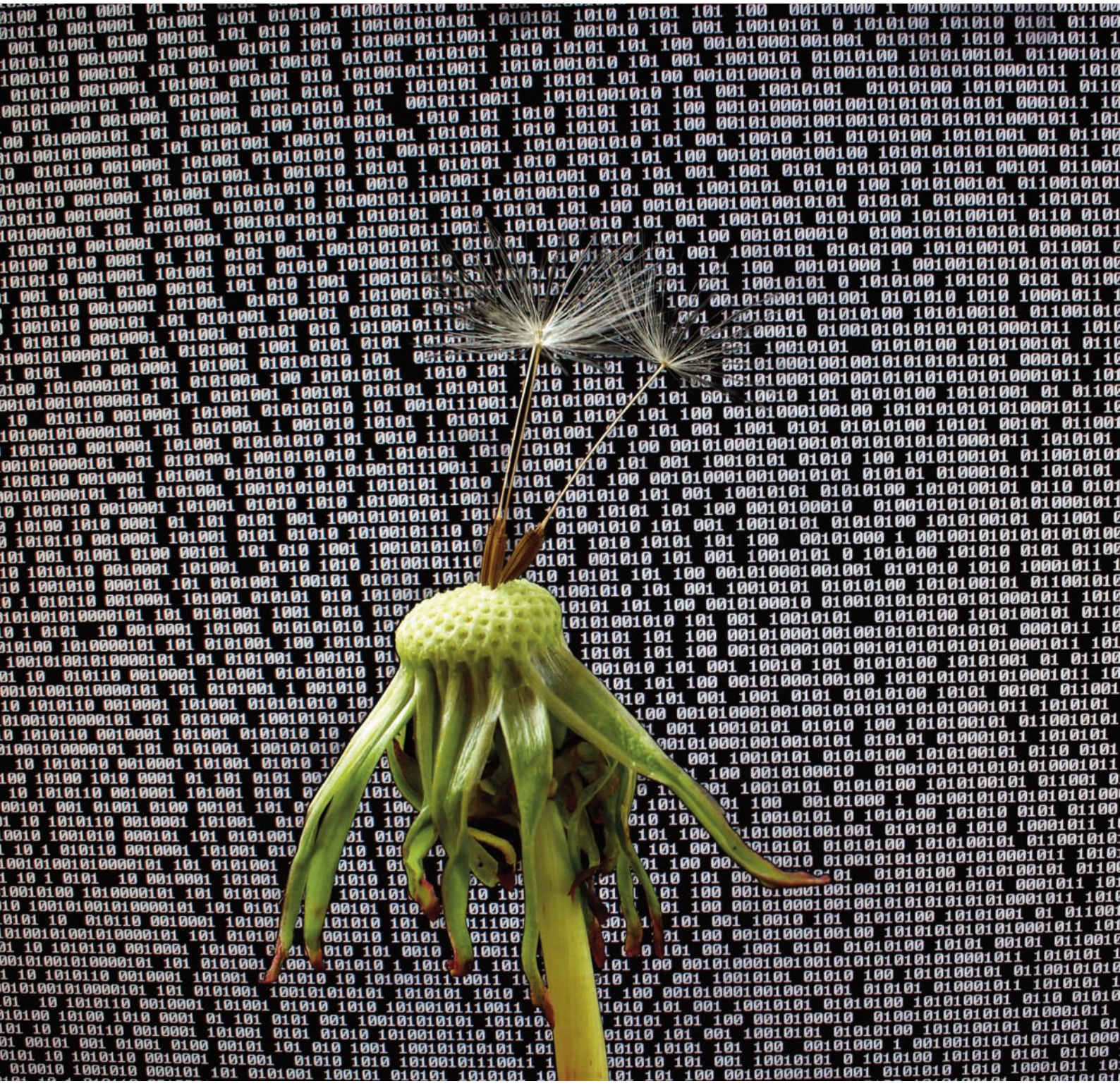
Making Data Accessible to All



GARNet and EGENIS Workshop Report

Making Data Accessible to All

Report from GARNet/Egenis Workshop

12-13 July 2012, University of Exeter, UK

Sabina Leonelli and Ruth Bastow

Contents

Introduction: Data Sharing in Plant Science

Scientists have long developed practices for the publication of selected datasets to accompany specific claims, via the publication of papers in journals. Publishing data in this way, however, means that only a small fraction of data produced is publicly accessible. The recent Open Data movement is geared at changing this situation, by encouraging scientists to publish the entirety of data that they produce, regardless of whether or not these data are associated with a publication.

Plant scientists are increasingly encouraged, and often required, to donate data to open access databases (for instance by the BBSRC data management policy; http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf). They are also encouraged to make use of these databases in order to boost their research and speed up discovery. The latest report on Open Science by the Royal Society specifically points to the urgent need for 'intelligent' data access (Royal Society 2012), which involves investing resources, time and effort into making data publicly available, findable, interpretable, re-usable and citable. There are several drivers for this requirement: increasing the transparency and reproducibility of research; speeding up research by facilitating cross-consultation and comparison among existing datasets; making the best of available resources by reducing duplications in the research process; introducing new methods for discovery, based on the partly automated mining of large datasets; and improving teaching and collaborative research in both the developed and the developing world, by making data produced through expensive and/or rare instruments and materials widely available for query and analysis.

However, despite the clear demand for data sharing and the strength of the motivations for it, its implementation is still limited. This is, at least in part, because data sharing raises several unanswered questions and challenges to current research practices. It is still unclear whether it is feasible and useful to store and disseminate all data; how decisions should be made about which datasets are most useful for dissemination, or which types of data should have priority when setting up databases and curatorial standards; who should maintain and financially support structures to host data; how responsibilities and related duties to data curation, such as the authorship of data and the efforts spent in posting them online, need to be allocated and rewarded within the scientific system; how such responsibilities need to be policed or enforced, and by whom (universities, institutions, publishers, funding bodies and national governments); and, more generally, how to go from efficient dissemination to intelligent re-use.

These issues are ever more pressing within the field of plant science, which is undergoing an era of great change driven by genomic scale technologies, theory-based approaches and informatics. New

technologies such as next generation sequencing are overcoming many barriers in research in economically important plants and allowing in depth studies of model species (as in the case of the Arabidopsis 1001 genome project; http://www.1001genomes.org/). This in turn is generating a plethora of new data, databases, and resources, such as the Wheat Initiative (http://www.wheatinitiative.org/) and the new UK phenotyping platform (http://www.phenomics.org.uk/). Although these advances are providing substantial opportunities for futhering scientific understanding, they are also generating significant challenges. For example The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org) was established in 2000 as a portal centred on a single genome (of accession *Columbia 0*) and associated data and was therefore not set up to deal with the current data deluge. This has led the community to plan the establishment of a new Arabidopsis Information Portal (AIP) (IAIC 2010, 2012) which will build upon the expertise of TAIR but provide many additional layers of functionality. TAIR is just one of many examples where the underlying data infrastructure is not adequate to deal with the needs of researchers, a problem which is being tackled in the US by the iPlant initiative (http://www.iplantcollaborative.org/) and in the EU by ELIXIR (http://www.elixir-europe.org/). However managing the data mountain is not the only issue that researchers face. There are also significant challenges to be overcome in the areas of data integration, encouraging researchers to use adequate standards and undertake data appropriate curation and management.

To discuss the issues surrounding data donation, publication and use in the plant sciences, GARNet (UK Arabidopsis Research Network) and EGENIS (ESRC Centre for Genomics and Society) led a workshop on 12-13th July 2012 which brought together researchers, data curators, publishers, and funders to assess what we can all do to facilitate sustainable and intelligent data dissemination. Discussions at the workshop were geared around three key questions:

1. Where do researchers who produce high throughput data store their results, and how do they make them publicly accessible (if at all)?
2. What types of data are most fruitful as a public resource?
3. What are the roles and responsibilities of funders, researchers, institutions, governments and publishers in facilitating intelligent data sharing, and how are they likely to evolve in the near future?

In this report, we summarise the presentations and debates undertaken at the workshop and discuss where responsibilities should lie in the pipeline from data generation to data dissemination.

1. Storing and Disseminating Data: Challenges

The eScience programme in the UK (http://www.nesc.ac.uk/), which was funded by the UK government between 2001 and 2011, has been instrumental in promoting two key ideas about data sharing. The first is 'going the last mile,' the need to go all the way to where the user wants to use the data, which means engaging meaningfully with user communities and understanding what uses specific data types will be put to when circulated widely. The second idea, an 'intellectual ramp,' would build accessible, usable databases so that users would not have to spend a considerable amount of time learning how to use specific databases. Both of these ideas constitute important features which scientists expect of the databases they use; however at present it is difficult to establish these requirements as database developers have struggled to get a clear idea of which uses data might be put to, and users are often required to acquire extra skills in order to be able to access, retrieve and re-use data online (Howe and Rhee, 2008; Leonelli, 2010).

To try and assess how we might begin to improve the situation, workshop delegates discussed what motivations scientists have to share data, including the possible uses that data might be put to when widely and freely accessible. It was agreed that the sharing of data and knowledge is central to science as data provides a global intellectual capital, especially since some data, for example time-series data collected over a week from multiple samples, are expensive or difficult to replicate in the absence of highly specialised experimental setups/instruments, which many labs do not have access to, especially in developing countries. Furthermore, the publication of data constitutes yet another opportunity to publicise the research of the scientists who produce them. It was also agreed that large datasets are essential to modelling efforts in systems biology and data-intensive science; and that accessing and comparing data across multiple projects facilitates the identification of subtle patterns and variations among species, environments, and methods employed in data gathering, as well as the identification of experimental errors.

Data sharing is also attractive to science funders: it makes their investment in data generation more worthwhile, as data can be underutilised if they are not widely shared; and increases the longevity of data, as data are likely to be more securely stored in a public repository than on an individual's hard drive. Indeed, data sharing is now required by most public funding agencies. In the UK, this effort has received a substantial boost as a result of reports published by the Royal Society (2012) and the UK Government (http://www.researchinfonet.org/publish/finch/). The data sharing requirements of major UK funders are summarised in this table by the Digital Curation Centre: http://www.dcc.ac.uk/webfm_send/873.

A survey undertaken by PARSE.Insight in 2009 assessed the digital preservation of research output in Europe. The survey showed that 91% of researchers viewed the ability to re-analyse existing data as the most important driver for preservation of research data, and that 96% of publishers thought that preservation of research data was key to stimulating the advancement of science. Yet despite this overwhelming agreement that research data should be stored and shared, only 25% of researchers surveyed actually make their data openly available (http://www.parse-insight.eu/project.php).

So what are the major barriers to data sharing? Workshop delegates noted that the rapid development and establishment of 'omics' and 'sequencing' technologies has moved biology (particularly molecular and cellular) from an experimental to data-intensive science, a leap from the 1st to 4thparadigm according to Jim Gray's view of the evolution of science (Hey et al., 2009). Many biological disciplines have therefore not had the time, space, nor adequate expertise and infrastructure, to consider how to deal with the 'data deluge' and build adequate and effective solutions.

Two important obstacles to data sharing are the availability and long-term support of data repositories, as well as the scepticism that many researchers still harbour about the quality of these repositories and of the data retrieved through them. In the PARSE.Insight survey 80% of respondents regarded the lack of sustainable hardware, software, or support of computer environment that may make the information accessible as the most important threat to digital preservation. Despite the development of ways to track specific datasets digitally, in many cases it is still unclear how, and on which criteria, the quality of data and data repositories are to be checked and evaluated (if at all).

To try and deal with the data mountain many researchers, groups and institutions have built their own solutions. An example is BioDare (Biological Data Repository) at the University of Edinburgh (http://www.biodare.ed.ac.uk), which was established to store, share and analyse rhythmic time series data within a collaborative project between four UK Universities. Due to its usefulness, BioDare was soon extended to include data from other projects. So, like the data themselves, the solutions to data sharing challenges can also quickly move up the 'data-sharing pyramid' (see figure 1 below). Local solutions can, with the right support and care bestowed upon them, become universal ones.

Box 4.1 The Data Pyramid – a hierarchy of rising value and permanence

Details of examples given in appendix 3.

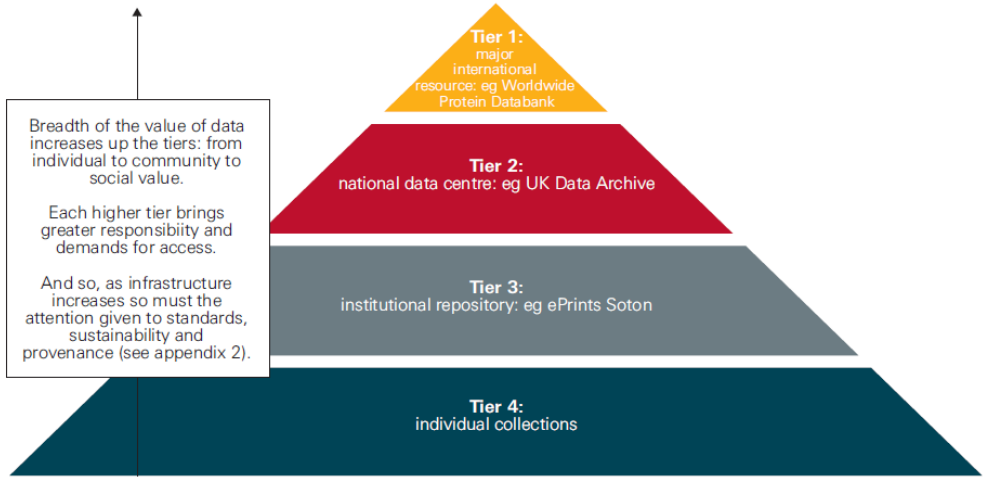


Figure 1: Data sharing pyramid from Royal Society report (2012).

The example of BioDare is not unique; other large-scale systems biology projects such as PRESTA (http://www2.warwick.ac.uk/fac/sci/lifesci/research/presta/group/) have also built in house data storage and sharing solutions. Although the proliferation of local databases is viewed by some as a hindrance to the development of national/internationally recognised databases, many researchers observe that the development of modes of data circulation within individual projects is central to encouraging data sharing. Local data curation, annotation and sharing with colleagues can often be the first big step to widening accessibility. It is important to note that guaranteeing that data are available, usable, compatible and durable at a local level is very labour-intensive and it is difficult to maintain the diversity of computational and biological expertise needed to effectively maintain these 'in house' generated solutions, especially after a grant/project has finished.

Local data sharing solutions can also provide useful insights into how data sharing can be promoted. For example the experience of a number of the UK plant systems biology centres and projects indicates that researchers are willing to actively submit data if 'in return' they are provided with useful tools and resources that enrich the user experience to make data analysis and data interpretation easier. At the same time, the experience of those utilising OMERO, the open microscopy environment that was built out of an initiative from the University of Dundee

(http://www.openmicroscopy.org/site), suggests that the process for the user needs to be as simple as possible, customizable and rewarding to use.

International initiatives to create mechanisms of data sharing are also crucial for the establishment of specialised data repositories. The international scientific research community has developed standards for the annotation and curation of data submitted to online databases (e.g. the MIBBI initiative); web services to help sharing data quickly and efficiently; and computer readable formats in which data of various types can become widely accessible.

Another very important obstacle discussed at the workshop was the culture of data sharing. Participants expressed the concerns that even if data sharing was made as easy and simple as possible, it is likely that not all researchers will be inclined to share their data. For some there are concerns over the release of primary data sets in case someone spots an error. Others might not want to release data until they think they have got all the publications they can from it. Publications are so critical to career progression in academia that many researchers end up hoarding data in readiness for 'what if' scenarios. Also, hierarchies and ethos within certain laboratories affect how individuals within the lab operate (for instance, if the PI does not wish to share data, young scientists will not be trained to do so and will probably not learn what the right procedures are). Further, some research that is partly funded by industry might raise intellectual property concerns. These sociological barriers are very hard to overcome and are only likely to change with 'top down' pressure from funders and journals and 'bottom up' pressure from researchers themselves to generate a community-wide ethos for data sharing and dissemination.

2. Types of data: What counts as which data, and for whom?

Not all data are created equal. Some data are easier to share, reuse, deposit and integrate than others. This is usually due to the variability of data themselves, the materials (specimens, tissue cultures, etc) and experimental set-ups used to collect the data, and the availability of standards and instruments through which data can be formatted and annotated. For transcriptomics data, there is a global standard for annotation (MIAME http://www.mged.org/Workgroups/MIAME/miame_1.1.html), numerous internationally recognised databases for submission and storage, including GEO (http://www.ncbi.nlm.nih.gov/geo/) and Array Express (http://www.ebi.ac.uk/arrayexpress/), and a variety of tools for integrating and analysing data, including NASC Arrays (http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl) and

Gene Investigator (https://www.genevestigator.com/gv/). Other data types are more problematic however, for example metabolomics data is harder to standardise (though there have been several initiatives including the metabolomics society and ArMet), which makes it difficult to assess data quality and reuse data. Some data types, such cell biology images and phenotypic data, have no generic or standard repository into which they can deposited, no associated standards, and no well defined ontologies. Additionally, some data are more valuable than others, e.g. large data sets collected under standard conditions with good quality controls such as the mass of Arabidopsis transcriptomic data generated by NASC, which are viewed by both 'wet' and 'dry' researchers as key data sets for data-intensive approaches.

Despite the variety of ways to deposit and store data as outlined in Table 1, it is possible to get agreement amongst researchers on which data types are more 'core' or 'salient' than others, e.g. the well curated Arabidopsis genome. There is less agreement on which data should be stored and what to do about data obtained through obsolete technologies. Many would argue that raw sequence data will soon be so cheap to produce that there will be no need to store the data: one can just repeat the experiment as and when it is needed. The same cannot be said for data or samples that are costly to produce or just very rare, for example herbaria and long term field studies. Further, it is impossible to know what questions researchers will want to ask in the future. For example, people are still using Darwin's samples to find answers to new questions in evolution and development.

Participants at the workshop could not agree on an answer to the difficult question of which data are worth storing. Some thought that the biological community does not want or need every scrap of data ever produced, and that trying to store and disseminate without discrimination is counter-productive and wasteful. Rather, researchers should disseminate enough details about their data generation methodology as to enable others to get the same data if they wish to generate equivalent datasets (thus following the classic idea of data experimental reproducibility). Also, data sharing does not necessarily promote transparency and quality of research: statistical analysis is often enough to show whether data have been cherry-picked or misread, thus making it unnecessary to submit whole data bulks. Others disagreed, both because reproducing data is itself often a wasteful effort, which few can afford, and because emerging fields and techniques such as systems biology and data-intensive methods need very large datasets to obtain results.

Data type	Main repository	Plant repository	Data standards, minimum information guidelines	General journal requirement	Other repositories
Genes and gene nomenclature	Genbank	TAIR	Genbank, TAIR	Submission to Genbank, or TAIR for plants	
Genome sequence	Genbank	TAIR	MIGS (Field et al., 2008)	Submission to Genbank	EMBL
DNA barcodes	EMBL		Barcode of Life standards		BOLD
RNA sequences	EMBL		BCB RNA-seq		RefSeq
Chip sequencing	GEO		Minseq (draft), GEO	Submission to GEO	Sequence read archive, IRC
Transcriptomics	GEO	NASC, TAIR	MIAME	Submission to GEO	Array Express
Protein structure	PDB	Plant PDB		Submission to PDB	
Proteomics	GEO	Plant PDB	MIAPE (Taylor et al., 2007)	Submission to GEO, SwissProt	PRIDE, Protein, pep2pr, EMBL
Metabolomics	BMRB	PMN	MSI	Submission to BMRB	Metabolome Express
Epigenomics	NCBI Epigenomics		MINSEQE(draft), NCBI		Chromatin.csl
Interactions	IntAct	TAIR	IntAct		Reactome (no Arabidopsis data yet)
Mathematical models	Biomodels.net	PLASMO	MIRIAM	PLoS one	Biomodels
Pathway information	BioCyc	TAIR: AraCyc			Reactome (no Arabidopsis data yet)
Synthetic biology	Parts registry		Parts registry		SBOL

Table 1 – List of online resources for sharing and depositing data

If researchers do wish to keep all data that is being generated, is it feasible to do so? At present the size of datasets being produced is continually increasing, such that the percentage of 'older' data in comparison to 'new' data is small. It is therefore not considered an onerous task to store old data if we have to cope with the new data deluge. However this is likely only to be true for sequence-based data. For image and phenotype data the size of datasets can be so large that decisions will have to be made on what can be stored. In general it was agreed by workshop participants that decisions should be left to the researchers who generated the data, as they are best placed to know what should be kept and what should be destroyed. One way to encourage scientists to give thought to data sharing is to ask for data management planning from the outset of any specific project, as recommended by current BBSRC guidelines on good practice (http://www.bbsrc.ac.uk/web/FILES/Policies/good_scientific_practice.pdf). All delegates agreed that as a general principle, if data were to be kept it would only be useful to others if they were interpretable, reusable and citable.

3. Scientific publication, a changing landscape

In the last 50 years scientific publishing has undergone a substantial change as it tries to keep pace with the explosive growth of data, and with new technologies for sharing information in the digital age. In 1953 Crick and Watson published their landmark paper on the structure of DNA, which had just two authors, one figure, no data, and was only a page long. In 2001 the human genome was published in a seminal paper that included 150 authors, 49 figures, 27 tables, and was 62 pages long. In 2010, the 1001 Genomes Project was published as an open access paper available online, and involved 76 institutions, 12 145 SRA run IDs, and covered 12 pages. The ENCODE consortium published 30 papers this September from their efforts to describe all the functional elements in the human genome. Each paper from ENCODE can be read as a unique unit however the papers also exist as an online package (http://www.nature.com/encode/) in which users can select a particular interest or 'thread', for example it is possible to look at just DNA methylation data. This is just one illustration of the changing landscape of scientific publishing, and the rate of change will only increase in coming decades.

However, the amount and types of data are not the only issues publishers have to deal with. Recent shifts by the government and funding bodies to open access policies, such as the new policy enforcing open data for all research funded by UK research councils (http://www.rcuk.ac.uk/documents/documents/RCUK%20_Policy_on_Access_to_Research_Outputs

.pdf), are increasing pressure on publishers to ensure that the underlying data on which a scientific paper is based can be accessed by all so that the scientific process can be scrutinised, results reproduced and research built upon. Yet, as noted above, only 25% of researchers currently make their data available, partly because of the scarcity of adequate repositories, but also as a result of fear that their data will be misused, that errors will be found, or that they will lose their scientific lead.

So how can publishers help to encourage and perhaps even enforce data sharing? Many publishing houses facilitate the discoverability of datasets by connecting the research article to the underlying dataset. It is standard for many journals to either link to the data entity via a DOI or re-direct readers to a repository such as the Protein Databank (http://www.wwpdb.org/ , Gen Bank (http://www.ncbi.nlm.nih.gov/genbank/), PubChem (http://pubchem.ncbi.nlm.nih.gov/) and GEO. In addition, some journals have integrated datasets within the paper via APIs or webservices, for example data from the PANGAEA database (http://www.pangaea.de/) and genome data provided by TAIR. Connecting data in this way helps to increase data discoverability, keeps the data in context with the research paper it is associated with, and improves online readability.

However this approach for integrating and linking to data is only feasible for data that are stored in well-established repositories. It is not possible for datasets that do not have an established home, or for small sets of data such as those as regularly used to produce a figure or table within in a paper, to be accessed in such a way. These data are therefore not available to users and cannot be accessed or reused as there is no clear/agreed mechanism or workflow as to how researchers publish or access this data. Yet answers will need to be found if publishers and researchers are to adhere to open access policies such as the one currently endorsed by RCUK. One possible solution is the use of generic data stores such as Dryad (http://www.datadryad.org/) and Figshare (http://figshare.com/).

Many publishers are struggling with where and how supplementary information fits into the new online, open access models of publishing. Supplementary information (SI) was created to deal with data that could not fit within a paper in journals that are limited to print space such as Nature and Science, and to allow publication of large datasets or data in media that is not accessible in print e.g. videos. This seems to have been successful, as SI submissions have increased exponentially in recent years. However, SI is hard to access and reuse, as it is often only available in pdf format and not xml or html format, which makes it hard to index and search. In addition, SI may not undergo the same rigorous review process as data in the main body of the paper, resulting in lack of quality control and associated metadata. This is partly dependent on each journal's policies, and partly on the difficulty to find referees to assess data quality (which constitutes yet another demand on referee's time).

Last but not least, several journal editors manifested worries about referees abusing the peer review system by asking for too much SI. The Journal of Neuroscience, for instance, stopped taking SI altogether, because reviewers were asking for too many additional experiments. This was regarded by some workshop participants as the wrong move, since it is the prerogative of a journal editor to step into these disagreements and draw the line between helpful referee demands and outrageous requests which do not add to the paper. At the same time, it was argued that if SI are really essential to understanding and evaluating claims in a paper, they should be included in the paper itself rather than being singled out as SI from the start.

To try and deal with these problems, from April 2012 Nature Neuroscience has begun a trial to encourage authors to submit one seamless paper, incorporating all of the essential information. The editor, in conversation with the referee and authors, will then decide what should be included in the paper and what should be SI. It is hoped this will make authors, reviewers and editors think more about data in publications in general, and more specifically to carefully consider which data is integral to the paper and needed to support the claims made. It will also ensure that all data undergoes the same rigorous review process, an important issue especially since journals are still regarded, by researchers as well as funding bodies, as responsible for assessing and policing the quality of research, including the quality of data produced. Another solution is provided by 'extended' or 'open' articles, which seamlessly link to several layers of information.

Finally, citation systems need to change to take account of data publication as an important component of research output. Examples of this shift are the new citation index released by Thomson Reuters (http://www.reuters.com/article/2012/06/22/idUS109861+22-Jun-2012+HUG20120622) and the new policy from STM (http://www.dcc.ac.uk/blog/joint-statement-data-citation-stm-publishers-and-datacite). Even more notable are efforts to develop 'generic data repositories' through which curated data can be stored, shared, and importantly, cited. Two such repositories that attracted a lot of attention and debate at the workshop were Dryad and Figshare.

Dryad (http://www.datadryad.org/) is an international digital repository, which aims to help researchers preserve all the underlying data reported in a paper at the time of publication. Dryad provides researchers with a repository in which to place all published data, not just those data sets that can be deposited in a specialised repository. To ensure researchers are credited for re use of the data deposited in Dryad, all data files are assigned a DOI and Dryad also promotes adoption of its best-practice data citation policy and traceability of data citations. Dryad is developed by the National Evolutionary Synthesis Center and the University of North Carolina Metadata Research

Center. It is supported by a number of journals and learned societies, particularly those specialising in ecology, taxonomy and evolution. Dryad provides a useful 'safety net' for those data sets that do not have a home elsewhere and prevents them being lost when personnel leave a lab, or a hard drive fails. However there are concerns amongst researchers that it does not contain enough metadata to properly allow in depth re-use or re-analysis of data.

Figshare (http://figshare.com/) is part of Digital Science (http://www.digital-science.com/), and promotes data sharing, data citation and data discoverability by providing a platform for researchers to submit any data, published or unpublished. Since its inception, Figshare has focussed on the needs of researchers, with a very few barriers to uploading data and instant rewards for submitting data, such as the ability to share it on social media sites and view metrics that provide information on how others are using the data. Figshare can accommodate static images, media, data spread sheets, data sets in all digitalised formats, posters, dissertations, grant applications, and pre-prints of articles of any size. Every object placed in Figshare is a citable entity, allowing users to get credit for unpublished datasets, figures, videos and posters. Allowing submission of data and resources outside a publication results in data, which are easier to find than when embedded in a paper, as they are directly accessible to search engines. Figshare is likely to be a transformative platform for data sharing, and as such Figshare has recently teamed up with F1000 to enable users to preview, download, cite and share data in accompanying datasets at the click of a button.

In addition to generic repositories there has also be a recent emergence of data journals such as Ecological Archives, ZooKeys, F1000Research, GigaScience, Database, and Earth System Science Data. In general no conclusions are drawn from data published in theses journals so the focus is on improving data interpretability and reuse and encouraging the concept that data is a research output in its own right. These journals also enable data producers and curators who may not qualify for authorship on a traditional journal paper to be credited for their work. Data journals are unlikely to replace specialised repositories, but may be useful for data types without specific repositories, and could also assist in providing a clear summary of methods and metadata to enable re-use of data.

4. Responsibilities in data publication

The final session of the workshop was devoted to a discussion about the work and responsibilities involved in data sharing, and how they should be divided up among relevant stakeholders.

Researchers

It was agreed that researchers have several key responsibilities in data sharing. First of all, they are responsible for generating data in the first place, promoting a data sharing culture among colleagues and students, and developing and encouraging the widespread use of community standards for data sharing. It is important for researchers to understand that their data might be valuable to many others; therefore they should devote time and resources to organising and publishing data in a reasonable length of time. Researchers are also in the best position to decide which format the data should be disseminated in, which data types should be prioritised when sharing, and how data needs to be organised in order to facilitate re-use by colleagues (e.g. through adequate choice of metadata). Furthermore, researchers are best placed to know which data should be released to the public, which should be stored in specialised databases, and which databases are likely to provide the best service.

The above tasks involve considerable labour, resources, and expertise in addition to the already extensive demands placed on researchers. This needs to be recognised by all the institutions and funding bodies involved in supporting and promoting science, which also need to commit to providing guidance on best practice.

Universities

Universities constitute the first port of call for providing substantial support to researchers wishing to share data. A commitment to data sharing involves more than hiring a limited number of personnel dedicated to dealing with data, as some universities have attempted. It involves providing training in data management, for instance by university libraries or specialised departments responsible for providing research support; shifting the system of credit attribution so that data sharing is seen as credible research output and can be used in promotion applications; adequate IT provision, including servers and technical assistance with data formatting; support for the move towards open access, for instance by providing financial support since open access publishing is expensive and is not always covered by research grants; support for the development of new databases, which as outlined above are often born out of specific research projects; adequate time allocation for research projects to take account of data sharing practices, for instance by inserting data sharing as a component of individual workload assessments; and providing clear guidelines concerning what is expected of staff when it comes to data publication.

Some research funders, such as the EPSRC, put emphasis on institutions as essential contributors to data storage, and encourage universities to develop their own repositories for data generated by their researchers. This is problematic in several respects. Often data produced by researchers is best stored in international databases that specialise in that data type. In general universities are not well-placed to develop in-house expertise in data storage, and are likely to produce a number of different storage systems that will only increase problems with the interoperability of databases available online. Further, most research is international in nature, involving different researchers from a variety of institutions so researchers, particularly at the postdoctoral level, tend to move frequently across institutions. It is not clear how specific universities could claim ownership of data in these situations, which brings many workshop participants to question the wisdom of university repositories altogether.

Public funders

Funding bodies play a crucial role in providing incentives, structures and resources for researchers to engage in data sharing. They should provide stable funding for long-term storage and curation of data, if possible through a funding stream that is separate from hypothesis-driven research and explicitly targets the development and support of permanent databases. Particular attention should be given to funding the development of tools and resources to help integrate and visualise data; to providing clear guidance on which funds are devoted to open access and data sharing; and to encouraging researchers who submit grants to consider their data sharing plans as early as possible when planning their future work. Also, there should be clear financial support for open access publications, and funding bodies need to act as a link between the research community and the government concerning the importance and future implications of sharing data.

The important role of funding bodies in fostering data sharing is clearly illustrated by the history of the production and dissemination of sequencing data. NSF and BBSRC have strongly supported the release of sequence data as a free, open resource, and funding bodies have also been instrumental to the establishment and implementation of the Bermuda Rules. The role of funding bodies in promoting good data sharing activity has recently been formalised thanks to the recent release of several reports and position statements outlining the responsibilities and future challenges of open data for funders (Finch 2012; Royal Society 2012; Wellcome Trust position statement on data management and sharing 2010, http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm).

However, relatively few resources and funding streams are currently explicitly set aside to foster data curation and sharing (Bastow and Leonelli, 2010). Also, it is not clear how public funders will police researcher compliance with their policies. This is an important issue, since it has been demonstrated that very few researchers sponsored by funding bodies with supposedly strict and clear data release policies actually release their data (Alsheikh-Ali et al., 2011).

Government

Governments in the western world have held policies supporting data sharing for many years, as expressed for instance in the OECD Report *Principles and Guidelines for Access to Research Data from Public Funding* (2007; http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf). Aside from supporting the work done by funding bodies, national governments need to actively monitor the broad impact and long-term significance of data sharing in science and beyond. This includes the ways in which data sharing might help to meet societal needs in the future and how data sharing might increase returns in investment in UKPLC, as well as its future capability to tackle major societal challenges (such as climate change and food security).

Publishers and journals

In their role as the principal means for science communication, and of coordinators and facilitators of peer review processes, journals act as guardians of the scientific record. They therefore can play a key role in facilitating data sharing practices, as their policies and practices have a huge influence on the activities and behaviour of researchers, who need to publish to promote their own research.

This role may involve taking responsibility for publishing data themselves, alongside papers. Most journals do not support this option however, and prefer to facilitate access to data underlying specific publications without committing to storing data themselves. In this way, journal editors and publishers see themselves as the conduit between data repositories and data producers/users.

One of the most important and relatively simple things journals can do in this role is to encourage their authors to cite data, for example via DataCite (http://www.datacite.org/), which establishes two way links between articles and archived data. Citations of specific databases should also be encouraged, especially when data used within a piece of research has been retrieved from a curated resource. A helpful guide to data citation is provided by the Digital Curation Centre at Edinburgh (Ball and Duke, 2012).

Another important role for journals is to provide clear policies on data sharing, tailored to the needs of the research communities served by the journal in question. This means that journals should understand, and provide guidance on, the community norms at play among their readers and authors. This includes their attitude towards data sharing and their preferences for specific repositories already available, the type of data available, the data needed for future research, and confidentiality issues that might be linked to the publication of specific data types.

A more difficult challenge, which researchers view as very important, is for journals and publishers to cooperate directly with data repositories so as to optimise the flow between data sharing and paper publication. This might mean that journals take on the additional task of monitoring the quality of data in publications, as mentioned above, for instance through the management of SI; and devote attention to enhancing the discoverability of data in the papers that they publish, for instance by improving links between data and paper.

Finally, journals can and should provide clear guidance to peer reviewers about what they should look for in terms of data submission and access; and to researchers, by explicitly stating how peer review procedures for data actually work.

Conclusions

Incentive, policing measures, and shifts in culture are needed in order for data sharing to take hold and bear fruit within biology as a whole. Funding bodies, universities, and publishers and journals can provide important 'sticks and carrots' by shifting priorities and attitudes to support the practice of data sharing, with all its demands. Incentivising data sharing will be a complex process, and each step needs to take in to account the needs of the community of researchers that is being targeted. This calls for a mixed model in regulating and supporting data sharing, such as provided by consortia of different institutions (as seen in the case of the Bermuda Rules).

At the same time, researchers need to seriously commit to data sharing by making it part of their principal aims and outputs. In most cases, community involvement matters much more than the availability of technology. Better curation is achieved through bottom up approaches. For example, if a community agreed that the baseline for data sharing should be that all the data produced through public funding is made accessible via a generic system like Dryad or Figshare, this would substantially shift data sharing principles. Although this level of data sharing will probably not be sufficient for 'intelligent' data access and data mining it would encourage an ethos of donation, and provide an important step toward more sophisticated forms of data sharing necessary for the data-intensive

science of the near future.

One important distinction to be made in terms of data sharing is that between data storage and data curation. It is possible to store some datasets with minimal curation, or to curate data only at the point of inclusion into a repository, without need for updates and tailored maintenance. Workshop participants referred to these low-input databases as 'bulk storage'. For instance, a well curated Affy or RNA seq experiment will generate a dataset, which if associated with the appropriate meta data and pre-analysis, will not need to be curated again and will be re-usable by others. However, in other cases data storage requires a sophisticated level of regular curation in order for data to be, and remain, re-usable. For example, complete genomes or pan genomes are consistently being updated and modified as new gene models are discovered, new accessions are sequenced or new technologies become available. In such cases, data curation and data analysis overlap considerably, and therefore maintaining the data so that is useful and reusable require considerably more support and attention from the research community and scientific institutions than in the case of general storage within a repository or database.

Whether publishers should be held responsible for data storage, as well as publication, needs careful consideration. On the one hand it might be viewed as unfeasible for journals to be involved in this, given the related costs and their lack of expertise in data curation. However, it is not clear why publishers should not be held responsible for the maintenance of primary databases, which hold essential data used as evidence in published papers. Smaller, sophisticated databases might well be best maintained by researchers themselves, but it is not clear who should take responsibility for linking those databases with each other and with bigger, more general databases.

Publishers indicated they would rather build collaborations with existing repositories than build in-house databases. This puts much of the financial burden of data sharing on public funders, as they constitute the main source of financial support behind the vast majority of credible databases that are currently freely available online.

Another possible solution is offered by the emergence of data-only journals. Wiley and BMC are starting to propose such journals. Recent initiatives include Ecological Archives, ZooKeys, F100Research, GigaScience and Database. The focus of these journals is to facilitate the re-use of data. Yet questions remain about their potential usefulness: are they the most feasible or the best way to store data? Will they make data too dispersed? How do they deal with quality control of data? More broadly, these journals raise a central question about the aims of data publication. A key

goal is the opportunity for data producers and curators to be credited for their work. Also, data publication enhances the perception of research data as a research output in itself. At the same time, it is not very clear what the difference will be between a data journal and a well-curated data repository. In both cases, it is essential to have adequate metadata detailing the provenance of data. Data journals seem to be particularly useful as vehicles for data that are not yet covered by large repositories.

Acknowledgments

The workshop was generously sponsored by the ESRC, GARNet and The Company of Biologists (Development, Journal of Cell Science, The Journal of Experimental Biology, Disease Models & Mechanisms and Biology Open). SL is funded by the ESRC as part of the ESRC Centre for Genomics in Society; RB is funded by BBSRC via grant BBG0214811. We would like to thank Charis Cook for providing the information for Table 1 and designing the front cover image. Image credits: *Dandelion* by Johnny Nyberg and *1/0 Session II – Lost Bits 5* by Carsten Mueller, both via stock.xchg.

Bibliography

Royal Society (2012). *Science as an open enterprise*. Report available on http://**royalsociety**.org/policy/projects/science-public-**enterprise**/report

International Arabidopsis Informatics Consortium (2012). Taking the Next Step: Building an Arabidopsis Information Portal. *Plant Cell* 24: 2248-2256

International Arabidopsis Informatics Consortium (2010). An international bioinformatics infrastructure to underpin the Arabidopsis community. *Plant Cell* 22: 2530–2536.

Hey,T., Tansley, S., Tollle, K., *Eds*. (2009). The fourth paradigm: Data Intensive Scientific Discovery. Microsoft Research, Redmond, WA, USA. Available online: http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Alsheikh-Ali, A.A., Qureshi, W., Al-Mallah, M.H., Ioannidis, J.P.A. (2011). Public Availability of Published Research Data in High-Impact Journals. *PLoS ONE* 6(9): e24357.

Ball, A. & Duke, M. (2012). 'How to Cite Datasets and Link to Publications'. *DCC How-to Guides*. Digital Curation Centre, Edinburgh, UK. Available online: http://www.dcc.ac.uk/resources/how-guides

Bastow, R. & Leonelli, S. (2010). Sustainable digital infrastructure. *EMBO Reports,* 11(10): 730-735.

Howe, D. & Rhee, S.Y. (2008). 'The Future of Biocuration.' *Nature* 455: 47-50.

The Finch Group (2012). *Accessibility , sustainability , excellence : how to expand access to research publications*. Report available on http://www.researchinfonet.org/publish/**finch**/

Leonelli, S. (2010). Packaging Data for Re-Use: Databases in Model Organism Biology. In Howlett, P. and Morgan, M.S. (eds) *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*. Cambridge University Press.

Appendix 1 – Workshop Programme

THURSDAY 12 JULY

12:00 – 13:00 Lunch and registration

13:00 – 13:30 Introduction (Sabina Leonelli and Ruth Bastow)

13:30 – 15:30 Session 1: Data donation, analysis and use (Chair: Ruth Bastow)

Andrew Millar (Edinburgh): “Creating, leveraging and sustaining public data (and more) with uncertain funding”

Nick Smirnoff (Exeter): “Accessing and using metabolomics data”

Jay Moore (Warwick): “From bench to web, via spreadsheets: practical data sharing in research groups”

Jacob Newman (UEA): “Sharing Data with Omero”

15:30 – 16:00 Tea/coffee

16:00 – 17:30 Discussion session: How are publicly accessible data being used? (Chair: Cathie Martin - editor of Plant Cell)

FRIDAY 13 JULY

9:00 – 10:40 Session 2: Curating and publishing data (Chair: Steve Hughes)

Mary Traynor (editor of Journal of Experimental Botany): “Providing more actionable data associated with articles”

Gilles Jonker (Executive Publisher for Agronomy at Elsevier): “Connecting Scientific Articles with Research Data”

Ruth Wilson (Nature Publishing Group): “Integrating Research Data and Publications”

Claire Bird (Senior Publisher, Life Sciences, Oxford Journals) “What role can publishers play in managing data?”

10:40 – 11:00 Tea/coffee

11:00–12:40 Session 3: Data curation and management (Chair: Sabina Leonelli)

Sean May (NASC): “NASC: Reciprocal CTRL-ALTruism”

Mark Hahnel (founder of Figshare): "Getting credit for all of your research"

Peter Burlinson (BBSRC): “Data sharing: a perspective from the BBSRC”

12:40 – 13:30 Lunch

13:30 – 15:00 Final discussion: The impact of data dissemination on plant science research

Appendix 2 – Speaker Biographies

Ruth Bastow has been the GARNet Coordinator since 2004 and is responsible for the day to day running and management of GARNet. Previously, she worked as a postdoctoral researcher in the labs of Prof Caroline Dean (John Innes Centre) and Prof Julie Gray (University of Sheffield). She carried out her PhD in the lab of Prof Andrew Millar (University of Edinburgh).

Claire Bird is a Senior Publisher in Life Sciences for Oxford Journals. Claire joined Oxford Journals after graduating in molecular and cellular biochemistry from Oxford University in 2000. She has responsibility for plant science and computational biology journals, including the Journal of Experimental Botany, Annals of Botany, Bioinformatics, and Database. She is particularly interested in tackling challenges in the research publishing environment, and played a central role in launching Oxford Journals' open access initiatives.

Peter Burlinson works in the Genomics, Data and Technology team at BBSRC as a strategy and policy officer. His main areas of responsibility are the Bioinformatics and Biological Resources Fund and Call Two of the Tools and Resources Development Fund. He also contributes to the delivery of BBSRC strategy relating to the computational requirements of the biosciences, in particular in the area of high performance computing, as well as data sharing. Prior to joining BBSRC he was a postdoc at INRA-Nancy in France and at the University of Toronto in Canada, involved in projects in the general areas of molecular microbiology, microbial genomics and plant sciences. He undertook his doctorate at Oxford University, applying forward and reverse genetics to dissect Pseudomonas interactions with C. elegans and a variety of other eukaryotes.

Mark Hahnel is the founder of Figshare, an open data tool that allows researchers to publish all of their data in a citable, searchable and sharable manner. He's fresh out of academia, having just completed his PhD in stem cell biology at Imperial College London, after previously studying genetics in both Newcastle and Leeds. He is passionate about open science and the potential it has to revolutionise the research community. For more information about FigShare, visit http://FigShare.com. You can follow him at @figshare.

Gilles Jonker has over 20 years experience in international STM Publishing. He joined Elsevier in 1998 from Kluwer Academic Publishers and has held Publishing Editor and Publisher positions in Chemistry and Physical Sciences. In his current role of Executive Publisher he leads Elsevier's

activities in agriculture, plant and soil sciences. In addition to the management of 34 international journals, his current role includes the organisation and direction of ancillary activities like conferences, seminars, journal launches and online information services. Gilles has established successful partnerships with academia, industry, NGOs and stakeholders in the fields of agriculture, plant and soil sciences.

Sabina Leonelli is a Senior Lecturer in the Department of Sociology and Philosophy and a Fellow of Egenis at the University of Exeter. Her research spans the fields of history and philosophy of biology, science and technology studies and general philosophy of science, currently focusing on data-intensive science and its impact on existing practices of knowledge production and scientific governance, particularly within plant science. She has edited a book on Scientific Understanding (2009, Pittsburgh University Press) and a special issue on 'Data-driven research in the biological and the biomedical sciences' (2012, Studies in the History and the Philosophy of the Biological and Biomedical Sciences), as well as several publications on the sustainability and curation of biological databases and bio-ontologies, and is currently working on a monograph on the philosophy of data-intensive biology. As an elected member of the Global Young Academy, she coordinated the writing and release of the GYA position statement on Open Science.

Sean May works for NASC, which is involved in the provision of biological resources as the primary European Arabidopsis Seed Centre, and in the provision of information relating to its stocks and Arabidopsis genome information in general. NASC operates in parallel with the Arabidopsis Biological Resource Center (ABRC) at Ohio State, providing a safe repository for nearly a million accessions and distributing this material to the scientific community. For details of the stocks distributed, take a look at http://arabidopsis.info. NASC currently distributes around 130,000 tubes of seed per year. From 2002 onwards, NASC also generated and released thousands of Affymetrix GeneChip datasets into the public domain, thereby making the first large global plant transcriptomics dataset (and for the first few years, outstripping the public animal datasets). It has continued this project as a cost-recovery service for the sake of continued public release of data. NASC was also involved in the development of genomic information resources to support other elements of the ongoing Arabidopsis research project. NASC originally developed the Arabidopsis Genome Resource (AGR) as part of UKcrop.net in the late 1990s and then produced atensembl.arabidopsis.info as its replacement. The focus of the database was to integrate Arabidopsis Genome Initiative (AGI) sequence data with physical and genetic maps of Arabidopsis as well as transcriptomics and germ plasm data (due to funding, this database is now only archival and is not actively developed).

Andrew Millar holds a Chair of Systems Biology at the University of Edinburgh, where he was the founding Director of the Centre for Systems Biology at Edinburgh (CSBE), now SynthSys Edinburgh. He was previously involved in the Scottish Universities Life Sciences Alliance (SULSA) and GARNet, the UK's Arabidopsis research network. Andrew grew up in Luxembourg and studied Genetics at Cambridge University. He began working on biological rhythms in 1988 during his Ph.D. with Nam-

Hai Chua at The Rockefeller University, New York. After postdoctoral research with Steve Kay and Gene Block at the NSF Center for Biological Timing at the University of Virginia, he worked from 1996 to 2004 at the University of Warwick. There he collaborated with Matthew Turner on mathematical models of the plant clock, and with David Rand on model analysis, leading to his current interests in Systems Biology. Research in CSBE, and also in Andrew's own group, used Systems Biology to understand dynamic biological systems, mostly at the intracellular level. This approach integrated experimental biology, mathematical modelling, the development of new theory, and of informatics infrastructure. The data infrastructure grew from work on the BBSRC-funded ROBuST and EU-funded TiMet projects. These distributed projects needed online data sharing among multiple collaborators, which was sustainable only because the same system also provides specialised data analysis. The infrastructure is now used for public data sharing, with data used for experimental comparisons and machine learning. Linked infrastructure supports the sharing of models in XML formats (www.plasmo.ed.ac.uk) and facilitates modelling tasks, notably model optimisation on high-performance computers (www.sbsi.ed.ac.uk). SynthSys Edinburgh builds on CSBE's research success, adding Synthetic and Chemical Biology methods in order to validate and implement novel applications of networked biology.

Jay Moore is currently working within Warwick Systems Biology Centre specialising in data integration, palaeointeractomics and evolutionary systems biology. He is working on analysis of deep-sequencing and protein interaction data and is interested in developing integrations of pre/historical, geographical, biodiversity, and omics data. He is developing bioinformatic approaches to understanding response to environmental stress in Arabidopsis and collaborating on genome assembly and development of a SNP platform for Brassica oleracea analysing genome resequencing and transcriptome sequence data. He is on the PALS group of the SYSMO programme, developing common approaches to bioinformatics for the systems biology of microorganisms. He is an honorary member of the Allaby Research Group which specialises in plant evolution and crop domestication, with whom he developed the TreeMos application to identify and visualise phylogenetic mosaicism, and collaborate on plant archaeogenomics. He maintains an ongoing interest in developing comparative models of genome organisation in Brassicaceae.

Jacob Newman is a computer scientist who gained a BSc in Computing with Electronics in 2007, and in 2011 completed a PhD entitled "Language Identification Using Visual features", both from the University of East Anglia. He is currently a senior research associate at the UEA, investigating the behaviour of microtubules in plant cells. His research interests include machine learning, natural language processing, and more recently, computational biology.

Nicholas Smirnoff is a researcher with an interest in the synthesis and functions of vitamin C (ascorbic acid) in the growth and stress tolerance of plants. This is investigated in the wider context of the role of reactive oxygen species (ROS) and antioxidants in plants. He uses biochemical and molecular genetics techniques as well as transcriptomic and metabolomic approaches. He is part of a systems biology research network with Prof Murray Grant and colleagues at Warwick and Essex

Universities, which aims to identify gene networks involved in plant response to stress. They are also using interspecific crosses combined with metabolite and transcript profiling to identify genes that control metabolic traits. His teaching focuses on biochemistry (particularly metabolism), plant diversity and plant physiology.

Mary Traynor is Managing Editor of the Journal of Experimental Botany. Mary is keenly interested in maximising the accessibility of scientific data and has played a key role in driving the JXB's unique open access policy.

Ruth Wilson is a publisher at Nature Publishing Group where she is responsible for the Nature branded physical science journals, she also has a remit for working across NPG on initiatives linking research data and publication. Prior to being a publisher she held a number of positions in NPG's web development teams including heading up the New Product Development group. Ruth's research background was in chemical physics and she received her PhD from the University of Sussex in molecular dynamics.