercise were asked about their requirements for services and tools to help them manage and work with their datasets. The results of this can be broadly group into the three main themes presented below: secure storage, data transfer and metadata.

### Secure storage

Most researchers interviewed from the three groups mentioned the need to have access to secure storage. Although secure storage meant different things to the different researchers, there was a general feeling that their data needs to be kept safe. Some of the requirements in this area included:

- Researchers' desktops and project servers, like the NAS system, require to be backed- up regularly and having off-site copies of the data.
- Researchers involved in the image processing, mesh generation and simulations require version control capabilities as well as access controls within the storage environment to keep track of changes in data and being able to restrict access to in-development data and share the final data outputs.

### Data transfer

Some of the data produced as part of this BBSRC project is large in size and researchers interviewed mentioned the necessity to be able to transfer these large amounts of data fast and reliably. More detailed requirements included:

- Fast access to the histology data to work with it, in some cases it is a matter of visualizing part of the histology data to look at it.
- A way to make the data available, to collaborators before the data are ready for publication but also more broadly once the data are published, that allows fast retrieval of the large data files.

### Metadata

The creation, processing and re-purposing of the data involves using a variety of tools and methodologies that are recorded in disparate ways by the researchers using them. A widespread requirement amongst researchers participating in the interviews is a metadata recording and management tool that would allow them to:

- publish articles based on their data and not having to go back to log-books, notes and files to find the information that allows reproducibility;
- access information about experiments and simulations remotely and
- search the data stored on the servers;

## Appendix 1. Questionnaire

The following questionnaire is based on the questionnaires developed for the IBVRE and eIUS projects.<sup>4</sup>

### Introduction

Give brief introduction to the EIDCSR project including overall aim and objectives. Provide an overview of the questions that will follow and remember the interviewee about the iterative nature of the interview, the intention of taking notes, record the interview (with permission), produce transcripts and to publish findings.

### Interview questions

I am interested in learning more about the research tasks that involve some form of data management in its widest sense. I would like to do this by understanding at what stage within the BBSRC funded project “*Technologies for 3D histologically-detailed reconstruction of individual whole hearts*” workflow process, described in section 4, your contribution comes into place and look at it in the context of a generic “*research life-cycle*”, starting from the data acquisition, all the way to publishing to understand to what extent the following elements fit in your average working day.

- a. Could you explain briefly at what stage in the research process shown in section 4 your research takes place describing your area of research and the types of research questions, with examples, that you try to answer?
  
- b. Data acquisition – Could you provide details about the process of acquisition providing details about:
  - the sorts of data (MRIs, histology, meshes) do you generate;
  - the amount of data are you producing and the growth rate;
  - who owns the data that you produce;
  - who will be using these data within this two research groups;
  - the instruments and software used to do this and the formats of the data;
  - and for the data generated by others that you may use, how are these found and accessed?
  
- c. Storage and description of the data – Could you explain what happens when the data have been acquired describing:
  - where are the datasets that you produce stored;
  - the security measures that are taken to preserve the integrity of the data;
  - who is responsible for the management of these datasets;
  - the information about how the data was generated that is collected and how;
  - and how are the files organized?
  
- d. Data manipulation and visualization – Could you describe how are the datasets manipulated and visualized providing information about:
  - the tools/software are used to manipulate and visualize the datasets;
  - and the resulting outputs of this and how are they version or organized?
  
- e. Sharing and publishing – Could you provide details of how are the datasets made available to others including:
  - the tools that are used to made the data available to collaborators or the wider research community;

---

<sup>4</sup>Integrative Biology Virtual Research Environment (IBVRE) Project, <<http://www.vre.ox.ac.uk/ibvre/>>; e-Infrastructure Use Cases and Service Usage Models (eIUS) Project, <<http://www.eius.ac.uk/>>.

- and how are the datasets or derived versions included and cited in published research papers?
- f. What tools or services would help you with your research activities that involve acquiring, storing and describing, manipulating and sharing data ?
- g. Is there anything else that you would like to add?

#### De-Brief

- h. How do you think the interview went?
- i. What are the benefits you believe you get from participating?
- j. Could you suggest anyone you know that could participate in these interviews?

#### Data Asset Register Form

The following data asset register form is based on the Data Audit Framework Methodology, a tool developed by the Digital Curation Centre and the Joint Information Systems Committee to enable research departments to carry out audits of data collections and data management practices helping them to find out what data they hold, where is located and who is responsible for it. The form below has been designed to capture this information. After the interviews the Analyst will fill the form for each data asset that:

- are still being created or added to;
- are used on frequent basis in the course of research work;
- underpin scientific replication e.g. revalidation;
- play a pivotal role in ongoing research;

<b>Dataset Name</b>	<i>Official name</i>
<b>Description</b>	<i>A description of the information contained the data asset and its spatial, temporal or subject coverage</i>
<b>Owner</b>	<i>Name or position</i>
<b>Author</b>	<i>Person, group or organization responsible for the intellectual content of the data asset</i>
<b>Subject</b>	<i>Information and keywords describing the subject matter of the data</i>
<b>Date of Creation</b>	<i>The date on which the data asset was created or published</i>
<b>Purpose</b>	<i>Reason why the asset was created, intended user communities or source of funding / original project title</i>
<b>Source</b>	<i>The source(s) of the information found in the data asset</i>
<b>Updating frequency</b>	<i>The frequency of updates to this dataset to indicate currency</i>
<b>Type</b>	<i>Description of the technical type of the data asset (e.g., database, photo collection, text corpus, etc.)</i>
<b>Formats</b>	<i>Physical formats of data asset, including file format information</i>
<b>Rights and restrictions</b>	<i>Basic indication of the user's rights to view, copy, redistribute or republish all or part of the information held in the data asset. Access restrictions on the data itself or any metadata recording its existence should also be noted</i>
<b>Usage frequency</b>	<i>Estimated frequency of use and if known required speed of retrieval to determine IT infrastructure and storage needs</i>
<b>Relation</b>	<i>Description of relations the data asset has with other data assets and any any DOI ISSN or ISBN references for publications based on this data</i>
<b>Back-up and archival policy</b>	<i>Number of copies of the data asset that are currently stored, frequency of back-up and archiving procedures</i>
<b>Management to date</b>	<i>History of maintenance and integrity of the data asset e.g. edit rights /security, and any curation or preservation activities performed</i>

## Appendix 2. Research Workflow Process

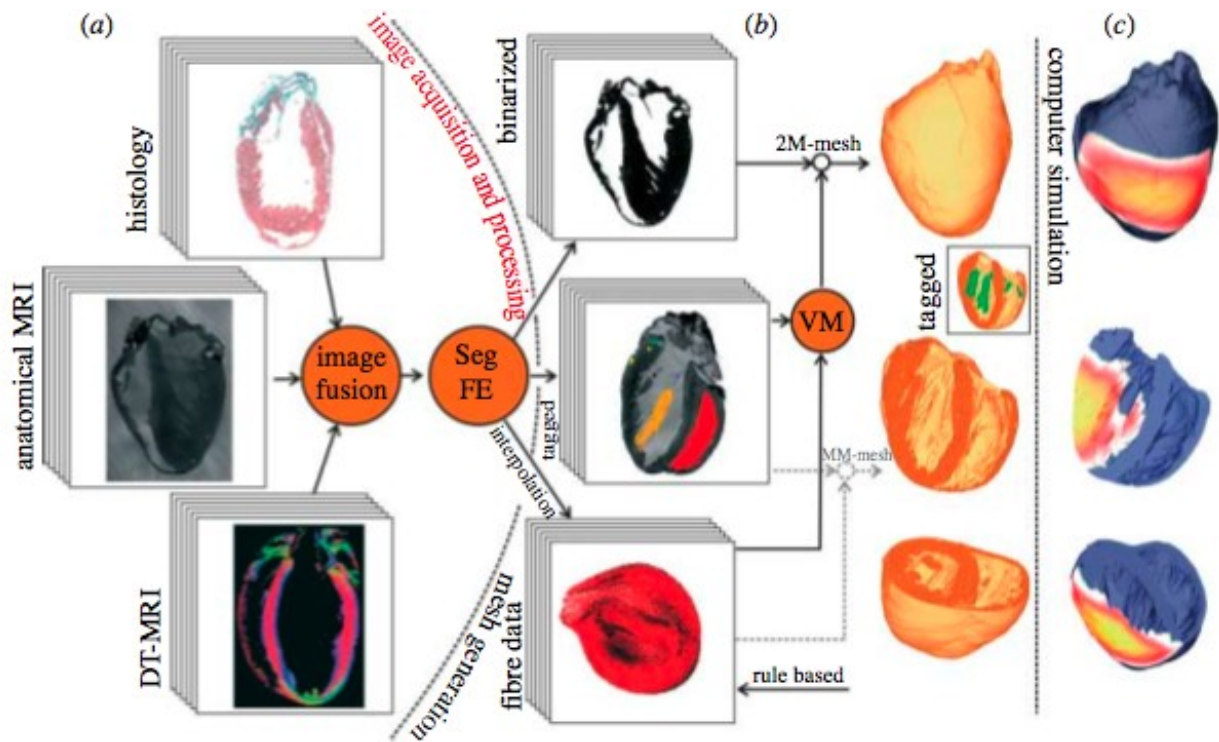


Figure 1. Workflow of an automated cardiac histo-anatomy processing pipeline. (a) Image stacks obtained from different ‘wet’ experimental modalities are fused into a single consistent three-dimensional image stack. Subsequently, the image stack is segmented (Seg) and relevant features are extracted (FE). (b) A binarized input stack is fed into the mesh generator. A voxel mapper (VM) transfers ‘tags’ that were assigned during the segmentation process, over to the final ‘tagged’ mesh. Meshes may have an arbitrary number of tags, while the mesh itself is a ‘two-material’ (2M) mesh only. Alternatively, the tagged image stack could be fed directly into the mesh generator to generate ‘multi-material’ (MM) meshes, i.e. the mesh generation strategy can be varied as a function of the tags assigned. Similarly, data on the tissue anisotropy (e.g. ‘fibre data’), either obtained from imaging data interpolated onto the mesh or generated on a per rule base, are mapped onto the final grid by the VM. In the inset, those regions of the mesh which were tagged as the endocardial surface of the RV are shown in red, the left atrioventricular valve in blue, a LV papillary muscle in orange, etc. (c) After assigning electrophysiological properties to the mesh and choosing electrode locations for stimulation, meshes are passed on to computer simulation. In the given example, the spread of electrical activation from an epicardial stimulation site is shown.

The research workflow process figure and the explanation has been gathered from the following research article:

Gernot Plank, Rebecca A.B. Burton, Patrick Hales, Martin Bishop, Tahir Mansoori, Miguel O. Bernabeu, Alan Garny, Anton J. Prassl, Christian Bollensdorff, Fleur Mason, Fahd Mahmood, Blanca Rodriguez, Vicente Grau, Jürgen E. Schneider, David Gavaghan, and Peter Kohl

**Generation of histo-anatomically representative models of the individual heart: tools and application**

Phil Trans R Soc A 2009 367: 2257-2292.

### Appendix 3. Sample methodology descriptions

#### Wetlab (from BBSRC funding proposal)

Hearts will be isolated from female Guinea pigs (350-400g) after cervical dislocation, and swiftly connected to a constant-flow (7- 8 mL C min<sup>-1</sup>) Langendorff perfusion system for 2 min wash with 'normal tyrode' solution (NT, see table 1 for composition of all solutions; time from sacrifice to perfusion 75-90 s). An incision is made in the pulmonary artery to avoid build-up of hydrostatic pressure. For MR scans on living tissue in Langendorff mode, a modified Krebs-Henseleit solution (KH) is used instead of NT to improve long-term tissue viability, and BSA is added to reduce tissue oedema. All solutions are checked for pH (7.4) and osmolality (300±5 mOsm; Knauer AG, Berlin), and maintained at 37°C.

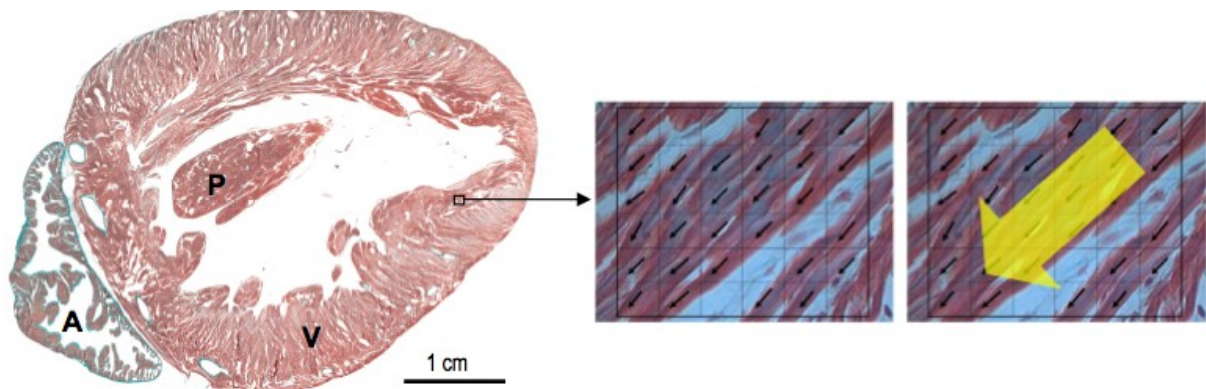
For establishment of the reference library, hearts will be fixed with minimal delay after either introduction of contracture (sodium- replacement by lithium, Li solution), cardioplegic arrest (using high potassium, HiK), or cardioplegic arrest plus volume-overload (using a left-ventricular fluid-filled balloon, inserted via the tied-off left atrium). Tissue fixation is by coronary perfusion with 50 mL of the fast-acting Karnovsky's fixative (2% formaldehyde, 2.5% glutaraldehyde mix), containing 2 mM gadodiamide (contrast agent).

During fixation, hearts are positioned in a glass container so that they become fully immersed in fixative. Fixed tissue is stored in Karnovsky's for 2-3 days before further use/processing. For MRI/DTMRI of fixed tissue, hearts are stabilized in an NMR tube, using low melting 1% agar with 2 mM paramagnetic contrast agent. For histological follow-up, hearts are rinsed in cacodylate buffer (3x) and then dehydrated using rising alcohol concentrations (8h each in 20/30/50/70% alcohol, 48h in 90%, 4x 6h in 100% alcohol). The alcohol is subsequently replaced by Xylene (five changes over 48 hours), and then the tissue is gradually infiltrated with wax (48h 25%; 12h 50%; 12h 75%; 1-2 weeks in 100%, depending on tissue size).

Whole heart tissue is serially sectioned (10 µm thickness), and every section is collected on APES-coated slides. Every fifth section is Trichrome stained to identify collagen (bluish green), myocytes (pink), cytoplasm (orange, highlighting non-myocytes), and nuclei (blue-black). Stained sections are mounted in DPX (Sigma), and imaged using a Leica QWin workstation and Leica QGO software (requested on this grant) to obtain whole cardiac cross-section mosaic images with a resolution of ~2.5 µm (Fig 1). Remaining sections are stored for back-up or subsequent additional analysis using other imaging modalities and/or staining protocols and labels.

	NaCl	LiCl	KCl	MgCl <sub>2</sub>	CaCl <sub>2</sub>	Glucose	K-glutamate	BSA	HEPES	NaHCO <sub>3</sub>	Gassed with
NT	140.0		5.4	1.0	1.8	11.0			5.0		O <sub>2</sub>
Li		140.0	5.4	1.0	1.8	11.0			5.0		O <sub>2</sub>
HiK	4.0		10.0	1.0	1.8	11.0	130.0		5.0		O <sub>2</sub>
KH	130.0		5.4	1.0	1.8	11.0		0.4%		10	CO <sub>2</sub> /O <sub>2</sub>

**Table 1:** Listing of solutions. BSA: bovine serum albumin. CO<sub>2</sub>/O<sub>2</sub>: carbogen [95% CO<sub>2</sub> / 5% O<sub>2</sub>]



**Figure 1:** Left: Mosaic montage image of longitudinal cardiac cross-section (trichrome), consisting of >380 individual frames (pilot work, rabbit). A: left auricular appendage; V: left ventricle; P: papillary muscle. Fibre orientation, cleavage planes, myocardial sheets or laminae, coronary vasculature, and myocyte / non-myocyte tissue can be clearly identified. Right: Schematic representation of regional fibre orientation extraction.

**Magnetic Resonance Imaging (from Gernot et al *Generation of histo-anatomically representative models of the individual heart: tools and application* Phil Trans R Soc A 2009 367: 2257-2292. )**

Prior to the inherently destructive histological processing, MRI was used to non-invasively provide three-dimensional datasets of all the ex vivo hearts. Imaging was carried out on a 9.4 T (400 MHz) MR system ( Varian, Inc., Palo Alto, CA, USA), comprising a horizontal magnet (bore size 210 mm), a VNMR5 DirectDrive console and a shielded gradient system (1 T mK<sup>-1</sup>, rise time 130 ms). A birdcage coil with an inner diameter of 28 mm ( Rapid Biomedical, Würzburg, Germany) was used to transmit/receive the MR signals.

(i) Anatomical imaging

Anatomical MRI scans were performed on all ex vivo hearts, using the three- dimensional gradient echo technique described by Schneider et al. (2004). The highest resolution datasets acquired were of 21.5 mm isotropic resolution for rat (data reconstructed as described before in Schneider et al. 2004) and 26.4!26.4 mm in plane, 24.4 mm out of plane for rabbit (as described by Burton et al. 2006).

(ii) Diffusion tensor imaging

A three-dimensional fast spin-echo pulse sequence was developed to provide diffusion-weighted MR images, and applied to ex vivo rat hearts. During acquisition, a pair of unipolar trapezoidal diffusion gradients was applied on either side of the first 180° pulse, and eight echoes were collected per excitation to reduce scan times. All diffusion and imaging gradients (including cross terms, to account for the interactions between the two) were included in the calculation of the diffusion weighting factor (or b-value, typically 700–870 s mm<sup>2</sup> ). Diffusion gradients were applied in six non-collinear directions, according to an optimized gradient scheme, based on the analogy of electrostatic repulsion described by Jones et al. (1999). Here, we collected diffusion-weighted three-dimensional datasets at an isotropic resolution of 101.6 μm in the ex vivo rat heart.

Following data acquisition, the diffusion tensor was calculated on a voxel- by-voxel basis via a weighted linear least-squares fit method (Kingsley 2006), using in-house software developed in Interactive Data Language ( ITT Corporation, CO, USA). The diffusion tensor was then decomposed to obtain its primary, secondary and tertiary eigenvectors.

**Segmentation (from forthcoming article Bishop et al. (2009) “Role of anatomical complexity in cardiac activation”)**

Segmentation Details

This Section contains specific details of the level-set segmentation methods including parameter values, where appropriate. For specific details of the nature of individual image processing filters, please refer to the ITK User Guide or the extensive online documentation ([www.itk.org](http://www.itk.org)). It should be noted that all parameter values given below were optimised based on extensive testing and visual inspection of the resulting segmentations.

Threshold Level-Set Filter

The first stage of the segmentation pipeline involved the use of the Threshold Level-Set Filter from the ITK libraries. The Threshold Level-Set Filter takes as an input an initial level-set surface generated by the fast marching method (set, 2002). The ITK Fast Marching Image Filter requires the specification of initial seed points in order to compute a distance map. Initial selection of seed points throughout the image volume was chosen based on local intensity values. Specifically, this involved moving a 3 × 3 × 3 voxel kernel throughout the image volume and comparing both the mean intensity of all 27 voxels within the kernel as well as the intensity of the central voxel, to user-defined limits (160 and 165, respectively). If both the mean and central voxel intensities were above these limits, a seed point was placed at the center of the kernel. The kernel was moved in steps of 2 voxels in each direction to avoid placing seed points within neighboring voxels. A user-provided initial distance (in this case 3) was subtracted from the distance map in order to obtain a level-set in which the zero set represented the initial contour. This initial level-set was given as input to the Threshold Level-Set Filter, along with the feature image itself (MR data set). The other user-defined limits in the Threshold Level-Set Filter are the upper (U) and lower (L) threshold limits which govern the propagation of the evolving contour. The propagation term P in Equation 1, at a given point in space with intensity level g, has a functional form rendering P positive for L < g < U , and negative otherwise. The threshold limits used here were L = 180 and U = 230. Scaling parameters were then used to balance the relative influence of the propagation (inflation) and curvature (surface smoothing) terms from Equation (1). Respective values of 1.0 and 5.0 were used for the parameters β and γ in Equation (1). The advection term was not used in this filter. The Threshold Level-Set Filter was run for maximum of 2000 iterations with an RMS tolerance of 0.02.

Geodesic Active Contour Level-Set Filter

The output of the Threshold Level-Set Filter, described above, acted as an initial level-set for the Geodesic Active Contour Level-Set Filter. This filter uses an advection term to attract the level-set to object boundaries. In addition to the initial level-set from the Threshold Filter, the Geodesic Filter takes as input the raw MR data stack. However, prior to its input, the raw MR image stack was firstly smoothed with a Curvature Anisotropic Diffusion Filter with a time-step of 0.0625, number of iterations 5 and conductance 9.0. An edge potential image was then produced by

passing the smoothed image through a Gradient Magnitude Recursive Gaussian Filter (using a sigma of 0.05), followed by a Sigmoid Image Filter with parameters  $\alpha = -1.0$ ,  $\beta = 2.0$ . The edge potential image along with the initial level-set was then used directly in the Geodesic Active Contours Level-Set Filter. The relative weights of the propagation, curvature and advection terms of the Geodesic Active Contours Level-Set Filter were set to  $-10$ ,  $1.0$  and  $1.0$ , respectively. The filter was run for a maximum of 2000 iterations with RMS error 0.001.

#### Laplacian Level-Set Filter

Finally, the output of the Geodesic Active Contours Level-Set Filter acts as an initial level-set input for the Laplacian Level-Set Filter. In this filter, the propagation (or speed) term in Equation 1 is constructed by applying a Laplacian image filter onto the raw MR image stack. The Laplacian filter calculates the second derivatives of the image and thus amplifies noise. To minimize this effect, we applied an initial smoothing anisotropic diffusion filter to the MR data set. In this case, the Gradient Anisotropic Diffusion Filter is used with 10 iterations, a time step of 0.0625 and conductance of 2.0. In addition to the smoothed raw MR image, the second input to the Laplacian Level-Set Filter is an initial level-set, which in this case is the direct result from the Geodesic Level-Set Filter above. The Laplacian Level-Set Filter is then run with the curvature scaling term (in Equation 1) set to 10.0 and the propagation scaling term set to 1.0 (the advection term is not used in this filter). Finally, the filter is run with a maximum RMS error of 0.002 and a maximum number of iterations of 250.

### Mesh (from forthcoming article Bishop et al. (2009) “Role of anatomical complexity in cardiac activation”)

#### Surface Discrimination Algorithm

Identification of different surfaces within the final ventricular finite element mesh was achieved by performing a secondary segmentation of the MR data using the same segmentation pipeline. However, in this case seed-points were manually placed only within the two ventricular cavities and outside the volume of the heart (i.e. excluding the blood vessels and the extracellular cleft spaces). In addition, following segmentation, the atria were left attached to the ventricles and a ‘cap’ put on top of the aorta to ensure the cavities were fully isolated from the outside space and from each other. For this particular application, the segmented voxel stack was meshed using the meshing software Tetgen ([www.tetgen.berlios.de/](http://www.tetgen.berlios.de/)), following prior computation of an STL surface representation using a marching cubes algorithm. Tetgen is capable of producing tetrahedral finite element meshes, where the elements corresponding to separate enclosed regions are automatically tagged with different numerical tags; a feature which is unfortunately not currently available in Tarantula. However, due to the reduced level of detail present in the segmented voxel stack, the results produced by Tetgen in this case were sufficiently robust. Meshing of the resulting segmented data set produced elements with three separate numerical tags representing the myocardial tissue volume, the LV cavity and the RV cavity. Note that in this case the ventricular cavities were directly connected to the atrial cavities as well. Examination of the numerical tags of each pair of elements bordering each bounding triangle face allowed discrimination between the three individual surfaces to be made. For example, a triangle which is part of the LV (or RV) endocardium, also forms part of two tetrahedral elements: one in the myocardium (green) and one in the LV cavity (red) (or RV cavity, blue). A triangle which is part of the epicardium only forms part of one tetrahedral element.

A list of all node points (and their spatial coordinates) was then formed for the epicardial, LV-endocardial and RV-endocardial surfaces in the secondary segmentation. The spatial coordinates of these surface node points were then mapped onto the final ventricular mesh using a nearest neighbor mapping algorithm to tag nodes in the final ventricular mesh as being one of the three surfaces. To avoid mapping nodes associated with the atrial tissue, all nodes with x, y, z positions residing above the plane used to remove the atria.

Parameter	Description	Value
ISO_REFINEMENT_LEVEL	Refines surface	0.5 (isosurface), 3 (levels)
INTERVAL_REFINEMENT_LEVEL	Refines inner mesh volume	0.5, 1.5 (isosurface lower & upper), 3 (levels)
BASEH	Maximal cell width	8
NO_SHRINK_ITERATIONS	Number of smoothing steps	1
SMOOTHING_ITERATIONS	Defines smoothing steps	5
TAGGING	Maps tags from segmentation to mesh	ON
TETRAHEDRALISE	Constructs purely tetrahedral mesh	ON

Table 1: Meshing parameters used in Tarantula.

## Appendix 4. Tools

Name of tool	Description	Purpose	Requirements	Comments
Leica MZ 95	The Leica MZ95 high-performance stereomicroscope features a 9.5:1 zoom ratio and magnifications of up to 480x. The high resolution up to 300 line pairs per millimetre with extremely high image contrast. <a href="http://tinyurl.com/mvdg2l">http://tinyurl.com/mvdg2l</a>	Imaging low resolution whole tissue		
Leica QWin	Leica QWin is a versatile image analysis and processing solution for quantitative microscopy which provides complete control of Leica microscopes, macroscopes and Leica digital cameras. <a href="http://tinyurl.com/n4hmhm">http://tinyurl.com/n4hmhm</a>	Imaging sections		
BMP Viewer	Developed by Alan Gamy. BMP Viewer allows to look at the histological data (24 bit BMP files). The software creates a thumbnail that is automatically generated when selecting a BMP file. <a href="http://mef.physiol.ox.ac.uk/Software/BMPViewer.exe">http://mef.physiol.ox.ac.uk/Software/BMPViewer.exe</a>	Visualization of histological images	Windows OS	Developed in-house Freely available
Insight Toolkit (ITK)	ITK is an open-source, cross-platform system that provides developers with an extensive suite of software tools for image analysis. Developed through extreme programming methodologies, ITK employs leading-edge algorithms for registering and segmenting multidimensional data. <a href="http://www.itk.org/">http://www.itk.org/</a>	Segmentation and registration	Cross-platform	Open source
Seg3D	Seg3D is a free volume segmentation and processing tool developed by the NIH Center for Integrative Biomedical Computing at the University of Utah Scientific Computing and Imaging (SCI) Institute.. <a href="http://www.sci.utah.edu/cibc/software/42-seg3d.html">http://www.sci.utah.edu/cibc/software/42-seg3d.html</a>	Segmentation	Input files accepted include DICOM, VFF, META, NRRD	Open source
Tarantula	Tarantula is a volume-mesh generator for voxel based data. It uses directly the MRT/CT data. The design is especially made for "huge" applications. <a href="http://www.meshing.at/Spiderhome/Tarantula.html">http://www.meshing.at/Spiderhome/Tarantula.html</a>	Mesh generation		Commercial software One licence available
Meshalzyer		Mesh		

		visualization	
Tetgen	<p>TetGen generates the Delaunay tetrahedralization, Voronoi diagram, constrained Delaunay tetrahedralizations and quality tetrahedral meshes.</p> <p><a href="http://tetgen.berlios.de/">http://tetgen.berlios.de/</a></p>	<p>Mesh generation</p>	<p>TetGen is written in ANSI C++ easy to compile and run on all major computer systems (e.g., Unix/Linux, Windows, MacOS, etc.).</p> <p>Formats                      .node; .poly; smesh;                      .ele; .face; .edge; .vol;                      .var; .neigh</p> <p>Open source</p>
CARP	<p>A virtual heart simulator capable of addressing clinical/experimental problems regarding cardiac function of an electrical or electromechanical nature.</p>	<p>Simulator</p>	<p>Commercial software</p>
CHASTE	<p>Chaste (Cancer, Heart and Soft Tissue Environment) is a general purpose simulation package aimed at multi-scale, computationally demanding problems arising in biology and physiology. Current functionality includes tissue and cell level electrophysiology, discrete tissue modelling, and soft tissue modelling..</p> <p><a href="http://web.comlab.ox.ac.uk/chaste/">http://web.comlab.ox.ac.uk/chaste/</a></p>	<p>Simulator</p>	<p>Developed in-house</p>

## Appendix 5. Storage resources

Type of storage	Operating System	Backup	Transfer protocols	Comments
Lab computers used by CMEFG	Windows based	N/A	N/A (unless data is being transferred between the lab computer's hard disk to on the of our NAS systems, in which case a simple Ethernet cable is used)	
Lab computers in DCM NAS systems		Uses RAID 5 technology to protect data against lost	FTP with username and passwords	Dr. Alan Garry is the person responsible for their management
Comlab Heart server				
Desktop computers at CMEFG	Windows based	Varies, but includes the University backup system (i.e. HFS)	NA	
Desktop computers in DCM		Varies	NA	
Desktop computers in CBG	Unix based	Varies	NA	