Combining and Uniting Business Intelligence with Semantic Technologies

Acronym: CUBIST
Project No: 257403

Small or Medium-scale Focused Research Project
FP7-ICT-2009-5
Duration: 2010/10/01-2013/09/30

WP7 Requirements Document

Abstract: Based on the directives provided by D.1.1.1, this report describes the requirements for the biological use case.

Type	Other
Document ID:	CUBIST D7.1.1
Workpackage:	WP7
Leading partner:	HWU
Author(s):	Kenneth M^{c}Leod and Albert Burger (HWU)
Dissemination level:	CO
Status:	Final
Date:	06 April 2011
Version:	1.0

Versioning and contribution history

Version	Description	Contributors
0.1	First draft	Kenneth M^{c}Leod (HWU)
0.2	Extended introduction, personas & utilisation scenarios	Kenneth M^{c}Leod (HWU)
0.3	Added paragraph on pathways	Kenneth M^{c}Leod (HWU)
0.4	Added requirements	Kenneth M^{c}Leod (HWU)
0.5	Edited following feedback from Albert Burger	Kenneth M^{c}Leod & Albert Burger (HWU)
0.6	Edited requirements following feedback from group; amended paragraph on pathways	Kenneth M^{c}Leod & CUBIST Consortium
0.7	Edited following review from FD	Kenneth M^{c}Leod
0.8	Edited following review from AS	Kenneth M^{c}Leod
1.0	Final version	
	Reviewer:	Frithjof Dau (SAP)
	Reviewer:	Alex Simov (Ontotext)

Table of contents

List of figures

List of tables

1 Introduction

This document describes the biological use case, focusing on the EMAGE gene expression database, and then provides the requirements for the use case centric aspect of the CUBIST project.

The contents of this document are, essentially, specified in a previous deliverable – D1.1.1, “Directives for the requirement analysis in the use cases”. That document defines what information is required from each of the use cases, and describes the communication methods to be employed. These mechanisms include “personas” to describe typical users; “utilisation scenarios” to chronicle how the representative users typically employ EMAGE currently; and, a “volere snowcard” to capture the atomic requirements of the model users.

Before the fruits of these mechanisms are presented some biological background is required - this is provided in Section 2. Section 3 extends the initial discussion, providing a detailed description of the data central to this use case. Four typical classes of EMAGE user are detailed through the use of fictional personas in Section 4, before Section 5 provides typical before, and after, CUBIST usage scenarios. The atomic requirements are summarised in Section 6.

2 Biological background

This chapter provides the background information necessary to understand the rest of the document. It quickly summarises the notions of gene expression information, *in situ* gene expression information, and the developmental mouse before describing the resource at the centre of this use case.

2.1 Biology 101

A reprise of the basic biological terminology used in this work is provided here.

Species is the name given to a group of similar individuals that are members of the same close biological family. Members of the same species should be able to reproduce as should their offspring. An individual member of a species is called an *organism*. Organisms are comprised of *systems*, e.g. the respiratory system. These systems contain *organs* (for example the lungs), which are in turn composed of similar groups of *cells* known as *tissues*. Cells are the basic unit of life and contain distinct structures (so-called *organelles*), e.g. the nucleus.

In mammalian cells the nucleus contains *chromosomes* - a chromosome is a long DNA (Deoxyribonucleic Acid) molecule. The chromosome may be split into a series of small units called *genes*. Each gene provides the information needed by the cell to function or develop. Every cell in an organism contains the same DNA; however, for an individual gene to be used by a cell it must be `switched on'. Genes that are switched on are said to be *expressed*. Similarly, genes that are not switched on are *not expressed*.

Genes enact changes in a cell by being converted into *proteins*. Proteins are molecules that alter the behaviour and function of the cell, in addition to carrying material out of a cell and around the body. Consequently, proteins have a wide range of functions ranging from catalysing chemical reactions, and switching off genes, to being the building blocks used to support cells and create structures such as hair.

Unwanted or abnormal features such as cleft lips are a result of certain genes having the `wrong' level of expression, e.g. genes that should be expressed are not expressed. To gain an understanding of these features, the set of genes expressed in normal healthy structures must be compared to the set of genes expressed in abnormal or unhealthy structures. The difference between the two sets provides an indication of the root cause of the abnormality and a basis for further research into a cure or prevention.

Additionally, gene expression information may be used to provide context for higher-level biological processes. For instance, a *pathway* documents the biochemical reactions of one particular (sub)process. Special proteins called *enzymes* facilitate the chemical reactions that form the pathway. It can be useful to map these enzymes back to their corresponding genes, and then explore the expression information for those genes.

A more in-depth discussion on the basics of biology is provided in standard textbooks such as (Campbell, et al. 2008).

2.2 Gene expression

As mentioned in Section 2.1, genes are instructions that control the development of an organism by affecting the type and number of proteins produced at any one time.

Genes are small units found in the chromosome of every cell. They are built from DNA. Initially this DNA is *transcribed* to form RNA (Ribonucleic Acid). The RNA is then *translated* into proteins. When a protein has been produced, the gene may be described as *expressed*.

Gene expression experiments are interested in determining which genes are active (expressed) in a specific type of cell within a particular organism at a precise time. Measuring the type and number of proteins present gives a true picture of which genes are expressed. Not every instance of RNA is translated into a protein, so looking at RNA levels only allows an estimate of the gene expression to be made. Examining DNA indicates whether or not a gene is present - it says nothing about the expression level of that gene.

In this case gene expression is studied with respect to the developmental mouse.

2.3 The developmental mouse

It is not possible to experiment on humans for moral, ethical, and legal reasons. Consequently, scientists use substitute organisms. These are known as *model organisms*. A wide range of plants, animals, fish, and insects are studied. This work concentrates on one of those, the mouse. “Mouse” is the common name for the animal with the Latin name *Mus*. There are many species ranging from *Mus musculus* (the house mouse) to *Peromyscus maniculatus* (the deer mouse). In addition to knowing the species of mouse used, it is necessary to know whether or not the mouse is a *mutant*. So-called *wild type* mice have a normal, or natural, set of chromosomes. Whereas mutants are bred to insure they have some particular trait, for example cancer.

The mouse *develops* from a single cell into a mammal with a complex anatomy comprising countless cells. This process of development was studied by Karl Theiler (Theiler 1989). He split the development of the house mouse into 28 distinct stages, called *Theiler Stages* (TS). The first 26 Theiler Stages deal with the unborn mouse. The final two describe the newborn and then postnatal adult mouse. Current convention creates a split between the unborn mouse and the final two stages. The former group being called the *developmental mouse* and the latter the *adult mouse*.

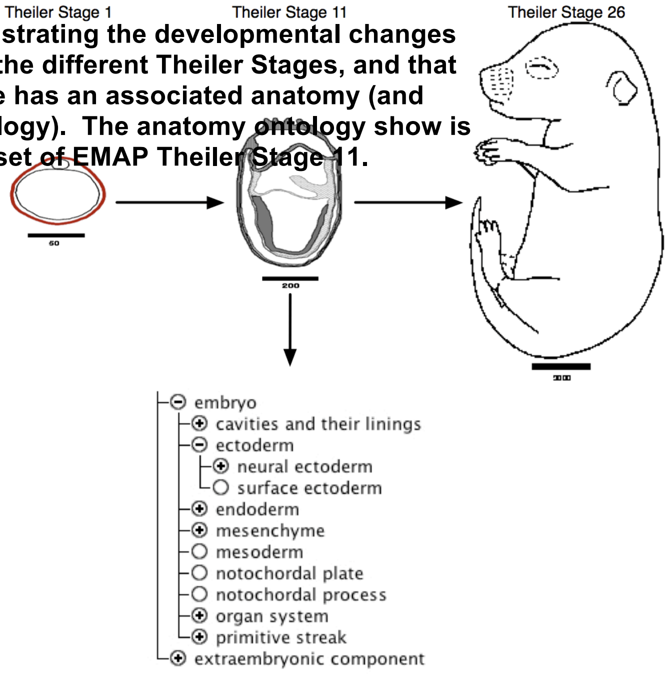


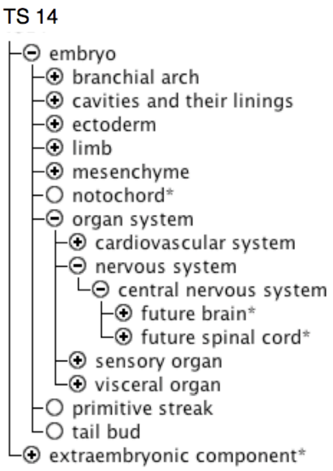
Figure 1 - Illustrating the developmental changes captured by the different Theiler Stages, and that each stage has an associated anatomy (and anatomy ontology). The anatomy ontology show is a subset of EMAP Theiler Stage 11.

A Theiler Stage is accompanied by an approximate time since conception measured in days, called *Days Post Conception* (DPC). It also includes a description of the anatomy at that stage, and highlights what has changed from the previous stage. Figures illustrate these changes. A summary and outline of these stages can be found at the Edinburgh Mouse Atlas Project (EMAP) website[1].

Each Theiler Stage has an associated anatomy, and corresponding anatomy ontology – see Figure 1. Although multiple anatomies exist, the main anatomy used in this work is EMAP.

2.3.1 EMAP Anatomy Ontology

The name EMAP is confusing because it is overloaded. It applies to both the Edinburgh Mouse Atlas Project[2] and the anatomies developed as one part of that project. There is one anatomy (ontology) per Theiler Stage of the Developmental Mouse - part of the anatomy ontology for Theiler Stage 14 can be seen in Figure 2. These ontologies describe the anatomy of the developmental mouse using a series of *part of* relations, thus looking at Figure 2 it is obvious that the FUTURE BRAIN is part of the CENTRAL NERVOUS SYSTEM, which is in turn part of the NERVOUS SYSTEM and that is part of the ORGAN SYSTEM.



```
TS 14
├⊖ embryo
│ ├⊕ branchial arch
│ ├⊕ cavities and their linings
│ ├⊕ ectoderm
│ ├⊕ limb
│ ├⊕ mesenchyme
│ ├○ notochord*
│ ├⊖ organ system
│ │ ├⊕ cardiovascular system
│ │ ├⊖ nervous system
│ │ │ └⊖ central nervous system
│ │ │   ├⊕ future brain*
│ │ │   └⊕ future spinal cord*
│ │ ├⊕ sensory organ
│ │ └⊕ visceral organ
│ ├○ primitive streak
│ └○ tail bud
└⊕ extraembryonic component*
```

Figure 2 Part of the EMAP anatomy ontology for Theiler Stage 14.

Each structure in the ontology is given a unique identifier in the form EMAP:*number*, e.g. EMAP:152. The structure also has a name, for example FUTURE BRAIN. This is the structure's *short name*. Its *full name* would be the entire path from the root node of the ontology to its short name, e.g. MOUSE.EMBRYO.ECTODERM.NEURAL ECTODERM.FUTURE BRAIN.

The same structure can appear in multiple stages, and can have the same short and full names in these stages; the above example applies equally to Theiler Stages 11, 12, and 13. The one unique feature of the future brain in these different stages is its ID: in TS11 it has EMAP:152, in TS12 EMAP:235, and in TS13 EMAP:441.

More details on the EMAP anatomy ontologies can be found in (Baldock and Davidson 2008).

[1] www.emouseatlas.org/emap/ema/theiler_stages/StageDefinition/stagecriteria.html

[2] www.emouseatlas.org

2.4 Gene expression experiments

There is a wide range of techniques to determine the expression level of a gene. These experiments differ not only in their method, but also in their precise focus. Primarily, experiments can concentrate on the location of expression or on the quantity of the gene expressed.

Often experiments rely on the close association between individual genes, RNA, and proteins - a gene is transcribed into a RNA, which may be translated into a protein. This means a protein or RNA can be mapped onto a corresponding gene. Therefore, gene expression can be evaluated by examining the proteins or RNA contained in a sample.

One key principal relied on in many of the different experimental techniques is the notion of *hybridisation*. This is the idea of chemically bonding a *probe* to the DNA or RNA. Probes are designed to bond with a particular gene sequence - ideally this sequence will correspond to one gene but this is not always the case. Probes are highly visible - for example they may be highly coloured or radioactive - the visibility of the probe provides an insight into where genes are expressed and the quantity of the expression found in that area.

The experiments that feature in this use case are all of examples of *in situ* hybridisation, hence this technique is given special consideration next.

2.4.1 *In situ* hybridisation gene expression experiments

In situ is the Latin for “in place”, accordingly *in situ* experiments focus on identifying precisely *where* a gene is expressed.

They produce images that are either an entire mouse embryo (a so-called *wholemount*), or a slice/section of that mouse (e.g. Figure 3). The areas of intense colour indicate where the gene is expressed – this is the so-called *expression pattern*. This visual result is commonly mapped to an anatomy, such as EMAP, thus documenting the link between the genes and the location of their expression.

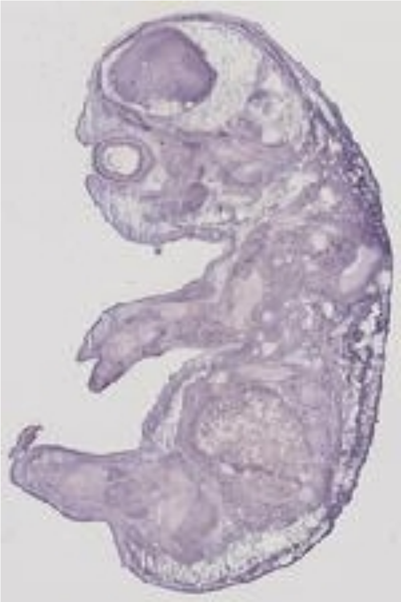


Figure 3 - Result of an *in situ* gene expression experiment (EMAGE:16950).

2.5 EMAGE

EMAP is the Edinburgh Mouse Atlas Project. It is the umbrella name for a range of activities.

The first of these activities is the creation and maintenance of an anatomy (and corresponding ontology) for each stage of the developmental mouse. This too has the name EMAP, and is discussed in Section 2.3.1. In addition to the anatomies, the project has produced a series of 3D computer models of the mouse, e.g. Figure 4. There is one 3D model for each Theiler Stage. Each model comprises a number of *voxels* (volumetric pixels) that are stacked in a 3D space.

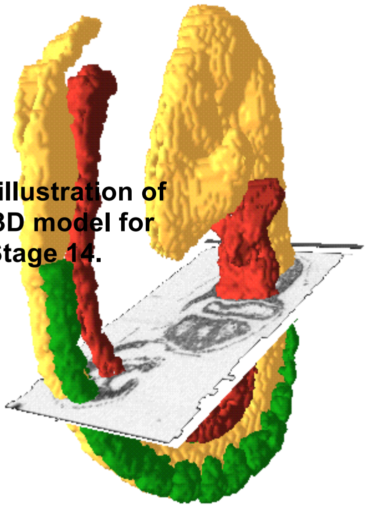


Figure 4 - An illustration of the EMAGE 3D model for Theiler Stage 14.

EMAGE[3] is the Edinburgh Mouse Atlas of Gene Expression, another element of EMAP. This is a gene expression database that (re)publishes *in situ* gene expression data for the developmental mouse. When researchers perform an experiment they may publish it in a traditional journal, and then have it republished in EMAGE (or a similar resource). Alternatively, the researchers may directly submit their data to a resource, such as EMAGE, by-passing the orthodox scientific journals.

EMAGE publishes two forms of results (see Figure 5): those tied to the anatomy ontologies (so-called *Textual Annotations*); and results tied to the 3D models (*Spatial Annotations*). EMAGE is unusual in this respect, because most gene expression resources still do not use 3D models, and thus cannot produce spatial annotations.

In terms of content, EMAGE contains the following types of experiment: *in situ* hybridisation against mRNA, immunohistochemistry, and *in situ* reporter. In February 2011, it contained details of over 38,000 procedures.

More details on EMAGE can be found in (Venkatarman, et al. 2007).

[3] www.emouseatlas.org/emage

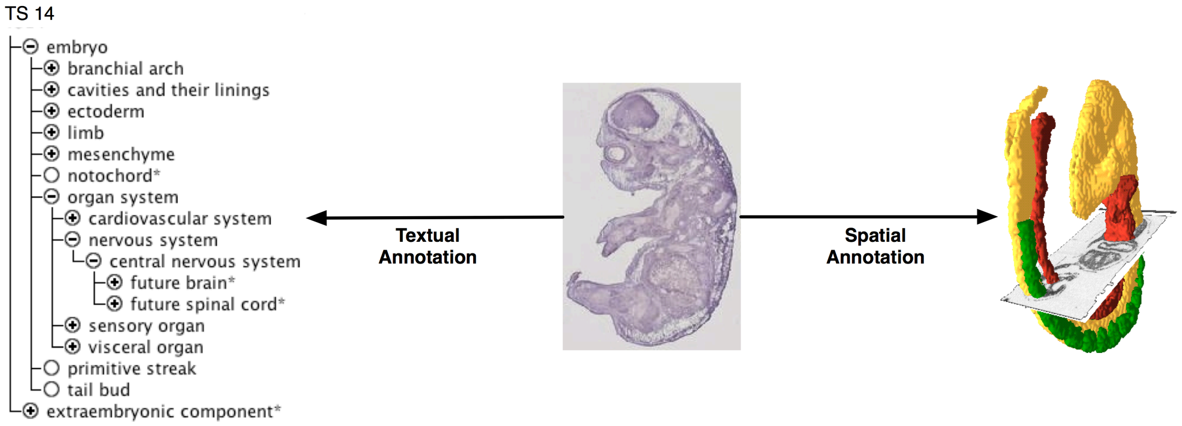


Figure 5 - Textual annotations versus spatial annotations; spatial are linked to one of the 3D models whereas textual are linked to one of the anatomy ontologies.

2.5.1 Search mechanisms

Via its homepage[4], EMAGE provides both programmatic and web based interfaces to its data. Here all these mechanisms will be quickly explored.

2.5.1.1 Search by gene

Lets a user to search for gene expression information by starting with the name (or symbol) of a gene or genes. The user is forced to restrict the answer to a particular level of expression (expressed, possibly expressed, or not expressed) and a series of Theiler Stages (it is possible to limit the results to a single stage if desired).

In natural language, the user is asking something like:

Where is the gene bmp4 detected in stages 15 - 17?

This is the most popular form of query, with over 90% of users employing it.

2.5.1.2 Search by embryo space

Allows the user to manipulate a 3D model of the mouse, and thus choose a point or region in space (a region is a series of points). As the user must first select a Theiler Stage, each point has four dimensions (time, X, Y, and Z). Again the user is required to restrict the expression level to one: detected, possible or not detected.

A natural language example of this form of query could be:

What genes are detected at location -11.8, 23.4, 112.3 in stage 17?

2.5.1.3 Search by anatomy name

Enables the user to specify a term from one of the 26 anatomy ontologies, and a level of expression in which they are interested.

A typical query may be:

What genes are expressed in the future brain in stage 15?

[4] www.emouseatlas.org/emage

2.5.1.4 Search by Biomart

Biomart[5] is a tool developed by the EBI[6]. It enables bulk download of specific data from a resource. Commonly, output is in the form of comma-separated files. This interface is intended for bioinformaticians rather than biologists.

2.5.1.5 Search by Gene Ontology

The Gene Ontology[7] (GO) is an ontology that describes genes according to three dimensions:

- Cellular component;
- Function;
- Processes a gene is involved in.

The genes contained in EMAGE are annotated with terms from GO, therefore a user can search for gene expression information by starting with GO terms; the terms are mapped onto the genes, and then EMAGE simply queries to discover expression information for the appropriate genes. Once again, the user is forced to restrict their query to a particular level of gene expression.

2.5.1.6 Programmatic access

EMAGE stores its data in an IBM DB2 repository. Programmatic access to the data is through one of the following means:

JDBC direct, and full, read-only access;

RMI access through JAVA RMI technology - provides almost full access to the entire database;

DAS Direct Annotation Server[8] - provides access to a subset of the EMAGE data encoded in XML;

Web service SOAP web service - limited data available, again encoded in XML;

URL API a number of parameterised URLs that can be used to generate standard EMAGE HTML pages.

2.5.2 Search results

If using programmatic access the output from EMAGE will depend on the actual technology employed - this shall not be considered further. All the web based interfaces, with the exception of Biomart, provide the same style and form of output. In each case, output is a web page (see Figure 6) that contains a table. Each row in the table is an individual assay, (i.e. experiment). This experiment is considered relevant to the original query.

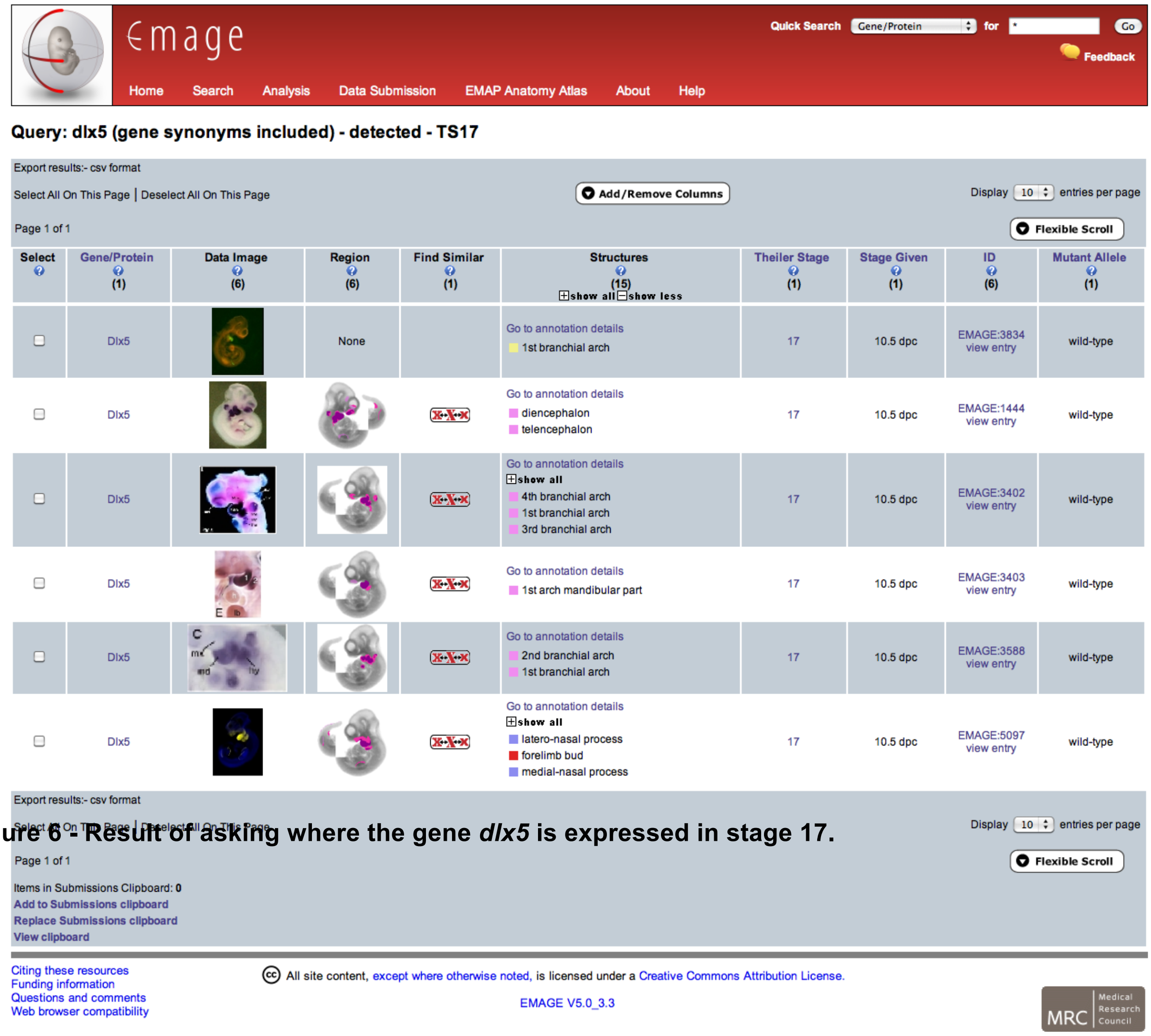
Figure 6 displays the result of asking *where is dlx5 expressed in Theiler Stage 17?* There is no summary, or higher-level analysis, purely raw information. Looking at the second row, and clicking on the link “EMAGE:1444 view entry” changes the web page to that captured in Figure 7. In order to have a greater understanding of the information stored for each experiment, Section 2.5.3 will walkthrough Figure 7.

[5] www.biomart.org

[6] European Bioinformatics Institute (www.ebi.ac.uk)

[7] www.geneontology.org

[8] www.biodas.org



Query: dlx5 (gene synonyms included) - detected - TS17

Export results:- csv format

Select All On This Page | Deselect All On This Page

Display 10 entries per page

Page 1 of 1

Select ?	Gene/Protein ? (1)	Data Image ? (6)	Region ? (6)	Find Similar ? (1)	Structures ? (15) ⊞show all ⊟show less	Theiler Stage ? (1)	Stage Given ? (1)	ID ? (6)	Mutant Allele ? (1)
□	Dlx5		None		Go to annotation details 🟨 1st branchial arch	17	10.5 dpc	EMAGE:3834 view entry	wild-type
□	Dlx5				Go to annotation details ■ diencephalon ■ telencephalon	17	10.5 dpc	EMAGE:1444 view entry	wild-type
□	Dlx5				Go to annotation details ⊞show all ■ 4th branchial arch ■ 1st branchial arch ■ 3rd branchial arch	17	10.5 dpc	EMAGE:3402 view entry	wild-type
□	Dlx5				Go to annotation details ■ 1st arch mandibular part	17	10.5 dpc	EMAGE:3403 view entry	wild-type
□	Dlx5				Go to annotation details ■ 2nd branchial arch ■ 1st branchial arch	17	10.5 dpc	EMAGE:3588 view entry	wild-type
□	Dlx5				Go to annotation details ⊞show all ■ latero-nasal process 🟥 forelimb bud ■ medial-nasal process	17	10.5 dpc	EMAGE:5097 view entry	wild-type

Export results:- csv format

Select All On This Page | Deselect All On This Page

Display 10 entries per page

Page 1 of 1

Items in Submissions Clipboard: **0**

Add to Submissions clipboard

Replace Submissions clipboard

View clipboard

Citing these resources
Funding information
Questions and comments
Web browser compatibility

All site content, except where otherwise noted, is licensed under a Creative Commons Attribution License.

EMAGE V5.0_3.3

Figure 6 - Result of asking where the gene *dlx5* is expressed in stage 17.

2.5.3 EMAGE walkthrough

In order to illustrate the information provided by a typical gene expression resource, a single experiment from EMAGE will be examined. The experiment's web page can be seen in Figure 7; there is nothing remarkable about this particular page or experiment, it was chosen because it is a typical example.

At the top left of the page, in bold font, is **EMAGE:697**. This is the accession identifier, a unique ID for the experiment.

Below that is information on the **Gene**. The gene's unique symbol (in this case *Fgf5*) precedes the gene's name (fibroblast growth factor 5). At the end of the line is a link to a page containing more details on this gene. That page is provided by a different resource, The Mouse Genomics Institute[9] (MGI). The link is presented as the accession ID for the gene in that resource.

The next line indicates the **Theiler Stage** on which the experiment was performed, i.e. the age of the mouse when the experiment was conducted. Theiler Stage 8 corresponds to approximately 6 days after conception.

Data source indicates whether or not a screening program performed the experiment, in this case it was not. Screening programs are large projects that are optimised for throughput and thus perform large volumes of experiments. The alternative is for the experiment to be conducted by a small research lab, possibly one specialising in the structure or gene being experimented on.

The subsequent five sections provide the experimental results. The first provides the actual result of the experiment - a series of images illustrating where the gene is expressed. The images are of a slice of a particular mouse. The following **Notes for interpretation** provides a key for the annotations used in the images.

By examining the images, an expert is able to identify which structures the gene is expressed in. The researchers do this for their publication based on this experiment by creating Textual Annotations. This analysis is reported in the segment called **Sites of Gene Expression Annotated Manually**. Three are provided, along with an indication of the level of expression, e.g. Strong. Notice that one of the results states that a gene is not expressed (not detected).

A manually produced mapping of the above images to the 3D models produces the **Spatial Annotation** section. This starts by showing the actual spatial mapping performed.

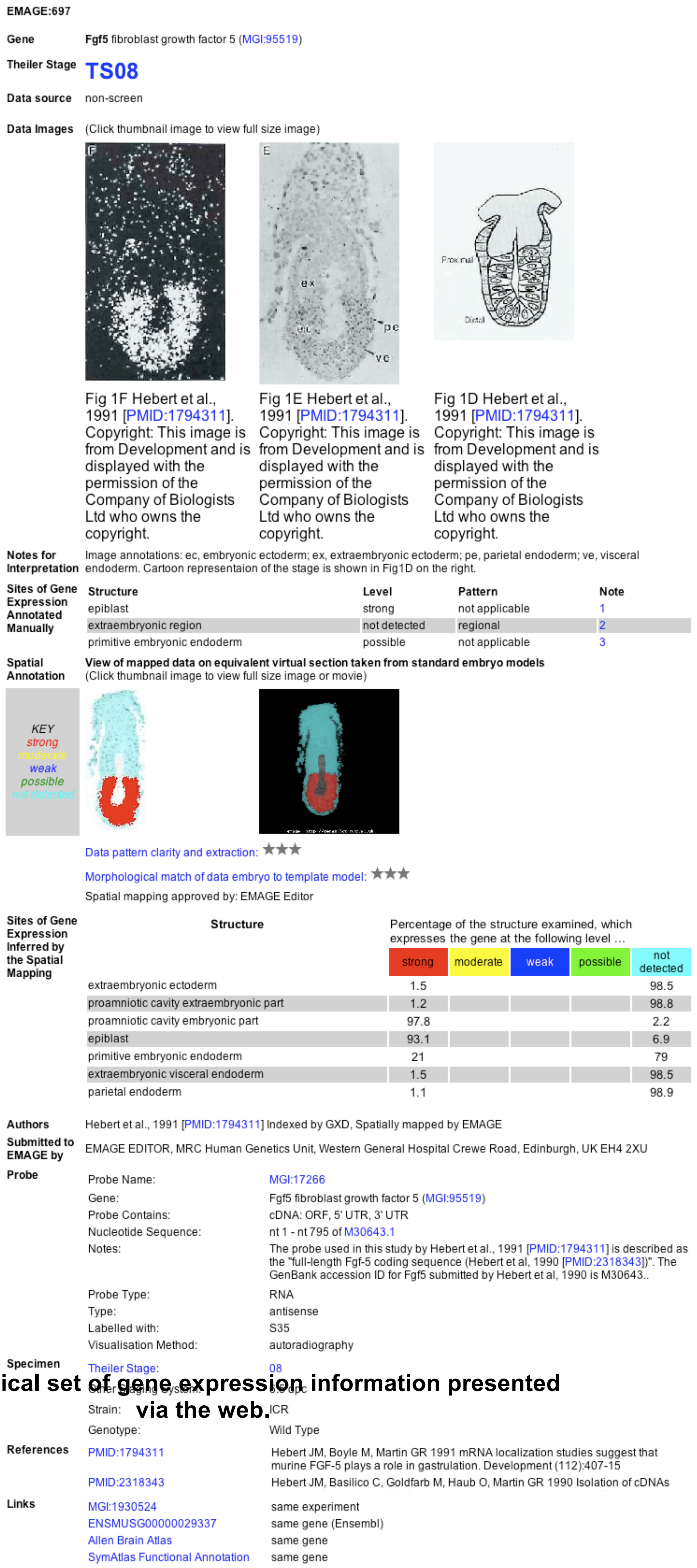
Subsequently, a list of quality measures is provided. “Data pattern clarity and extraction” indicates how clear the experimental image is. “Morphological match of data embryo to template model” indicates how well the subject in the image relates to the standard 3D model used for spatial annotations. Lastly it states who approved the spatial mapping: it can be either the researcher or the EMAGE editors.

Then comes **Sites of Gene Expression Inferred by the Spatial Mapping** - this is the list of textual annotations derived from the spatial annotations. A list of relevant structures is provided, as is the volume of that structure that has the gene expressed in it. This is shown as a percentage, for each level of expression. For example, 1.5% of EXTRAEMBRYONIC ECTODERM has *fgf5* strongly expressed, and the remainder of the structure (98.5%) does not contain the gene.

Authors lists the people responsible for the different information on the page. Firstly, the researcher is credited and their research paper cited (if one exists). “Indexed by” indicates who took the experimental result and mapped it to the EMAP anatomy ontology. It can be the researcher, the GXD editors, or the EMAGE editors (GXD[10] is EMAGE's “sister” resource; however, it only provides textual annotations). Finally, the reader is informed who created the spatial mapping, normally this is the EMAGE editor, but it can be the researchers.

[9] www.informatics.jax.org/mgihome/

[10] http://informatics.jax.org



EMAGE:697

Gene **Fgf5** fibroblast growth factor 5 (MGI:95519)

Theiler Stage **TS08**

Data source non-screen

Data Images (Click thumbnail image to view full size image)

Fig 1F Hebert et al., 1991 [PMID:1794311]. Copyright: This image is from Development and is displayed with the permission of the Company of Biologists Ltd who owns the copyright.

Fig 1E Hebert et al., 1991 [PMID:1794311]. Copyright: This image is from Development and is displayed with the permission of the Company of Biologists Ltd who owns the copyright.

Fig 1D Hebert et al., 1991 [PMID:1794311]. Copyright: This image is from Development and is displayed with the permission of the Company of Biologists Ltd who owns the copyright.

Notes for Interpretation Image annotations: ec, embryonic ectoderm; ex, extraembryonic ectoderm; pe, parietal endoderm; ve, visceral endoderm. Cartoon representaion of the stage is shown in Fig1D on the right.

Sites of Gene Expression Annotated Manually

Structure	Level	Pattern	Note
epiblast	strong	not applicable	1
extraembryonic region	not detected	regional	2
primitive embryonic endoderm	possible	not applicable	3

Spatial Annotation **View of mapped data on equivalent virtual section taken from standard embryo models**
(Click thumbnail image to view full size image or movie)

Data pattern clarity and extraction: ★★★

Morphological match of data embryo to template model: ★★★

Spatial mapping approved by: EMAGE Editor

Sites of Gene Expression Inferred by the Spatial Mapping

Structure	Percentage of the structure examined, which expresses the gene at the following level ...				
	strong	moderate	weak	possible	not detected
extraembryonic ectoderm	1.5				98.5
proamniotic cavity extraembryonic part	1.2				98.8
proamniotic cavity embryonic part	97.8				2.2
epiblast	93.1				6.9
primitive embryonic endoderm	21				79
extraembryonic visceral endoderm	1.5				98.5
parietal endoderm	1.1				98.9

Authors Hebert et al., 1991 [PMID:1794311] Indexed by GXD, Spatially mapped by EMAGE

Submitted to EMAGE by EMAGE EDITOR, MRC Human Genetics Unit, Western General Hospital Crewe Road, Edinburgh, UK EH4 2XU

Probe

Probe Name:	MGI:17266
Gene:	Fgf5 fibroblast growth factor 5 (MGI:95519)
Probe Contains:	cDNA: ORF, 5' UTR, 3' UTR
Nucleotide Sequence:	nt 1 - nt 795 of M30643.1
Notes:	The probe used in this study by Hebert et al., 1991 [PMID:1794311] is described as the "full-length Fgf-5 coding sequence (Hebert et al, 1990 [PMID:2318343])". The GenBank accession ID for Fgf5 submitted by Hebert et al, 1990 is M30643..
Probe Type:	RNA
Type:	antisense
Labelled with:	S35
Visualisation Method:	autoradiography

Specimen

Theiler Stage:	08
Other Staging System:	6.5 dpc
Strain:	ICR
Genotype:	Wild Type

References

PMID:1794311	Hebert JM, Boyle M, Martin GR 1991 mRNA localization studies suggest that murine FGF-5 plays a role in gastrulation. Development (112):407-15
PMID:2318343	Hebert JM, Basilico C, Goldfarb M, Haub O, Martin GR 1990 Isolation of cDNAs

Links

MGI:1930524	same experiment
ENSMUSG00000029337	same gene (Ensembl)
Allen Brain Atlas	same gene
SymAtlas Functional Annotation	same gene

Figure 7 - Typical set of gene expression information presented via the web.

Submitted to EMAGE by informs a reader who submitted the experiment for inclusion in EMAGE. EMAGE supplies tools that allow researchers to do this directly, and the EMAGE editors read published papers and submit data themselves. Historically, they have shared data with GXD, and verified it before it is included in the database. In this instance, the "Indexed by GXD" from the previous **Authors** section tells a reader that GXD read the published paper, mapped the results to the EMAP anatomy, and shared the data with EMAGE. The "Submitted to EMAGE by ... EMAGE Editor" indicates that the Editorial team have reviewed and accepted the data from GXD.

The next two sections provide provenance information for the experiment. Ideally the **Probe** section provides enough information for the same probe to be used by someone else. The information presented here will be taken from the article in which the experiment was published.

Specimen gives details of the experimental subject. The age of the mouse is given in two forms: Theiler Stage; and days post conception. Its "genotype" indicates that this is the standard mouse with no genetic mutations. The "preparation" informs that the mouse was cut into slices (called sections) and placed in paraffin.

The penultimate section provides a reference for the paper in which this experiment was published. In this case there are two publications, because the second one contains details of the probe used in the current experiment (the first citation).

Links presents connections (URLs) to related information in other resources, which can be considered complementary to information offered by EMAGE.

3 Data

This section focuses on the data stored in the EMAGE and EMAP databases. The anatomy ontologies and 3D spatial models are associated with EMAP, whilst EMAGE contains the gene expression information.

To begin with, this section discusses the unstructured data in this use case, in Section 3.1. Section 3.2 explores an abstraction of the notion of gene expression information, before Sections 3.3 to 3.5 delve into more concrete details. Finally, Section 3.6 considers how to access the information, and the data models a user will encounter.

3.1 Unstructured data

This use case does not contain any unstructured data. All data to be used in this use case will come from a structured data repository.

The only unstructured data items that may prove pertinent to this use case, are the journal publications that often precede an experiment being published in EMAGE. Such publications regularly contain more information than the resource. Yet, this information is hidden inside electronic articles, only some of which are public domain. As this data is not central to the use case, it shall not be discussed further.

3.2 Gene expression information: an abstract view

This section will consider the notion of gene expression information from a so-called “kite” level view, before subsequent sections consider the separate elements in more depth.

Effectively gene expression is a triple: *gene*, *level of expression*, and *location of expression*. Various representations of each element exist in the biological domain: in some cases, EMAGE itself has multiple representations too.

EMAGE contains two different descriptions of gene expression information: textual annotations and spatial annotations. The textual annotation is a triple, e.g.

fgf5	strong	EPIBLAST TS8

Although this triple forms the core of the EMAGE textual annotation data set it is possible to extend beyond the triple as EMAGE contains far more information - as described in Section 2.5.3.

In contrast spatial annotations are, in their raw form, images that show the *pattern* of expression. After the spatial annotations have been created (i.e. the results are mapped to a 3D model) the gene expression information can be viewed as a slice (2D image) through the 3D model. For example see Figure 8.

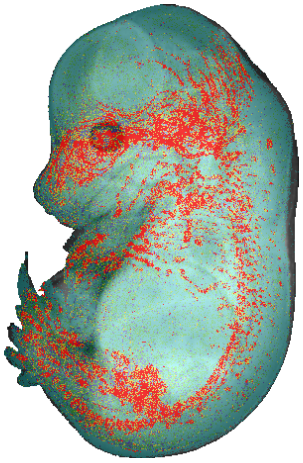


Figure 8 - 2D representation of spatial annotations from experiment EMAGE:6809.

The 2D representation is created by defining a plane in the 3D model, and then constructing an image that represents the data in that plane. Each 2D image contains a significant amount of gene expression information.

Textual annotations record expression level at the granularity of individual structure (terms in the anatomy ontology). Spatial annotations record gene expression information at a voxel (3D pixel) level. Accordingly, there is considerably more information in the 3D models than the textual annotations provide. The ability to query this richer data source would be very valuable.

3.2.1 Important aspects of gene expression information

It is important to note that a single experiment can produce multiple annotations; they may all be textual, spatial or a mixture of both. Furthermore, because multiple experiments can examine the same gene and location, it is possible for there to be conflicting annotations, e.g.

bmp4	strong	EPIBLAST TS8
bmp4	not detected	EPIBLAST TS8

These two annotations deal with the same gene and the same structure (at the same point in time), ideally they would have the same level too. Yet, this is not the case: the first annotation suggests the gene is expressed, whilst the second suggests it is not.

Depending on the biological task being undertaken the following textual annotations may be conflicting or in agreement:

bmp4	strong	EPIBLAST TS8
bmp4	weak	EPIBLAST TS8

Although the level of inconsistent information in EMAGE is minimal, it does exist. Precise figures are difficult to provide because they depend on which data set is used. Values can be based on textual annotations or spatial annotations. The textual annotations may, or may not, have propagation (see Section 3.5.1) included. Additionally, there are two definitions of inconsistency: binary (expressed versus not expressed) and analogue (e.g. strong expression is distinct from *weak* expression despite both levels suggesting a gene is expressed). Moreover, because the database is continually growing the values are constantly changing. However, to provide some indication of level of inconsistency, in January 2011 the CUBIST consortium was provided with approximately 61800 textual annotations (without propagation). This data set included approximately 340 conflicting annotations (binary conflicts without propagation).

EMAGE is not the only resource with this difficulty; many biological resources suffer from similar issues. Furthermore, when the EMAGE dataset is integrated with complementary resources, there is the possibility of conflict between the resources too.

The final point worthy of note is that none of the gene expression resources can be considered complete. In the case of EMAGE this means that it does not contain data for every gene in every location.

3.3 Gene

Although EMAGE contains some information concerning genes, EMAGE cannot be described as a source of gene information.

EMAGE takes its knowledge of mouse genes from its sister resource at the MGI[11]. The information stored in EMAGE is a subset of the information located at the MGI.

Each gene at the MGI is given a unique identifier of the form MGI:*number*, e.g. MGI:88180[12]. Additionally each gene has, at least, the following information associated with its record:

[11] Mouse Genomics Institute (http://informatics.jax.org)

Name e.g. bone morphogenetic protein 4;

Symbol a.k.a. a short name, e.g. *bmp4;*

Synonyms e.g. *bmp-4*, *bmp2b*, *bmp2b-1*, *bmp2ba*;

GO terms from the GO ontology that classify the gene, e.g. cytoplasm, BMP receptor binding, … there are 202 terms associated with *bmp4*;

Links to related information in other resources, e.g. information about the corresponding protein from InterPro[13].

EMAGE duplicates the above information, and stores it locally.

3.4 Level

The terms used to describe the *level* of gene expression differ from resource to resource. In this section the focus is on EMAGE, and thus only its terminology will be reviewed.

Broadly speaking there are three categories of gene expression:

detected a.k.a. expressed;

possible

not detected a.k.a. not expressed.

Additionally, there are three subcategories of positive expression that may be employed: **strong**, **moderate**, and **weak**. In total that provides four levels of positive expression utilised within EMAGE; the term **detected** is used too.

Possible is to be used when the gene experiment's result (photo) is not clear, and therefore it is not easy to determine whether or not the gene is actually expressed. A second category of possible exists: **not examined** - this term should be self-explanatory.

Both of the **possible** terms indicate that the level of expression is not known, this is distinct from **not detected**, which informs the reader that the gene is not expressed in a location.

3.5 Location

As illustrated in Section 3.2 EMAGE has two possible means of describing location: a term from an EMAP anatomy ontology, and the EMAP 3D spatial models. Both will now be considered in more depth.

3.5.1 EMAP anatomy

As mentioned in Section 2.3.1, there is one anatomy (and corresponding ontology) for each of the 26 Theiler Stages of the developmental mouse.

Each term in the ontologies corresponds to a particular biological structure, and has its own unique identification of the form EMAP:*number*, e.g. EMAP:63. Gene expression information can be mapped to any of the anatomical terms; however, it is commonly mapped to middle and lower level terms. For example, ideally a gene would not be described as being expressed in the brain as that is too coarse grain. Instead one of the brain's substructures, including the direct descendants, would be used. The term chosen would depend on the biologist's confidence to accurately determine the structure(s), which would, in turn, depend on the quality of the image.

[12] To see the full MGI page for this gene go to

www.informatics.jax.org/javawi2/servlet/WIFetch?page=markerDetail&key=605

[13] www.ebi.ac.uk/interpro/IEntry?ac=IPR001111

The EMAP ontologies were developed using *part of* relationships. As a consequence of this, gene expression information needs to be propagated up or down the anatomy. To illustrate, consider the following two (inconsistent) textual annotations:

bmp4	strong	EMAP:64
bmp4	not detected	EMAP:64

a section of TS8 is reproduced here (indentation indicates a structure is *part of* the structure above it):

mouse (EMAP:25773)

 embryo (EMAP:57)

  ...

 extraembryonic component (EMAP:63)

  cavities (EMAP:64)

   extraembryonic component of the proamniotic cavity (EMAP:65)

   yolk sac cavity (EMAP:66)

  ectoderm (EMAP:67)

EMAP:64, cavities, is part of EMAP:63. EMAP:63 is part of EMAP:25773. Likewise, EMAP:64 has two substructures: EMAP:65 and EMAP:66.

The first annotation reveals that *bmp4* is expressed in EMAP:64. The *part of* relationships mean that positive expression must be propagated up the anatomy. Hence, the following annotations can be added:

bmp4	strong	EMAP:63
bmp4	strong	EMAP:25773

The second annotation suggests that *bmp4* is not expressed in EMAP:64. This time the *part of* relationships force the expression level to be propagated down the anatomy. This results in the following annotations:

bmp4	not detected	EMAP:65
bmp4	not detected	EMAP:66

The database only contains the basic, non propagated, annotations. The propagation is done at run time whenever a user queries EMAGE through the HTML interfaces.

3.5.2 Spatial models

When designing the RDF triple representation for the 3D spatial models it is important to appreciate the volume of information available. For example a simplistic pixel-based description of the spatial relations information of a 2MB image of 1920x1200 resolution and four directional relations (above, below, left and right) would result in over 5 trillion triples for a single image.

Furthermore, this image represents only the data for one of the 2000 genes EMAGE has information on. Additionally, the image is merely 1 slice through the 3D model; theoretically hundreds more exist. Finally, this discussion only relates to one of the many EMAGE models.

In total, EMAGE contains over 2TB of image data that may be converted for use in the CUBIST triplet store.

The complexity of the EMAGE 3D spatial models is considerable. Furthermore, the textual annotations will be the initial focus of CUBIST. Accordingly, this document will postpone the in-depth discussion of the models.

3.6 Data access and models

The data model encountered by a user will depend on the mechanism the user employs to extract data from EMAP/EMAGE. A series of alternatives seem, at least initially, viable for CUBIST. Each shall be discussed in turn.

3.6.1 IBM DB2

EMAP and EMAGE have their own instances in the repository. Design documents, and the actual schemas are available online:

EMAP

design document	www.emouseatlas.org/Databases/atlas/atlasDesign.html
schema	www.emouseatlas.org/Databases/atlas/atlas.mysql.ddl.txt

EMAGE

design document	www.emouseatlas.org/Databases/emage/design.html
schema	www.emouseatlas.org/Databases/emage/emage.db2.ddl.txt

Before a user can connect directly to either database instance they must register online:

https://www.emouseatlas.org/emage/secure/register.html

3.6.2 JAVA RMI

EMAP and EMAGE data can be accessed through the one Remote Method Invocation (RMI) interface. The interface is described online at the following URL:

www.emouseatlas.org/emage/help/emageAPI/index.html

In order to use the RMI interface, the user requires the EMAGE RMI jar file to be placed on their classpath. This is obtained by registering online[14]:

https://www.emouseatlas.org/emage/secure/register.html

3.6.3 Biomart

Biomart is a mechanism designed to provide bulk information quickly. It is accessed via the web interface, and unlike the previous mechanisms does not require pre-registration. Unfortunately, it does not provide access to the same breadth of information that the direct SQL and RMI methods do. Despite this, in the initial phases of the CUBIST project Biomart may provide sufficient data, and is thus worthy of consideration.

EMAGE's Biomart interface can be found at: http://biomart.emouseatlas.org/biomart/martview

The output of this tool can be a comma separated value file, or a tab separated value file. This makes pulling data through this mechanism considerably easier and quicker than the previously discussed routes.

Unfortunately, the interface to Biomart was not designed with usability in mind. Furthermore, a lack of online help necessitates that the user must already have a good knowledge of both Biomart and EMAP/EMAGE to use the tool with any degree of success.

Despite the apparent awkwardness of the interface, it is easy to learn. Additionally, Biomart provides access to the most commonly used datasets. Resultantly, Biomart can be a very powerful access mechanism.

[14] The one registration allows both RMI and direct SQL access.

4 Personas

Deliverable D1.1.1 (CUBIST Consortium 2010) described ”personas” as:

Personas are fictional persons, which represent typical users of the CUBIST system… They can be understood as imaginary characters with actual behaviour and goals, which are representative [of] larger user groups.

This chapter discusses the personas for the biological use case. Broadly speaking the users will be classified into personas according to their academic background – see Table 1. Not all personas will interact with CUBIST through its user interface, see Figure 9.

Persona	Background in …		
	Mathematics	Biology	Informatics
Biologist		Undergraduate & Ph.D	
Computational biologist	Undergraduate or Ph.D	Undergraduate or Ph.D	
Bioinformatician		Undergraduate or MSc.	Undergraduate or MSc.
Software developer			Undergraduate

Table 1: Typical academic background of EMAGE’s regular users and the way in which this affects the persona they are assigned to.

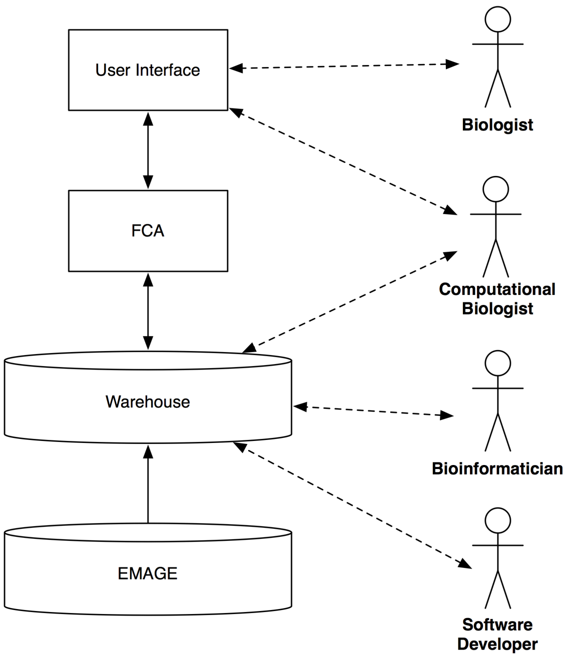


Figure 9 - Illustrating that the different personas will interact with different layers of CUBIST.

There are four classes of regular user, and each class will have its own persona. Each persona will be discussed in turn, starting with the biologist in Section 4.1

4.1 Biologist

There are effectively two classes of biologists associated with EMAGE: those that maintain the resource (so-called *curators*) and those that use it during their research. This persona focuses on the later group as it is considerably bigger than the former.

In the domain of *in situ* gene expression for the developmental mouse, biologists can either work for large scale organisations that conduct industrial scale research, or academic groups that conduct specialist experiments on focusing on a particular area of the body, or a particular family of genes. Again, the later group is selected.

In this scenario, it is assumed that the group is lead by an experienced academic. Her team comprises of two research fellows, a post-doc researcher, a couple of Ph.D students and a bioinformatician. This fictious persona will be based on one of the research fellows, rather than the group leader.

Katy works predominately in the so-called *wet labs* - she conducts actual biological experiments. Her research is directed by the group leader, a position she one day wishes to hold herself. Her university education is solely in biology; she has both an undergraduate and postgraduate degree in the subject. After finishing her Ph.D she worked as a post doc for her supervisor, before taking her current position as a research fellow.

Katy conducts *in situ* hybridisation gene expression experiments, thus she uses EMAGE in a number of ways.

Although Katy regularly publishes in journals, she uses EMAGE to publish her research online, as it enables a wider audience to become familiar with her work. Katy normally leaves the group bioinformatician to actually submit the research, merely checking over the information before it is finally submitted.

EMAGE allows Katy's peers to keep track of her research, and likewise helps her follow her peers. Katy is able to browse the resource via its web interface and thus examine what has been researched, and by whom.

Additionally, Katy uses EMAGE along with other, complementary, resources to answer specific biological questions. Existing data must be explored before the group can commit resources to performing experiments.

Katy's group focus on the process of metabolism. Therefore Katy uses resources that contain *pathway* information to explore the various different sub-process involved in metabolism. Katy prefers to use KEGG[15] for this initial exploration. When she has identified genes of interest, she drills down to discover more about them using ENSEMBL[16], ArrayExpress[17] and EMAGE.

ENSEMBL provides detailed low level information about the gene, for example its position in the chromosome. ArrayExpresses provides gene expression information, concentrating on the level of the gene expressed in a coarse-grain area of the body. EMAGE provides precise location information that complements the information from ArrayExpress.

Katy interacts with all these resources through their web interfaces; she realises that programmatic access is available, but she does not have the skill set necessary to utilise those interfaces. Katy finds the EMAGE interface usable, but wishes it provided greater analysis and summary features. However, her previous experience of ”fancy computer programs” has not been positive. Katy has

[15] www.kegg.com

[16] www.ensembl.org

[17] www.ebi.ac.uk/arrayexpress

noticed that computing people tend to make interfaces they can use, but seldom consider the needs of non-computing people, thus she wonders whether a more powerful interface would be usable.

4.2 Computational biologist

Computational biologists are researchers who explore biological relationships using computer science and mathematics. Steve studied biology to undergraduate level at university, and then took a Ph.D in mathematics. During his mathematical career he learnt about Hidden Markov Models, Evolutionary Algorithms and other forms of Computer Science.

Following his Ph.D Steve undertook several post doc positions before finally being offered his own group. He now leads a small team of three computational biologists, including one Ph.D and one post doc.

His group explore the relationships between various biological entities. Conventionally, Steve's team use the EMAGE data as part of an integrated data set. They process and analyse large volumes of data attempting to identify statistically meaningful patterns.

Although Steve enjoys active research, as group leader he is responsible for directing the research, identifying new areas of study, and obtaining funding. Consequently, he finds that he is doing less and less hands on work. Often he is now restricted to preliminary data discovery where he attempts an initial exploration of a hypothesis using whatever standard tools a resources provides. Unfortunately, he then has to leave the detailed work to his team. Unlike Steve, the team will connect to a resource using its programmatic access.

Steve has no problems using the online tools resources provide, but often wishes they were more flexible and offered a wider range of information. He craves raw data and the ability to manipulate and explore it in a wide variety of views. Steve is excited by CUBIST for two reasons.

Firstly, he believes it will provide him with a more detailed and more flexible way to navigate the data in EMAGE. He has often lamented the inability of EMAGE to help him identify anomalies in the data set, and the lack of a mechanism to study the expression level, location and pattern trends of a set of genes over a series of Theiler Stages.

The second reason for Steve's enthusiasm for CUBIST is that he believes the semantic representation of the EMAGE data will make it easier to integrate that data with complementary and competing data sources. This increased power should enable his team to be more productive, and combine resources in ways currently impossible.

4.3 Bioinformatician

Bioinformaticians commonly start out with an interest in either informatics or biology, and then take a conversion course that provides them with the necessary background in the other discipline. In Mike's case, his undergraduate degree was in biology, and he followed that up with a masters degree in bioinformatics. This degree taught him a broad range of informatics skills including software development.

Directly after his time at university, Mike joined the same research group as Katy (see Section 4.1) where he assists five biologists. In his view, both the informatics and biological communities are too insular, and neither are able to see a problem from the other community's perspective. Mike considers himself as a bridge between the biologists and the world of informatics.

His job involves doing any computing task that the biologists are unable, or willing, to do for themselves. This commonly involves the integration of data from different data sources.

Whilst the biologists are able to integrate resources via the web interfaces, that technique is time consuming and inefficient. Mike is able to automate that workflow for the biologist by creating an *in silico* experiment[18] using the workflow workbench Taverna[19]. Mike often finds that once a workflow is

[18] *In silico* experiments are experiments performed on a computer; see http://en.wikipedia.org/wiki/In_silico for more information.

created, he then has to run it too. The biologists simply send him some genes to feed into his *in silico* experiment, and tell him to email them the results once he has finished. Mike finds this part of the job tedious; however, he enjoys the challenge of creating the workflow because linking the resources often turns into a problem solving exercise.

Creating a workflow is difficult partially because the web services are badly documented, and partially because none of the services were designed to be linked. Regularly, Mike finds he needs to write a convertor to change the output of one resource so that it can be used as input to a second. This allows Mike to maintain his software development skills, another aspect of his job he enjoys.

Mike hopes that the CUBIST semantic warehouse will make it easier for him to understand, and thus integrate the EMAGE data with other data sources.

4.4 Software developer

The INCF[20] helps organise Neuroinformatics researchers across the world by providing a framework for collaboration. Although most people involved with the INCF are researchers (or research groups) that are funded by government grants or philanthropy, a small number of people are directly employed by the INCF - Brian is one of this number.

Brian took a Bachelor of Engineering degree in software engineering at university. As part of his degree he did a final year project; the supervisor of this work was involved in the INCF program. Following graduation Brian took a job as a research associate, working for his project supervisor. Although he has no biological training, Brian enjoyed this realm, and worked hard. After a successful spell, Brian was asked to join the INCF as a software developer. He is now based in Stockholm but collaborates with bioinformaticians, and computational biologists all over the world.

Brian's current work centres on the INCF's Digital Atlasing[21] activity. This aims to integrate various mouse brain 3D atlases including the Allen Brain Atlas[22] (ABA), and EMAGE (although EMAGE is a whole body atlas, just the brain is included in this work).

Ultimately, the goal is to be able to write one query that will be able to produce results for all underlying atlases. To achieve this, the underlying resources must be mapped. Currently there are two ways of doing this:

> **spatial tansformation** converts an X,Y,Z coordinate in one spatial reference system (e.g. EMAP) to an X,Y,Z coordinate in a second spatial reference system (e.g. ABA);
>
> **ontological mapping** links the terms in one anatomy ontology to the terms in a second anatomy ontology.

Unfortunately, neither technique is reliably accurate. The possibility of a third mechanism, a semantic spatial description, interests Brian and his colleagues. Unfortunately, none of the resources they work with describe their location information semantically. However, the forthcoming SPARQL endpoint for EMAGE, created as a side product of CUBIST, provides the possibility of investigating semantic mappings for the first time.

[19] www.taverna.org.uk

[20] International Neuroinformatics Coordinating Facility (www.incf.org)

[21] www.incf.org/core/programs/atlasing

[22] www.brain-map.org

5 Utilisation scenarios

Utilisation scenarios are designed to present a common task for a single persona. Deliverable D1.1.1 (CUBIST Consortium 2010) described them as follows:

Utilization scenarios in turn represent typical days of these personas, thus they help the developers to understand the environment and processes the personas usually deal with. They are fictitious days in the life of the personas and describe in a story-like manner typical daily activities from the viewpoint of the personas.

In this use case the standard usage is not applicable. The justification for this claim is the overlap between a number of personas. For example, the computational biologist and a regular biologist would use EMAGE in the same way to answer the same questions: both use EMAGE to determine if it is worth allocating the resources to do in-depth research. The difference occurs in the background of these users and in the interfaces they expect. The same will be true for the software developer and bioinformatician too.

The differences in background were recounted in Section 4, and the interface requirements that distinguish these users will be dealt with during the mock up and atomic requirements (see Section 6). Accordingly, in this section, there will only be one scenario for both the biologist and computational biologist - the scenario will be written for Katy (the biologist), but is equally applicable to Steve (the computational biologist). Likewise a single scenario for Mike (the bioinformatician) will represent both him and Brian (the software developer).

For each persona grouping there will be a short introduction that leads into the following pre and post CUBIST scenarios. The ”before CUBIST” scenario describes what each persona currently does, and what issues they encounter when undertaking their tasks. Subsequently, a ”post CUBIST” scenario will present an idealised version of events following the introduction of CUBIST.

5.1 Biologist and computational biologist

Katy's research focuses on a particular subprocess of metabolism. Following an investigation of KEGG she has identified a set of genes which she feels merits further investigation. Querying ArrayExpress she discovers these genes are expressed somewhere in the ALIMENTARY SYSTEM. Katy wants more precision regarding the location. For this purpose Katy likes EMAGE. She is able to see *exactly* where the genes are expressed; after all, seeing is believing.

Katy reduces her list of genes to include only those genes that ArrayExpress suggests have high levels of expression in the ALIMENTARY SYSTEM. Then she goes to EMAGE and enters her list, and the range of Theiler Stages she is interested in. After a few minutes she receives her first page of results…

5.1.1 Before CUBIST

The standard EMAGE results page is displayed: a table, where each row is a single experiment that deals with a single gene and a single stage. As she scans the page her heart sinks when she realises it is the first of 20 pages. With 10 experiments per page, that is 200 experiments.

EMAGE only provides the ability to sort the table, for example by Theiler Stage or gene symbol. With no mechanism to group or filter the results, Katy will have to go through each experiment one by one.

Katy sorts the results by gene, and starts to go through the experiments for the first gene. She looks at the result image displayed for each experiment to determine if it shows the gene being expressed in appropriate location. She ignores those experiments that do not show the gene being expressed in approximate region of the alimentary system. Not for the first time, Katy wonders why you cannot search jointly for a gene and tissue.

After reading through the results, and noting down the results on her pad, Katy runs a second query asking where her list of genes are not expressed. Ideally one query would provide all this information, but unfortunately EMAGE does not provide that functionality. Again Katy runs through the results, making notes on her pad.

Once completed, she realises that one gene is shown to be expressed in, what she thinks is, the DORSAL MESENTERY in stages 16 and 18, but is not expressed in stage 17. This so-called ”flip flop” is extremely unusual, and is most likely a sign that the experiment on TS17 is incorrect. Katy quickly inspects the relevant experiment and, noticing that the probe sequence is missing, decides to ignore both it and its annotations.

Moving onto her notes for the next gene, Katy discovers that it is described as being expressed and not expressed in the same location at the same time. Sighing, she starts to read the web pages for both experiments in order to determine which is accurate. Once her decision is made, she makes a note of the correct annotation on her pad.

After hours of reading web pages, and updating the notes on her pad, Katy has determined the exact location (in the alimentary system) of strong expression for her entire set of genes. With this information safely recorded on her pad, alongside the list of relevant experiments, Katy goes to discuss her findings with the rest of the group.

During the short weekly group meeting, someone asked about a gene they were not familiar with. Unfortunately, Katy was unable to answer the questions and now returns to EMAGE to determine where (and when) else the genes co-expressed[23] in the alimentary system are co-expressed. This time Katy uses the ”search by anatomy name” facility, and enters the term ALIMENTARY SYSTEM. She is presented with a table which contains a row for every experiment that suggests a gene is expressed in the alimentary system. It occurs to Katy that the ability to search by experiment identifier would be useful, unfortunately, this functionality is not available[24].

Instead, Katy sorts the table by gene symbol, then turns the pages until she arrives at the pages containing information on her genes. She quickly scans the experiments, looking for those she earlier recorded on her pad. When she finds a match, she examines the experiment in more detail. Studying the result, she records a list of other locations in which the gene is expressed. After an hour of study she is able to spot a pattern, and goes off to discuss it with her colleague.

5.1.2 After CUBIST

When EMAGE has processed Katy's query, rather than presenting a ream of raw information, EMAGE presents Katy with a number of methods for examining the data further. Some provide a broad analysis of the data, whereas others drill down into specific details. Furthermore she is able to filter the results in a number of ways to reduce the data she needs to sort through.

Katy likes to filter the result once she sees how large it is. From experience Katy knows that if she adds the filter to the query she is likely to received a very limited, or possibly empty, result set. On this occasion, over 200 answers are returned, so Katy decides to filter the data further. She decides to exclude experiments that have a ”data pattern clarity” of less than 2 and spatial annotations that have a ”morphological match” of less than 2. She decides to play with the reduced data set before deciding if further filters are necessary.

The first thing Katy does is analyse the trend information for each gene. Katy's particular focus is the expression level: initially she sets the visualisation to display binary gene expression information (on or off). She is checking to see if CUBIST managed to identify and automatically correct the flip flops. Pleased to see that the ”magic” worked, she quickly browses the list of errors (flip flops) CUBIST detected.

Seeing nothing wrong, Katy switches to the analogue ”expression level view” and scans the visualisation. Identifying the points of high expression, Katy drills down to discover which locations they correspond to. It occurs to Katy that she did this process in reverse last week, and smiles as she realises how much easier her life is with the flexibility of the new interface.

[23] When genes are expressed in the same location they are said to be co-expressed.

[24] It can be done programmatically, but Katy does not operate at this level.

The ”location view” makes it easy for Katy to concentrate on the genes found in the alimentary system. Looking at one particular location, CUBIST asks for her help - apparently there is an inconsistency that cannot be automatically resolved. CUBIST wishes Katy to resolve the inconsistency; she does so happily knowing that her input will assist anyone who runs a similar query.

With her work done, Katy glances at the clock - over an hour till the next group meeting. She decides to investigate a particular gene further. She has the option to explore the gene by: the location of expression, the GO terms it is annotated with or the genes it is co-expressed with. Initially, she looks at the GO terms the gene has, then asks for genes with similar terms - she sets a filter to ensure that the returned set of genes have at least 3 terms in common with the original gene.

Momentarily the query is complete and a set of 12 genes is displayed. Katy could analyse these individually but instead chooses to treat them as a group. She asks where they are strongly co-expressed, setting the minimum number of genes in each co-expressed group to 4. On receiving a disappointedly low number of results, Katy weakens the expression level so it can be strong or moderate. She finds that four of these genes are expressed in the stomach's EPITHELIUM. As her colleague Bill walks past, she hurries off to discuss her findings with him.

5.2 Software developer and bioinformatician

The INCF framework for neuroinformatics integration effectively uses a central hub, and standardised spatial model, to mediate across a range of alternative resources. Mike has been asked to implement the EMAP hub, according to a specification drawn up the by INCF Digital Atlasing Taskforce[25].

As the textual annotations report gene expression information at the structure level, they are too coarse for the current task. Resultantly, Mike requires access to EMAGE's spatial annotations …

5.2.1 Before CUBIST

Mike examines the online EMAGE documentation before deciding to use the RMI interface to access the spatial annotations. As a rule he prefers direct access to the database where possible; however, in this case the image analysis and manipulation techniques built into the RMI interface mean he does not have to process the spatial annotations himself.

After familiarising himself with the RMI interface Mike begins to create the operations specified in the interface document. His colleagues at EMAGE have already provided him with a spatial transformation for EMAP Theiler Stage 23; a mathematical model for mapping between the INCF's standardised spatial model and a particular EMAP model. Following a lengthy test process, Mike declares the EMAP hub ready for outside testing.

During the weekly INCF Digital Atlasing Task Force teleconference Mike gives an impromptu demo of the EMAP hub. One of the neuroscientists on the call queries the results of an operation. Exploring the details of the output in more depth, the neuroscientist states that the spatial transformation requires further investigation.

In the subsequent few days Mike collaborates with a few of the neuroscientists that attended the teleconference in an attempt to precisely determine what the issue is. Eventually, the group arrive at the realisation that the transformation model has variable accuracy - it works extremely well for certain regions of the brain, less well for other regions, and badly for the remaining regions.

Reading the available literature on the field, it transpires this a common flaw with spatial transformations. The only solution is to develop multiple models, with each model mapping a specific region(s) of the brain. Unfortunately, the process of creating transformations is both time consuming and expensive. Mike accepts that he will have to wait to see if EMAGE are able to produce a series of alternative transformations.

In the meantime, Mike attempts to extend his hub to work with Theiler Stage 16; however, he quickly stops when he realises that he does not have the spatial transformation for that stage. He sends an

[25] www.incf.org/core/programs/atlasing

email to his counterpart at EMAGE requesting they investigate this transformation. Knowing it will be sometime before EMAGE can respond he moves onto another project.

5.2.2 After CUBIST

During the CUBIST project raw and processed EMAGE data was stored in the project's semantic warehouse. Additionally, a SPARQL endpoint to EMAGE was developed.

The pre-processed spatial annotations stored in the warehouse are immediately attractive to Mike. Studying the resource's online documentation he can immediately see how to gain access to the data, and develops suitable client code.

Mike uses the spatial descriptions of EMAP space to map directly into the similarly described INCF spatial model. After testing his code to ensure it works, he contacts a group of neuroscientists to verify the accuracy of his new semantic spatial mappings. Although it is clear that the spatial transformation is superior in areas where it works well, overall the semantic mappings are more reliable.

Mike updates his EMAP hub to use this new mapping method, replacing the spatial transformations in areas where they are less reliable. Additionally, Mike extends the hub to work with stages, such as Theiler Stage 16, where there are no spatial transformations.

Pleased with his work, Mike publishes his client to the web, and send an email to mailing list to tell everyone about the improvements to the EMAP hub.

6 Formal requirements

As discussed in D1.1.1 (CUBIST Consortium 2010), the atomic requirements for this use case were captured using a cut-down version of the Volere Snowcard[26].

In particular, only eleven classes of requirement were considered:

Purpose – very high level

Mandated – absolutely necessary;

Functional – what the product should do;

Data – what data (sources) are required;

Look and feel – appearance;

Usability and humanity – ergonomics;

Performance – response time, volumes of data to be handled, etc.;

Operational – partner with applications and platforms;

Maintainability and support – how the product should be maintained;

Security – how should data integrity and safety be maintained;

Legal – standards system must adhere to.

Although the full details of the atomic requirements are contained within the Volere Snowcard, they will be summarised in this document.

In each of the subsequent tables the first column is the requirement ID, the second contains a description, an explanation is found in the fourth column, and the final column suggests how to test that the requirement is met.

ID	Description	Explanation/justification	Fit criterion

The requirements will now be presented, separated by their class.

6.1 Purpose of project

HWU001	Provide analytical features on top of existing data	Lots of data, yet no analytical features	CUBIST provides a series of analytical features for EMAGE data

6.2 Mandated constraint

HWU010	Must be web based	Many biological users will not control their computer & thus cannot install software	CUBIST runs in a web browser

[26] www.volere.co.uk

HWU011	Must not require installation of plug-in, extension etc. [Note: modern browsers often come pre-packaged with Flash[27], using this is fine]	Many biological users will not control their computer & thus cannot install software.	CUBIST runs perfectly in a newly installed browser

6.3 Functional constraint

HWU021	Results include notion of expression level propagation for textual annotations	Propagating textual annotations (up & down) is standard practise	Textual annotations not in raw data are included in results
HWU022	Locate & fix errors/inconsistencies in underlying data	Biological data is naturally inconsistent & incomplete. Users find it hard to deal with these issues	Suggests inconsistency "solutions" to user
HWU023	Display trend information for expression level	Level changes over time	Ask human expert to verify three examples
HWU024	Display trend information for expression pattern	Pattern changes over time	Ask human expert to verify three examples
HWU025	Display trend information for co-expression level	Co-expression changes over time	Ask human expert to verify three examples
HWU027	Identify genes with similar expression patterns	Provide spatially orientated co-expression information	Ask human expert to verify three examples
HWU028	Describe similarities of genes involved in same process	Indicates what other processes, functions etc. the genes may be involved in	Ask human expert to verify three examples
HWU030	Queries & results may be saved & reloaded later	Allows users to return to previous state	A test query can be exported, then later imported successfully
HWU031	Filter results	Users may find it helpful to deal with a subset of the full result set	Result set can be reduced by setting a range of parameters
HWU032	Flexible presentation of results	Different personas have different computing background & experience	User can switch to different visualisation of same information
HWU033	Link back to main EMAGE page	If users wish to drill down, they should be referred to EMAGE web pages	CUBIST links out to EMAGE web pages
HWU034	Expanded means to investigate genes	Currently users forced to query in specific ways only	Can look by gene, gene + location… with or without a time restriction
HWU035	Expanded means to investigate by location	Currently users forced to query in specific ways only	Can look by location, gene + location… with or without a time restriction
HWU037	Expanded means to investigate by co-expression	Currently users forced to query in specific ways only	Can look by co-expression alone, co-expression in set of particular locations, co-expression for particular genes, co-expression for particular genes in a set of locations, co-expression for particular GO terms… all with(out) a time restriction

[27] http://www.adobe.com/products/flashplayer/

HWU039	Compare multiple experiments	Useful to go back and look at a range of experiments; comparing and contrasting their contents	Can enter multiple IDs and get the same presentation as HWU038 for each one

6.4 Data requirement

HWU040	CUBIST warehouse must include textual annotations	Core dataset for CUBIST use case	See description
HWU041	Repository should include EMAP anatomy ontologies	Required for textual annotations – HWU040	See description
HWU042	Repository should contain EMAGE controlled vocabulary	Required to understand data in EMAGE; e.g. distinction between moderate & present	See description
HWU043	Any inconsistencies/errors identified should be reported to EMAGE curators	Curators are responsible for maintaining accuracy of resource	See description
HWU044	Warehouse should be updated inline with regular EMAGE updates	EMAGE contents constantly changes; new versions are released quarterly	
HWU045	Repository should include spatial annotations	Core dataset for EMAGE	See description
HWU046	Repository should include spatial representation of mouse	Core dataset for EMAGE	See description
HWU047	Repository should include EMAGE's pre-computed spatial clusters	Required for HWU032	See description

6.5 Look and feel

HWU051	Interface should be clean and uncluttered	Makes interface easier to use	
HWU052	Looks like prominent web app	Mimicking resource will make CUBIST seem friendlier & easier to use	See description

6.6 Usability and humanity requirement

HWU060	Use biological metaphors when designing interface	Biologists understand biological metaphors but not computing ones	
HWU061	Provide constant feedback to users	Biologists are often not computer experts & need lots of support. Feedback improves the transparency too.	Users are able to describe what CUBIST is doing whilst using the system
HWU062	Flag data when unsure of auto correction (HWU022)	System needs to be transparent and provide constant feedback to users	Users can tell when CUBIST has changed the underlying context (e.g. with fault tolerance)
HWU064	Users will be able to use system once they have watched the screencast	Users are unwilling to read manual or work through a tutorial	Users can use CUBIST when their only introduction is the screencast
HWU065	Users should only need to watch screencast once	Users are unwilling to use tools that seem hard to use	Users can use CUBIST 2 weeks later without any help or

			reminders
HWU066	Use standard web metaphors	Users are familiar with web	
HWU067	Make users aware of progress	Users must realise that system is working	
HWU068	Interface (around visualisations) should be simple	Biologist are not comfortable with complex interfaces, and traditionally ignore them	Evaluation shows biologists describe the interface as "simple"
HWU069	Visualisations should be detailed and convey large amounts of information (persona: computational biologist)	Computational biologists want quick access to lots of information; and have background in mathematics so are used to dealing with complex graphs	Computational biologists indicate all "key" information is in visualisation
HWU070	Visualisations should be simple (persona: biologist)	Biologists are not used to dealing with lattices and other forms of visualisation CUBIST is likely to employ	Biologists can interpret visualisations

6.7 Performance requirements

HWU080	80% of queries must have results displayed to user in less than 50 seconds	Web based interfaces need quick response times, or they appear to be broken	See description
HWU081	When results cannot be displayed in less than 50 seconds, should provide option to email results to user	Many biological tasks are computationally expensive. Standard procedure is to email user when finished	See description

6.8 Operational requirements

HWU090	Must not rely on windows technology	Significant number of linux & mac boxes in use	5 test queries can be run using linux & mac machines
HWU091	Works & looks the similar in major browsers (with Flash support). No mobile browsers targeted	Significant number of users use firefox, chrome, safari & IE browsers	CUBIST looks the same in all 4 browsers; works for 5 test queries

6.9 Maintainability and support requirements

HWU100	Pipeline to important more data automatically	EMAGE data is continually expanding	CUBIST contains latest version of EMAGE data set
HWU101	Pipeline to automatically import new data will only import new/amended data	Pulling entire dataset every time is too expensive	See description
HWU102	Maintenance tool should be online	Allows management of EMAGE specific aspects by EMAGE people, even if run on server elsewhere	An EMAGE admin can control the system remotely
HWU103	System should provide full manual describing maintenance options	Allows management of resource by EMAGE staff	An EMAGE admin can change the system

6.10 Security requirements

HWU110	Only system admin & CUBIST developers have write access to data in CUBIST warehouse	Data should not be changed by anyone other than trusted admin	See description

HWU111	CUBIST should not attempt to change original data source	Data should not be changed by anyone other than EMAGE admin	CUBIST will only have read access anyway

6.11 Legal requirements

HWU120	Information provider (i.e. researcher & journal) must be credited somewhere on screen	Standard practise to acknowledge knowledge creator; also, some experimental results (images) are copyright	A user can determine who performed the experiment
HWU121	CUBIST software should be open source to allow it to be used/extended by EMAGE after project ends	Technology in CUBIST may prove useful, and EMAGE wish to be able to use and extend it after project ends	All code and documentation is publically available, and licensed for public use
HWU122	Open access to EMAGE SPARQL	EMAGE is a public resource funded by the UK tax payer	Any interested party can use the EMAGE SPARQL query mechanism
HWU123	Open access to EMAGE data in CUBIST warehouse	Registration provides access to the warehouse for interested parties	Once registered, users have full read access to EMAGE data in warehouse

7 Bibliography

Baldock, R, and D Davidson. In Anatomy ontologies for bioinformatics: principles and practise, by A

Burger and R Baldock, 249 - 265. Springer Verlag, 2008.

Campbell, N A., et al. Biology. 8th Edition. Pearson Benjamin Cummings, 2008.

CUBIST Consortium. D1.1.1 - Objectives for the requirements analysis in the use cases. Project Deliverable, CUBIST Consortium, 2010.

pathway. Dictionary.com. Merriam-Webster's Medical Dictionary. Merriam-Webster, Inc. http://dictionary.reference.com/browse/pathway (accessed: February 18, 2011)

Theiler, K. The house mouse - atlas of embryonic development. Springer Verlag, 1989.

Venkatarman, S, et al. “Edinburgh mouse atlas of gene expression: 2008 update.” Nucleic Acids Research 36 (2007): D860 - D865.