G Health Sciences

Johanna McEntyre and Alma Swan

1 Introduction and methodology

1.1 Introduction

This chapter provides an overview of research data management in the health sciences, primarily focused upon the sort of data curated by the European Bioinformatics Institute and similar organisations. In this field, data management is well-advanced, with a sophisticated infrastructure created and maintained by the community for the benefit of all.

These advances have been brought about because the field has been data-intense for many years and has been driven by the challenges biology faces. Science in this area cannot be done on a small scale: it is effectively a collaborative effort where data must be shared for any advances to be made. This has long been acknowledged. The HUGO (Human Genome Project) set the standards, because the demands of that project were so great that only a concerted effort across the whole genome science community would enable the achievement of that goal. It established new norms of scientific behaviour in this discipline and has influenced cultural developments in the discipline ever since.

The human genome is now long-decoded, but today's scientific questions in health sciences are no less challenging. The infrastructure, practices, standards and norms established in the life sciences can be viewed as good practice markers for those who wish to learn from what has gone before. Not everything practised in the life sciences will read across to other fields and disciplines, but many basic principles of research data management practice have been established that will transfer readily elsewhere. Perhaps most importantly, the life sciences have now reached the stage where the issues of long term planning, organisation and sustainability are now being tackled. The answers to these things are only partially worked out as yet, but some

fundamental principles are being elucidated and these will be useful in a more general sense.

1.2 Methodology

The material in this chapter was developed by the following means:

- Literature review and analysis
- Semi-structured interviews with experimental and theoretical scientists
- Observational studies of experimental biologists working at EBI, The Sanger Institute and in a number of universities in the UK

2 The European Bioinformatics Institute: an overview

EBI (the European Bioinformatics Institute) is part of the European Molecular Biology Laboratory (EMBL). It was established in Hinxton, UK, in 1994 to build on EMBL's pioneering work in providing publicly-available biological information in the form of databases to the scientific community.

Such information was beginning to accumulate rapidly as molecular biology technologies created increasing amounts of data. New skills and resources were required to collect, curate and store these data and present them to the research community through reliable, professionally managed channels.

From small beginnings, EBI has grown and now provides data resources across all molecular biology domains. It hosts a number of public databases, most through collaborative initiatives with partner organisations throughout the world, particularly in Europe, the US and Japan. Services include Ensembl (a genome database), ENA, the European Nucleotide Archive (containing DNA and RNA sequences), UniProt (containing protein sequences) and PDBe, the European arm of the Protein Data Bank. The expression data is captured by ArrayExpress Archive which is a database of functional genomics information and the Gene Expression Atlas, which contains expression data from Array Express that has been re-annotated for particular purposes. More recently, the EBI has developed CiteXplore, a database of biomedical abstracts from research articles and patents, and UKPMC (UK PubMed Central), a full-text article database. This source is used actively for gathering information from literature to create information-rich databases such as GOA (Gene Ontology Annotation) and IntAct, and for text mining information.

These and other services, such as ChEBI (Chemical Entities of Biological Interest), Reactome, and InterPro, offered mean that EBI is the primary

provider of biological information in Europe, and one of the major global providers.

Integration and distribution of information from and to various sources requires standardisation of the data input, storage and distribution. EBI scientists have been active in developing or contributing to the community efforts towards the development of international standards for use in bioinformatics. Two examples involving EBI are the MIAME standard for microarray experiments (Minimal Information About a Microarray Experiment) and the Human Proteome Organisation's Proteomics Standards Initiative (PSI).

EBI plays a coordinating role in Europe with respect to bioinformatics work, such as the ELIXIR (European Life sciences Infrastructure for Biological Information) objective to fund and maintain a world-class infrastructure for life science information management in Europe; ENFIN (Experimental Network for Functional Integration), an initiative to bring together experimental and computational biologists to develop the next generation of informatics resources for systems biology; IMPACT (Experimental Network for Functional Integration), developing infrastructure to improve protein annotation; and the SLING Integrating Activity (Serving Life Science Information for the Next Generation) which aims to bring together a wide range of information sources and services and help them to keep pace with scientific developments.

EBI also has a substantial programme of research. Research groups collaborate in experimental areas such as genomics, developmental biology, protein structure, evolutionary studies and cellular interactions, amongst others. EBI also contributes to research in the areas of computational biology and the development of simulation and modelling systems and of mark-up standards for biological data.

A further area of operation for EBI is training in bioinformatics, providing very active international PhD and postdoctoral training programmes for young researchers aiming to become bioinformaticians.

Additionally, EBI runs training programmes for users of biomedical data to equip experimental biologists with the skills needed to best use the information resources that EBI provides. An e-learning programme is being developed to complement the face-to-face training programmes.

EBI also coordinates the Bioinformatics Roadshow, a mobile training programme run in collaboration with the Swiss Institute for Bioinformatics, the European Patent Office and the BRENDA project (BRaunschweig *ENzyme* Database, an initiative of the Technical University of Braunschweig).

Up-to-date information on the Institute's activities can always be found in its Annual Report[1].

3 The health sciences

3.1 The scope of research activities in health sciences

Health science is a very broad area. It spans some elements of environmental science at one end of the spectrum through biomedicine, clinical medicine and veterinary science to medical physics and mathematical biology. Health-related questions and issues are studied at multiple levels.

At the molecular level, researchers study biomolecules and their activities and interactions in fields such as genomics (the study of the genetic complement of organisms), transcriptiomics (the functional transcript of the genetic component), proteomics (the study of the structure and function of proteins), metabolomics (the study of small molecules, metabolites, that are generated by living systems), macromolecular structures and interactions, and bioinformatics (the application of computer methodologies and statistical analysis to the study of molecular systems). The above resources are used to study network biology and regulation of biological systems, to eventually give us an understanding of systems biology.

At the next level – the study of cellular processes and behaviour – research is aimed at elucidating the ways in which cells and tissues interact and influence others and how cellular systems are regulated. Scientists also work at understanding how these systems relate to known molecular pathways and events and what can lead to cell and organ dysfunction. Included in this area of research activity is the study of mechanisms that form the basis of disease. *In vitro* (experimental) model systems are developed for human and animal diseases and malfunctions in order to try to understand what processes are aberrant in disease conditions. In addition, *in vitro* or computer model systems are used to research potential therapeutic agents and to test for toxicity.

At whole-organism level research areas include infection and immunity (encompassing the areas of immunology, microbiology and pharmacology); the wide-ranging clinical medicine disciplines; therapeutics and translational research; transplantation and regeneration; toxicology and environmental health; public health; and aging and wellbeing research.

With such a broad-scope area as health sciences, research activities are necessarily hugely varied. Research can be, at one end of the scale, devel-

[1] http://www.ebi.ac.uk/Information/Brochures/pdf/Annual_Report_2010_low_res.pdf.

oping advanced instrumentation for clinical therapies to, at the other end, sequencing the mutated gene responsible for a very rare disease. At each of these steps, however, a consolidated source of information is very useful and this can be provided by various sources made available at EBI.

3.2 Types of research activity and the main experimental and theoretical methodologies used

For the purpose of this exercise we are not attempting to cover the whole gamut of health science research areas. Instead, we focus on the research that looks at biochemical and sub-cellular processes, along with any allied approaches this may entail.

The areas of focus therefore include: genomics, proteomics, and metabolomics; macromolecular structures and interactions; cellular structure, function and signalling; and bioinformatics.

A major component of the above approaches is the analysis and production of data on a large scale, which, in the biomedical sciences is a process facilitated by the availability of public databases for both data deposition and retrieval for analysis. The data resources at the EBI, for example, have grown dramatically in recent years (see Figure G.1 below), This has led in some cases to "data-driven science" (PMID: 14696046) in which analysis of large public datasets gives rise to new hypotheses.

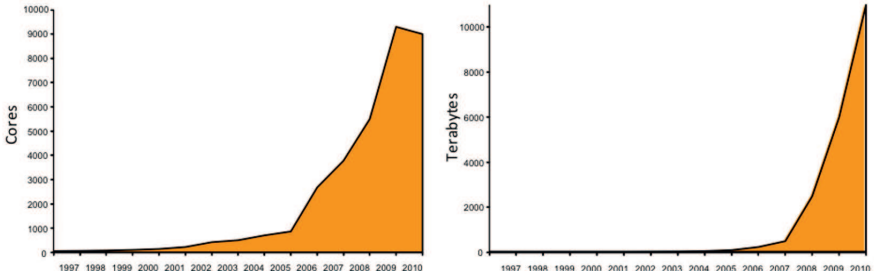

Figure G.1 Growth of compute and storage of data at EBI

Other experimental approaches that may be used in biomedical investigations include light and electron microscopy techniques (which produce images), scanning techniques (which also mainly produce images), and biochemical analytical techniques such as nuclear magnetic resonance and chromatography (which produce text or data files after computation of the machine analysis.

Theoretical approaches to research in this field include second-level informatics and modelling. These use computational techniques to further process data from experimental procedures. Models may be of a number of types, including mathematical models, computer simulations, computer models and 3-dimensional models. The development of integrative technologies is an area of considerable research focus: researchers are creating a wide range of data-integrative algorithms that enable combined analysis of diverse data sources.

These methodologies are described in more detail in section 4.1.

3.3 Types of research output and the way they are used

The research activities that are described here produce outputs of the following types:

(i) *"Big data".* The deposition of data in public databases such as those provided by the EBI, are the end-point of some experiments but increasingly provide starting points for others. The deposition of data in public databases is increasingly a requirement of journals for publication. Storing data in centralised databases with uniform format and structure allows the development of computational tools for comparative data analysis (e.g. BLAST) and in-depth search and display mechanisms[2]. The figure below shows records available in some key public data resources maintained at the EBI:

(ii) *Research-lab generated datasets.* Experimental data are exploited first by their creators (researchers) and may:

a. remain in the possession and care of those creators. In these cases, researchers may elect to share datasets with enquirers or with the research field at large, perhaps via a public website or service (see section 6).
b. get deposited in the public databases (often a requirement of journals) (see section 6.2) or
c. appended as supplemental data files attached to research articles published in peer reviewed journals (see section 6.3)

(iii) *Research articles in journals.* In health sciences, journal articles are the primary output type for research findings, contrasting with some other fields like computer science and engineering where peer-reviewed conference proceedings are the main dissemination channel. There are several thousand peer-reviewed journals covering biomedicine so finding an outlet for publication is not especially difficult.

Keeping up with the literature in this field is, however, challenging given the volume of papers published in each year. Probably half the total research

[2] See, for example: http://www.ebi.ac.uk/Tools/sss/

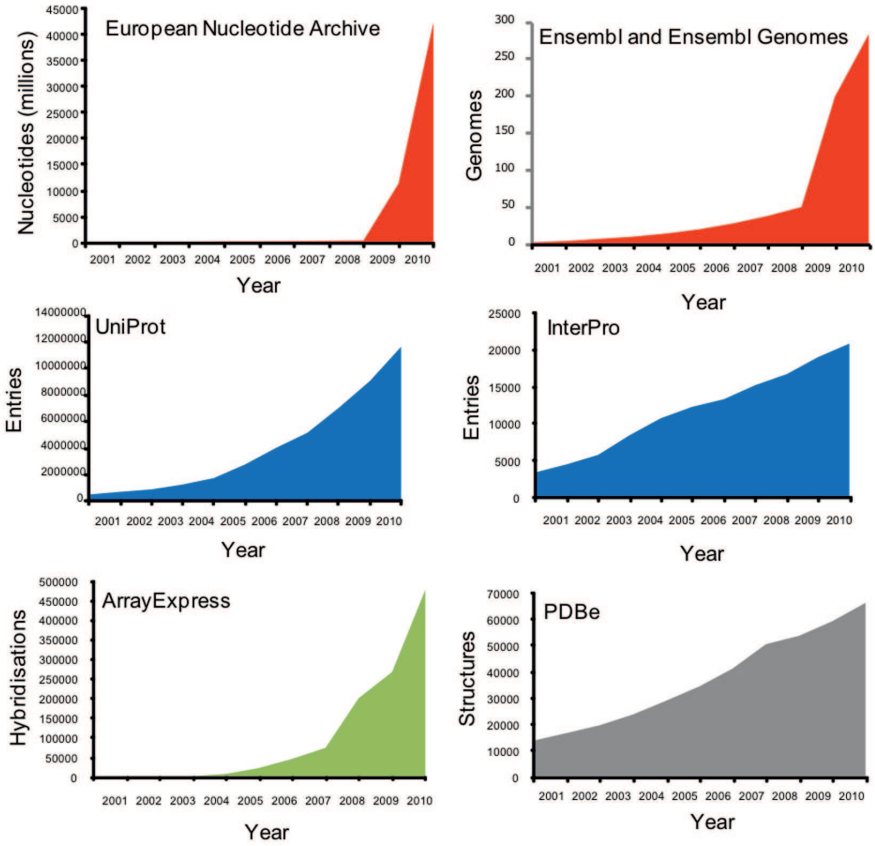

Figure G.2 Growth of key resources at EBI

literature is in health sciences, reflecting the priority that research in this area has for society and the levels of funding received from governments and other research funders.

Journals in the discipline are published by commercial publishers, medical charities, learned societies, medical institutions, and universities and research institutes.

(iv) *Conference papers.* Peer-reviewed conference proceedings are not common in health sciences, but they form an occasional outlet for research findings.

(v) *Outputs through more informal channels such as blogs, wikis, open notebooks and similar.* In recent years there has been growth in the use of

online sites for the management of projects and the dissemination of materials from them. Especially in fields where the research cycle is short and progress is very rapid, such informal channels may be the best way to alert the community to new developments and findings.

Occasionally, laboratories do disseminate pre-publication results from analytical machine runs via wikis or blogs, for early communication through these routes. Access to data generated by other people's work may therefore be facilitated in this way and some bioinformaticians hunt for and harvest data from such websites for their own meta-analysis, sometimes by screen scraping. The data so obtained , however, do not have the parsable format that would make them more amenable to re-use and distribution.

One of the most successful and well-known examples of a community 'Web 2.0'-type facility is Open WetWare[3], a site that hosts individual laboratory websites and on which users share results, protocols, details about materials and so forth.

3.4 Workflows in life science research

3.4.1 Genomics

Gene sequencing experiments

- Experiment planning: Define experimental goals; identify source of sample(s); agree experimental conditions; plan and prepare for use of experimental machines; plan data handling procedures
- Experimental process: prepare samples and carry out machine run (on one or multiple samples)
- Data production: machine produces raw data (traces) and processed data (text-based outputs)
- Data storage and preservation: Optionally, store preliminary data from machine ('raw' pre-base called data) for future analysis or further processing if ever necessary: routinely, store the text-based base-called data
- Data quality checking: carry out manual quality check; discard datasets with obvious errors
- Data processing and enrichment: process data; annotate datasets where appropriate
- Data publishing and storage: complying with agreed standards, (e.g. MIARE, Minimum Information About an RNAi Experiment or MIAFGE, Minimum Information About a Functional Genomics Experiment) deposit representative dataset(s) in public databank (e.g. Genbank)

[3] http://openwetware.org/wiki/Main_Page

- Data analysis: capture relevant datasets from Genbank for computational analysis using preferred software
- Re-submit processed datasets if the data have been improved in some way
- Susceptibility or resistance to infection can be provided by SNP (single nucleotide polymorphism) or variation analysis or epigenetic analysis of the genome

Microarray experiments

- Experiment planning: Define experimental goals, identify source of sample(s); agree experimental conditions; plan and prepare for use of experimental machines; plan data handling procedures
- Experimental process: prepare samples and carry out machine run (on one or multiple samples)
- Data production: machine produces raw data (images)
- Data processing and enrichment: normalise raw data; statistically analyse data
- Data publishing and storage: complying with agreed standards (e.g. MIAME, Minimum Information About a Microarray Experiment), deposit representative dataset(s) in public databank (e.g. ArrayExpress)
- Data analysis: capture relevant datasets for computational analysis using preferred software
- Re-submit processed datasets if the data have been improved in some way

3.4.2 Proteomics

- Experiment planning: Define experimental goals, identify source of sample(s), agree experimental conditions, plan and prepare for use of experimental equipment; plan data handling procedures
- Experimental process: prepare samples and carry out experimental procedure (on one or multiple samples)
- Data production: collect results from experiment
- Data processing and enrichment: process data; annotate datasets where appropriate
- Data publishing and storage: complying with agreed standards, (e.g. MIAPEgelDB, Minimum Information About a Proteomics Experiment [gel electrophoresis]) deposit representative dataset(s) in public databank (e.g. PRIDE, PRoteomics IDEntifications database)
- Data analysis: capture relevant datasets for computational analysis using preferred software

– Re-submit processed datasets if the data have been improved in some way

3.4.3 Metabolomics

– Experiment planning: Define experimental goals, identify source of sample(s), agree experimental conditions, plan and prepare for use of experimental equipment; plan data handling procedures
– Experimental process: prepare samples and carry out experimental procedure (on one or multiple samples)
– Data production: collect data from experimental process
– Data processing and enrichment: process data; apply statistical analysis or visualisation techniques; annotate datasets where appropriate
– Data publishing and storage: complying with agreed standards
– Data analysis: capture relevant datasets for computational analysis using preferred software
– Re-submit processed datasets if the data have been improved in some way

3.4.4 Computational bioinformatics

– Experiment planning: Define experimental goals; identify source of data: agree experimental conditions; plan and prepare for use of experimental machines (if relevant); plan data handling procedures.
– Experimental process: prepare samples and carry out machine run (on one or multiple samples) if generating primary data: or, process previously-created data
– Data production: machine produces raw data (traces) and processed data (text-based outputs)
– Data storage and preservation: Store preliminary data from machine ('raw' pre-base called data) for future analysis or further processing if ever necessary
– Data quality checking: carry out manual quality check; discard datasets with obvious errors
– Data processing and enrichment: process data; annotate datasets where appropriate (for example, combine microarrays for analysis as a group, align gene sequence data to the genome, etc); convert data to appropriate commonly-used file formats (e.g. SAM/BAM for sequence alignment data)
– Data publishing and storage: complying with agreed standards, (e.g. MIARE, Minimum Information About an RNAi Experiment) or MIAFGE, Minimum Information About a Functional Genomics Experi-

ment) deposit representative dataset(s) in public databank (e.g. Genbank)
- Data analysis: capture relevant datasets from Genbank for computational analysis using preferred software
- Re-submit processed datasets if the data have been improved in some way. Options include adding data to the UCSC Genome Browser[4] for use by a larger audience

3.4.5 Microscopy

- Experiment planning: Define experimental goals, identify source of sample(s), agree experimental conditions, plan and prepare for use of experimental equipment; plan data handling procedures
- Experimental process: prepare samples and carry out experimental procedure
- Data production: collect data from experimental process (micrographs)
- Data processing and enrichment: manipulate and analyse image data by computational techniques
- Data publishing and storage I: store locally on hard drives or transportable media, or submit to public database (e.g. the Mouse Brain Library), depending on the type of project and collaborative nature of the work
- Data analysis: capture relevant datasets for computational analysis using preferred software
- Data publishing and storage II: store data derived from analytical/processing step locally on hard drives or transportable media, or submit to public database, depending on the type of project and collaborative nature of the work

3.5 Case studies: short description of typical use cases in health science research

Case study 1:
The biology of Trypanosoma brucei, the parasite that causes sleeping sickness

The research is aimed at obtaining a better understanding of the parasite's biology and virulence. It involves examining the genomic sequences that encode components of the flagellum (the parasite's organ of motility) and the molecular processes that drive the motor functions. The antigenic surface pro-

[4] http://genome.ucsc.edu/

teins on the parasite's coat and the behaviour of the parasite's chromosomes at cell division are also studied.

Methodologies used include comparative genomic analysis, protein analysis and cytological studies of chromosomes using the FISH technique (see section 3.1). Mass spectrometer data are generated from proteomic studies. Protein-protein and protein-small molecule data are obtained from co-purification studies and provide information about molecular interactions within the parasite and between host and parasite.

Curated databases like UniprotKB, Ensembl and others gather additional information and help the meta analysis. Data are also available via cross-referencing to the homologous genes from other organisms. These data may provide indicative evidence for the role of the proteins.

Meta-analyses in the form of curated databases are also produced when looking at potential homologies between genes. This work involves the use of data from other laboratories: these data are obtained from public databases or from datasets supporting journal articles.

Software for data analysis is either written in-house or an existing package is tailored to suit the research group's needs. Data are made available through Ensembl Genomes[5]. Annotated datasets are substituted each time the dataset is updated, so although original gene sequences are archived the annotated datasets are continually updated.

Published data are always processed and annotated to an extent, though effort is made to publish data in as useful a form as possible (this is not necessarily the norm throughout this research community). Software tools produced in the laboratory may also be made available on the research group's website if they have potential use outside that laboratory.

Case study 2:
neuroimaging in psychiatric diagnosis and therapy

This case is about research into the prevention of psychosis and the role of neuroimaging methods, and data curation and sharing, in this effort.

Isolated or transient symptoms of psychosis are common, but the development of a clinically-defined psychotic state only follows in a minority of patients. Patients with high risk of developing schizophrenia (risk based on genetic factors) undergo MRI (magnetic resonance imaging) scanning procedures for diagnostic evaluation, which is supported by genomic analysis. Many of the phenotypic characteristics are shared between many distinct genetic diseases. Also many of the diseases have multiple causes with similar manifestations. The MIM (Mendelian Inheritance in Man) database, a de-

[5] http://www.ensemblgenomes.org

scriptive database, is a major source of information on these various diseases. Two major developments are underway as part of the research, and these will provide publicly-available community resources for the future.

A collaborative effort to integrate image data from multiple sources is being undertaken involving clinical teams, imaging experts and e-scientists, to create a Grid-based network of neuroimaging centres and a neuroimaging toolkit. The aim is to share data, experience and expertise so as to facilitate the archiving, curation, retrieval and analysis of imaging data from multiple sites and enable large clinical studies. The process involves: collecting retrospective data to help develop ways of harmonising scans from different machines; integrating existing datasets (scans and other clinical information); and developing a generic ontology for a psychosis database that can be used in future studies and clinical management.

A second collaborative effort is a further health informatics initiative that aims to develop a functioning "e-community" and build a secure electronic database to hold anonymised clinical data about people presenting with first-episode psychosis. The focus is on working with the network of research centres to create a shared metadata model and ontology of terms for clinical and biological data relevant to psychosis. Decipher, a Database of Chromosomal Imbalance and Phenotype in Human diseases, uses Ensembl resources to help pinpoint chromosomal sources of imbalance in patients, though its data are not openly available.

For this project a formalised risk assessment process for digital repositories (DRAMBORA[6]) has been applied, along with the OAIS functional model for archival information systems, to consider recommended activities for a data archive for all these data.

Dissemination of findings from this case is through journal articles in basic and clinical neuroscience journals, and also in the form of image and numerical datasets that can be shared via the Grid-based system being created.

Case study 3:
the mechanics and dynamics of cell division

The focus in this project is on the way in which the products of chromosome duplication are separated and moved equationally and simultaneously into the two daughter cells at cell division. If this process is faulty, the resultant daughter cells will be non-viable or malfunctioning.

The research questions are mainly of a molecular/mechanical nature and relate to the mass, speed, distance and timing of the main events of cell

[6] Digital Repository Audit Method based on Risk Assessment: http://www.repositoryaudit.eu/about/

division. Four components are involved: the chromosomes; a system of microtubules making up the mitotic spindle; the site to which the microtubules attach to the chromosome – the kinetochore; and the structure to which the microtubules are anchored at the poles of the mitotic spindle – the centrosome.

The research involves various kinds of approach: first, the use of advanced techniques of light microscopy (confocal microscopy, fluorescence microscopy) coupled to the use of immunofluorescent labels that attach specifically to defined components of the chromosome, the spindle, the kinetochore and the surrounding cell matrix; second, the use of mutant organisms in which cells lack specific proteins that are important for spindle function; third, the study of shifts in cell chemistry that coincide with microscopically observable events; and, fourth, the study of physical properties, such as visco-elasticity, and physical strain in relation to dimension and the energy required to move chromosomes quickly and directionally through the inside of a living cell.

Such studies have highlighted, amongst other things, the astonishing reliability of the system and the extensive fail-safe redundancies that have evolved to reduce the level of failure – quite important in a human, for example, where up to 100 million cell divisions are taking place at any one time.

Recent research in this area has involved a high level of interdisciplinary collaboration, with physicists, mathematicians, statisticians and engineers working alongside microscopists, geneticists, and molecular biologists. The experimental set-up, though microscopy-based, is heavily computerised, with software drivers for cameras, microscopes and analysing and manipulating the results. This analysis can be carried out during the experiment if required.

Dissemination of results is almost exclusively through journal articles summarising the work and which include the photomicrographs from microscopy. Image data are stored locally.

Case study 4:
quantitative models for simulating neuronal cell signalling

This research project develops computational models of cell signalling (interactions between cells that involve a signal-response effect) at multiple scales. The tools and technologies used are modelling environments and simulation software.

The group is also in charge of the world reference database of such models. As custodian of such things, one of the areas of work the group undertakes is the development of good practice procedures and standards. Data are readily shared with other scientists: standard XML formats are used which are richer

and more easily re-usable and exploitable than text-based data or spreadsheet formats.

A large toolkit of standards, formats and ontologies has been created to describe, annotate and interface the models created by the community. As an example, one of these, Minimum Information Requested In the Annotation of biochemical Models (MIRIAM) is a set of rules defining minimum standards for encoding the models for the biochemical modelling community and was developed by the group in the mid-2000s.

Because such standards guide metadata/annotation practice and encourage the use of standardised terminology, searching for datasets of interest is facilitated, the community's confidence in found datasets is maximised and they can be re-used with precision. In all, the value is greatly increased as a result.

Data and models from the research group are published in journal articles and on the group's websites as well as in relevant web-based public databases. Wikis are used to perform the work in the laboratory but not so far to exchange or annotate datasets.

Case study 5:
databases for mouse embryonic development

The research project is to develop a publicly-available resource, a detailed model of the mouse embryo through development from conception to birth. The database provides information about the morphology (shape, size and structure) and histology (tissue structures at cellular level) of the embryo at different stages of development. It also provides the framework for adding information from genetic studies about gene function and expression in the embryo.

The database differs from other, tabular databases by enabling different types of information to be mapped onto the 3-dimensional organism. An extensive anatomical ontology has been developed to aid in the understanding of the relationships between the embryo's anatomy and other spatial, temporal and genetic information. The ontology of anatomical names was mapped to successive developmental stages of the mouse embryo.

A sister database of gene expression data has also been produced by the same project team. Data for this database were sourced from published reports, datasets in public databanks, original data from laboratories, and datasets from large-scale projects. Together, these two databases – of embryo structure data and gene expression data – provide an overall resource rich in detail and functionality that can be used in research and teaching[7].

[7] http://www.emouseatlas.org/emap/home.html

The research team consists of both biologists and computer scientists and has a full-time curator for the databases. Scientists in the community are invited to submit data for inclusion in the databases. Datasets that are accepted for inclusion are curated by the project team.

The databases are made publicly-available through a website. Users can view histological sections through an embryo in different planes, 3-D models and reconstructions, and videos of a rotating embryo prepared in different ways including by differential staining, to show its complete external morphology. The genetic data are analysable by text-based and spatial-based methods.

4 Current status of research infrastructure, workflows and life cycles

The research infrastructure should enable scientists to:

- access the physical resources, materials, and services necessary for their research, at the point at which they are needed
- access the information resources and the networks that transmit them, at the point at which they are needed
- have the means to access these resources at the time of need and have the skills required to use them to best advantage
- have confidence in the quality and integrity of these resources
- access the technologies they need for advanced or collaborative work
- have the means and skills to exploit these technologies to maximum effect
- have the analysis tools to include their data and analyse them with respect to the data already contained within the database

The following section describes the workflows and research systems and processes that operate in the areas of focus in biomedical research.

4.1 The experimental infrastructure: approaches and protocols

Experimental approaches vary between the fields that are the focus of this report.

4.1.1 Genomics

Genomics is the study of the genetic make-up of living organisms. Genes are composed of long runs of nucleotides (bases) that form the backbone of DNA. Their sequence along the DNA determines the function of the gene, since this

sequence is transcribed into functional or messenger RNA and thence to proteins in the cell. The expression of genes at spatial, temporal and intensity levels is also studied.

(a) Genome sequencing

The sequencing of DNA is therefore at the core of genomic investigations, and since the human genome was sequenced in 2003, the technology for DNA sequencing has been dramatically advanced. The most recently developed massively parallel processes, referred to as "next generation sequencing" can output hundreds of millions of short DNA fragments of around 50 bases long in a matter of days. The bottleneck of making use of these data has switched from data generation to data analysis, with huge computational power now required to assemble these short strings of bases into complete genome sequences[8]

DNA sequencing, while still carried out in many research laboratories around the world in pursuit of specific research questions, is becoming an industrialized process, with many companies now offer next-generation sequencing services or related products. These technology changes are ensuring that sequence submissions to public databases continue to increase at exponential rates (see Figure G.3 below).

The availability of these powerful sequencing technologies are allowing genomic scientists to undertake comparative genomic experiments on a mass scale, giving rise to efforts such as the '1000 genomes project'[9], that have huge potential benefits for human health and well-being.

In addition, relatively cheap genome sequencing methodologies will have the power to revolutionise taxonomy and ecogenomics studies. Taxonomical relationships between organisms will be much easier and quicker to elucidate, and the application of molecular sequencing techniques on a genome-wide scale will have massive benefits for research that is aimed at better understanding ecological and evolutionary processes.

Figure G.4 shows a record of a short-read sequence from a nucleotide database.

(b) Gene expression

The other most commonly used methodology for looking at genetic activity is microarray technology. Microarray technology is a way of studying gene

[8] See for example: http://www.nature.com/nmeth/journal/v7/n7/fig_tab/nmeth0710-495_T1.html

[9] http://www.1000genomes.org

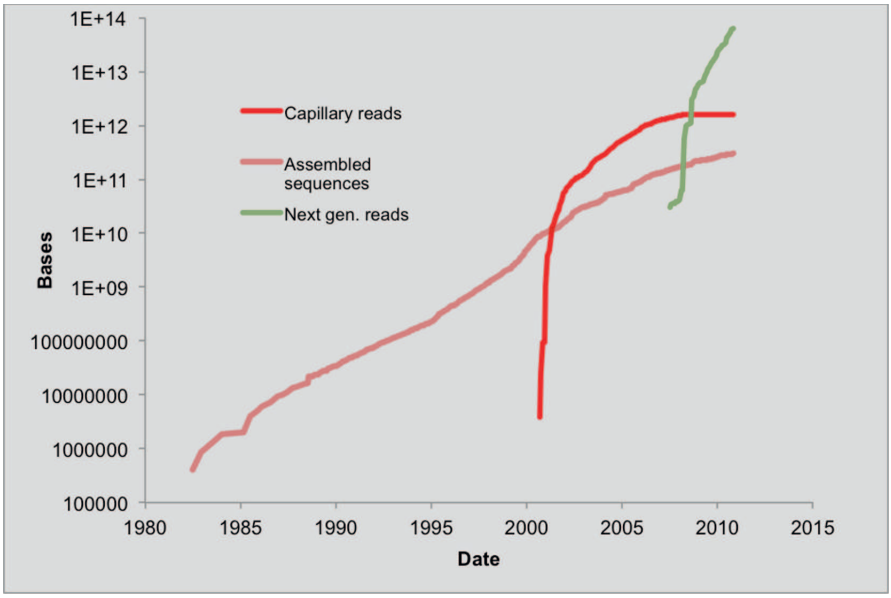

Figure G.3 public domain nucleic acid sequence data (kindly supplied by Guy Cochrane, EBI)

expression. In microarray work, thousands (sometimes millions) of genes or gene fragments can be assayed at once. This makes the work of looking at the expression of genes much quicker than it used to be, but the volume of data generated in the experimental process is very large.

The products of gene expression are messenger RNAs (mRNAs). In this process, thousands of genes or parts of genes are bound to a substrate on microarray plates. The mRNAs of interest (usually in the form of cDNA) are added to the plates and hybridise with (attach to) their complementary DNA sequence. The mRNAs are labelled, usually with a fluorescent dye, so that where they attach to a gene or gene fragment they can be visualised. The microarray plates are then scanned and the luminous dots of the fluorescent probes, which indicate where an mRNA has bound to a particular gene fragment, are recorded (Figure G.5).

Image analysis software can be used to process these findings, but very large datasets are produced as a result, sometimes hundreds of gigabytes in size. Downstream processing of such datasets requires considerable computing power. Microarray data also can be produced as text files that consist of rows and columns, sometimes in vast numbers (millions).

SRA Experiment: ERX009450 : Illumina Genome Analyzer II paired end sequencing; Sanger_zebrafish_sequencing

View: XML — Download: XML
Attributes — Send Feedback

Submitting Centre	Platform	Model	Read Count	Base Count
The Wellcome Trust Sanger Institute	ILLUMINA	Illumina Genome Analyzer II	28,157,772	4Gb

Library Layout	Library Strategy	Library Source	Library Selection	Library Name
PAIRED	RNA-Seq	TRANSCRIPTOMIC	cDNA	RNA from Zebrafish adult swim bladder

Description
Sanger_zebrafish_sequencing

Navigation | **Fastq Files** | Submitted Files | Attributes

This table contains the fastq files for experiment ERX009450 only. Please go to the study ERP000447 to see all files for the study.
Please note that submitted files are available in the Submitted Files tab.

View: TEXT — Download: TEXT
Select columns

Study	Sample	Experiment	Run	Organism	Instrument Model	Library Layout	Run Read Count	Run Base Count	ftp	Aspera
ERP000447	ERS017859	ERX009450	ERR023148	Danio rerio	Illumina Genome Analyzer II	PAIRED	28,157,772	4Gb	file#1	not installed
ERP000447	ERS017859	ERX009450	ERR023148	Danio rerio	Illumina Genome Analyzer II	PAIRED	28,157,772	4Gb	file#2	not installed

For Aspera download, please download and install Aspera Connect

Figure G.4 a short-read sequence record in the European Nucleotide Archive at EBI

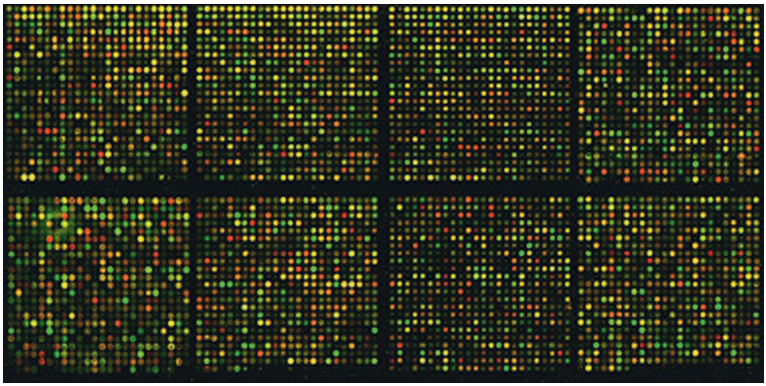

Figure G.5 Microarray plate showing gene expression differences between two muse tissues (red and green dots indicate which genes are turned on or off, yellow dots indicate that gene expression is unchanged. (Photo: Dr Jason Kang, National Cancer Institute, USA)

Genomics data are shared predominantly through a mature infrastructure of public databanks (see section 6). Researchers deposit datasets from ma-

chines as soon as is practicable, and this may be directly from the machine to the database. Large groups tend to deposit all their data while small groups are more likely to deposit a typical representative trace or other dataset because they do not have the resources to deposit the hundreds they might generate from each sample.

Summary findings are written up as articles and published in one or other of the many journals that cover health sciences. Functional genomics is a collaborative effort with application of standardised data integration, storage and dissemination policies and practices.

4.1.2 Proteomics

As well as studying the structure and function of genes, molecular biologists are interested in proteins, their molecular composition and how they function in regulating cellular processes. This field is nowadays called proteomics. Proteins are the final product of gene expression: they are composed of amino acids, the order of which in the molecule is determined by the sequence of nucleotides in the mRNA from which they were translated, and which in turn is determined by the sequence of nucleotides in the DNA of the original gene. To give a sense of the scale of the challenge, the 35,000 genes in the human genome can code for ten times as many proteins: in an extreme example, one gene can code for 1000 different proteins (in the case of genes expressed in the immune system).

Proteins can be sequenced (that is, the amino acids that compose them can be determined) by either of two methods. The most commonly used one is the Edman Degradation, a now-automated derivative of the original method developed by Edman in the 1950s. Proteins are degraded (broken down into constituent amino acids) and the individual amino acids released are assayed by high performance liquid chromatography (HPLC). The amino acids can be marked by compounds that produce a colour, enabling the presence and the amount of each amino acid to be determined.

The other method of sequencing proteins is mass spectrometry. This process is also automated. An electric current degrades the protein into its constituent amino acids or into small peptides (protein fragments) and these are identified by their individual mass. The spectra are expressed in terms of numeric data (i.e. peak intensities), as text-based data such as lists of protein IDs, or graphically. Considerable computational power is required for this process, but with improvements in this and in data storage, this technology is becoming common in protein studies. The process can be carried out very quickly – within seconds as opposed to many hours for earlier, manual methods.

Protein interaction data are becoming prominent as the functions of proteins are better analysed at the genomic or proteomic scale, and are becoming available as standardised databases. Pathways can be predicted, based on these data, to give a better picture of the combined functionality of the gene products in cells.

For most data types generated by proteomic studies, there existing public databases available for data deposition.

4.1.3 Metabolomics

Mass spectroscopy is also used for analysis of small molecules in metabolomics research, where metabolite levels in tissues or fluids are assayed. Techniques employed may be nuclear magnetic resonance spectroscopy and gas or liquid mass spectroscopy. All these tools are fully automated processes with sophisticated computational analysis at the other end of the process. For most data types generated by metabolomic studies, there are existing public databases available for data deposition.

4.1.4 Computational bioinformatics

Although the broad term 'bioinformatics' applies across all the technologies so far described, it can also be used in a more narrow way to describe the specific application of computer technologies to data integration, mining and other analytical practices. Such computational approaches to health research are usually termed bioinformatics, medical informatics or health informatics. Research in this area includes work on how to store, retrieve and use research data and findings, with a strong focus on manipulation of data, often collected from disparate sources, to derive conclusions or further data for further analysis.

Skills required are those of information science and software engineering as well as biomedical knowledge. Bioinformaticians may be biologists who train in computational technologies, or computer scientists or information scientists who gain knowledge and understanding of biological systems. The former pattern is more common. Either way, the field is interdisciplinary and is evolving rather fast.

4.1.5 Microscopy

Finally, there are visualisation methodologies. These include light microscopy (bright-field, phase-contrast, differential interference, fluorescence and confocal microscopy), electron microscopy (transmission and scanning types) and scanning technologies such as magnetic resonance imaging or computerised

tomography. These technologies are used to study structure and function of tissues, cells or sub-cellular organelles, or to localise entities.

One example of the latter is the ability to use microscopy to map genes to specific locations on the whole chromosome set by the FISH (fluorescence *in-situ* hybridisation) technique. There are variations of this, but in essence it consists of attaching a fluorescent marker to the mRNA of interest and allowing the mRNA to hybridise to a chromosome set attached to a substrate. The mRNA shows up in fluorescence microscope images as luminous dots or bands. These may be computer-enhanced to improve the images or to enable easy discrimination between different genes where more than one mRNA has been used.

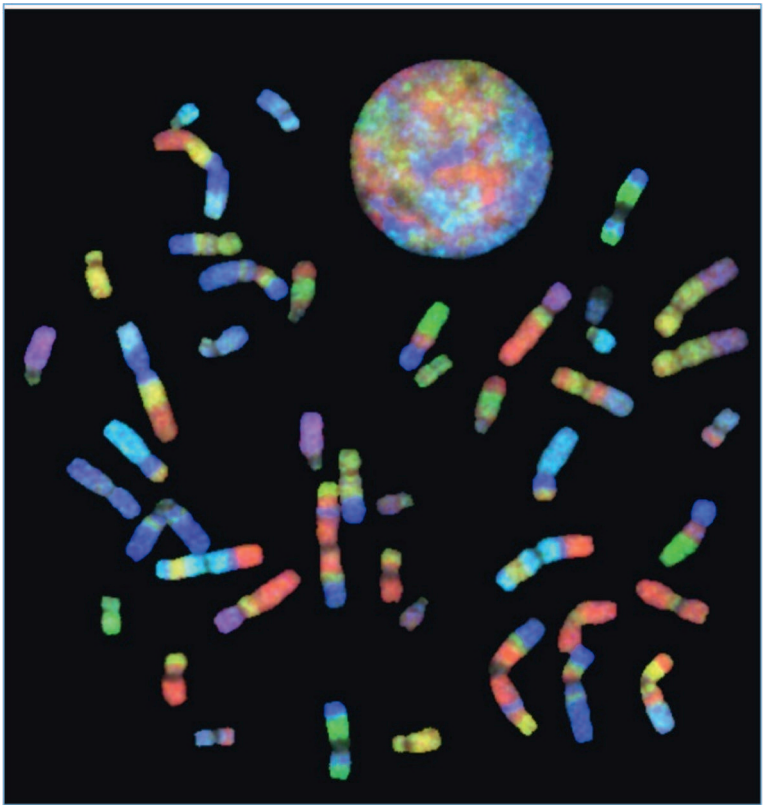

Figure G.6 In situ hybridisation of seven chromosome specific-paint probes derived from a gibbon to a set of human chromosomes (source: picture kindly provided by Dr. Fengtang Yang, The Sanger Centre, Cambridge UK)

If research is about a clinical condition, then additional techniques may be employed, such as MRI or CT (computerised tomography) scanning. Clinical imaging technologies produce large datasets delivering considerable storage and archiving challenges. Figure G.7 shows an example, a series of MRI images from a study of schizophrenia.

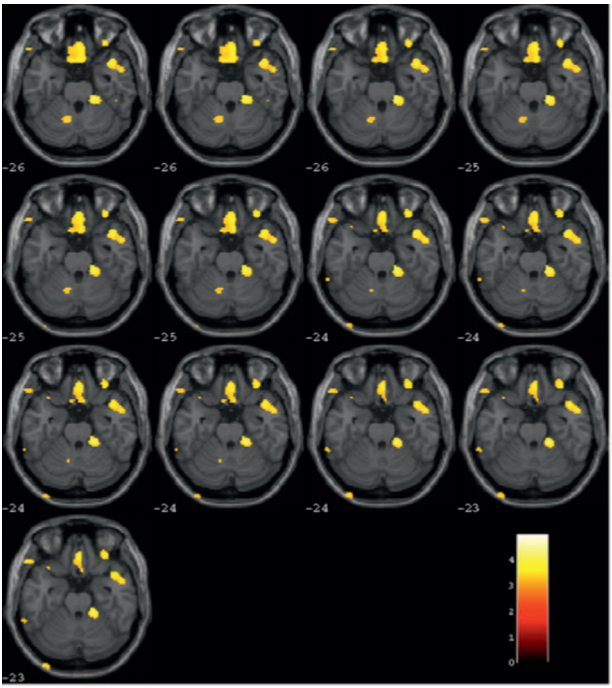

Figure G.7 group average difference map showing grey matter density in subjects in a schizophrenia study (courtesy of Professor Stephen Lawrie, University of Edinburgh)

4.2 The community infrastructure: collaborative research

The molecular biology community enjoys an extremely well-organised system for dissemination and curation of research results and through that system for connecting scientists and research groups with one another. The European Bioinformatics Institute is a focal point in the information infrastructure underlying research efforts in this discipline, providing the technologies and structural components required to collect, hold, curate and preserve research data outputs.

4.2.1 Interdisciplinary collaboration

Interdisciplinary collaborations are increasingly necessary in many fields. The growth in the application of computational technologies to bench-generated experimental data means that informatics experts from information science and software engineering are needed to complement the analytical skills of biologists.

Many larger groups now employ someone specifically dedicated to research data management, a role that requires a high level of skill and expertise: it encompasses not only data are properly stored and easily retrievable for further analysis but also of preparing data management plans when a new round of experimentation is planned, a didactic role in ensuring bench scientists understand and get optimal results from machines, tracking down data from other laboratories that may be needed for data-mining by the local research team, and writing scripts that enable best use of datasets from these machines or other research groups.

Data managers ensure best practice in the care, preservation and sharing of data, maximising confidence in biomedical data and encouraging data re-use and exploitation for new knowledge creation. This skill area is rapidly growing in importance as research becomes more data-intensive and as funders introduce formal data requirements to their funding process, and is becoming a career option for scientists with aptitude in this area.

Interdisciplinary approaches are also needed to tackle the challenges of experimental work in many areas. The increasing sophistication of light microscopy, scanners and other imaging techniques, for example, and the approaches needed to answer some of the questions at the cutting edge of cell biology and medicine, may draw on the skills of physicists, mathematicians, chemists and engineers as well as those with expertise in computational applications. This 'systems biology' paradigm, where scientific study in the life sciences is approached through synthetic, rather than reductionist, approaches is increasingly appropriate in responding to the scientific questions and challenges faced in the health science arena today.

4.2.2 Large-scale research and e-science

As well as interdisciplinary efforts, collaborations between research groups or laboratories are becoming more common in health sciences as major fields of research in this discipline become more data-intensive and e-science methodologies are applicable. Examples are the 1000 genomes, International Cancer Genome Consortium (ICGC) and the International Human Epigenome Consortium, in all of which the EBI are one of many global collaborators. It is useful to provide a short description of each of these to demonstrate the

nature of the initiatives and to illustrate that collaborative approaches are necessary to deliver such ambitious and labour-intensive results.

The 1000 Genomes Project[10] is aimed at sequencing the entire genomes of 1000 people, in order to find most genetic variants that have frequencies of at least 1% in the populations that these individuals represent. The project has a steering committee of twenty-four scientists and a panel of several hundred scientists contributing to the laboratory work.

The goal of the International Cancer Genome Consortium[11] is to obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumour types and/or subtypes which are of clinical and societal importance across the globe. It coordinates a number of projects (35 at the time of writing) across the world with the aim of developing comprehensive catalogues of genomic abnormalities in these tumours and will make the data available to the entire research community as rapidly as possible, and with minimal restrictions, to accelerate research into the causes and control of cancer.

The International Human Epigenome Consortium[12] coordinates epigenetic mapping projects (projects that study the organisation of the genetic material in the cell and how this organisation affects gene expression and the control of cellular functions) worldwide. The aim is to prevent redundancy and duplication of effort and to implement high data quality standards, to coordinate data storage, management and analysis and to provide free access to the epigenomes produced.

In some areas of life science research, grid technologies are now warranted and used. This is particularly the case in fields where imaging technologies are intensively used. The neuro-imaging case study described in section 3.4, for example, has become a collaborative effort involving six or seven laboratories across the UK. The project has amassed clinical and demographic data on a terabyte scale, in the form of millions of individual files. Data management on this scale is a major undertaking.

Collaborations may also arise where large amounts of funding are needed for one piece of work, where individual expertise or technologies need to be pooled to answer a research question, or where large volumes of data need to be gathered for meta-analysis. In such cases, collaborative efforts may be transient, lasting only for the length of time needed to achieve that immediate goal, or they may persist for many years with repeat funding being attracted for further collaborative work. An example is the Bloodomics project[13].

[10] http://www.1000genomes.org/home
[11] http://www.icgc.org/content/icgc-home
[12] http://www.ihec-epigenomes.org/
[13] http://www.bloodomics.org/

4.3 The temporal infrastructure: research life cycles

While in some areas of health science (for example, epidemiology) the research life cycle is a long one, life cycles in many of the fields of focus here are relatively short. A DNA sequence run can be completed and the results deposited in a public databank for others to use within a few hours. The other biochemical analytical techniques described can also be carried out in a matter of hours or days.

Cytological experiments may require some days of specimen preparation and microscopy, plus more for computer manipulation of the resulting images to maximise the usefulness and clarity of the results. Large-scale clinical studies necessarily take much more time, though, sometimes requiring years to collect sufficient data for analysis. And computational (informatics) research is variable depending upon the complexity of the questions to be answered and the data to be manipulated.

It should be noted that for all these data-intensive research activities considerable curation, technical or computational/algorithmic expertise and effort are also required to ensure that datasets are usable by the research group that produced them and by others, and that these datasets are accessible and re-usable in time to come.

The research life cycle model related to knowledge creation and dissemination developed by Charles Humphrey[14] is useful here (Figure G.8).

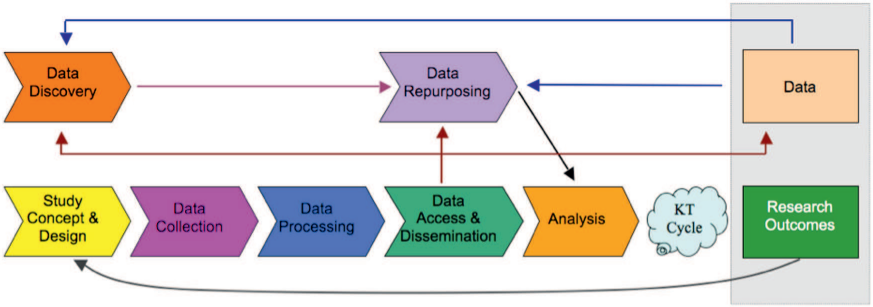

Figure G.8 The life cycle model of research knowledge creation (Humphrey, 2008) ['KT Cycle' is the Knowledge Transfer Cycle]

Experimental research is largely represented by the bottom set of activities in the diagram. The top set represents informatics/e-science approaches,

[14] Humphrey, C (2008) e-Science and the life cycle of research. http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

where experimental data are re-purposed and analysed to create new data (which may themselves be re-purposed and analysed).

The key issue with respect to data services is that the community both contributes to and interacts with them. The community creates the databases, and EBI is the custodian of those resources, curating the data to some degree as part of that custodial role. Some databases get only light curation (such as the sequence databases, where a fairly simple metadata check suffices), but others are heavily-curated (such as UniProt, where EBI curators search for articles about a gene, find evidence on it and add that to the UniProt database). It is worth saying at this point that data curation at this level has become an established career option, emphasised by the fact that there is a now professional society of curators.

As well as curators, the rest of the community accesses and uses the databases in specific ways. Users of different types interact with the established databases differently. Essentially, there are:

- Power users: those who routinely download a database in bulk to carry out, for example, whole-genome analysis. These users work using FTP and web services to access the volumes of data they need
- Biologists: those who need to access and use relevant, small-volume data for their work; for example, someone who is working on a particular disease and needs to check on sequencing data for genes that are implicated in the disease
- Occasional users: for example, teachers, students, editors, citizen scientists and so forth, who may on occasions wish to view or analyse a dataset

In summary, the life science databases are growing, developing entities. They form a hub for the community's activities, but that hub is dynamic. Users not only take and use the data, they provide feedback on the service, their requirements change, research develops in new directions, and the services evolve accordingly.

4.4 The skills and training infrastructure

Basic and clinical research practices and technologies are acquired in the usual way through postgraduate and postdoctoral training. Technologies may be complex to master and require significant intellectual effort as well as practical skills. Some cytological techniques require exceptional dexterity.

The importance of informatics expertise in many areas of health science research imposes a new requirement on researchers who have been trained in conventional biological methodologies and approaches.

Many biologists learn 'on the job' and pick up coding skills where needed. The past decade has, however, seen rapid growth in masters level training in

informatics (bioinformatics and cheminformatics, especially) and many young health scientists are entering the field with these qualifications. In addition, some library and information science schools are offering new modules or courses in data management, something that is expected to become a career option for librarians in the future.

Data management roles are also becoming more common within research groups in response to the increasing data-intensity of research and the requirements of funders for curation and preservation of datasets. Where these posts exist, they may be occupied by senior researchers who have shown interest in and aptitude for the role. These people are normally called 'data scientist' or similar, and may be distinguished from data managers by differences in their role and position. The terminology is extremely fuzzy at the moment, but we distinguish between data scientists and data managers simply on the basis of the set of tasks that they carry out and the overall objective of their job:

- Data science: the conceptualisation, creation, use and appraisal of data, the selection of data for re-use, and the application of tools to re-use and exploit data
- Data management: a specialist area of computational science – database technology – which focuses on ensuring that data produced and needed by the researchers are properly stored, curated and preserved. Included here (but not exclusively) is the work carried out in data centres or professional databanks. Often computer scientists with biomedical research background not required

Where there is not a discrete data scientist role within a research group or laboratory, various members of the research group may undertake data-related tasks. They may or may not be formally trained in the skills required, but there is growing attention to this within the community and summer schools and short training programmes covering specific areas of health science data manipulation or informatics are becoming more common. Almost all biologists working in informatics-based fields are able to write scripts that enable them to use datasets from other laboratories or to mash together datasets from different machines. Computer scientists may be employed to bring their software or databasing skills to biomedical research teams. In this case they must assimilate the biological domain knowledge that they need 'on the job'.

5 Current status of Open Access to the research literature

5.1 The policy foundation for Open Access to the biomedical literature

The first Open Access policy that covered any biomedical/health science literature was the institutional mandatory policy at Queensland University of Technology, Brisbane, Australia, in 2004. This was followed the same year by a second institutional mandate at the University of Minho in Portugal.

These institutions began a trend that has continued. At the time of writing there are 117 institutional policies on Open Access to journal articles and 30 sub-institutional ones (departments or schools within universities or research institutes), including Harvard Medical School.

Several research funders have also adopted Open Access policies and mandates, lead by the NIH in the USA and the London-based Wellcome Trust, and it is this that has affected significantly the proportion of the health sciences literature that is now openly available. Of the 47 current mandatory policies from research funders, 22 are from funders of health-related research. The list includes national research councils funding health research in Australia, Canada, Ireland, UK and USA, and around a dozen medical charities in these countries plus Italy. Researchers supported by these funders are required to deposit their articles in institutional or subject-specific repositories (such as PubMed Central and its growing national/regional variants).

In addition, the European Research Council has a mandatory policy requiring Open Access to outputs from research that it funds and some of this falls under the health sciences banner. The European Commission has a 'pilot' mandatory policy covering 20% of the current FP7 research programme, and the health research programme is included in this 20%. This will expand to cover 100% of EU-funded research.[15]

The National Institutes of Health (NIH) in the US, the largest funder of research in the world, alone covers some 90,000 journal articles published each year from research that it funds. The NIH policy originally began as a voluntary one that requested grant-holders to deposit copies of their journal articles in the PubMed Central repository. After two years under this policy, only 5% of relevant articles were deposited by their authors voluntarily. Publisher deposits raised the total to 19% but this was still disappointing.

In 2008, the US Congress instructed the NIH to make the policy mandatory, the result of which is that the percentage of articles being deposited climbed to over 70% in 2009. This was the strongest possible evidence that a policy

[15] http://www.youtube.com/watch?v=GIU14-3hYto

must be mandatory to work effectively: since that time new policies from health funders have been mandatory and existing policies based on voluntary action by authors have been revised to make them mandatory.

There is considerable evidence from other quarters that also supports the necessity for the mandatory nature of a policy on Open Access. The earliest study to produce data on this was by Sale (2006), who looked at the repositories of three Australian universities. Sale showed that the accumulation of Open Access articles at the University of Tasmania, where there was no policy but there was some active advocacy on OA, was extremely slow. At the University of Queensland, where there was a policy encouraging authors to deposit their work, supported by active advocacy from the library, the accumulation rate was higher. A fast rate of accumulation of content was seen, however, at Queensland University of Technology, where there was both active advocacy and practical support from the library *and* a mandatory policy.

This evidence that only mandatory policies work effectively was compounded by a recent study by Gargouri *et al* (2010) that compared, amongst other things, the level of OA articles in university repositories that have mandatory policies with the general level of accumulation of articles in non-mandated repositories. This 'control' level of accumulation is around 15% of total outputs from the institution, whereas mandated repositories are collecting 60% of total outputs.

Most policies accommodate a short embargo period, usually 6 months but in some cases 12 months, to enable publishers to continue to operate their subscription sales model. The argument is ongoing about whether even a short embargo is detrimental to research progress: certainly the speed at which research moves in health sciences means that a delay in access by medical researchers and medical practitioners outside of research institutions, patients, therapists and research-based small companies that need this research to innovate may all be disadvantaged by having to wait for access to new research findings.

Further policies from other funders are expected as the benefits of opening up the biomedical literature become more apparent. Additionally, there is continued growth in institutional policies: these help to boost the Open Access corpus in biomedicine incrementally.

5.2 Open Access to the research literature

5.2.1 Open Access repositories

Health sciences is one of the few disciplinary areas of research where extensive, subject-based repositories of Open Access material exist. The fact that there is such centralised infrastructure reflects available funding, the criticality of

research in this discipline and the ability of the discipline to organise around coordinating bodies.

PubMed Central (PMC) was established in the US in the year 2000, with the contents of just two journals in the repository. Within two years it covered 55 journals and numbers have been growing ever since. The database currently has around 2 million full-text journal articles and receives the full contents of 600 journals as well as manuscripts deposited by authors. All are free to access and read, but only about 11% fall under the strictest definition of Open Access by being distributed under a Creative Commons (or Creative Commons-type) licence that permits more liberal re-use.

The NCBI (National Center for Biotechnology Information), which manages PMC, has added many features over the decade, including a good search function, linking between articles, and between articles and other types of content such as commentaries and books. More features are planned. Such features enhance the user experience and utility of the database.

In 2007, the first international PMC (PMCi) was established in the UK by the Wellcome Trust and a consortium of other research funders. This repository, UKPMC, collects articles in biomedicine from UK scientists and shares content with PMC itself. This is the first of what may be many PMCis: already a Canadian site has been announced, with discussion of additional sites in other regions, including the possibility of transforming the UK site into a European PMC.

UKPMC launched with the intention of providing cutting-edge services to researchers and has already broken new ground by creating XML-marked-up texts that are amenable to data-mining and text-mining. UKPMC is already being used by the UK's National Text-Mining Centre (NaCTeM) and the European Bioinformatics Institute (EBI) to extract facts, concepts and relationships from the literature within UKPMC and combine them to create new knowledge. UKPMC has also developed collections of other types of content, such as clinical guidelines and project grant details. It also enables users to cross-search PMC alongside CiteXplore, an indexing service that covers a number of other large research databases.

This informatics work is laying the foundations for a future where interoperability is truly achievable between Open Access research collections. Indeed, UKPMC is also developing ways to support UK research institutions in their quest to fill their own institutional repositories. The policy from many of the UKPMC funders is to require deposit of research outputs directly into UKPMC itself, thus conflicting with policy requirements of many UK universities that require deposit into the local institutional repository. To obviate the need for researchers to deposit in both services, UKPMC will serve arti-

cles to the institutions from which they originate for population of the local repository.

A well as these centralised collections of Open Access content, health science literature is accumulating in institutional repositories. Mandatory policies on these, of course, cover all research carried out in those institutions and are thus essential in 'sweeping up' outputs from unfunded research and from research not covered by funder mandates.

5.2.2 Open Access to journal articles

Health sciences are also well-represented in Open Access journals.

The largest open Access publisher, BioMed Central (now part of the Springer science publishing organisation), specialises in biomedical research, as is obvious from its name, though it does also now cover some chemistry and mathematics too. It publishes some 210 journals, most of which are in biomedicine. BioMed Central deposits all its journal articles in PMC at the time of publication as well as hosting them on its own website.

The Public Library of Science, another leading Open Access publisher, has not only developed some very high quality journals in biology and medicine (*PLoS Biology* and *PLoS Medicine*, plus others) but has changed the shape of publishing through *PLoS ONE*. This is a journal that covers all the natural sciences. It introduced a new system of quality control, still based up on peer review, where referees are asked to judge an article purely on the basis of whether the work has been carried out in a sound scientific manner. Judgments about its relevance, significance and impact are made through community response post-publication. The model has proved very successful and has recently been emulated by the Nature Publishing Group with the launch of *Nature Scientific Reports*[16].

The Scielo (Scientific Electronic Library Online), a collection of peer-reviewed Open Access journals published mainly from South American countries in Spanish or Portuguese, covers over 800 journals. Of these 45 are in biological sciences and 261 in health sciences, representing a large part of the Latin American biomedical literature.

Bioline International, a service that provides a free electronic publishing platform for small publishers wishing to publish Open Access journals in the biosciences, has over 50 journals in its collection, all from developing and emerging countries, covering biomedicine and agriculture.

The Directory of Open Access Journals[17], a listing of Open Access journals from around the world, currently details 709 journals in its 'health sciences'

[16] http://www.nature.com/srep/marketing/index.html
[17] www.doaj.org

category (covering medicine, dentistry etc) and a further 217 in its 'biology' category. This list overlaps with the specific services mentioned above.

In addition to the fully Open Access journals, many publishers have now offered a so-called 'hybrid' Open Access option, whereby authors can pay a publication fee and have their article made Open Access within an otherwise subscription journal. Take-up on these options is not high, largely because of the level of fee, and it should be noted that many journals offering this option do not make the articles available under a liberal licence, meaning they are free to access and read but often not to re-use in other ways, including computing upon them.

5.2.3 The proportion of Open Access literature in health sciences

Some attempts have been made to measure how much research in total is available in Open Access, and to break this down by discipline in some cases.

Björk and co-workers estimated that in 2008 using a sample of almost 1850 articles, that 20.4% of the total literature was available in some form of Open Access (in OA journals, in repositories or on author websites) (Björk *et al*, 2010). This compares to a previous study by the same authors of the situation in 2006 (Björk *et al*, 2009) that found a total of 19.4% of the literature to be Open Access. The difference is within confidence limits.

Hajjem *et al* (2005), using a sample of 1.3 million journal articles, found that the proportion of Open Access articles varied between disciplines from 5% to 16%. A later study by [Gargouri *et al*, 2010] from the same group found the OA share overall to be 20%, with biology scoring 21% and clinical medicine 3%. Note that this study used only 'green' OA (articles self-archived by their authors in repositories, including PubMed Central, not those published in 'gold' OA journals). The latest estimate by this group of the percentage of research openly available through repositories is 20-22% and, if Björk's estimate of the percentage available through journals ('gold' Open Access) is added, the total is currently about 30%[18].

Matsubayashi *et al* (2009) studied the discipline of biomedicine specifically and found the OA availability of articles to be 26%.

The findings from studies so far are summarised in Table G.1.

The more recent of the two studies by Björk *et al* showed that for the fields of medicine, biochemistry/genetics/molecular biology, and 'other areas related to medicine', the proportion of OA articles in Open Access journals was higher than that in repositories. This position is reversed for all other fields in this study, presumably reflecting the domination of the 'Gold' Open

[18] Stevan Harnad and Yassine Gargouri, personal communication (to be published shortly)

Table G.1 Open Access availability of journal articles

Study	Overall % OA	Specific fields % OA
Björk *et al* (2009)	19.4	n/s
Björk *et al* (2010)	20.4	Medicine 21.7% (13.9% OA journals, 7.8% OA repositories) Biochemistry, genetics, molecular biology 19.9% (13.7% OA journals, 6.2% OA repositories)
Matsubayashi *et al* (2009)	n/s	Biomedicine 20%
Hajjem *et al* (2005)	n/s	Biology 15% Health 6% Psychology 7%
Gargouri (2009)	20.0	Biology 21% Clinical medicine 3% Health 18% Biomedicine 11% Psychology 25%

(ns = not studied)

Access journal-publishing arena by biomedical journals. The data from Björk *et al* (2009) on this point are shown in Figure G.9.

5.3 New developments in dissemination in health sciences

The vision of enhancing the traditional scientific article (or book) has been developing over the past few years. The Web provides the opportunity to link an article written and presented in the traditional format with supporting data, commentaries, similar articles, datasets in public databanks and so on. Semantic technologies now hold the promise of creating a scientific Web that is linked by meaning and context, with all research outputs fully and meaningfully linked to one another.

Traditional forms of peer review, sequential publishing, the way rights and ownership of knowledge are managed, and the emphasis on disseminating scientific findings in the form of reports 'crystallised' in time are likely to metamorphose into a system that makes maximum use of the opportunities offered by the Web and optimises research communication.

An FP7-funded project, Liquid Publications[19], has been investigating options and developing 'liquid journal' and 'liquid conference' use cases. A position paper defines some of the conditions and the benefits of liquid publishing:

[19] http://liquidpub.org/

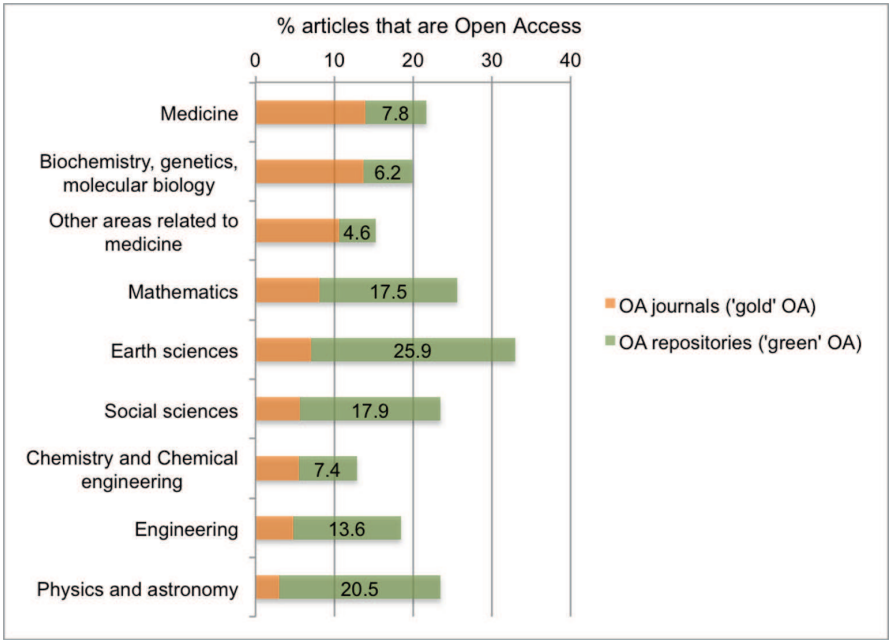

Figure G.9 Percentage of Open Access articles by discipline and mode of dissemination (data from Björk et al, 2010)

real-time dissemination of findings, encouragement of early sharing of results and ideas with reward systems in place to benefit researchers who maximise this behaviour, the concomitant increase in collaboration that will arise from early and widespread dissemination, and lightweight and real-time assessment and evaluation of findings.

While such as system remains to be achieved, there have been steps taken towards developing an improved, linked scientific knowledge base. The far-seeing work of UKPMC in establishing a system that ensures material ingested into the repository is in XML and marked-up for semantic data-mining and text-mining tools is one important advance, and it is taking place in the health science discipline. That repository also provides the means for supporting datasets to be deposited and linked to articles, thus taking another step towards a properly-linked scientific corpus.

Publishers have also been active in this area. In 2007, the Public Library of Science launched PLoS ONE, a broad-scope Open Access journal covering the whole of science. PLoS ONE is interactive, providing the means for readers to post comments and discuss articles, and it also incorporates various additional

features including a range of article-level metrics that inform the author about the usage and impact of their paper. Two years ago, Elsevier Science released prototypes of what it called 'the article of the future'[20], which was hyperlinked (mainly to other articles) and contained embedded video and audio files, and some integrated social media tools.

These are small steps and ones still bounded by the limitations of the traditional model of a scientific paper. The concept of true liquidity is a different level altogether. Nonetheless, the experimentation is commendable and the incremental advance is welcome. These moves signal something that will be of fundamental importance for the efficacy of the future scientific communication system.

In this scenario, biological databases are, and will be, also critical. Each database release is equivalent to 'publication' and most databases are highly fluid in the sense that sequences are modified as more alignments become available. These modifications change the data for all downstream databases and this realises/synchronisation process plays an important role in maintaining data consistency.

5.4 Open notebooks

Open notebooks – the open dissemination of the day-to-day experimental activities in the laboratory – have become fairly common in certain fields. Scientists record their experimental procedures and results and publish them on the Web, usually in blog form[21].

The idea of this form of communication is to speed up scientific endeavour in a field, to gather feedback from the relevant community, to engage the community in general discussion about the ongoing work, and to capitalise on the collective wisdom of the crowd.

The discipline where the use of open notebooks is furthest advanced is chemistry. This is partly because early-adopters of the concept were chemists and partly because chemistry is largely free from the issues of concern that health scientists may have about the practice.

There are two main areas of concern expressed by researchers in various fields of health science. First, there is concern about patient confidentiality and the imperative to safeguard patient anonymity. Second, there is the risk of releasing early data or information that subsequently turns out to be inaccurate but which, if used in the meantime, may have damaging consequences for people.

[20] http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_01279

[21] For example, the chemist Cameron Neylon's open notebook:http://biolab.isis.rl.ac.uk/camerons_labblog

As a result, open notebook science may not be a concept that translates well to some areas of health science research. Nonetheless, there are areas, such as the molecular biology fields, where the two concerns have little relevance. In these cases, open notebook work or something akin to the concept, are used in practice (see section 5.3).

Implications for OpenAIRE

- Mandatory policies from the European Commission and from a growing number of health research funders and institutions will increase the amount of health science literature accumulating in Open Access repositories across Europe for harvesting by OpenAIRE
- Continuing development of OpenAIRE policy on content acquisition may wish to consider whether to enrich the resource by harvesting health science material from OA journals and repositories, or by linking to that content in its original locations
- OpenAIRE will need to consider whether to mark-up and enhance the content it harvests from institutional and other repositories so as to provide the functionality for the future that UKPMC (and the European PMCthat UKPMC is planned to be transformed into) is delivering.
- OpenAIRE should keep a watching brief on how the concept of open notebooks and similar initiatives develop in health sciences. There is an opportunity for enriching OpenAIRE content by linking to these things but the implications of that in management overheads could be significant

6 Current status of Open Access to research data

6.1 The policy foundation for Open Access to biomedical data

Policies on research data in biomedicine have been accumulating for some years now. The discipline is fairly well-advanced in this respect.

A requirement for the inclusion of a data management plan, including details of how data will be stored and managed for a future period after the cessation of funding, has been part of the policy of a number of large research funders for some time, although there are many funders with a mandatory policy on Open Access to the literature that do not have an accompanying one on data.

There are several permutations on funder position. They may have an OA policy on both literature and data, or just literature. The data policy may include the requirement for a data management plan (including how data

may be shared and how they will be cared for in the longer terms as well as through the lifetime of the project) or not. The funder may provide guidelines or rules about data sharing and curation, or not. And the funder may specify detail, such as the period of time within which data must be deposited in an Open Access location, or not.

The European Research Council (ERC), for example, has a data archiving/sharing policy that requires grant-holders to make their data available for others and that this occurs within a period of 6 months after the completion of the project[22]. The data (such as 'nucleotide/protein sequences, macromolecular atomic coordinates and anonymised epidemiological data') must be placed in an appropriate public databank. Examples of appropriate databanks are given as GenBank and PDB (Protein DataBank).

In the UK, the Biotechnology & Biological Sciences Research Council (BBSRC) has a similar policy to this, as does the Fonds zur Förderung der wissenschaftlichen Forschung (Austrian Science Council). Other European biomedical funders – the Wellcome Trust, Cancer Research UK, the Medical Research Council (UK), and the Országos Tudományos Kutatási Alapprogramok (Hungarian Scientific Research Fund), along with funders outside Europe such as the NIH, National Science Foundation (NSF; US), Canadian Institutes of Health Research, Gordon & Betty Moore Foundation, Heart & Stroke Foundation of Canada, Michael Smith Foundation for Health Research and the Ontario Institute for Cancer Research, have data access policies without any stipulation of how much time must elapse before researchers make their data available.

To an extent, this is to accommodate a variation in needs between disciplines: full data exploitation by their creators takes much more time in some fields, such as epidemiology or certain clinical areas, than in genomics or proteomics, and funders wish to allow researchers sufficient time to carry out all the analyses they want before sharing their data openly. There is, however, an argument for reasonableness and most funders would not expect data to be withheld for a decade or more.

6.2 Formal infrastructure for sharing research data

In the life sciences, data sharing is mature in many areas of health sciences, notably those focused on in this chapter. This situation has evolved because an open approach is the only one that could enable the challenges of modern life science research to be tackled, based as it is on analytical and/or comparative approaches and intensely data-rich. Without the development of a

[22] http://erc.europa.eu/pdf/ScC_Guidelines_Open_Access_revised_Dec07_FINAL.pdf

formal infrastructure for data sharing, this research simply could not happen. Public databanks for various types of biomedical data are, as a result, long-standing and the organisational infrastructure to support these in the long term is established in many cases.

6.2.1 Large public databanks

Funding for these organisations and their ongoing work is from national and regional funders. The main players are as follows:

- NCBI (National Center for Biotechnology Information), established in 1988 as a division of the National Library of Medicine at the NIH, Bethesda, USA
- EBI (European Bioinformatics Institute), part of the European Molecular Biology Laboratory, based at Hinxton, UK
- Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ), established in 1987 at the National Institute of Genetics, Yata, Japan

These three organisations curate and store biomedical data in a number of individual databases. They exchange and share data and researchers typically upload their data to, and download data from, the nearest site geographically. The original formal model for data exchange between sites was for nucleotide data resources[23], but this was such a successful example of formal data sharing on an international scale that it has now been followed by others.

Access is not enough in many cases, however. Tools for accessing data are also needed and the data custodians play a role here, too. They may carry out data assembly (put together multiple datasets to make a whole genome to save individual researchers having to do this), provide tools for searching and analysing datasets, and integrate data from other sources into the database (such as information on genes from journal articles). The outcome is a rich resource composed of standardised data elements, maximising the value to users. Figure G.10 shows an example of the multiple-view facility offered by EBI for a gene called Tpi1.

Curating life science data is not only about collecting and storing datasets and the operation of these large public databanks is sophisticated.

The metadata requirements are often demanding, ensuring that re-usability of the datasets is optimised. Researchers must follow strict rules on data structure and metadata entry when uploading datasets. The databanks employ a body of professional, highly-skilled developers and curators who check entries and will correspond with depositors if there are errors or inconsistencies in the metadata.

[23] For example http://insdc.org/

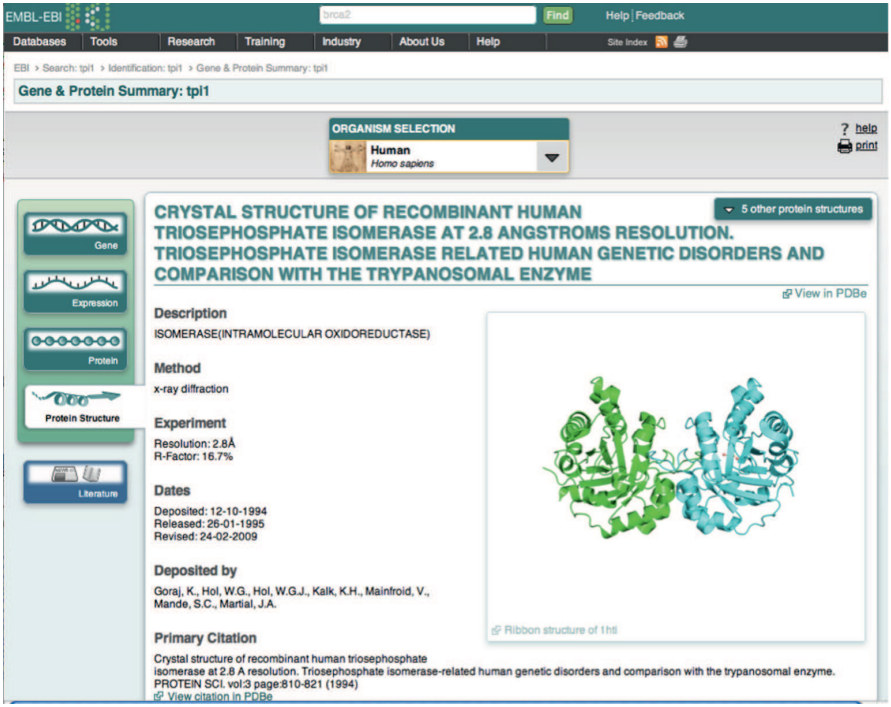

Figure G.10 Search results from the new EBI site search. Introductory information for genes (in this case, Tpi1) is shown in gene-, expression-, protein-, structure-, and literature-centric page views.

In some cases – such as microarray data – there is a further problem in that the data are meaningful only in the context of the particular individual sample used. Annotation therefore requires details of the experimental conditions and the gene name, but gene names are not yet fully standardised. Ambiguities in expression data and the possibility of many-to-many relationships between genes compound this problem. There are international efforts to establish ontologies and other standards that will resolve this. The MIAME standard (Minimal Information About a Microarray Experiment), developed at EBI and others, is a major advance here.

Moreover, primary data are often not sufficient for the type of work that needs to be done. The data curators therefore add value – considerable value in many cases, making data more accessible and more usable for the different constituencies that will use them: the curation process may include data

integration, which is an area where much effort is expended and which still presents many challenges.

The process can be thought of as tiered, with some databases undergoing more curation than others, and curation can be by human or machine:

- Preliminary data sets: examples of these are genomic short-reads (short fragments of DNA whose sequence is ascertained)
- Primary data records: for example, a whole-gene sequence
- Computationally-annotated data records (for example, assembled-sequence records, where a computer has put together separate sequences to construct something much large, perhaps even a whole genome of an organism)
- Curated data records (for example UniProt or RefSeq databases where human curation in the form of the searching out and adding of further contextual information, perhaps from journal articles or other sources, has taken place)
- Curated 'views' (for example the Reactome database, where expert curators construct biological pathways using data from many sources)

6.2.2 Small public databases

As well as the large databanks described in the previous section, there are thousands of small, specialised databases that are made openly available. Most of these are hosted on the websites of research groups or specific research projects. Examples are databases containing sequences from the genome of a single organism or information about single genes or gene families.

The journal *Nucleic Acids Research* (an Open Access journal), publishes a list of databases in molecular biology each year: the latest one features over 1300 of them[24]. This listing only covers molecular biology: outside of this field there are many more public databases covering diseases, therapies, diagnostics and so on. The discipline overall is rich with information.

The problem that can arise is one of sustainability. Many of these small databases are supported by project funding and when the project comes to an end, the database may no longer be updated or curated and may even disappear.

6.2.3 Journals

A number of major journals, particularly in molecular biology, have policies that require authors to make their data freely available to others when a paper is published.

[24] http://nar.oxfordjournals.org/content/35/suppl_1#EDITORIAL

The usual way of ensuring that this is done for data that should be in a public database is for the journal to require the accession number of the dataset, proving that it has been deposited in a database and providing the direct link to the dataset.

This is not a foolproof system, however. One study looked at the level of compliance with journal policies by checking for datasets that should be in GenBank. It showed that 9% of articles did not cite accession numbers of the datasets, even though the datasets were in GenBank (Noor *et al*, 2006). A further 7% had not submitted the datasets to GenBank.

Nonetheless, these are small percentages. The norm in molecular biology fields, at least, is to make data available for sharing. In these fields data sharing is relatively non-contentious because the purpose of the experiment is often to determine a genetic sequence and thus the scientist has achieved his/her objective by doing this and publishing the result.

For other types of research data and outputs, such as software, the journal itself may host the outputs on its website. This may apply quite widely: data from PubMed Central show that 25% of articles published in 2009, for example, have supplemental datasets attached to them. In such cases, reviewers may reject a paper if the supporting material does not accompany it.

One reward that researchers may enjoy from sharing their data is increased impact for their research. Piwowar *et al* ((2007), examining citations to microarray clinical trials, demonstrated that articles where supporting datasets had been made publicly available enjoyed an average 69% increase in citations compared to articles with no available data.

In fields where data take considerable time to analyse and exploit, there tends to be more reluctance on the part of researchers to relinquish their data to the community within a short period. Funders understand that there are significant cultural differences between research communities on this issue and generally word policies to accommodate these differences.

6.3 Informal infrastructure for sharing research data

Informal data sharing also goes on in health sciences, most commonly in fields outside of molecular biology. Research groups will usually supply data if asked by another group (though the data may not always be in a usable format). Such negotiations may also lead to more fruitful engagement in terms of formal collaboration or joint publication.

In many instances, though, data remain stored locally and never shared. Or there may be an attempt at sharing, by including some data in published articles, though in practice this makes access and re-use by third parties extremely difficult. A table of transcriptomics data in the published PDF file

of a journal article is effectively unusable without a great deal of work in manually transcribing the table contents into a software programme that can manipulate and analyse the data or integrate them with other datasets for further analysis.

What is clear is that large amounts of data that are unsuitable for the big public databank services (either because the datasets are too small or they are not an appropriate type) remain stored locally by their creators on hard drives or portable media. Discovery of these datasets is almost impossible when metadata are not made available on the Web and so they languish unused when they might be exploitable by others.

Institutions are rising to this challenge to a degree. There has been considerable work in the library community to scope and analyse the needs here, and some studies have gone some way towards providing a cost analysis and guidelines on practice for institutional efforts to preserve and curate research data[25]. There is, however, a clear need for this situation to be clarified and acted upon at a European level, and OpenAIRE may take a leading role here.

The diagram below shows the overall picture with respect to the literature, and data creation, manipulation and management, in the biomedical domain.

Implications for OpenAIRE

- Mandatory policies from the European Commission and from a growing number of health research funders and institutions will increase the amount of health science data accumulating in databases and repositories across Europe
- Continuing development of OpenAIRE policy on content acquisition may wish to consider whether to enrich the resource by harvesting health science data from institutional repositories, or by linking to that content in its original locations
- OpenAIRE might consider providing a storage and curation service for research datasets that are not suitable for the professional databanks but that should be made openly available. This service should be offered for all publicly-funded research, not only Framework Programme research

7 Challenges and opportunities

As mentioned in the Introduction to this chapter, research data management in the life sciences is comparatively advanced. Many basic principles of good practice and infrastructure development have already been established. The

[25] For example, one guideline for preserving and curating data is the Data Sea of Approval: http://www.datasealofapproval.org

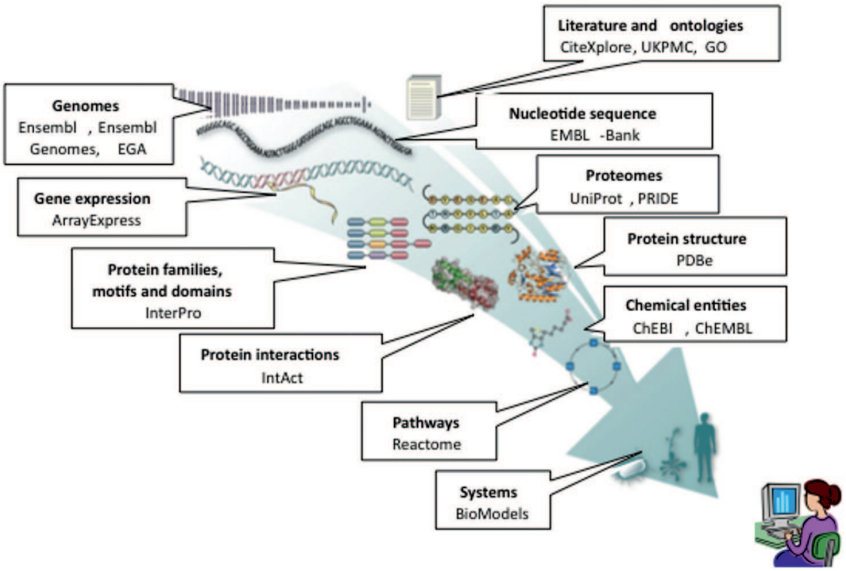

Figure G.11 The biomedical data arena

field is well-regarded by many as a worthy example of how to organise on a community level and put in place solutions that work for all. That does not mean that all problems have been solved: many remain, not least how to plan for and deal with the longer term challenges of data management.

The main information-related opportunities and challenges in health science research that are now receiving attention are as follows:

- Managing the increasing volumes of data generated. This introduces challenges in terms of providing access, storing and preserving the data
- Making cost / benefit assessments of data storage and preservation processes, so that decisions on what to keep and how are arrived at objectively
- What to do with 'small' datasets that are too small for the professional databases and will require manual curation
- Cost / benefit considerations for data curation: what do we want to keep / document?
- How to link data sets to core databases and the literature to create additional value

- How to resolve a number of generic issues around standardising metadata for discovery, data documentation and packaging of related files (including documentation)
- Persistence of datasets: there are initiatives, such as DataCite, that are attempting to address long term findability but the persistence of the actual datasets through the long term is still an area of concern, particularly for 'small' data
- Sustainability of data curation services in health sciences. The ELIXIR initiative, where European funding is provided for core data services in life sciences, is certainly part of the answer but the problem is bigger than this and growing
- Effecting behaviour change in areas where sharing is not the norm
- Attribution and intellectual property rights when datasets are 'stacked' (created from datasets that were in turn created by many others)
- Sharing of systems that include components that were provided by third parties. There are currently many barriers to sharing because third-party components are licensed in ways that prevent this
- Integration of data curation and data exchange facilities in the workflows of research groups, both technically and organizationally
- Problems of the use of data from secondary sources that could be better annotated
- Incentives for data curation and sharing by researchers: there is currently no career-advancement advantage in sharing data or putting effort into curating data to enhance their value to others and biologists do these things altruistically. A more formal reward system, akin to that traditionally offered for publishing research articles, would help here
- Text-mining the literature for material that enables data enrichment: this is a technical challenge, partly, but even more so a process challenge for database curators
- Privacy and data protection issues: this is an issue of major importance in health sciences, but will be particularly so with respect to 'personal genomes' (where an individual's genome is sequenced: what are the implications in terms of privacy, insurance and so forth?)

8 List of figures

9 List of tables

10 Bibliography

Björk B-C, Roos A & Lauri M (2009). "Scientific journal publishing: yearly volume and open access availability" *Information Research*, **14**(1) paper 391.

http://InformationR.net/ir/14-1/paper391.html

Björk B-C, Welling P, Laakso M, Majlender P, Hedlund T, et al. (2010) Open Access to the scientific journal literature: Situation 2009. PLoS ONE 5(6): e11273. doi:10.1371/journal.pone.0011273 http://www.plosone.org/article/info:doi/10.1371/journal.pone.0011273

Sale, AHJ (2006) Comparison of IR content policies in Australia. *First Monday*, **11 (4).** http://eprints.utas.edu.au/264/

Gargouri Y, Hajjem C, Lariviere V, Gingras Y, Brody T, Carr L and Harnad S (2010) Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. *PLOS ONE*, 5 (10). e13636 http://eprints.ecs.soton.ac.uk/18493/

Hajjem, C., Harnad, S. and Gingras, Y. (2005) Ten-Year Cross-Disciplinary Comparison of the Growth of Open
Access and How it Increases Research Citation Impact. *IEEE Data Engineering Bulletin*, 28 (4). pp. 39–47. http://eprints.ecs.soton.ac.uk/12906/

Matsubayashi M, Kurata K, Sakai Y, Morioka T, Kato S, et al. (2009) Status of open access in the biomedical field in 2005. Journal of the Medical Library Association 97: 4–11. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605039/pdf/mlab-97-01-4.pdf

Noor MAF, Zimmerman KJ, Teeter KC (2006) Data Sharing: How Much Doesn't Get Submitted to GenBank? PLoS Biol 4(7): e228. doi:10.1371/journal.pbio.0040228 http://www.plosbiology.org/article/info:doi%2F10.1371%2Fjournal.pbio.0040228

Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308 http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0000308