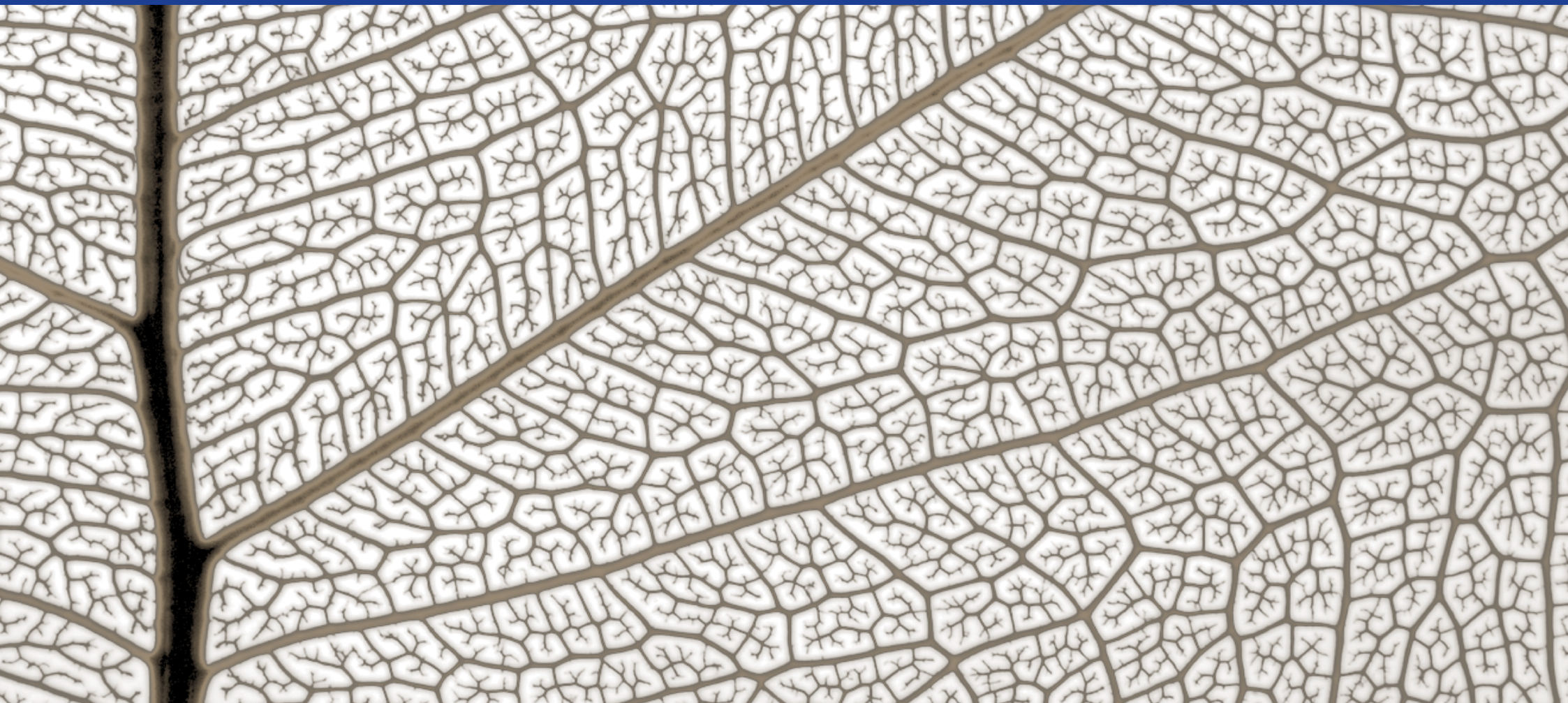
research information network

BRITISH LIBRARY

Patterns of information use and exchange: case studies of researchers in the life sciences

A report by the Research Information Network and the British Library

November 2009



The research on which this report was based was undertaken by the the University of Edinburgh's Institute for the Study of Science, Technology and Innovation and the Digital Curation Centre.

Acknowledgements

Robin Williams, Director of the Institute for the Study of Science, Technology and Innovation, and Graham Pryor, Associate Director of the UK Digital Curation Centre, at the University of Edinburgh wish to acknowledge the contribution of the researchers, Ann Bruce, Stuart Macdonald and Wendy Marsden and of the wider project team Jane Calvert, Marshall Dozier and Colin Neilson in carrying out and analysing the study reported here.

We are grateful to Michael Jubb and Aaron Griffiths of the Research Information Network, Allan Sudlow of the British Library, and Rob Procter of the Manchester e-Research Centre for valuable advice on the conduct of the study and feedback on this report.

We would like to express deep gratitude to our many life science respondents who made time in their busy research schedules to complete our research instruments and participate in interviews and focus groups.

Contents

Executive summary

Scientific advances, the availability of powerful new information and communications technologies, and new policies governing research funding have brought major changes for life science researchers. Together these developments have significantly altered both their needs and their practices in acquiring, generating and using information resources.

In this context, our key aim in the case studies we report on here has been to improve understanding of information use in the life sciences, and to provide a broader and deeper base of evidence to inform discussions about how information policy and practice can most effectively be supported and improved.

The starting point for each case study was the use of 'probes' - specially designed 'information lab books' – to chart individual researchers' information practices. We followed these up with detailed discussions in interviews and focus groups. The picture that emerges from our work is one not so much of transformation, but rather an uneven pattern as life scientists, individually and in research groups, grapple with the new functionalities and possibilities of use offered by emerging information policies, tools and services.

Above all, we found that the views and the practices of life science researchers are sharply divergent from the strategies adopted and promoted by policy-makers and information

service providers. Thus in seeking to identify relevant information resources, researchers use a limited range of services, and resort to informal advice from colleagues, rather than institutional service teams; and although sharing and exchanging information of many kinds is central to the ethos of life science research, individual researchers wish to choose what to share, with whom, and when. There is a significant gap between how researchers behave and the policies and strategies of funders and service providers. This suggests that the attempts to implement such strategies have had only a limited impact.

Our key conclusion, therefore, is that the policies and strategies of research funders and information service providers must be informed by an understanding of the exigencies and practices of different research communities. Only thus will they be effective in optimising the use and exchange of information, and in ensuring that they are scientifically productive and cost-effective.

Diversity: patterns of information use and exchange

There are marked differences in the patterns of information use and exchange both within and between different groups of life science researchers. There is much talk of 'big science', and our initial research design presumed that we would be studying large-scale formal collaborations. But we found that most research groups in the life sciences continue to operate on a relatively small scale, and we revised our plans accordingly. The groups we studied are thus rather unstructured and operate through plural associations, informal as well as formal, internal as well as external.

> "Life scientists, individually and in research groups, grapple with the new functionalities and possibilities of use offered by emerging information policies, tools and services."

Researchers may be part of larger projects and groupings, and mid-career and senior staff will typically be part of a number of more or less overlapping collaborations.

Differences in the intensity and character of the information practices we found *within* research groups reflect divisions of labour, expertise and responsibility. The ranges of information sources and exchanges we found among doctoral students are thus much more limited than those of more experienced researchers. Such variations are explained when individual researchers are located within the division of knowledge and labour within their particular research group. Thus any survey that examines responses from researchers divorced from their context and role can provide only an incomplete understanding of their information practices and needs. The complex methodology we adopted, by contrast, has enabled us to characterise the overall flows of knowledge and information in the groups we have studied, and provided important insights that might otherwise have been missed.

Our study also highlights substantial differences *between* the life science research groups we investigated, related to the research challenges they are seeking to address. Life science research encompasses a hugely intricate, indeed Baroque, range of formal and informal approaches to discovering, collecting, processing and disseminating information. Researchers do different things, even in apparently similar areas of study, as in the very different approaches to epidemiology displayed in two of our groups. There are of course some commonalities, and the groups we studied use similar processes that fit with well-established 'information life cycle' models. But each of them also shows a distinctive fine-structure of information use and production, including many intermediate information-cycles

Frameworks of support

Most life science researchers spend much of their time searching for and organising information. The groups we studied thus tend to manage information and data in an informal way. In most cases, no one had been identified to support access to and use of new resources and tools, to help with information services training, to advise on metadata creation, or assist with the curation of data and workflows. Most of them have very little contact with institutional library and information services.

But the groups we studied express a strong desire for information support, located close to them and if possible closely integrated with research teams and laboratories.

There are at least two aspects to this.

First, there is a need to re-establish a lively and sustained dialogue between information professionals and their research communities. Better engagement could add to the efficiency and effectiveness of research, with specialist support facilitating the use of new tools, and providing individuated professional advice, training and documentation on a subject or discipline basis. Such a strategy would have to be proactive, for researchers are reluctant to adopt new tools and services unless they know someone who can recommend or share knowledge about them. Support needs to be based on a close understanding of the researchers' work, its patterns and timetables. Integration is therefore key. In developing strategies to support researchers in data curation and sharing, for example, our research groups believe that only researchers themselves, or specialists located close to them, can have the subject knowledge necessary to curate their specialist data.

Second, there is a need for funders, institutions, and professional associations to work together to support the development of the emerging specialist roles – bio-informaticians in various guises, statisticians, modellers and curators – as well as strengthening broader skills as part of life science research training. Current career development and reward systems are not always effective in recognising and rewarding such roles. And project-based modes of funding make it difficult for small research groups in particular to sustain such roles, and the information tools and resources they generate. Policy-makers therefore need to work together to attend to the entrenched features of professional formation processes that inhibit effective information use and exchange. In particular, we recommend that an assessment should be commissioned of the national requirements for skills in research data curation and support, with a view to producing appropriate, effective and sustainable models for training and careers in managing information, and catering for the potentially diverging requirements of different domains of research.

Barriers to sharing data and information

The main goal of most of the researchers we studied is to add to our understanding of key aspects of the natural world – the mechanisms of a disease, an ecosystem, or the way in which certain molecules behave – and to communicate their findings. To achieve this they collect and generate various kinds of data and information. But for most researchers, collecting data and information is not a goal in itself.

Researchers communicate their findings – new knowledge, new methodologies and tools – primarily through conference proceedings and journal articles. These public activities have strong institutional and professional incentives in building reputations, securing promotion and so on. Incentives for other kinds of communication and sharing are weaker and indirect.

Most research councils have policies requiring researchers to set up formal mechanisms to manage created data, including provision for access and re-use. Moreover, the experience of sharing data such as gene sequences in high-profile research programmes in fields such as genomics or proteomics has come to be seen as something of a paradigm or model around which policies and practice will converge.

But our study suggests that such a model remains exceptional. Indeed, researchers highlight a number of barriers to sharing their research data, including concerns about potential misuse, ethical constraints, and intellectual property. Above all, they see data as a critical part of their 'intellectual capital', generated through a considerable investment of time, effort and skill. In a competitive environment, their willingness to share is therefore subject to reservations, in particular as to the control they have over the manner and timing of sharing.

Discussion of these issues has been hampered by confusions and inconsistent usage of the terms 'data' and 'information'. The current preoccupation with sharing research data has diverted attention from the diverse range of formal and informal information exchanges that take place in the life sciences. Given the limited current understanding of which forms of sharing and exchange are most effective and beneficial, and under what circumstances, we suggest that policy-makers need to engage in further discussions with researchers to identify and address the constraints, as well as to preserve the exercise of informed choice that is fundamental to science.

Narrowly prescriptive approaches are unlikely to be effective. We recommend rather that funders should adopt a more pragmatic and experimental policy that recognises the multiplicity of contexts, and the different approaches to information sharing; and which builds upon the informal sharing that is already taking place, based on the recognition of mutual needs. Such a bottom-up view is needed in order:

- to attend to the practicalities of data sharing: what makes information from other sources intelligible? Under what circumstances is such sharing useful and sufficiently beneficial to warrant the labour necessary to achieve it? and
- to address existing barriers and drivers for change, including the perceived self-interests and goals of researchers, and their need to sustain their intellectual capital and advance their careers.

A key message from our work, therefore, is that policy intervention and support systems for researchers need to be built around the many different and successful tools and practices emerging within life science research communities themselves.

1. Introduction

A combination of factors is commonly seen as heralding change in information practices and information service requirements for life science researchers.

These include:

- the availability of an array of powerful, and increasingly 'user-friendly' information and communication technology (ICT) tools and services
- the development of new analytic techniques and instruments, including the use of automated analysis facilities, yielding data – for example for gene sequencing – on an 'industrial' scale
- research funders' increasing interest – supported in some cases by new policies and incentives – in promoting the sharing of data, and
- the success of large scale collaborative research projects in areas such as genomics and proteomics, based around large-scale data repositories, with leading segments of the life science research community calling for the wider adoption of standards and protocols needed for such data sharing.

The last of these has come to be seen as something of a new paradigm in the life sciences, central to the emergence of new fields of enquiry such as bioinformatics, genomics, proteomics and systems biology. Some have seen these changes as bringing in an era of big science – as perhaps the equivalent for life scientists to the projects associated with the Large Hadron Collider at CERN, which bring together tens of thousands of distributed researchers around a single facility. Much recent social science research in this field has focused on laboratories in which new bioinformatics-centred approaches are salient. The focus on these new developments brings the risk that they may be taken as representative of life sciences as a whole, or as ideal types around which future life science research will converge.

Our case studies seek to provide a broader evidence base about information practices across life science research; and they provide an important corrective to this kind of vision. In contrast to current discussions of transformation, they reveal a more uneven pattern, as life scientists individually and in their research groups grapple with the changing 'affordances' of emerging information tools and services available for their diverse activities.

Affordances

The concept of affordances refers to the "quality of an object, or an environment, that allows an individual to perform an action" and is used in design and technology studies to highlight that the uses of an artefact are not a simple product of its functionality but arise through human exploration of its applicability to their purposes.

Through a wide array of cases, across differing areas of life science research, we have been able to assay the current state of the art of information practices. Our methodology allowed us to address the specificity of information practices of individual researchers and the groups within which they are located. This report represents our main findings, including an explanation of patterns and trends as they emerged across the entire sample of our case studies, and observations on the implications for funders and information service providers.

"Through a wide array of cases, across differing areas of life science research, we have been able to assay the current state of the art of information practice."

Aims and objectives

The broad aim of our project was "to enhance understanding of how researchers locate, evaluate, organise, manage, transform and communicate information resources as part of the research process, with a view to identifying how information-related policy, strategy and practice might be improved to meet the needs of researchers".

Our objectives were

- to analyse how researchers in the life sciences
    - make use of the information resources and services provided by publishers, libraries and other service providers to discover and gain access to information sources relevant to their research
    - analyse, evaluate and manage the information they acquire through such services
    - create, gather, manage and communicate new data and information in the course of their research
    - produce, present and disseminate new information resulting from their research
- to set out systematically the distinctions, commonalities and contrasts among the practices and needs of researchers in different fields, disciplines and institutional contexts
- to provide an indication of changes that researchers anticipate in information management practices and requirements in their fields
- to identify and explain the barriers to more effective performance in using, creating and managing information resources, and to information exchange across disciplines and fields
- to develop proposals, aimed at specific agencies, for general and specific measures to overcome these barriers, improve provision of information, and meet the changing requirements of researchers.

"The aim of our project was to enhance understanding of how researchers locate, evaluate, organise, manage, transform and communicate information resources as part of the research process."

Methodology

A challenge for this study was to obtain a reliable and detailed understanding of life science researchers' information practices and perceived needs across a wide range of contexts and within the project's time and resource constraints.

Seven case studies were completed across a diverse range of laboratories and research groups. They included groups studying humans, animals and plants, and covered different kinds of research context, encompassing analytical laboratory-based research, field research and in-silico research. The nature of the data used in the research process also varied, including quantitative, image, clinical, laboratory-derived and field data (including aquaculture and botanic collections). The full details of the seven case studies are available in the Annex.

The starting point for each case study was the use of self-administered 'probes' to chart individual researchers' information practices. Probes describe a cluster of approaches and tools with a range of applications that are used in social science research. In this study we used a 'diary', the information lab book (ILB), as a simple tool for collecting 'informational' data. Participants were asked to keep the ILBs for five days, recording the information that they were *using* (who or where was the information coming from and what kind of information was being used) and *creating* (what kind of information was being created and who/where would the information be shared with and how).

We collated this information into tables for comparison within and between groups. We also used it to produce flow maps of the research process, capturing and displaying the main information search, exchange, transformation and curation processes by integrating the data to produce a composite map for each group, highlighting the flow of information within the research life cycle.

The ILBs allowed individual researchers to record their information practices in detail over a short period, over which they could be expected to exercise reliable recall. They were followed by face-to-face interviews with a sample of researchers from each group, to elaborate upon and explain the data, and to show how these results formed part of the division of labour, expertise and knowledge flows within each group. We thus collected a large amount of information which we analysed using flexible tools and techniques borrowed from 'cognitive mapping', allowing us to map diverse information search, transformation and exchange processes against a model of the research information lifecycle.

Finally, we organised focus groups with each group. We used these first to verify empirical findings and our mappings of research processes and information flows, and second to encourage researchers to reflect and comment upon their practices and needs and more broadly upon the implications for information policies and strategies. A complete account of the methodology and phases of the analysis are included in the Annex.

By deploying a range of research methodologies and tools – including short-term ethnographic techniques and semi structured instruments – we were able to build a rich picture covering a wide range of kinds and contexts of life science research. For the purposes of this study the methodology was extraordinarily successful. Deeper insights would of course be possible from more extended ethnographic study.

2. Summary of findings

We present below brief summaries of our seven case studies, along with a flow map that seeks to characterise the main research and information processes undertaken in each case.

We present each study in turn and then, as a first step towards comparative analysis, provide a brief overview of the range of different kinds of data creation and usage activities across the whole set of cases.

Information flow maps

The maps consist of 'activities' or 'concepts', joined by 'links'. Activities are colour-coded so that a range of different processes can be distinguished on a single map.

Activities or concepts are expressed as short statements, each covering a single action or notion such that sources of information and types of information are linked but separate: for example 'on-line shared folder' as a source of information providing standard operating procedures (SOPs) as a second concept indicating the type of information produced.

Links: activities are linked by straight lines with arrow-heads indicating a relationship i.e. A 'may lead to' B. Links act in the direction of the arrow and are positive.

Different colours are used for different types of activity within an information cycle, adapted from a model developed by Charles Humphrey (2006).

We produced maps for each individual diary and then combined them into a single composite for the case study. These maps are a way of summarising the information given in the case studies and demonstrating clearly the links between various aspects of information activities, which enables comparison between the activities in different case studies. We discussed the composite maps with the focus groups and amended them as necessary.

Knowledge transfer activities (including grant writing, technical reports, journal articles, presentations and popular articles) were found to be common across the case studies and have been separated out from these diagrams and are reported in Section 5.

3. Case studies

1: Animal genetics and animal disease genetics

Our first group studies animal and animal disease genetics, falling broadly under the disciplinary categories of epidemiology and quantitative genetics/genomics. Their main information sources involve field data (farm and aquaculture), experimental field data and laboratory data (gene sequence information).

The processes carried out in the group are largely desk-based, the key one being the analysis of quantitative data created by industrial or research partners. A prerequisite to this is the ability to extract the desired information from data sets, to carry out quality control procedures, and to format the data appropriately for statistical analysis packages. Statistical analyses are carried out on an iterative basis; results are interpreted and published in reports and scientific papers. The computer processing power required for the statistical analyses depends on the type of analysis. Primary analysis and data production can be achieved within a desktop environment, whereas secondary analysis of large amounts of data is conducted in a grid/parallel processing setting.

Figure 1 provides an overview of the main information flows for this group (numbers in parentheses below indicate the numbers used in the Figure).

The main information sources used are:

- industry (24),
- experimental field data (23), and
- information from collaborators (28).

Scripts are used to extract data in suitable format (5). Extensive use is made of statistical software packages: ASReml (19), Survival kit (33), Grid QTL (22) and its predecessor QTL Express (20).

Analysis is an iterative process and some results may be displayed in graphical format.

Appropriate genetic data (such as genetic marker information) is likely to be posted on public databases e.g. the Australian Sheep Gene Mapping website, VISTA Genome (37).

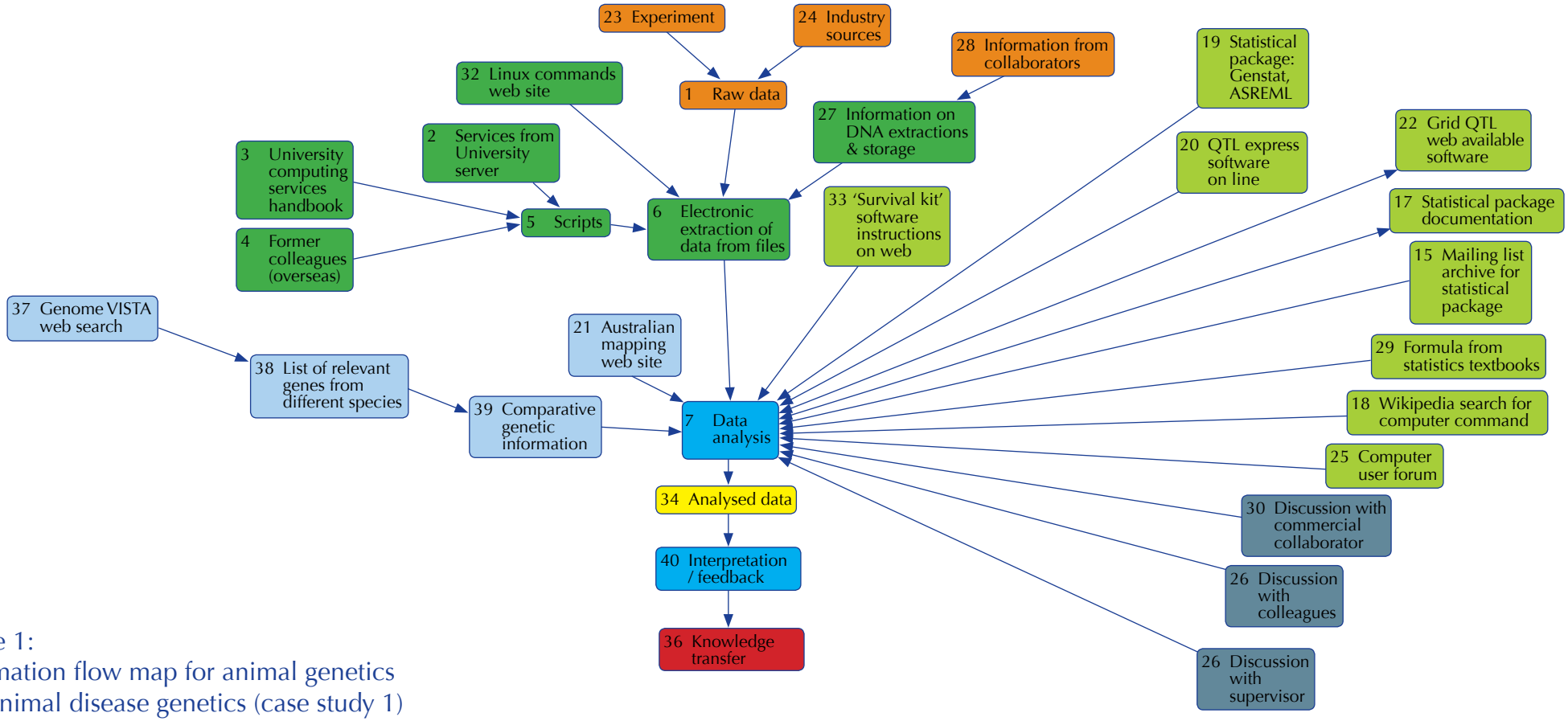


Figure 1:
Information flow map for animal genetics and animal disease genetics (case study 1)

2: Transgenesis in the chick and development of the chick embryo

This group applies transgenesis (the process of introducing an exogenous gene into a living organism) to chickens and uses this technique to study developmental processes and viral diseases; and to produce therapeutic proteins in eggs. Transgenic chickens carrying different marker genes that allow the study of developmental processes are distributed to other academic groups, enabling their research. The group is developing techniques for introducing various different genes, (e.g. as markers, or for specific function, or of size) into embryonic chromosomes to produce transgenic eggs which will develop into organisms, i.e. chickens. The main data sources involve lab-based processes.

Figure 2 provides an overview of the main information flows of this group.

Data is collected from experimental material (17) which is then cultured and evaluated in the lab (2).

Heavy reliance is also placed on information and instruction from biotech companies (32) on how to use their equipment and reagents e.g. how to use kits to purify DNA, information on enzymes – how they work and which loci they will cut; online protocols and instructions (33). Standard operating procedures (SOPs), both internal (15) and shared standards within the discipline (31), are important.

Results, in the form of gel images (34), cell data (38), and quantitative density information (35, 4) are obtained and made available for access through a Lab Book (24) which is shared with colleagues (23). Applicable genetic data (such as genetic sequence information) is obtained from and deposited in open access repositories, such as the National Centre for Biotechnology Information (NCBI) (19).

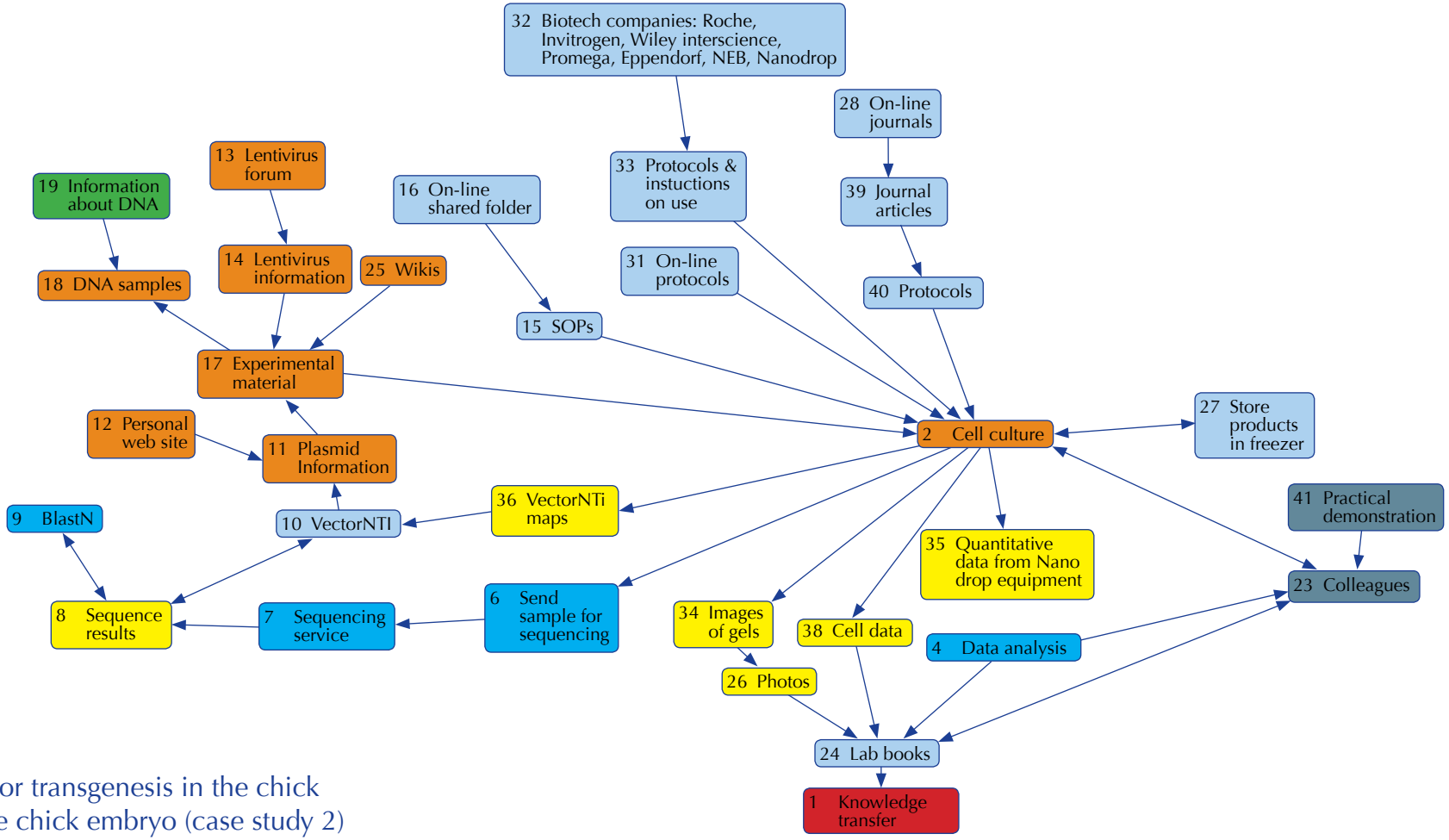


Figure 2:
Information flow map for transgenesis in the chick and development of the chick embryo (case study 2)

3: Epidemiology of zoonotic diseases

The third group is a specialist team in immunology and infection research and is one element of a larger collaborative group. The group studies a range of epidemiological issues involving domestic livestock and humans, in particular the epidemiology of zoonotic (diseases transmitted from animals to humans) and emerging/re-emerging diseases.

They work with a range of local and international partners, and are involved in research projects based both in Europe and in Africa. Their primary information sources are field data (animal and human), spatial data (including GPS tracks, GIS map layers, satellite imagery, data from government agencies) and numeric disease data. Field data are collected by members of the team or by collaborators, and are brought back to the institute for digitisation, processing and analysis. The processing carried out in the team is primarily desk-based.

The group also carries out lab work (but not at the time of this study) and they work in close collaboration with other groups at the institute who are working in related fields. The group has a focus on obtaining good quality data, which comes from a number of different sources.

Figure 3 provides an overview of the main information flows for this group.

The group gathers experimental field data from colleagues (15 & 16) and health and veterinary health surveillance agencies (50); raw data obtained from questionnaires or from colleagues (11, 12 & 35); data from published papers (13); and spatial data obtained from data services (3, 5, 6) or collected from the field (51).

The data are analysed using statistical packages (1), including GIS map layers (4) and a variety of different information sources used to inform the analysis such as text books (20), journal articles (online and hardcopy) for mathematical equations (for modelling) (43, 44) and other web sources.

A specialist wiki site (9) is used to share information with internal and external colleagues on a variety of topics including spatial epidemiology methods and tools, observational study design, and statistical analysis.

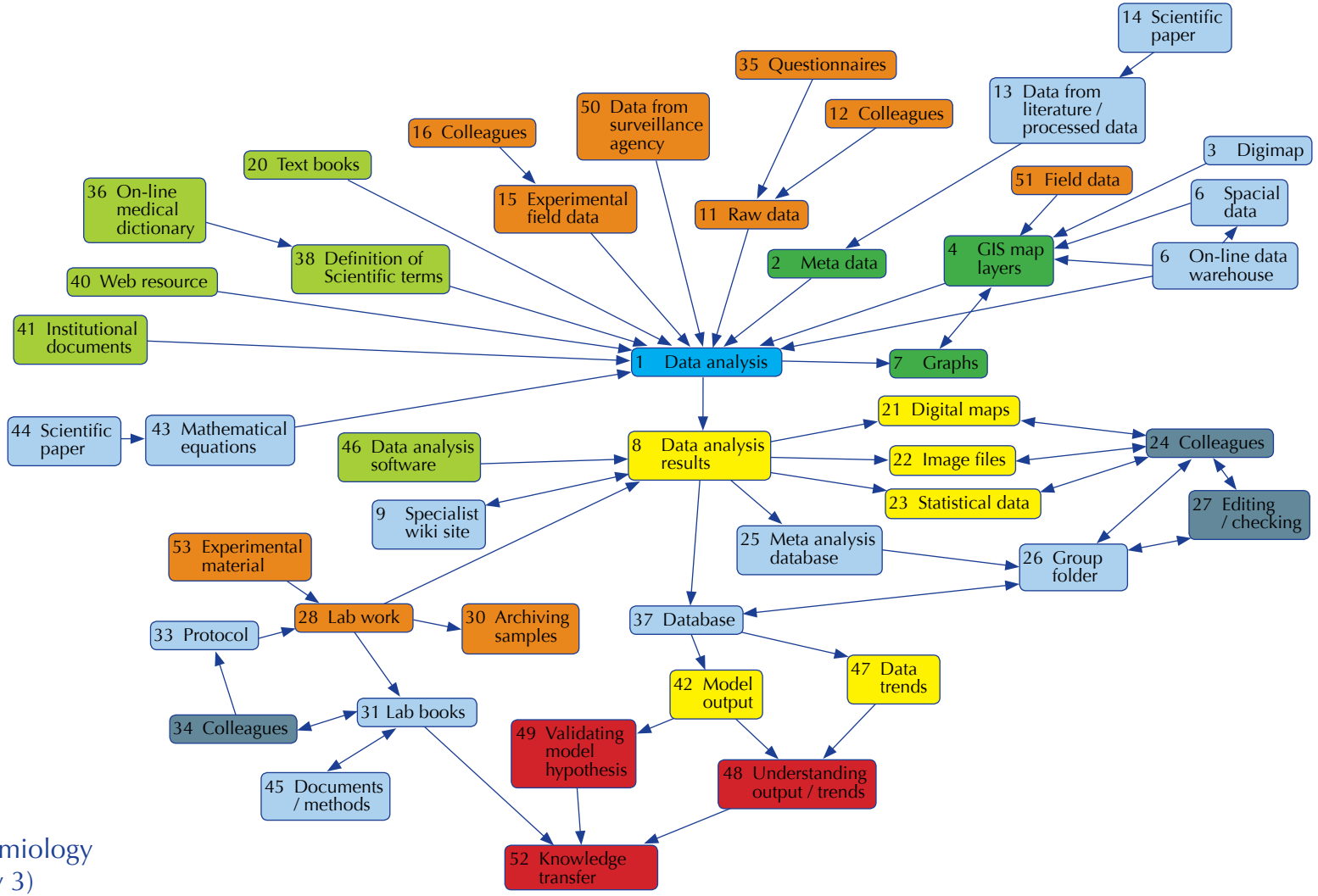
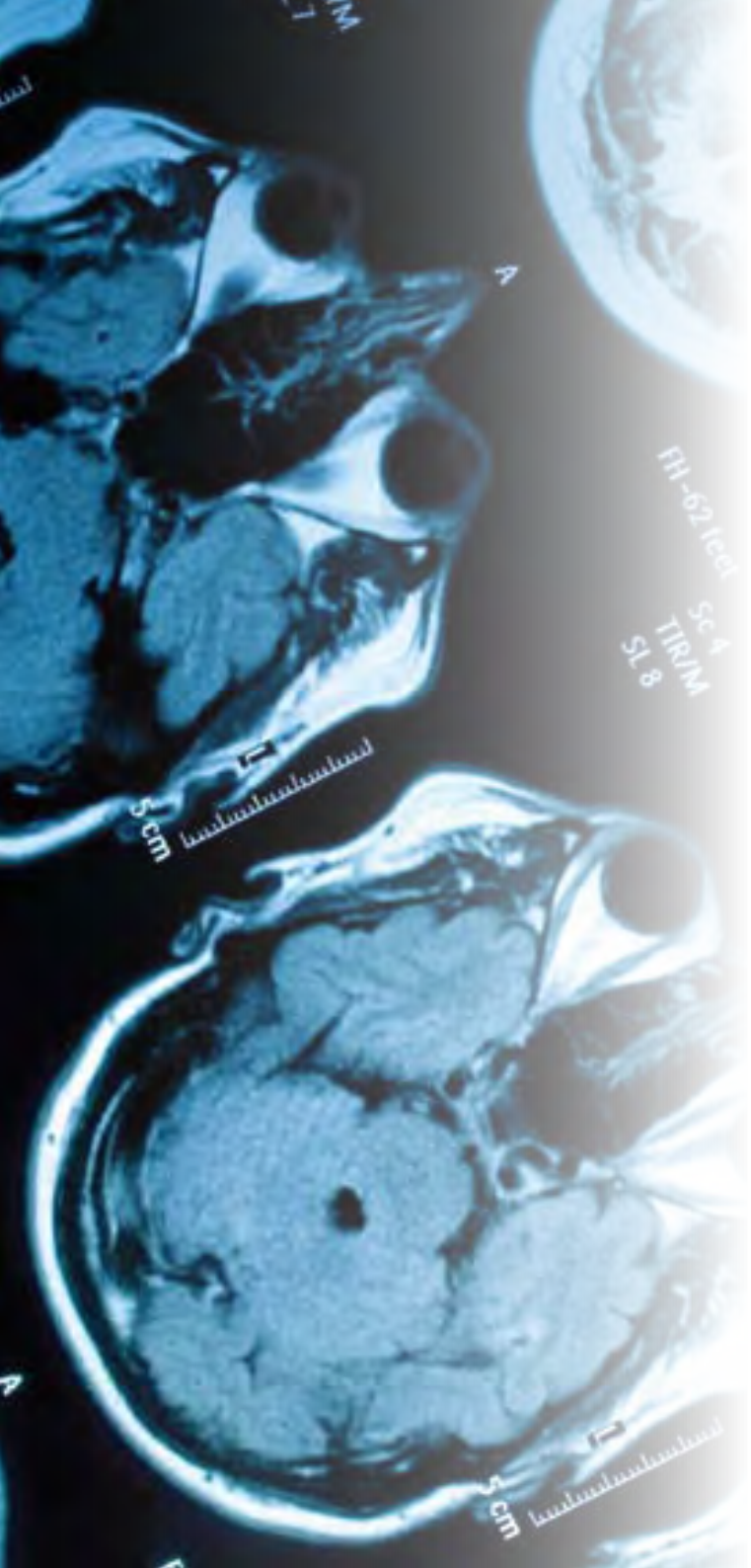


Figure 3:
Information flow map for epidemiology of zoonotic diseases (case study 3)



4: Neuroscience

The group works to develop research tools to improve the management of major psychiatric disorders. This primarily involves the application of neuro-imaging techniques to distinguish people with schizophrenia from their relatives and those with other disorders. Future studies along these lines will have to be done in multiple centres, so the group is developing the capability for multi-centre brain imaging studies by developing image analysis techniques for combining existing scans; harmonising metrics and calibration techniques for future studies; and establishing common image acquisition protocols of interest to psychosis researchers across five sites in the UK. All these data need to be carefully collected, annotated and curated – with the ultimate aim of being able to furnish a neuro-imaging database of normal and abnormal human brain development as a reference resource, such that risk estimates for future psychosis could be provided on an individual case basis. In addition, they are developing new acquisition and analysis techniques to be able to distinguish risks for different disorders.

The group is also working with a pharmaceutical company on functional magnetic resonance imaging (fMRI) biomarkers of cognitive enhancement, on combined genetic and imaging approaches, and on the development of novel therapeutic strategies using animal fMRI. The successful application of all of this work in clinical practice will depend on developing new treatments, and on patients being willing to try these new therapies.

Figure 4 provides an overview of the main information flows of this group.

Data consist of various scans to produce brain images (11, 42, 28) which are stored on a departmental server (21) for use by several colleagues (6). Additional data are derived from demographic data (12), behavioural data (17), genetic data (18) and animal imaging data (19).

Intermediate brain image data (44) are produced to identify specific features using various analytical tools (43, 47, 45). Various data are then used for specific analyses (1) which may include use of other statistical packages (40) as well as other published resources (27, 20).

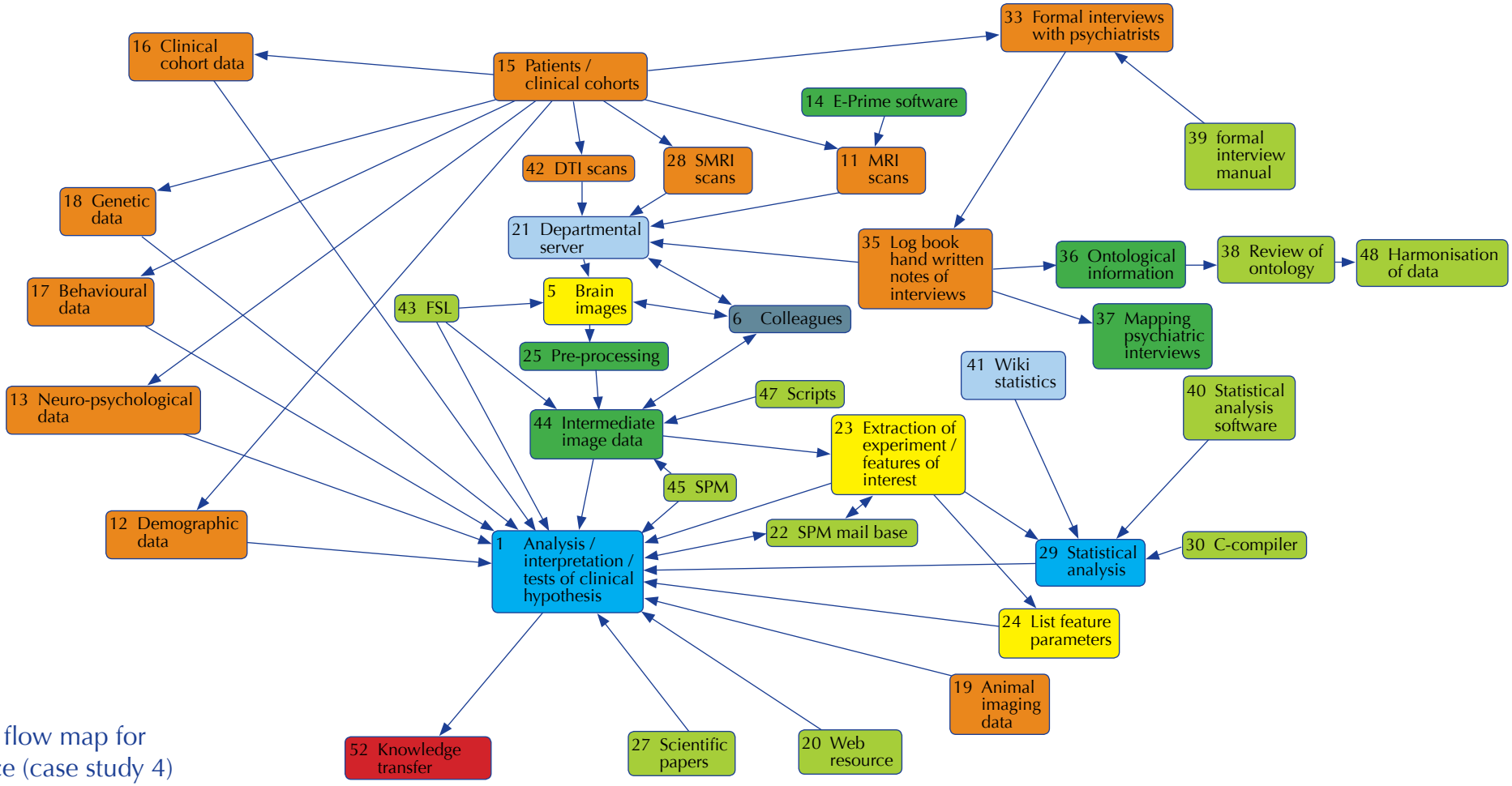


Figure 4:
Information flow map for neuroscience (case study 4)

5: Systems biology

Systems biology is an approach to the study of biological phenomena that integrates experimental and theoretical research using large-scale modelling. This group's research goal is to develop broadly-applicable methods and large-scale infrastructure for modelling biological phenomena. Modelling is central to systems biology. It involves taking existing knowledge, often in the form of large datasets and static and kinetic models, and using it to generate new knowledge.

The group is complex. There is a core of people in the centre, who work across a range of ten different projects. Their expertise includes: engineering, bioinformatics, mathematics, computer science, proteomics, cell biology, and biophysics. Some of them are interdisciplinary, combining knowledge from two usually distinct areas of work. There are three key research areas central to systems biology: experimental biology, mathematical modelling and informatics.

Figure 5.1 provides an overview of the main information flows of this group in computing-based processes and Figure 5.2 in lab-based processes.

The computing based process described in Figure 5.1 relies on statistical data from experiments (29) as input into the modelling process. Numerical algorithms are developed (21) and a range of modelling tools (25) is used and the results shared with colleagues (16) through a wiki (12). The wiki is also used to store meeting minutes (14) and presentations (13) as well as plans for future developments (15 & 17). A range of software is developed (1) to use in the modelling (21) or to extend standardised information exchange (20) and may be shared in publications (19) or via code repositories. The programming process (1) is informed by a variety of sources including documentation on programming tools (10), tutorials (8 & 9) and journal articles (6).

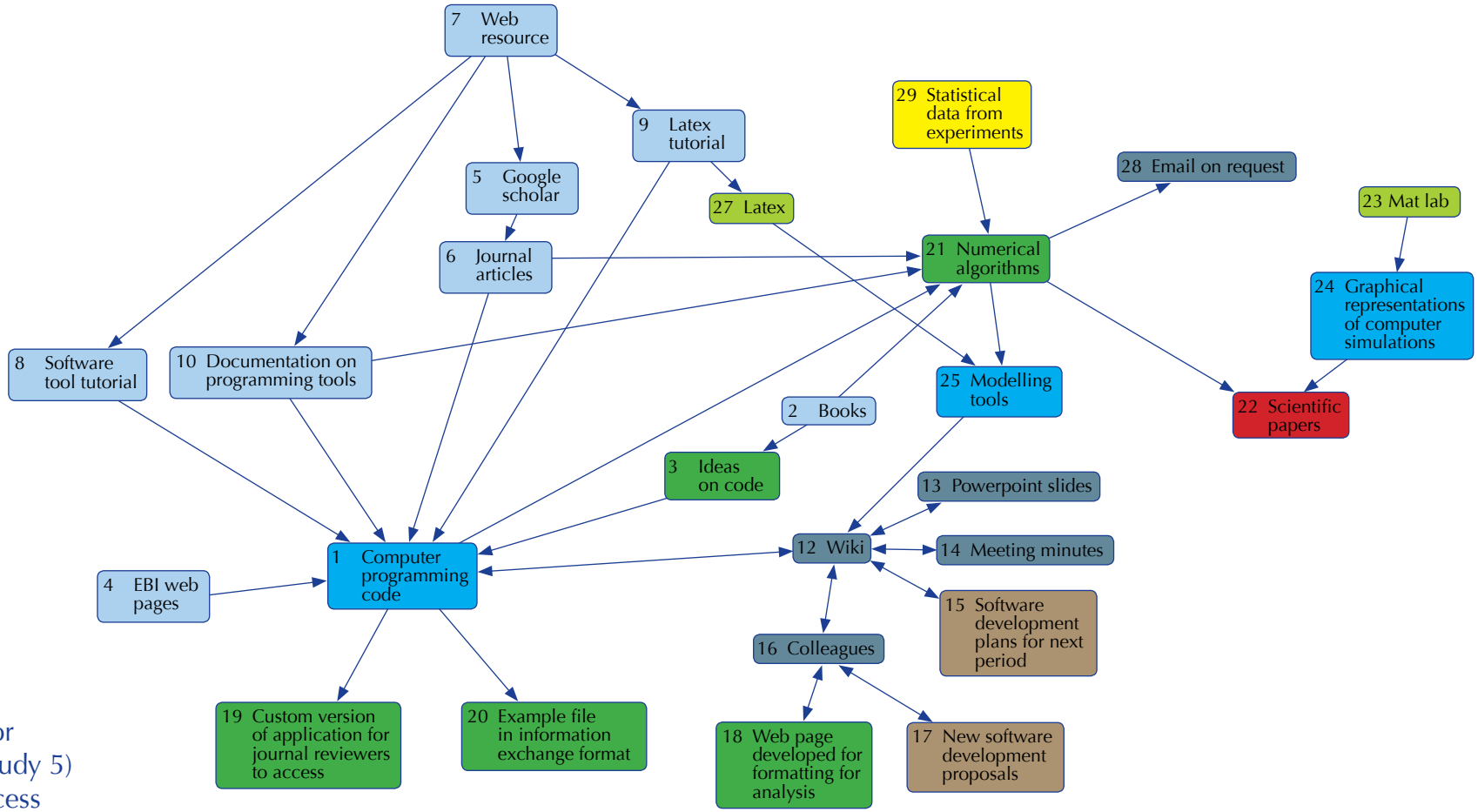


Figure 5.1:
Information flow map for systems biology (case study 5) – computing-based process

The lab-based process described in figure 5.2 relies on data from the experimental material (19) and experimental procedures (1) on which measurements are taken, in this case mass spectrometry (2). Information is stored in lab books (3) which may be shared with colleagues. The experimental procedures require knowledge of Standard Operating Procedures (SOPs) (6) and the running of specialist equipment such as the mass spectrometer (2) are informed by equipment manufacturers (8, 5) and may require specific information on chemical interactions obtained from a variety of sources including Google (21) and Science Direct (13)

> ‘There are three key research areas central to systems biology: experimental biology, mathematical modelling and informatics.’

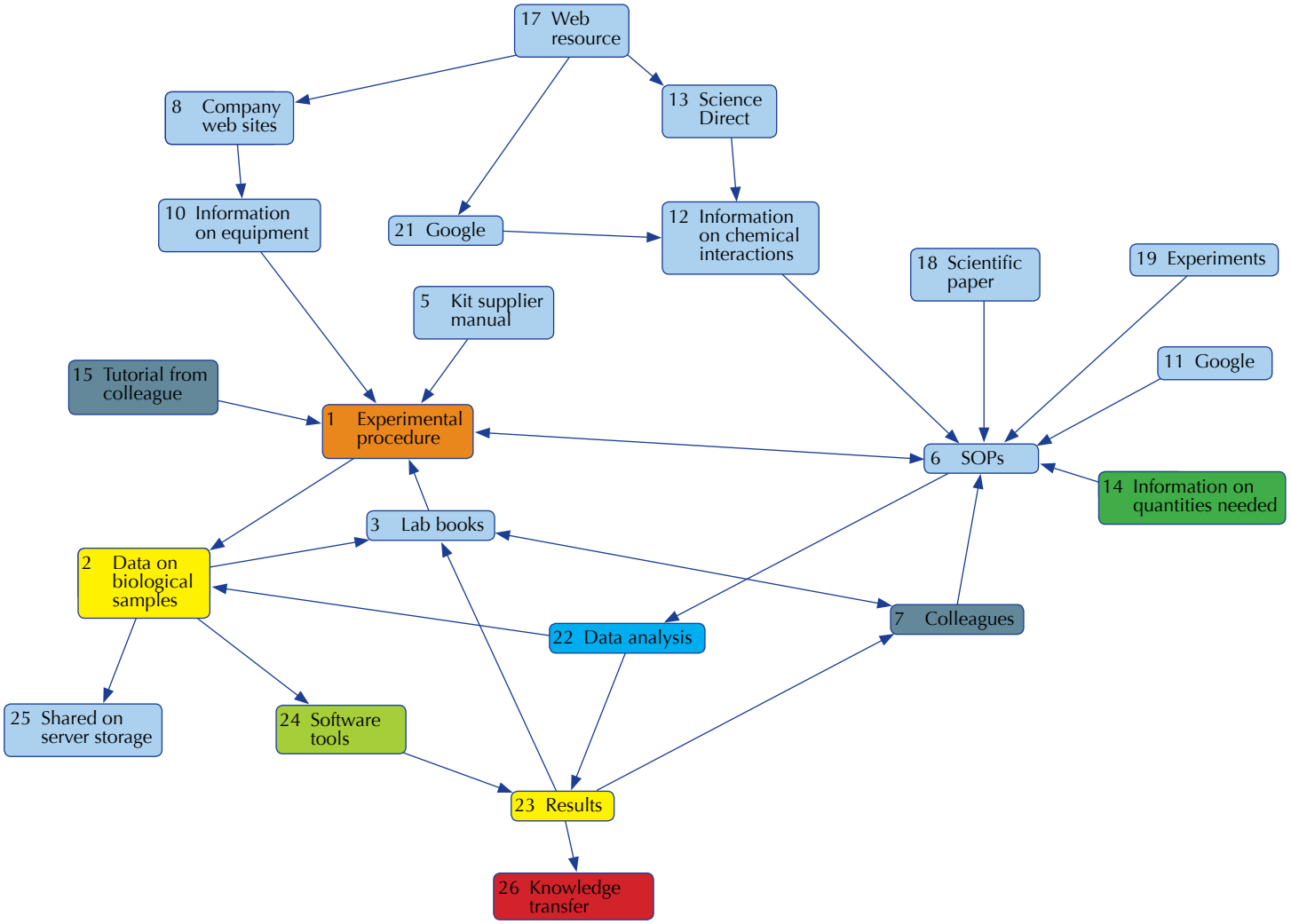
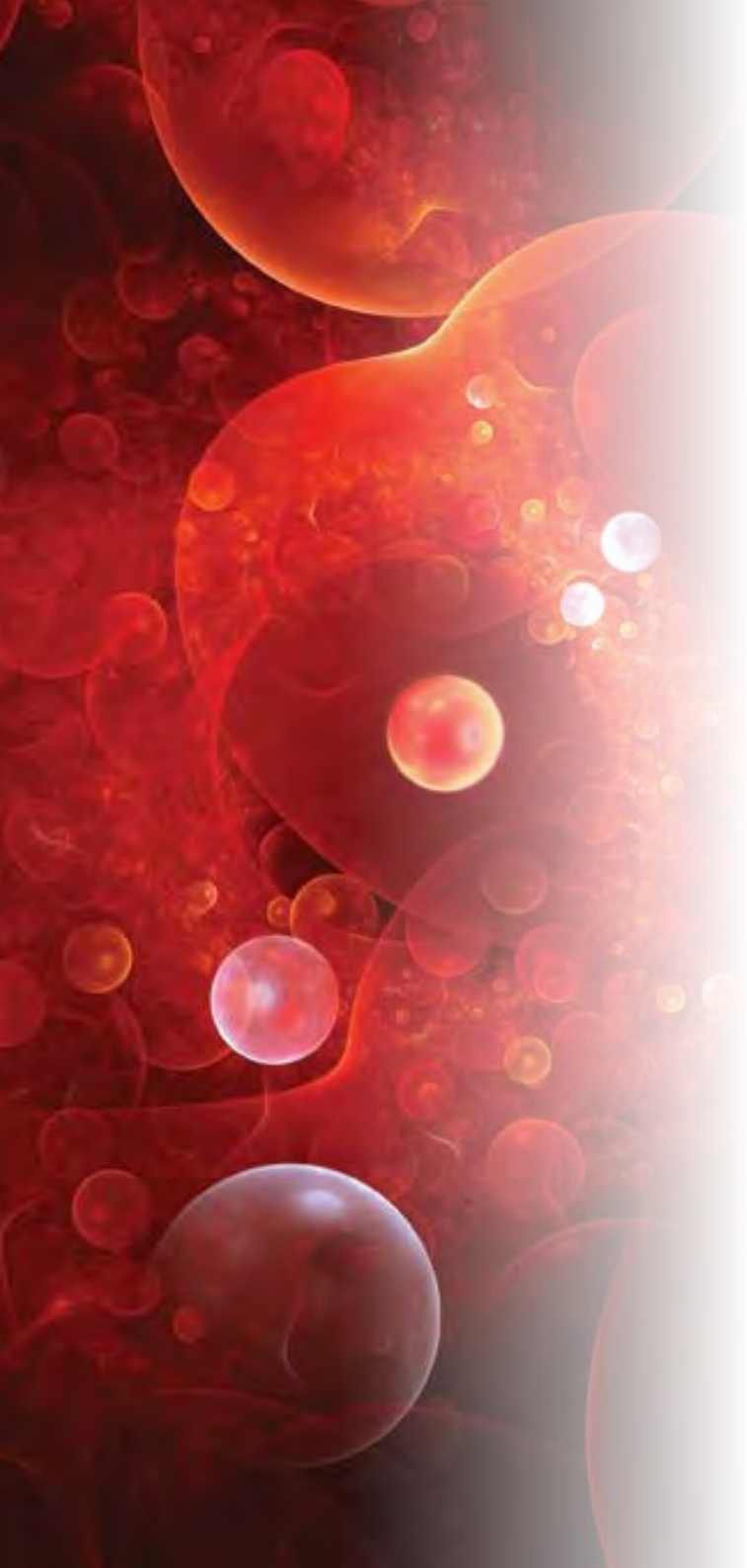


Figure 5.2:
Information flow map for systems biology (case study 5) –lab-based process



6: Regenerative medicine

This group works in the field of regenerative medicine. They have three key areas of research: the cellular basis of specific diseases; the influence of cells of the immune system on the repair process of specific tissue, and the production of cell-based therapies derived from both adult and embryonic human stem cell populations.

The group is currently working on a project to create a therapeutic 'product' which can be developed commercially. The potential to patent means the project work is not currently shared outside the group; nor do other groups working in the same field share data or information except by publication. Patenting has to be complete prior to publication. There are three areas of lab work, the cell cultures, protein and gene assays and the materials.

Figure 6 provides an overview of the main information flows of this group.

The main information source for this group is their own lab data (8 & 17) and work on materials (12). A range of information from external sources is accessed including technical product information (27), information on chemicals (11), techniques (10), journal articles (15) and reference books (32).

Various analytical methods are used to obtain research results: microscopic analysis (2), x-ray images (22), CT-imaging (34), immunocytochemistry (23), cell-counting (18) and QPCR data (28). Much of the data consists of large image files (1) which are primarily stored on individual PCs (35), although some may also be placed on a shared drive (3) which is accessible by all colleagues (5). Extensive use is made of the shared drive which stores information on protocols (24), information on projects in other labs (31), new research plans (19), and information from commercial collaborators (30). Data from current work may also be used for future work (29).

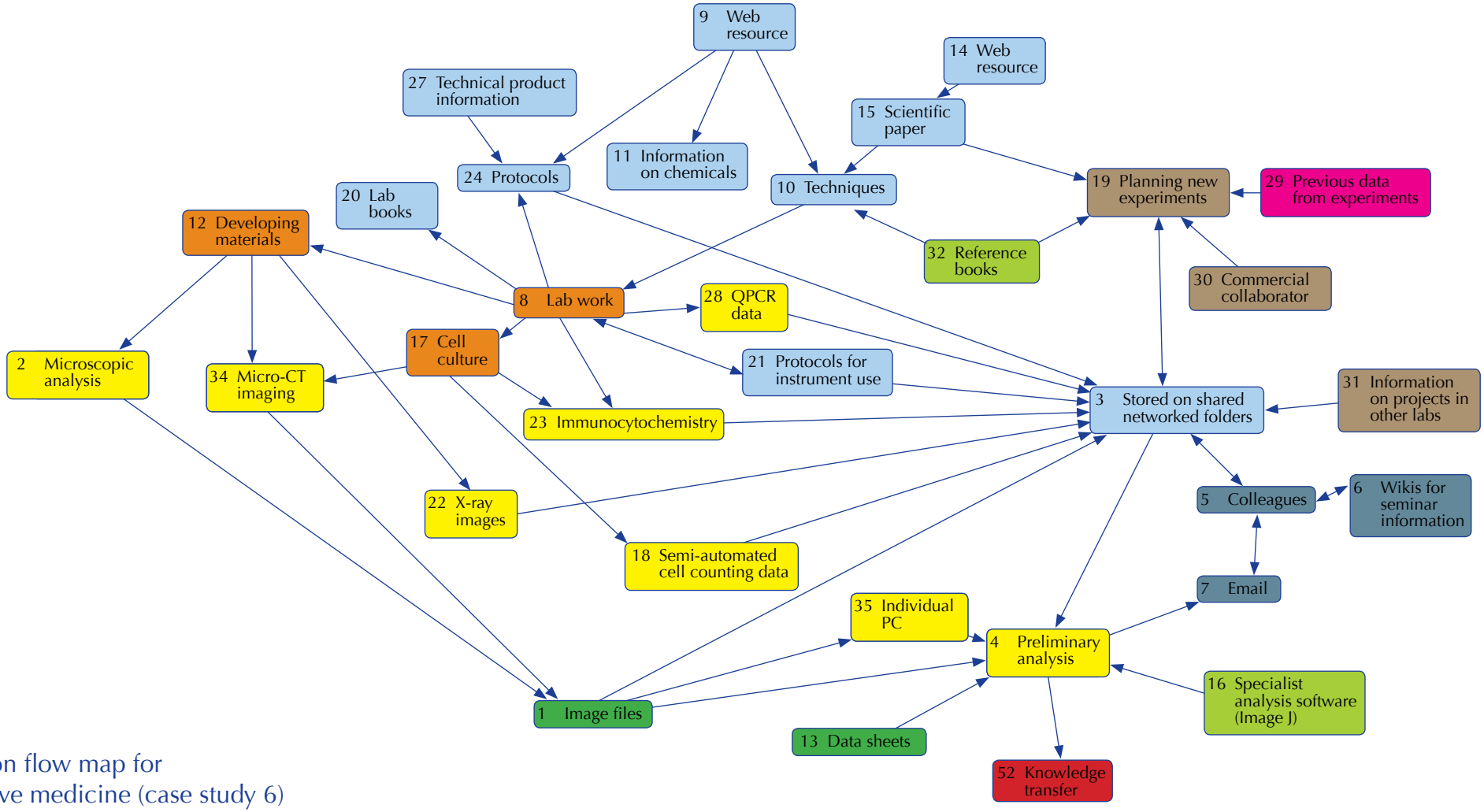


Figure 6:
Information flow map for regenerative medicine (case study 6)

7: Botanical curation

This is a small specialist group, working in a herbarium: a collection of preserved plants stored, catalogued and arranged systematically for study by both professional and amateur taxonomists, and by botanists. They undertake professional classification of organisms into groups and hierarchies on the basis of their structure (e.g. morphology), origin (e.g. genetics) and behaviour. The specimens in the herbarium are a working reference collection used in the identification of plants, the writing of Floras (a description of all the plants in a country or region), monographs (a description of plants within a plant group, such as a family) and the study of plant evolutionary relationships.

This herbarium numbers nearly 3.5 million specimens representing half to two thirds of the world's flora. It is considered a leading botanical collection, and every year many researchers from around the world visit to study specimens. The collection is actively used to support research at institutions around the world. They send out on average 4,000 specimens on loan per year, and add 10,000 new specimens to the collection each year. There are a number of digitisation projects focusing on the herbarium collections and other material relating to herbarium specimens.

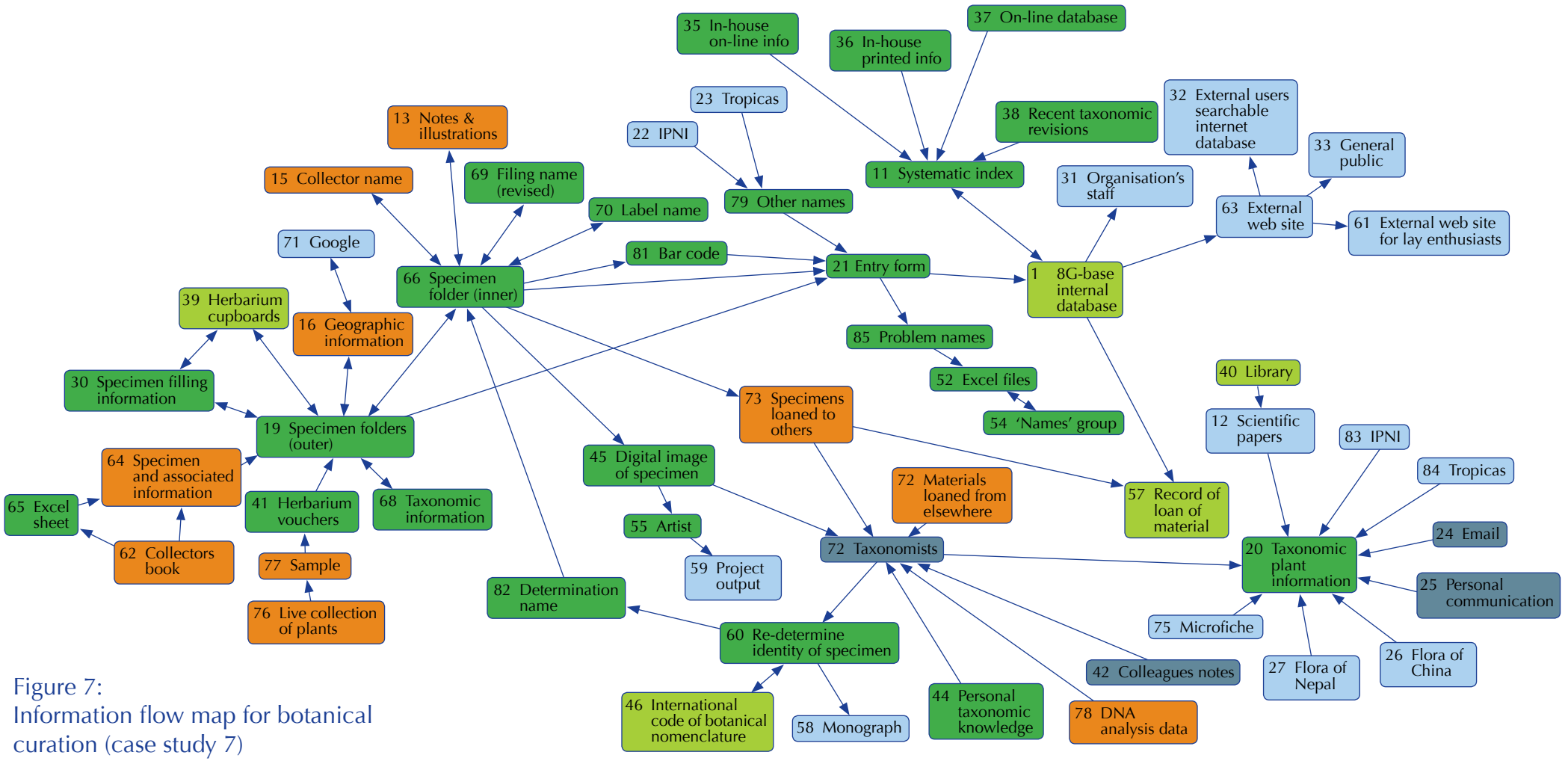


Figure 7:
Information flow map for botanical curation (case study 7)

Figure 7 provides an overview of the main information flows of this group.

The main information sources are the specimens and their associated information (64) held in folders (19 & 66) in the herbarium cupboards (39). These specimens have been collected in the past and are still being collected. New information is derived from collectors' books (62) which may already have been translated into Excel sheets (65).

Information from the folders (19 & 66) is either being digitised using entry forms (21) into a database (1) which can then be accessed by internal staff (31) or is made available to external interested people: other taxonomists (32), lay enthusiasts (61) and the general public (33).

Information stored in the herbarium (19 & 66) may also be combined with specimens from elsewhere (74) to be reviewed by taxonomists (72). The identity of specimens may be re-determined (60) and this re-determination communicated back to the herbarium (82) or if appropriate, the international code of botanical nomenclature (46). The re-determination of the identity of a species also relies on a range of other information including DNA analysis (78), personal knowledge (44), database information (83, 84), information collated in floras (26, 27) and information from journals (12) and micro-fiche (75). Specimens in the herbarium may have several names e.g. their original 'label' name (70) or revised 'filing' name (69). Additional names for digitisation purposes may be located from the IPNI (22) and Tropicos (23) databases. Problems around naming may be resolved by a 'Names' group (54).

‘There are a number of digitisation projects focusing on the herbarium collections and other material relating to herbarium specimens.’

4. Overall observations

Diversity of cases

Some immediate observations stand out from these empirical findings.

- First, there is an enormous range of information exchange and use activities involved in life science research, with information being exchanged through multiple informal as well as formal channels with many actors outwith as well as within the research groups.
- Second, each case reveals an intricately structured, indeed Baroque, pattern of information processes.
- Third, as the maps vividly demonstrate, each case presents a distinctive pattern. Though there are similar kinds of activities across the different groups, they are configured together in remarkably different ways – especially when the fine structure is taken into account. These structures emerged from the specific tasks being undertaken – and the analysis we present below seeks to uncover some of the broad factors shaping them.
- Fourth, although information life-cycle models are helpful in understanding the overarching flows of information, the fine structure of research and information activities does not conform to a simple linear or cyclical model. Instead we found multiple internal information sub-cycles leading to intermediate products (tools, methodologies, half-processed data etc.) that were inputs for other information sub-cycles.

This section seeks to explore the differences and similarities between the case study groups by developing a comparative taxonomic analysis. We start by looking at the marked differences in role and information use within the research groups.

Division of labour, expertise and roles

As already noted, the activities of individual members of research groups are strongly influenced by their role and level of seniority – differences between the various roles are as significant as differences between groups. For example, doctoral researchers tend to have a narrower set of relationships primarily but not exclusively with local collaborators. At the other extreme, principal investigators (PIs) are often involved in developing and running a wide variety of research and related activities, and have more intense links with a diverse range of players including local colleagues, wider groups of peers nationally and internationally, funding bodies and so on.

A primary objective of this study is to characterise different kinds of life science research and their information processes, rather than the division of labour of research – which was broadly similar across most of the groups we studied. Though not our primary analytical goal, it is nevertheless helpful to characterise the main categories of researchers represented in the study.

Most (but not all) groups consist of a PI together with one or more post-doctoral researchers (post-docs) and variable numbers of PhD students. In several groups technicians also completed information lab books. Some larger groups also include lab and/or project managers. In the analysis below we characterise the generic patterns of information exchange and use, broadly categorised under PI, post-doc, PhD student, technician and lab manager.

PIs have the most complex information needs. Not only are they leading the research activity in the group but they have additional responsibilities which may include teaching/examining activities, supervision of PhD students, clinical duties, preparation of new research proposals and reporting to funders on completed projects, knowledge transfer activities and travel planning as well as research. These additional responsibilities are reflected in the range of information resources they create and use.

“A primary objective of this study is to characterise different kinds of life science research and their information processes, rather than the division of labour of research.”

Post-docs are the key members of the groups devoting their efforts mainly and directly to research. Their information activities thus most closely follow the research activity in the groups.

PhD students tend to be focused on a specific aspect of the research. Their information diaries therefore tend to be less complex than those of others.

Technicians' information needs tend to focus around the functioning of experimental equipment and protocols. They may depend on commercial companies as well as their own experience for considerable amounts of information on equipment and reagents.

Only a few of the larger groups include a *lab and/or project manager*, and their information needs vary. Information-related activities include both internally-facing activities (managing meetings, meeting reports, staff project reports, visitors' agendas) and externally-facing activities (such as maintaining the public face of the research group through a web site).

A number of the groups also include staff at post-doc or (more likely) senior post-doc level who have specialistexpertise in a specific area. Examples include mathematicians, specialist modellers, statisticians and curators. The distinction between post-docs and these specialist staff is sometimes indistinct, since post-docs may also undertake these functions as part of their portfolio of activities where they have developed the necessary expertise. But it should be noted that staff who perform specialist roles are by definition untypical, and their information behaviours and usage cannot readily be grouped together.

Towards a taxonomy: differences and similarities between cases

Some form of taxonomic ordering is needed to facilitate a comparative analysis of the diversity of our cases. The original research design proposed some avenues for constructing a typology. In particular it proposed a simple two-by-two matrix along two dimensions:

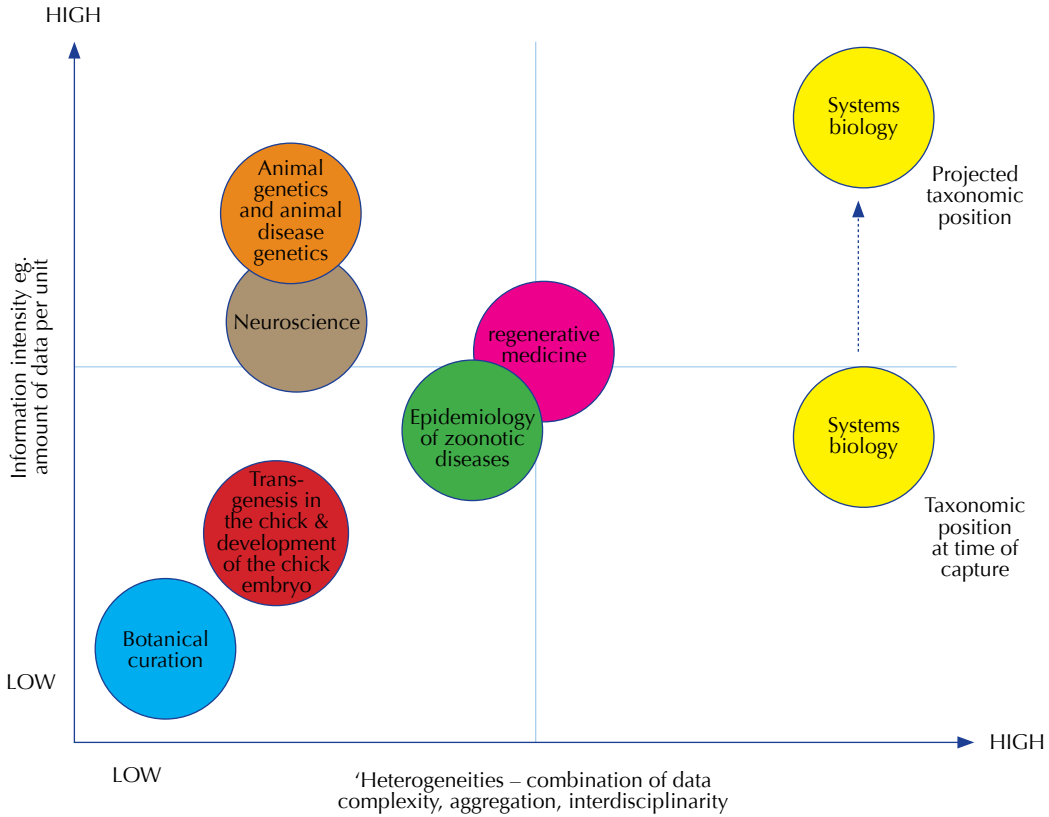
- the volume of data being handled
- the complexity or heterogeneity of that data

The international proteomic or genomic research programmes are characterised by sharing high volumes of largely standardised data. Systems biology, however, which attempts to pull together diverse data (genomic, gene expression, metabolic data and so on) is characterised by large-scale processing but of much more heterogeneous kinds of information. It therefore poses a challenge for researchers to integrate the different taxonomic structures that have emerged in these specialist domains.

This simple matrix is useful in our attempt to understand differences and similarities between different kinds of research endeavour. However, our empirical findings suggest that the heterogeneity of information comprises several elements that cannot readily be reduced to a single dimension. One source of heterogeneity is where judgement still needs to be exercised over diverse data. Another type of heterogeneity arises between different types of data. For example, many of our cases involve the accumulation of extensive collections of graphical images (e.g. of brain structures or of cell differentiation). These may be data intensive – but they involve rich data that can inform pattern recognition and analysis. Issues of the level of inter-disciplinarity are also at stake. On the other hand, homogeneity of data is often a complex achievement. Data from diverse sources may be treated as equivalent as a result, for example, of the standardisation of research practices or equipment, or the provision of high-quality standardised metadata.

For all these reasons, we cannot resolve to two simple dimensions the level of diversity of life science research based on how researchers create, manage and use data and other information resources. But we can make broad and consistent categorisations of our different cases. The 'impressionistic' taxonomy (below) provides an indication of approximate location of each case in terms of intensity and heterogeneities of research data exchange (see diagram 1).

Diagram 1: 'Impressionistic' taxonomy of case study research data



Note: this diagram conveys relative rather than absolute positions. Locations are approximate and might change. The arrow conveys how the systems biology group expected to migrate from its currently experimental focus towards re-use of existing data.

5. Information lifecycle

Information access

This section provides an overview of the information resources and utilities accessed and used by the groups we studied.

Researchers in these groups discover and gain access to the information they need predominantly through direct access to web-based resources, including bibliographic search and retrieval tools, on-line scientific publications, and dedicated websites that they trust. There was little evidence to suggest widespread use of the physical library except in case study 7 (botanical curation), where the curators make extensive use of their own specialist library.

Members of all the groups regard Google as the ultimate enabler. They like its ease of use, its word-search capability and its ostensibly large index: it not only indexes and caches web pages but also takes "snapshots" of other file types including PDFs, Word documents, Excel spreadsheets, Flash files, plain text files and videos. Thus searches often deliver serendipitous contextual information in addition to what is expected. Researchers are not unaware of Google's limitations but seem unconcerned that it yields a partial and potentially unmediated set of results. They also use other Google services (Google Scholar, Google Docs, Google Earth, GoogleMail) for a variety of purposes.

Researchers also see email alert services as useful in facilitating access to information; and they make heavy use of services such as PubMed which serve specific domains and are perceived as comprehensive and authoritative. However, researchers with an interdisciplinary or transdisciplinary orientation often see such resources as potentially restrictive:

'I'm originally trained as a biologist so I've always used PubMed, and it's always had everything I've needed...but with systems biology...I have started using Google Scholar more to get a broader [range of publications]'
(Systems biology, case study 5).

The researchers we studied are also positive about the concept of open access to scientific publications, and about the use of online forums and discussion lists which were seen as facilitating the exchange of expertise and information.

Nevertheless, researchers appear to have a limited awareness of the range of information services and resources available to them, and the number that they report using seems surprisingly small. They also show loyalty to particular resources or services that they like or trust, or that supply them with what they need. There are of course some fields of research which have specific requirements, and where there are few or no alternative options available.

A fundamental reason for this narrow, and often opportune, choice of information tools and resources is that researchers don't have time to review the whole information landscape. They also supplement their search strategies by seeking advice from colleagues (life scientists more than information professionals) as to the most appropriate and useful sources of information, treating them almost as intermediaries in identifying the information they need. Researchers see informal and local exchange of information (e.g. updates on experiments, results, methods, Standard Operating Procedures (SOPs), technical advice, journal articles) as conducive to increasing intellectual 'productivity'. This complements more formal mechanisms for information exchange, both internal (in-house meetings, lab group seminars, journal clubs) and external (conferences, reports to funders, scientific publication). There are also learning costs associated with identifying particular resources and developing the skills necessary to interrogate and exploit them. This may explain the tendency to favour generic services with easy-to-use interfaces.

"(Researchers) supplement their search strategies by seeking advice from colleagues...treating them almost as intermediaries in indentifying the information they need."

Some researchers see the training that is available to them from institutional information services as not specific enough for the kinds of refined resources or utilities they are using. Moreover, the distinctions that may be highly significant to information professionals – as to the differing status of information resources, for example – may not be paramount for researchers. Thus it is immaterial to most researchers whether they need to use an information portal, commercial website, publisher's web service or bibliographic database. Their orientation is primarily pragmatic: they simply require immediate and uninhibited online access to the resource of interest.

Researchers from all groups are concerned about access to information, and any potential barriers or inhibitors. Although most are happy with the range of online journals available to them, some have encountered access problems;

and they express frustration when, having used publicly available search mechanisms, they identify materials to which they do not have full text access online. This may be a particular problem in specialist areas where high subscription charges are a potential barrier. Some researchers use alternative methods in order to gain access to full text material otherwise unavailable, such as contacting colleagues from institutions that subscribe to the full text of the required publication.

Our evidence suggests that RSS feeds, podcasts, social bookmarks, blogs and other social networking tools are little used. Two primary reasons are that there is not the critical mass of individuals using such services to make it worthwhile for the purposes of enhancing research, and that the time required in order to become a proficient user is prohibitive. Moreover, many researchers regard the need to access multiple services and resources on different machines or systems as a major inhibiter to uptake and use. Those using grid technologies highlight the importance of being able to access an intricate array of analytical tools and utilities from within their desktop environment.

Information sharing and collaboration

In selecting our case studies, we deliberately excluded groups whose main function was information sharing and collaboration *per se*, such as specialised bioinformatics centres. We wanted to address the practices of the wider range of life scientists. Thus our focus below is on information sharing and collaboration among life science researchers for whom it is part, but not the major part, of their activities.

They are in principle in favour of sharing many kinds of information, and are remarkably willing to do so in order to facilitate each other's research, without any apparent formal reward. Thus information is extensively shared within research groups and laboratories, and informally across organisational boundaries through wider research networks, both before and after formal publication. This may include sharing, on request and depending on circumstances, standard operating procedures (SOPs), plasmids, computer programmes, scripts and statistical analysis tools that they have written. But researchers are much more ready to share methods and tools than experimental data, and members of our research groups identified a number of barriers to sharing of data.

Researchers experience competing pressures which influence their attitudes towards information sharing, depending on the context and the kinds of information involved. In identifying the particular exigencies surrounding information sharing, we address first of all factors favouring or inhibiting the sharing and re-use of experimental and field data. Then we explore the various other kinds of information and how the propensity to share varies between them.

Sharing data

Data can be difficult and expensive to collect. Collecting them can take years, especially when considerable time and effort may be needed to build relationships of trust among a range of stakeholders – such as commercial collaborators, clinicians or patients – to obtain the data in the first place. In addition, those working in highly innovative fields need to develop new techniques.

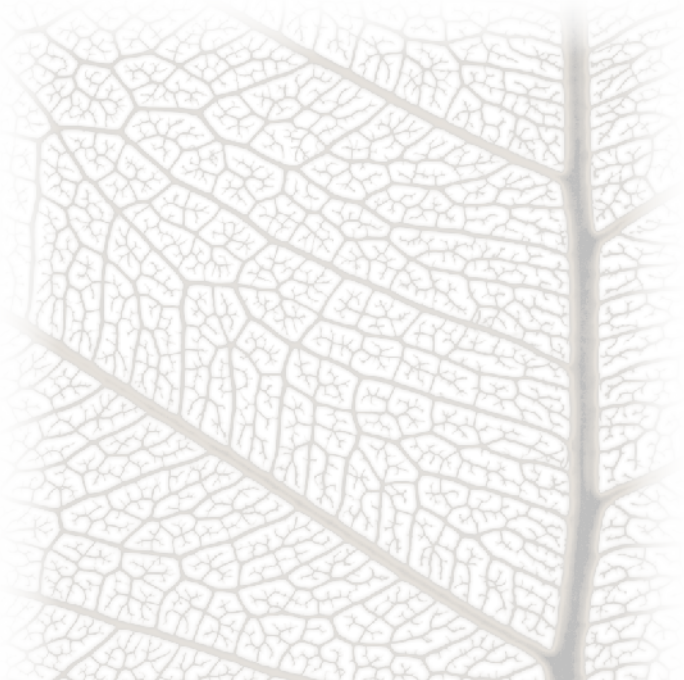
Researchers see themselves as securing career rewards for the research they do and the findings they achieve rather than the data they collect. They are reluctant to share the data that make up their 'intellectual capital'. In particular, they are wary of giving away their data for someone else to analyse and get the credit. There is a career imperative at

the heart of this resistance to open sharing of data. Many researchers thus seek to retain control over their knowledge and information, since this is important for both formal and informal recognition and reward processes. This imperative is deeply rooted in professional and institutional structures and the incentives experienced by both individuals and groups. And it is notable that some cross-disciplinary activities in data-sharing are not well recognised or rewarded by these professional and institutional structures. For example, those working at the boundaries of informatics and life science may find that their services to the other discipline is not well-rewarded by either discipline (Calvert and Williams, 2008).

Researchers are thus willing to share experimental data subject to two strong provisos.

- First they are concerned that they need sufficient time to complete the analysis and, in some cases, to explore intellectual property rights (though when asked, they are often reluctant to define exactly how long this period of time should be).
- Second they want to publish their results before or simultaneously to publishing the data – and they want to be the ones who publish the data.

Achieving a balance between open access to information and data and the need to protect intellectual capital for future use are therefore issues that continue to concern many researchers. Some kinds of data which do not constitute a source of 'added value' for the researchers concerned, such as geographic or gene marker information, may be readily shared. This is because researchers distinguish between two aspects of the data they collect: recorded 'facts' such as geographic positions, where sharing is unproblematic; and the attributes which they themselves provide and which add value to the data. Many researchers do not want to share the attributes at least until after a paper has been published.



The researchers we studied also express a keen sense of 'ownership' and protectiveness towards their data. They feel responsible for the data they have generated and express concern that someone might re-analyse them inappropriately. They want to know who is going to use any data and for what purpose, rather than make their data freely available. For when data are shared there is a perceived loss of control about how the data are subsequently used.

Some researchers therefore suggest that data for re-use should be made available only on application, and that those who collect data should play a key role in determining whether or not the data should be released for the specific re-use requested. Some also support the suggestion that portals should be established to highlight the existence of datasets that could in principle be shared, rather than making them routinely accessible in centralised data centres; and we heard one suggestion that this should be the responsibility of relevant professional associations.

It is also notable that most of our researchers do not wish to re-use research data collected by other researchers because there are so many differences in experimental design and data collection practice.

'I don't trust people so I don't know if they have done it to the same standards that I would have done it'
(Regenerative medicine, case study 6).

Researchers are also conscious of the lack of standardisation, for example of image formats; and of the intricacies of experimental design and data that are often not easy to understand. Moreover, experience of, for example, microarrays, has also shown that in a developing field it takes time for standardisation to be implemented and to work effectively. Thus while researchers support the sharing of genetic marker information in web-accessible repositories, they note that that there is no system for the standardised naming of these markers, which may have two to three different names. For all these reasons, researchers prefer collaborative arrangements and direct contact with potential users where differences and intricacies can be elucidated and understood, rather than making data freely available for re-use. This in turn raises questions about how much time should be spent on annotating data.

Thus while willingness to share is part of the ethos of life science research, individuals like to choose what to share, with whom, and when. Lack of trust in wider 'cyberspace' is pervasive. Some researchers raise concerns about posting data on the web; about the risks of making them available in 'the cloud'; and about possible misuse. Some kinds of information and data are therefore shared on a highly restricted basis, with privileged access being the rule; others may be more freely exchanged with peers. In biomedical areas, for example, there are particular sensitivities and issues of confidentiality surrounding the sharing of specific types of data such as brain images, as well as data and information derived from animal experiments. Similarly, where commercial organisations are collaborating in the research or where there is potential for patenting – an imperative that may come from university authorities as well as commercial partners – then protection of both data and information from premature disclosure becomes important. This has obvious implications for any concept of 'open science'.

'While science is all about sharing, there's a lot of things that we're not allowed to say, or information that can't really be shared and a lot of people are the same, they don't want to give you information from different groups'
(Regenerative medicine, case study 6).

Given the specific contexts in which they find themselves and the particular interests and incentives they encounter, researchers who do not make their data freely available are typically behaving quite rationally. Personal relationships are important and have a strong influence on whom a scientist might be willing to share with, as well as on the manner of sharing: through collaborations, joint funding bids, and so on.

There are exceptions, however, to these kinds of general considerations on sharing information. For example, the researchers in the systems biology group (case study 5), whilst sharing some of the concerns expressed by other groups, have fewer reservations about freely sharing their

data. For such sharing is part of rationale behind systems biology. However, because they are working in a highly innovative field, developing new techniques for sharing is very much a matter of timing. It is not a question of whether to share, but when and how. Initially it will be done through publication.

Similarly, the ethos of the botanical group (case study 7) is all about sharing information - 'open access and no payment' - with other herbaria around the world, and with other taxonomists, scientists, amateur botanists and the public. The group is participating in an international project with several hundred other herbaria, sending 'type' data to a foundation in New York; and they give information on request and loan specimens to other taxonomists. Finally, one group stated that in times of crisis everyone shares data freely. For example, on the outbreak of bird flu', all flu' data went to the World Health Organisation (WHO). Researchers did not withhold until publication release of data that could help with virus strain verification. WHO is trusted to ensure that credit is given if researchers' data contributes to its work.

Sharing other kinds of information

Researchers are generally more relaxed about sharing many other kinds of information, even when they are recognised as crucial to the successful conduct of experiments. They are thus generally happy if they are asked to share software, code, scripts and tools. Their preference is for this to be done through direct contact with colleagues through peer networks, as well as by formal publication. Codes and algorithms are seen as having no intrinsic value; their value lies in the data to which they can be applied. But codes and algorithms can be treated in two ways: 1) made open access and free to the user or 2) commercialised to make a profit. Statistical tools, which are not patentable, are widely shared.

Many researchers distinguish between sharing information and data internally and externally, which may depend on the perceived value to them of the information to be shared. This is not reducible to a distinction between formal and informal sharing; indeed, much of the sharing with external collaborators is quite informal. Experimental processes and methodologies are thus part of an ongoing, almost continuous internal discussion, where processes are carefully logged and recorded but rarely shared beyond the group. Sometimes researchers are keen to protect individual or team 'know-how'; and where a novel technique has been developed, details tend to be seen as privileged until after a paper has been published.

In general, however, protocols and tools are widely and readily shared, often before publication. The value that possessing particular kinds of information and experience presents to an individual or group is linked to their ability to trade possession to advantage in terms of reputation, funding and career development; and it may be asymmetrically distributed according to their institutional position – and especially to their professional or disciplinary location. Hence many researchers who need a new tool look for what others are using that might do the job and seek to customise it rather than design one from scratch. Such a pragmatic approach often gets quicker results and is cheaper in cost and time.

Mechanisms for sharing

Shared folders are commonly used as a simple means of sharing resources within a group. Lack of common file structures and naming conventions can, however, lead to problems in finding documents. This has prompted some to investigate using other tools, such as wikis. Some groups have also still to establish satisfactory online mechanisms to share large files.

Some researchers are aware of the potential of Web 2.0 and social networking tools but they do not them intensively. The reasons given include lack of time to invest in the learning curve of using the tools, the sheer number of tools and services, and the lack of a critical mass of people using them:

'It's a matter of getting everyone being prepared to use it ...Unless you had everyone fully committed to that sort of thing then it would be quite difficult to rely on it fully'
(Animal genetics, case study 1).

The tools are not necessarily 'fit for purpose' and their benefits are not immediately obvious: there is an apparent lack of models of successful use.

'What amazes me is how little lumpiness there has to be in the use of something for everyone not to want to use it'
(Regenerative medicine, case study 6).

There are also organisational or institutional restrictions on the use of 'cloud' computing and Web 2.0 tools and facilities, which are considered security risks to both systems and data.

Knowledge Transfer Cycle

Our case studies focused on the activities undertaken by our research groups over a period of five days. Hence we were unable to track all their communication and knowledge transfer activities from start to finish. Nevertheless, it is clear that all the groups undertake activities related to all the stages of the Knowledge Transfer Cycle outlined by Charles Humphrey (Figure 2).

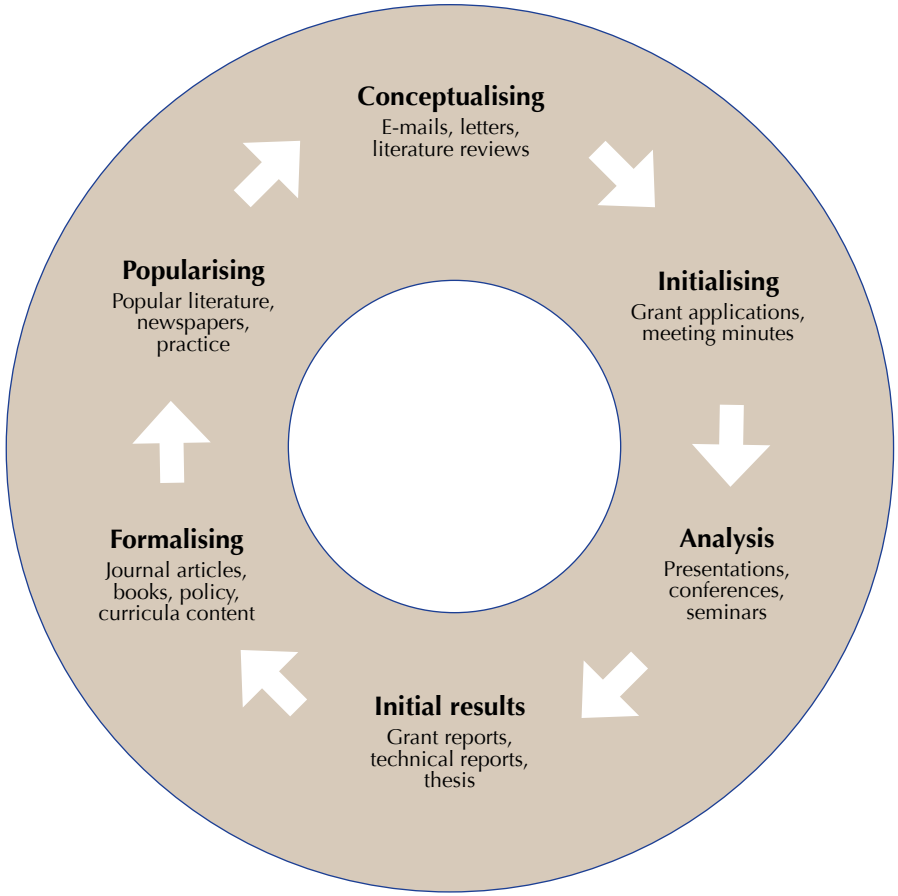


Figure 2:
Knowledge Transfer Cycle

Conceptualising new research:
this activity is informed by published literature, research presentations from colleagues and international collaborators as well as exchange of information among academic and industrial collaborators using a variety of mechanisms including email, Skype and shared Google documents.

Developing and initialising research proposals:
this may involve reading material on both domain-specific and funders web sites; undertaking 'environmental scans' of relevant research by reviewing project web sites as well as communicating with collaborators; studying policy manuals and government publications; and accessing information from both university and manufacturers' sources needed for costing, including the costing of equipment and reagents. Internal communication may be oral or written, formal and informal. It may also involve using web sites to disseminate information on matters such as forthcoming funding opportunities, current awards, and lectures and seminars.

'Pre-analysis' phase:
though not addressed in Humphrey's knowledge transfer cycle, our cases reveal a wide range of information exchange associated with running a project and a laboratory: referring to and writing SOPs; ordering consumables (requiring information from catalogues and product information sheets, as well as on how to order them); health and safety information (often relying on in-house web sites as well as specialised information on specific chemicals); equipment and reagent information from industry (from web sites, manuals, company literature, meetings with sales reps and industry stands at conferences); demonstrations by colleagues or in training sessions.

Analysis and presentation of research:
activities here range from in-house seminars to large international conferences. We noted some use of Second Life and teleconferencing as 'virtual' means to present and discuss research. PowerPoint is the primary medium for presentations, with other software used to provide summary graphs and images to aid presentation. PowerPoint slides are readily shared between colleagues to allow early circulation of results.

Formalising outputs:
this involves preparing project reports, journal articles and other publications.

Project reports are prepared with collaborators, using both face-to-face and electronic meetings. They are sent both to funders and, if appropriate, to industrial collaborators.

Journal articles draw on both the research results and previously-published literature. Drafts are prepared with or informed by comments from colleagues.

Other publications include book chapters requiring scientific knowledge of the subject, knowledge of the publisher's requirements, and in some cases input from other international authors; depositing code into open source repositories; producing ontologies and indexes; posting gene sequence data on public databases.

Popularisation of science activities:
in one of the groups, a scientist was using his own knowledge in combination with data from web sites to develop a podcast for secondary school pupils.

Researchers' perceptions of future challenges and needs

In the focus groups we directly elicited views about future challenges and the barriers and drivers for change in addressing them. Although there were some commonalities across the groups in the support they think they need to meet future challenges, there were also some areas where the help required was quite specific.

There is strong agreement about the need for **bioinformatics support**. In some cases, the groups recognise that this may have to be provided on a centralised basis, and they recognise the need to develop the skills required to interface with bioinformaticians and a range of other experts in areas such as statistics and modelling. But their preference is for local and easily-accessible support from bioinformaticians who know their projects and the work they do; are on-site to help with problems; and are located nearby so that relationships can be built, thus avoiding delays in waiting for responses from a central service.

Standardisation represents another area of common needs. Data can be highly variable, and different forms and formats can cause problems. The groups find ways to deal with these through screening, checking, or searches in order to make the data useful and useable. They often have to use multiple software and analysis programs in order to accommodate the variability of datasets. There are also instances where the highly specialist nature of the research means that specialist software and data might be useful only to a very small number of researchers, who probably already know, or know of, each other.

Members of several groups share concerns about receiving, generating, handling and managing **huge volumes of information and data**, in particular genome sequence data, as well as accessing, organising and analysing vast quantities of varied data, and then making sense of them. The sheer volume of data and information that is now being produced, and expected to be produced in the future, is a cause for concern. Researchers fear that there will be too much data to handle, process or even look at.

Some groups are also concerned about **technical barriers** such as lack of processing power, and congestion on (and occasional crashing of), the grid; and **cost barriers** associated with using commercial services and tools which may not be available through or supported by their university but which may be more efficient or effective for particular purposes than the established university services.

We also identified two related **concerns about funding**. First, researchers are concerned that support for the development of data curation should not be at the expense of funding for research. Thus, if there are competing demands between research and information needs, research should have the priority claim on resources. Second, they are concerned that the short term nature of funding may frustrate attempts to build and sustain data repositories. The research councils and the Wellcome Trust among other funders now allow research grant proposals to include resources for data curation. However, once a project is completed, the money stops and the people move, leaving the risk that the data are unsupported.

Our groups were able to set out a wish-list of developments they thought would help them in their work. These include:

- User-friendly tools, and fast data transfer. New technologies and new software programmes that are compatible with old ones, or preferably one software

programme to fit all, so (i) you don't have to spend time running different pieces of software and (ii) you can share any kind of information or data within the group. Likewise, specialist software for complicated and specialist work rather than several different programs.

- Better quality proteomic data and better annotated experimental data. Better funding of research on organisms that are not in the mainstream to bring them up to the same levels as, for example, human or mouse genetics.
- Translators with specialist knowledge, to translate papers published in other languages; of particular relevance is the increasing collaboration and contact with Chinese researchers and interest in their work.
- Easy-to-use tools to help find relevant information and resources. Help to deal with information and data overload (sorting and filtering). For example, in a fast-developing interdisciplinary field such as tissue engineering researchers want to get at the area/topic/sub-field they are interested in without having to plough through huge amounts of information or to learn the whole subject area.

Other possible developments were specific to the needs of specific groups, including:

- better searches and links in image databases
- help with determining from the plethora of software available for doing sequence analysis that which is fit for purpose (expensive licensed or freeware)
- tools to help with finding, linking and extracting information from plant indexes and moving data over the Internet for automated/semi automated data entry into the database and geo-referencing
- a software tool to compare ontologies e.g. 30,000 genes against 30,000 proteins to show if they match up or are removed from each other
- funding to create and maintain an international sophisticated brain imaging database, and
- being able to make the information and data held much more useful and accessible for research analysis by securing funding to digitise the whole herbarium.

Digital data management constitutes a major future challenge for the botanical curation case study. The group both manages the inherited generations of approximately 3 million non-digital specimens and records new information and data coming in at the rate of about 10,000 specimens each year. In addition to worries about the retrospective digitisation of hardcopy specimen records, researchers were also concerned about archiving and back up of data, storage, restoration, disaster recovery, moving large amounts of data (for archiving or backup) and remote access.

> ‘Researchers are concerned that support for the development of data curation should not be at the expense of funding for research.’

6. Implications for institutional information services

Evidence from our seven case studies has a direct bearing on the provision of institutional information services (IIS) for these specific subjects and perhaps others. Our findings contribute further to the debate about university library strategies for serving the needs of researchers as users and creators of a diverse and expanding range of digital information and data. Some of these issues were explored in depth in the RIN report *Researchers' use of academic libraries and their services* (April 2007).

We explored earlier in this report the considerable variances in practice for information use and exchange amongst the groups we studied. This represents a major challenge for IIS providers intent on supporting a broad range of research. We also stress how the bottom up view provided by life science researchers does not accord with the top down view taken by IIS providers. Hence, when considering the implications for IIS, we have focused on those particular aspects of the research process where, according to the views of our research groups, there is some common thread with respect to the enablers and inhibiters around which IIS may usefully reorient their strategies.

Identification of information sources and resources

In the main, when seeking to identify appropriate information resources, our researchers rely on advice from colleagues and other trusted sources. Their interest is largely pragmatic. They are relatively unconcerned about differences between the selectivity or completeness of the sources of information available to them. Information professionals, on the other hand, are more aware of differences between various resources. The challenge for them is to communicate these issues, and work with researchers in addressing them.

Researchers *are* concerned, however, about accessibility, especially about barriers to accessing published information online. One repeated source of frustration, not exclusive to the life sciences, is the experience of successfully identifying resources via a search engine or bibliographic portal, but finding that access to the full text is denied. This is usually because the institution does not subscribe to the journal or other resource in question. Researchers from our animal

genetics group explained the resulting predicament: while online abstracts may provide sufficient information to arouse interest, they give insufficient evidence to warrant purchase of the whole paper.

With IIS budgets under pressure this kind of experience is likely to become more common, unless libraries can increase their purchasing power by participating in regional or national consortia, or provide access through selective domain-based cooperatives. In current circumstances it is increasingly important, but also increasingly difficult, for libraries to sustain a dynamically-relevant subscriptions portfolio that supports the ever-evolving research programmes pursued in the university as a whole. Hence it is crucial that a lively and continuous dialogue is maintained between researchers and the IIS team. Otherwise, the evidence from our groups is that failure to include in the subscriptions portfolio the specialist journals they need can constitute a real barrier to effective research.

Use of the university library

Conventional university library facilities rank low as a vehicle for accessing published information. The traditional role of professional information intermediaries has been largely replaced by direct access to online resources, with heavy reliance upon Google to identify them. Given the limitations of generic search engines such as Google, measures to reconnect researchers with IIS professionals could bring improvements in information retrieval, and benefits to the research process.

Researchers also tend to use services that have been 'proven' by colleagues, or to interrogate websites they regard as authoritative and comprehensive in their field. When they use such services, researchers tend to take the results on trust: the specificity and the breadth of the information retrieved do not appear to require further enquiry.

The result of all these developments is that many life science researchers have removed themselves from the mainstream library user population. They do not even use the library catalogue. Library-based services can replace the services researchers do use only by demonstrating that they can improve retrieval capability, and deliver results within a timeframe that corresponds to researchers' own patterns of work. This is a significant challenge when researchers are driven by a desire for immediate online access to specific resources of interest, at a time convenient to them, and from a known and trusted source.

This journey away from use of the institutional library is given greater weight by a less predictable factor: despite the potential risks and limitations of Google as a search tool, one unexpected advantage can be its delivery of surprising sources or parcels of information. In this sense Google searches may have replaced the experience of serendipitous browsing the library shelves.

Accessing and using new resources or tools

Overall our groups use a narrow range of search engines and bibliographic resources, probably for three reasons:

- lack of awareness and time to achieve or build a broader suite
- the 'comfort' that comes from relying on a small set of familiar resources, usually endorsed by peers and colleagues, and
- the cost in time and effort needed to identify other resources, and to learn to use them effectively.

For our animal genetics group, for example, the time required to learn how to use a new tool is a major barrier: they are already 'too busy' with the conduct of research and when faced with a seemingly difficult information problem will 'just Google it'. The low take-up of new Web 2.0 tools arises for similar reasons. Unless new tools are easy to use and bring obvious, realisable benefits, they will remain unused.

Effective online search and retrieval requires knowledge and understanding of how each resource and service is organised; the search methods that will yield the best results; the level of access that will be given (abstracts, synopses or full text); and time to prepare and structure a search. There is little standardisation of structure between individual online databases; and without the understanding generally gained through repeated and frequent use, search and retrieval may at best be challenging and at worst prohibitive. Researchers consequently prefer generic services with easy interfaces.

The challenge for institutional information services is thus to develop and provide online services geared to the needs of their research groups and thereby to add value to the research process, facilitating the use of new tools, providing individuated professional support, as well as advice, training and documentation on a subject or discipline basis. Any such strategy would have to be proactive: as noted by our regenerative medicine group, researchers are reluctant to adopt new tools and services unless they know a colleague who can recommend or share knowledge about them.

Research data

Most major research funders in the UK with direct influence on our study cohort – for example, the BBSRC, MRC and Wellcome Trust – have policies that require grant holders to submit data management plans for formal approval, to manage their data in accordance with those plans, and to make the data available for re-use by third parties. The Digital Curation Centre (DCC) provides on its website an analysis of the policies of all the major UK funders.

One of the fundamental challenges in implementing such policies derives from confusion over terminology. Data may be described expertly as 'evidence supporting research and scholarship', (DCC, Charter and Statement of Principles). But it is generally assumed that data have no intrinsic meaning until converted to information through some process of analysis, interpretation and description: typically, the process by which experimental data becomes a research paper.

The distinction between data and information is conveniently summarised by Kock et al (1997): data are carriers of knowledge and information; a means through which knowledge and information can be stored and transferred. Both information and knowledge are communicated through data, and by means of data storage and transfer devices and systems. In this sense, a piece of data becomes information or knowledge only when it is interpreted by its receiver. In the same sense, information

and knowledge held by a person can be communicated to another person only after they are encoded as data.

We found some fundamental differences in understanding of the terms 'data' and 'information' in the groups we studied, with these terms and others such as 'digital object', 'database', and 'dataset' often used interchangeably. Such lack of definition can lead to serious consequences when researchers and information experts use the same terms with different meanings. One consequence is the frequent conflation of information exchange with the sharing of research data. The narrow focus in discussions on formal data exchange, and its attendant benefits and difficulties, diverts attention from the many different kinds of information being shared in differing ways by life scientists.

Data curation and sharing

Data curation is, of course, only one element in the research lifecycle, and did not feature prominently in our case studies. We found little evidence that planned data management has yet been adopted as standard practice. Where there has been direct contact with data management professionals we did find significant if isolated change, as illustrated by our neuroscience group, which has been the subject of an immersive study by the DCC's SCARP project. One of the outcomes of this relationship was described in discussion with the group as 'now we manage our data, whereas before we didn't'. The SCARP investigation demonstrated that effective curation needs human infrastructure, where researchers' heedful attention to each other's data underpins curation within the team. The dividend from this approach has clearly been recognised by the group, which has elected its own Data Management Group to oversee the storage of large amounts of image data being generated, processed and analysed. Members of the systems biology group, however, expressed the desire for 'a curation person' to organise and manage data in-house.

There are challenges here for IIS departments. They may need to consider from an institutional perspective the risks that may result from lack of compliance with funder policies; and in addressing that issue they may learn from the 'think local' approach adopted by our neuroscience group. But meeting the diverse curation requirements of a wide range of research groups would pose a formidable challenge for IIS, especially if such requirements were to translate into a demand for support from a central cohort of data management experts who were also expected to display a substantial level of subject knowledge. One possibility might be better (i.e. easy-to-use) tool-based support for practitioners to undertake their own data curation.

Nonetheless, universities are increasingly requiring that researchers deposit the final peer-reviewed and accepted versions of research outputs in institutional repositories. Researchers in the groups we studied want to know who is going to use the information or data they may be required to deposit and for what purpose. And while they may see depositing *publications* as a means to improve dissemination of research results, they are concerned that requirements to deposit *data* in institutional repositories will require safeguards significantly stronger than those provided for publications.

Thus if IIS departments are to support funders in achieving a step-change in data sharing practice, it will be essential to demonstrate not only that mechanisms for embargoing or providing restricted or closed access to data collections can be reliably achieved, but that researchers' strong desire to be consulted about the re-use of their data will be met. Active engagement between data producers and institutional data custodians is crucial for the agreement of workable processes, checks and balances.

7. Policy challenges and recommendations

This study has highlighted a number of gulfs between the practices and exigencies of life science researchers and current prevalent visions and policies that are projecting a radical shift towards large scale sharing and re-use of primary data. The case studies presented here point to the diversity of life science research and its associated information flows. They call into question conceptions of the convergence of life science around particular models of how information is collected and used. All the groups we studied engage in vigorous and extensive information sharing and re-use. This sharing takes very different forms, however, depending on the nature of the research activity under way. It includes publication and the exchange of methodological knowledge, as well as the sharing of data. But while sharing of derived data such as images, or of canonical and reference data may be relatively extensive, experimental data sharing takes place primarily within research groups and networks rather than through large-scale centralised repositories.

These diverse information sharing practices are driven primarily by the needs and benefits perceived by life scientists rather than the policies of external players such as research funders, universities or information service providers. Researchers are motivated more by reciprocal and altruistic exchanges within peer communities, so long as these do not cut across other incentives such as the need to exploit IP or publication opportunities. Thus there is a high level of informal exchange of important information and experience, and this contributes to improving experimental and analytical methods and overcoming often intractable problems in getting these to work. These forms of exchange are often overlooked by policies promoting formal procedures.

Our key conclusion is that the policies and strategies of research councils and information service providers must be informed by an understanding of the exigencies and practices of research communities if they are to be effective in optimising the use and exchange of information, and in ensuring that this is scientifically productive and cost-effective. Change is under way. But a single approach to the future of life sciences or a one-size-fits-all information policy will not be productive or effective. Moreover, a corollary is that information service policies and provision need to be brought closer to research groups and communities.

‘A single approach to the future of life sciences or a one-size-fits-all information policy will not be productive or effective.’

Research councils and other research funders

Sharing and exchange

Many of the current policies to promote and support the exchange of data and information tend to offer a generic recipe or template based on specific visions of how sharing and re-use are to be achieved and the benefits they will deliver. Blanket requirements are then written into the funders’ terms and conditions of grant. Doubtless recipients do what is necessary to be seen to fulfil these requirements. But notwithstanding an ethos in favour of sharing data and information, practical and human issues which serve to restrict exchange are likely to persist in the life sciences and also in other domains. These range from matters of recognition and reward to issues of confidentiality and the contextual entanglement of the data that is being produced.

Given the limited current understanding of which forms of sharing and exchange are most effective and beneficial, and under what circumstances, we suggest that **further engagement with researchers themselves is essential to identify and address the constraints surrounding information exchange, as well as to preserve the exercise of informed choice that is fundamental to science**. In current circumstances, we suggest, narrowly prescriptive approaches are unlikely to be effective. Instead, we believe, it would be more helpful for funders to **adopt a more pragmatic and experimental policy that recognises the multiplicity of contexts, and different approaches to information sharing, and which builds upon the informal sharing that is already taking place, based on the recognition of mutual needs**. Such a bottom-up view is needed in order:

- to attend to the practicalities of data sharing: what makes information from other sources intelligible? Under what circumstances is such sharing useful and sufficiently beneficial to warrant the labour necessary to achieve it?, and
- to address existing barriers and drivers for change, including the perceived self-interests and goals of researchers and their need to sustain their intellectual capital and advance their careers.

Funders whose policies prescribe the sharing of research outputs may need therefore to define more closely, in consultation with researchers from different communities, which data and information they expect to be shared, to what ends and under what circumstances.

Support for researchers

Funders also need to take account of the implications of researchers' strong call for locally-available information support, closely integrated within research groups and laboratories. We see a proliferation of hybrid roles here (and the term bioinformatician has been used to describe a range of research-oriented and support roles). These roles, and their attendant divisions of labour and expertise, are still evolving. Funders as well as institutional managers and professional bodies (see below) need to monitor these developments and their implications for training and career development systems.

In developing their strategies for the supply of trained people in the life sciences – as, for instance, through the BBSRC's Bioscience Skills and Careers Strategy Panel – it is important that funders should work with learned societies and professional associations to support emerging specialist roles, as well as strengthening such broader skills as part of life science research training. We recommend that **an assessment should be commissioned of the national requirements for skills in research data curation and support, with a view to producing appropriate, effective and sustainable models for training and careers in managing data throughout the lifecycle, and catering for the potentially diverging requirements of different domains of research**.

Higher education and research institution managers

Institutional career development and reward systems – for appraisal, promotion and so on – set the context for academic life, allocating incentives and establishing drivers and barriers to change. A key issue arising from our studies is how effective such systems are in recognising and rewarding the kinds of cross-disciplinary integration that are now emerging in the life sciences. The specialists involved experience acute problems, characteristic of cross-disciplinary collaboration, in securing professional recognition whether by their discipline of origin or where they find themselves now located.

Large, well-funded life-science projects now commonly encompass support from a range of specialists, including bioinformaticians, statisticians, modellers and so on. Current project-based modes of funding often make it difficult, however, to sustain such roles, and the information tools and resources they generate. Smaller groups find it especially difficult to secure and sustain this kind of infrastructure and support, though they may get informal access to such specialists working on other better-resourced initiatives.

There is thus a risk that current policies and systems may inhibit improvements both in the effectiveness of research and in information sharing. **Decision-makers in higher education and research institutions therefore need to work in conjunction with learned and professional bodies, to attend to the entrenched features of current professional formation processes – including training and career development, and professional recognition and reward structures – which currently inhibit the effective use and exchange of information**. They should consider in particular how to ensure

- that information sharing activities and competences are recognised, rewarded and supported, and
- that there are sustainable careers for those providing information services and computing skills within the life sciences.

Professional associations

Professional associations such as the Society of Biology, promote professional practice and standards through their Chartership and Continuing Professional Development (CDP) programmes. It is important that development and achievement for those in specialist roles – bioinformaticians, statisticians, modellers, biocurators and so on – should be included in such programmes. In some cases, support for recognition of newer fields such as biocuration may be provided through new professional fora such as the International Society for Biocuration, where practitioners can meet and discuss ideas with their peers.

> ‘Large, well-funded life-science projects now commonly encompass support from a range of specialists, including bioinformaticians, statisticians, modellers and so on.’

Library and information service providers

Information services for researchers are now provided on a range of models which may be more or less dispersed or centralised, generic or specialised. Library and information service providers in the higher education sector need to come to a clearer view of their structures and roles. Distributed solutions offer different trade-offs between costs and benefits than more centralised provision (easier start up; catering more readily to particular local needs, but risking potential compatibility problems and greater long-term alignment costs).

Given the current dispersal of life science resources, some of our groups expressed a desire for better portals and tools to identify the information resources relevant to researchers working in their domain. Some of the specialised repositories that are emerging (e.g. in neurophysiology) may help to develop such services. But the proliferation of facilities at the ‘meso-level’ calls for new kinds of support for standardisation and harmonisation, and for disclosure. The structures and mechanisms for such ‘meta-harmonisation’ efforts are not, however, well-established.

University information services

Re-establishing a lively and sustained dialogue with their research communities is a key challenge for the library and information services in many universities. Such dialogue is essential if libraries are to provide the publications, other information resources and services that their researchers need. Better engagement between information professionals and researchers could add to the efficiency and effectiveness of research, with specialist support facilitating the use of new tools, and providing individuated professional advice, training and documentation on a subject or discipline basis. Such a strategy would have to be proactive, for researchers are reluctant to adopt new tools and services unless they know a colleague who can recommend or share knowledge about them. And it would have to meet the challenge of delivering results that correspond to researchers' patterns and timetables of work.

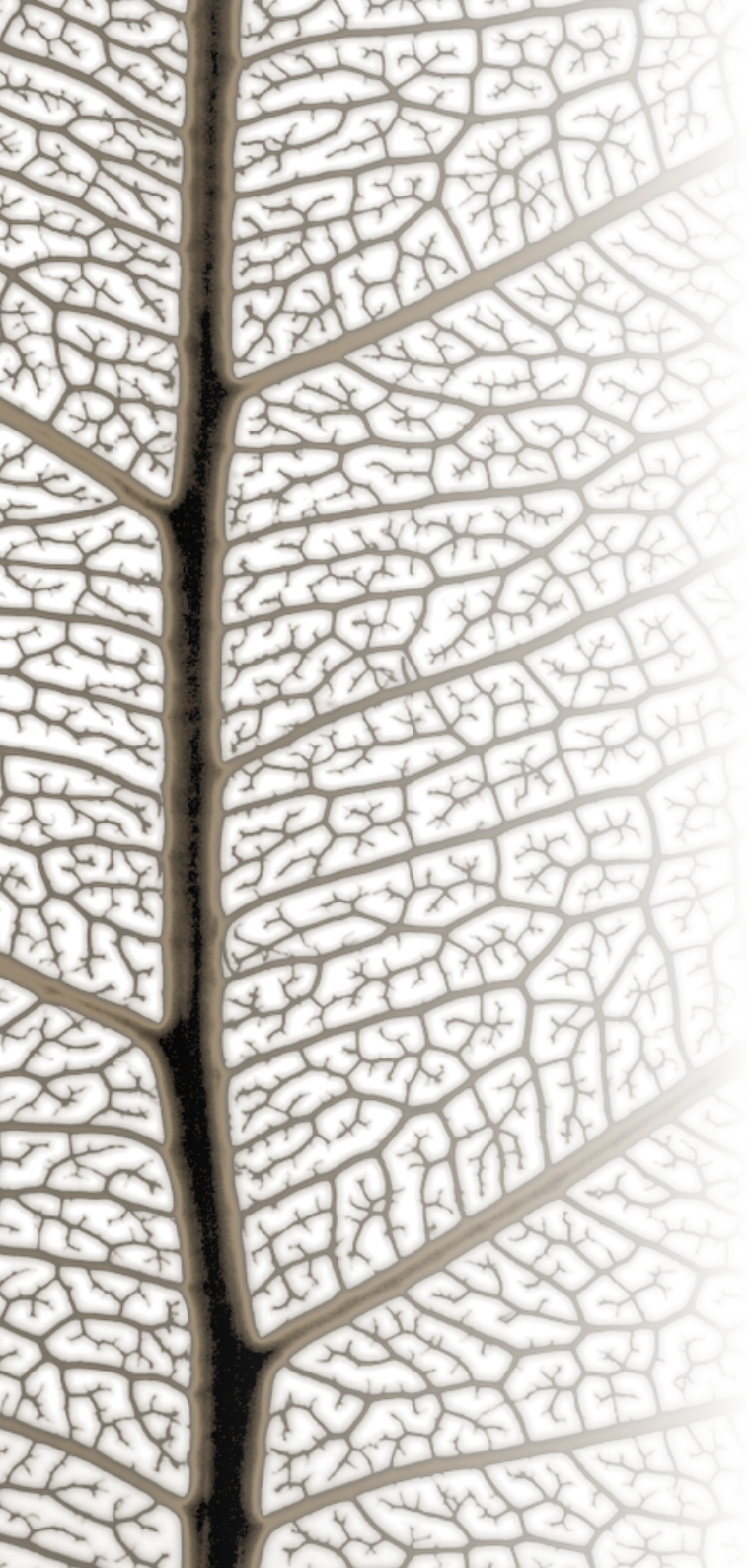
Research data pose particular challenges. Library and information services cannot be expected to provide support for all the research groups in the university from a central cohort of data management experts, particularly when a substantial level of subject knowledge is required. One possibility might be better (i.e. easy-to-use) tool-based support for researchers to undertake their own data curation.

If library and information services are to support research funders in achieving a step-change in data and information sharing practice, active engagement between data producers and curators will be essential for the agreement of workable processes, checks and balances. For it will be critical to demonstrate not only that mechanisms for embargoing or providing restricted or closed access to data collections can be reliably achieved, but that researchers' strong desire to be consulted about the re-use of their data will be met (e.g. DCC's Data Audit Framework, DRAMBORA and Data Management Plan).

Other bodies

As the context and practice of research continues to change, there is a need to improve understanding of information sharing practices and how they are and are not changing. Our studies have demonstrated the value of medium-scale qualitative studies, but have also highlighted areas where more detailed longitudinal and ethnographic research could contribute to the development of improvements in both policy and practice. They further point to a need for work which compares contrasting types of research within and across disciplinary domains. More studies of this kind should be commissioned by the RIN and other bodies to help ensure that policies and strategies are properly founded on the practices and exigencies of researchers themselves.

“More studies of this kind should be commissioned…to help ensure that policies and strategies are properly founded on the practices and exigencies of researchers themselves.”



References

Biotechnology and Biological Sciences Research Council. *Bioscience Skills and Careers Strategy Panel* http://www.bbsrc.ac.uk/organisation/structures/panels/skills_careers/index.html

Calvert and Williams (2008). *Report of BBSRC-funded workshop: Data sharing in the biosciences: a sociological perspective*. www.genomicsnetwork.ac.uk/innogen/publications/workshopreports/title,21207,en.html

Digital Curation Centre. *Charter and Statement of Principles* www.dcc.ac.uk/charter

Digital Curation Centre (2008). *The curation of neuroimaging research data for sharing and re-use* www.dcc.ac.uk/docs/publications/case-studies/SCARP_B4821_NeuroCase_v1_1.pdf

Digital Curation Centre. *Data Audit Framework (DAF)* www.dcc.ac.uk/tools/daf

Digital Curation Centre. *Data curation policies* www.dcc.ac.uk/resource/curation-policies

Digital Curation Centre. *Data Management Plan* www.dcc.ac.uk/docs/templates/DMP_checklist.pdf

Digital Curation Centre. *Digital repository audit method based on risk assessment (DRAMBORA)* www.dcc.ac.uk/tools/drambora/

Humphrey, C (2006). *E-science and the life cycle of research* http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

International Society for Biocuration http://www.biocurator.org/

Kock, N. F., R. J. McQueen, and J.L. Corner (1997). *The nature of data, information and knowledge exchanges in business processes: implications for process improvement and organizational learning* in The Learning Organization, Vol. 4(2) pp: 70-80

Research Information Network (2007) *Researchers' use of academic libraries and their services* www.rin.ac.uk/our-work/using-and-accessing-information-resources/researchers-use-academic-libraries-and-their-serv

Society of Biology http://societyofbiology.org

About the Research Information Network

Who we are

The Research Information Network has been established by the higher education funding councils, the research councils, and the national libraries in the UK. We investigate how efficient and effective the information services provided for the UK research community are, how they are changing, and how they might be improved for the future. We help to ensure that researchers in the UK benefit from world-leading information services, so that they can sustain their position as among the most successful and productive researchers in the world.

What we work on

We provide policy, guidance and support, focusing on the current environment in information research and looking at future trends. Our work focuses on five key themes: **search and discovery, access and use of information services, scholarly communications, digital content and e-research, collaborative collection management and storage.**

How we communicate

As an independent voice, we can create debates that lead to real change. We use our reports and other publications, events and workshops, blogs, networks and the media to communicate our ideas. All our **publications** are available on our website at **www.rin.ac.uk**

This report and the supporting Annex are available at **www.rin.ac.uk/case-studies** or further hard copies can be ordered via **contact@rin.ac.uk**

Get in touch with us

The Research Information Network
96 Euston Road
London
NW1 2DB
UK

Telephone **+44 (0)20 7412 7946**
Fax **+44 (0)20 7412 7339**
Email **contact@rin.ac.uk**