BioMedBridges

Principles of data management and sharing at European Research Infrastructures

Version 1.0, January 2014

This paper was jointly developed under the umbrella of the BioMedBridges project by the following research infrastructures:

- AnaEE
- BBMRI
- EATRIS
- ECRIN
- ELIXIR
- EMBRC
- ERINHA
- EU-OPENSCREEN
- Euro-BioImaging
- INFRAFRONTIER
- INSTRUCT
- ISBE
- LifeWatch
- MIRRI

Contents

1 Introduction

Recognising the importance of access to findings from publicly funded research, including research data, there is an increasing requirement from funders to publish and make data broadly available. There is also a societal aspect of data sharing that is particularly pertinent to the health sector, where the release of data that enables an open assessment of patient safety or treatment efficacy is increasingly seen as a basic ethical requirement.

An often implicit assumption in these requirements is that the data will be made available in a machine-readable and reusable format—a prerequisite to realising the value of the data produced within new contexts. Hence, the issue of data sharing and re-use is tightly connected with effective research data management and the presence of processes and infrastructure to support data coordination, meta-data curation and deposition in suitable archives.

It is also important to note that access—and even Open Access—to data does not automatically mean that data can or will be freely accessible to anyone at any time and for any purpose. In some cases this may be true, while in others very specific conditions concerning access to and use of the data may exist (e.g. data with intellectual property restrictions or, particularly in the case of data related to human health, sensitive data with ethics or privacy considerations). The specific circumstances and environment, as well as the necessary technical infrastructure, must be considered.

To ensure effective use of the research infrastructures (RIs) and the data they produce or hold, this document outlines the context in which data held by the RIs is produced, archived, shared and used, and describes how this may influence data sharing. It provides recommendations on how to facilitate and support good data management practices and encourage data sharing, and includes guidelines concerning how, depending on their focus on basic and/or applied research and the implications arising from this, data sharing within the RI user communities of the fourteen life- and environmental sciences RIs that have developed this document may be stimulated. Other RIs may join this initiative at a later point and the document may be updated at any time to address developments in terms of policy, funding, and/or strategy of the RIs.

2 Recommendations for data management and sharing

2.1 Summary of recommendations

1. The RIs encourage data sharing and reuse and support the notion that public funders should encourage Open Access to data from publicly funded research where possible.

2. Importantly, some data may only be shared under certain conditions and with appropriate safekeeping mechanisms in place, such as personally identifiable data, data subject to ethical or legal restrictions, or restrictions for intellectual property protection.

3. To encourage data sharing, systematic reward and recognition mechanisms are necessary.

4. Proposals for publicly funded research at RIs should include a data management plan concerning the deposition of data in long-term archives that addresses specific resources and activities (including standardisation of data production and curation/annotation).

5. Funding for tools and activities connected to data deposition must be available.

6. Systems, services and resources must be in place to facilitate straightforward data deposition by researchers, including support concerning the necessary data use agreements and consent forms for data with data protection or intellectual property requirements.

7. Systems are also needed to capture and track data provenance and use.

8. To ensure necessary trust by data providers or depositors, RIs must guarantee high standards of security and traceability.

2.2 Detailed recommendations

Overall, the **BMS RIs encourage data sharing and reuse** of data either generated or deposited at the RIs, regardless whether the funding was provided by the RI itself or by public (research) or private (commercial) funders, and concerning both raw and/or derived data. It is important to note that, in some cases, it may only be possible to make data available and accessible under certain conditions, such as **protection of intellectual property or safekeeping of personally identifiable data, or data underlying ethical considerations**. To ensure smooth operation of data integrating infrastructures or infrastructures providing data interoperability, good and agreed data sharing and management policies must be in place.

To encourage data sharing in general, **systematic reward and recognition mechanisms for the sharing of data and materials** need to be installed. These may include citation mechanisms for data and measurements of citation impact; however, it is important to investigate other mechanisms as well.

Data publishing journals are established in the earth sciences and are becoming more and more accepted also in the entire spectrum of the life and biomedical sciences. These journals provide new ways to disseminate, organise, understand and, in some instances, use data. One recent example is “Scientific Data[1]” from Nature Publishing Group, a new open access, online publication for descriptions of scientifically valuable datasets, which intends to make the data more discoverable, interpretable and reusable. Another example is the Journal GigaScience, which publishes 'big-data' studies from the entire spectrum of the life and biomedical sciences. The journal links standard manuscript publication with an extensive database that hosts all associated data and provides data analysis tools and cloud computing resources. If datasets currently serviced by large public repositories can be published or referenced in such journals, and these publications are citable and their impact measurable like that of traditional manuscript publications, publication of datasets and sharing of reusable research data would become a first-class objective of science.

There is also an emerging system of both institutional and community-driven data repositories such as DataDryad[2] and figshare[3], which complement large thematic repositories and data archives such as Ensembl[4], Pride[5] and UniProt[6]. As data discovery across such repositories remains a significant challenge, there are emerging efforts to address this such as re3data[7], which aims to support the identification and reuse of datasets.

The RIs very much agree with the European Commission that **public funders should encourage sharing of data from publicly funded research**. As outlined in the Commission’s recommendations from July

[1] http://www.nature.com/scientificdata/
[2] http://datadryad.org/
[3] http://figshare.com/
[4] http://www.ensembl.org/index.html
[5] http://www.ebi.ac.uk/pride/
[6] http://www.uniprot.org/
[7] http://www.re3data.org/

2012[8], this might be accomplished by crediting the researcher for their sharing and, in turn, helping them to secure future grants.

Proposals for publicly funded research at RIs should include a **data management plan**. The plan should describe how the proposal will conform to the funder's policy on the dissemination and sharing of research results and must specify resources and activities concerning deposition of data in long-term archives. Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data created or gathered in the course of work during the project as far as permitted by current legislation. Grant beneficiaries are expected to encourage and facilitate such sharing.

There must be **funding for activities connected to data deposition**[9] as these can be significant and expensive efforts (e.g. large data volumes, provision of high quality metadata that needs manual curation/annotation). Appropriate funding models have to be evaluated in liaison with different European national and international funding organizations.

Funding must also be available to cover **the cost of storing research data**. This is especially important in the case of large datasets. These costs could be minimized for academic researchers and SMEs if this is undertaken by the RIs themselves. Since access and even Open Access to research data does not mean free access, the costs of access might be different for users willing to provide data and those not willing to provide data. These options of curating and preserving data in exchange for Open Access have to be evaluated. Funding bodies might even provide resources to data managing RIs that enable them to support projects with funds to hire experienced personnel to prepare and deliver their data.

[8] http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf

[9] The Swiss SyBIT initiative may serve as an example of a possible approach to supporting data deposition by researchers.

Systems and services must be in place to **facilitate and support straightforward data deposition** by researchers themselves. Systems are also needed to **capture and track data provenance**, e.g. in the context of specific conditions for the use of data[10] or where attribution is requested by the data depositor[11]. Such systems require the development of new tooling. To ensure the necessary **trust by data providers or depositors**, RIs must guarantee **high standards of security and traceability**. They must administer data securely while ensuring that their origin is visible and that the originator is suitably acknowledged when his/her contribution is used e.g. for other published work or the development of foreground intellectual property. The development and agreement of common database structures may facilitate this and is worth consideration. In addition, data with any use restrictions (such as ethical or intellectual property requirements) must be deposited with the necessary legal framework in place, and RIs must ensure that use of those data is within the given restrictions. Any licenses and data use agreements must be compatible with the intended use of the data[12].

3 Background of data sharing at RIs

3.1 Roles of RIs and types of data

The type of data produced and/or stored by the RIs as well as the context in which the data is produced and/or stored are very diverse. For the purposes of this document it is important to recognise these distinctions.

[10] e.g. intellectual property restrictions, software licenses, or patient data with specific consent

[11] e.g. citizen scientists contributing data to environmental monitoring or census efforts may want to be acknowledged when data is analysed and/or published

[12] The BioMedBridges project is tackling this issue by providing templates for suitable data use agreements and consent forms via an online tool, a pilot of which was delivered in December 2013.

3.1.1 Type of data produced and/or managed by the RIs

- Medical and clinical data: individual-level, pseudonymous patient data, biosample data, clinical trial metadata
- Very diverse life science data, including population genetics, molecular phylogenetics, DNA barcoding, biobanking data, phenotyping data, genomic, proteomic and other –omics data
- Virulence / pathogenicity factors for highly pathogenic agents
- Biosafety (procedures, rules, instructions) and biosecurity (BSL4 personnel, site security) information
- Data from continuous environmental monitoring and observation and analytical chemistry data from sampled ecosystems
- Chemical screening data
- Biological and medical imaging data, metadata, and data derived from computational pipelines
- Structural biology data: raw data from various sites (synchrotrons, NMR centres, Electron Microscopy and imaging centres) and final atomic coordinates for biomolecules derived through the application of computational pipelines.

3.1.2 Roles of the RIs and context of data generation

It is important to note that there are data producing and data managing RIs, with some RIs having both roles. These roles, together with the context in which data is generated, have implications for the approach taken concerning data management and sharing. A distinction must also be made between raw data and processed and/or curated data.

<u>Data producing RIs</u>

Data producing RIs will have policies and processes for raw data and in some instances also for processed (annotated/curated) data.

RI users may be researchers who produce data and/or researchers who use data deposited at the RI. Data are produced by researchers in the context of funded research projects. Such data belong to the user and the RIs do not assert any IP rights on the data produced, but may impose policies before and/or after data is produced.

Data managing RIs

Data managing RIs may have policies for raw and processed (annotated/curated) data, including privacy and data security policies where applicable, as well as for results derived of deposited data (research results gained from the re-use of deposited data).

Data managing RIs manage data produced elsewhere. RI users may be researchers who have deposited their own data with the RI and/or those who are interested in re-use or repurposing of data deposited by other researchers.

Concerning the origin of research data within the RIs, the following four cases must be distinguished:

1. Data generated within the BMS RI where the whole or part of the project is funded through the RI (e.g. through collaborative research projects)
2. Data generated by a public research project carried out at the BMS RI and where the project covers the operating costs of the RI
3. Data generated by commercial / private entities covering operating costs
4. Data generated by Public-Private Partnership consortiums

Combinations of the above also exist.

3.2 Technical aspects of data sharing

3.2.1 Access and reliability

The following conditions must be met for data to be shared widely:

- Data must be discoverable
- Data must be accessible, i.e. it must be archived in a way that makes it possible for users to get to it in an automated way
- Data integrity must be preserved, i.e. partial data losses must be prevented (via backups) as well as unjustified alterations without traceability; correspondence between raw data and processed/ curated data must be guaranteed
- Data provenance information must be available and propagated through data integration layers
- Data archives/databases must be sustainable and reliable, i.e. they must have permanence so the user (as well as the data provider/depositor) can rely on being able to provide a permanent reference to the data in question (e.g. persistent identifiers). “Down times” must be minimal.

3.2.2 Standards/formats and discoverability

Common data standards and formats are required for data as well as meta-data to enable interoperability, wide sharing, integration and reuse in different contexts. Only when agreed standards are used can data be understood in terms of its syntax (format/structure), meaning (concepts/semantics) and, importantly, the context in which it was generated (provenance). Such standards are a key aspect for data interoperability and to making data discoverable for reuse in future research. Supporting the development and maintenance of standards, together with the research community, is a key role for the RIs.

3.2.3 Data production and provenance

Effective capture of metadata, implementation of data standards, and establishing data provenance are critical and could best be dealt with at the time of data production, e.g. within LIMS systems or clinical case record forms. Data should be accompanied by (or linked to) sufficient metadata to assess its scientific validity, including the capture of experimental detail which is sufficient to allow repeat of the experiment. The RIs strive to issue polices and publish best practices wherever feasible in order to avoid costly and time-consuming data curation efforts at a later stage.

3.2.4 Curation and maintenance

For research data to be made accessible to the wider scientific community it must be curated, with necessary metadata allowing the data to be understood and used. Data curation requires highly skilled specialists qualified in their scientific fields who understand the data and can use supporting software. Policies must encourage researchers and projects to deposit their data and associated “knowledge” (software, metadata, documentation etc.), as there is often a lack of resources to make this additional effort. The data managing RIs also have a key role in supporting this process through the provision of expertise and tools as well as, in some cases, provision of curational resource.

3.3 Data sharing constraints and obligations

External constraints and obligations may include policies imposed by funding bodies and those imposed during the publication process, or may result e.g. from the necessity to protect intellectual property or patient privacy.

3.3.1 Clinical trials

Due to the subject matter, data from clinical trials[13] underlies some special obligations, both before the start of a given trial and concerning publication of the results. These include:

- Trial metadata and trial summary results are published in clinical trial registries[14]
- The World Health Organization has formulated minimal criteria for registration of clinical trials
- The ECRIN transparency policy requires that any ECRIN-supported clinical trial is registered and published irrespective of findings (positive, negative and neutral results) and that, once a trial is completed, the raw, anonymised datasets must be made available to the scientific community upon legitimate request to the sponsor or principal investigator
- Defined standards and requirements for data publication exist, including the preparation of raw datasets for publication
- Trial protocols are shared with regulatory and ethics committees and often published in journals or uploaded to registries.

3.3.2 Data with Ethical, Legal or Societal Implications (ELSI)

To make data with Ethical, Legal or Societal Implications available (e.g. non-anonymised patient data including personally identifiable information), suitable systems must be in place so the data can be securely stored and only accessed if allowed by law or based on explicit consent by the data subject. Data services must be compliant with local, national and European regulations and privacy rules which take into account that data may be shared across national borders (i.e. deposited in one country and accessed from another, or generated in one and

[13] NOTE: investigator-driven clinical *studies* may have very different obligations and requirements than clinical trials

[14] e.g. www.controlled-trials.com; www.clinicaltrials.gov

deposited in another). Appropriate authentication and authorisation processes must be in place. The BioMedBridges project is developing such a system[15]; a pilot of the security framework will be implemented by December 2015.

It remains to be seen how the proposed EU data protection regulation[16] will affect the sharing of data with ELSI constraints.

3.3.3 Requirements from publication in journals

Essentially all journals publishing results from research in **structural biology** require deposition of atomic coordinates in the PDB and release of the information by the time the results of the research have been published. The situation is not widespread regarding EM 3D maps, but the general trend towards compulsory deposition is very clear.

Similarly, for **molecular data** (e.g. nucleotide sequences, proteomics, cheminformatics, metabolomics, genome wide association studies, etc.) most journals either require or strongly encourage deposition of data in a publicly accessible archive, such as those hosted at EMBL-EBI or in future by the ELIXIR nodes.

In the case of **clinical trials**, publication of clinical trial metadata is an obligatory prerequisite for journal publication. In addition, a number of high-impact journals request that depersonalised individual patient data become uploaded or accessible before they will publish results from clinical trials. Results of trials are usually published in aggregated form, but the additional publication of raw datasets is increasing.

[15] BioMedBridges “Secure Access” work package: http://www.biomedbridges.eu/workpackages/wp5

[16] http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm

3.3.4 Requirements of different funding bodies

In July 2012, the **European Commission** adopted a policy package containing a series of measures to improve access to scientific information produced in Europe. The recommendations of the Commission[17,18] encourage Stakeholder Organizations and Member States of the European Union to prioritize the dissemination of data. Improved policies on access to scientific publications and data, preservation and re-use of scientific information are suggested[19]. On 16 December 2013, the European Commission announced that it will continue and expand efforts started under its Framework Program 7 (such as OpenAIRE[20]) through a pilot scheme for Open Access to and re-use of research data generated by selected projects funded under the new framework program Horizon 2020[21].

In addition, funders in several **EU Member States** (e.g. the UK[22], NL[23]) have developed policies and requirements for data deposition and sharing.

US funding bodies require data generated with their funding to be made available under certain conditions and after certain periods of time. For example in structural biology, raw data generated from projects conducted at the **US National Institutes of Health** (NIH) must be made publicly available and are published in PubChem[24]. Other funding bodies (e.g. PSI) may require that coordinates are deposited in the PDB as soon as

[17] http://ec.europa.eu/euraxess/pdf/research_policies/era-communication_en.pdf
[18] http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf
[19] http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf
[20] OpenAIRE is an e-infrastructure to deposit and access peer-reviewed articles and datasets resulting from EU-funded projects
[21] http://europa.eu/rapid/press-release_IP-13-1257_en.htm
[22] See e.g. Digital Curation Centre: Overview of Funder's data policies http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies
[23] http://www.nwo.nl/en/news-and-events/dossiers/open+access
[24] http://pubchem.ncbi.nlm.nih.gov/

available. As of June 2013 it is now mandatory for US government agencies to make data openly available in machine-readable format[25].

3.3.5 Withholding of data

While most of the RIs either encourage immediate publication of data or even state this as their default mode, there are several scenarios in which data produced or held by the RIs may be withheld from publication (either in databases or scientific publications/journals) for a defined period of time (embargo period) or indefinitely, or where special terms of use for the data must be contractually agreed.

For example, the chemical screening data from **EU-OPENSCREEN** has an optional embargo period of 18 months, which allows protection of potential discoveries (and derivatives thereof) via patenting or further work for distinguished scientific journal publication.

In the case of **LifeWatch**, industry users have to establish a contractual agreement with a separate company that deals with risk-taking commercial arrangements. Conversely, industry users of **ELIXIR** may freely access openly available data resources and are under no obligation to publish results derived from those data.

To get access to data with data protection and privacy requirements (such as personally identifiable data held e.g. by **BBMRI** biobanks, ELIXIR), the data user must submit a data access request for the dataset in question. The request is then reviewed and, if appropriate and based on the availability of the necessary informed consent by the data subject, approved by the responsible data access or ethics committee.

Translational research projects such as those facilitated by **EATRIS** consist of a framework of partners, all of whom may have their own data

[25] "Open Government" following an executive order by the US Administration: http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

access and sharing policies. Typically, data will only be shared within the project and not outside, unless there is an explicit decision to do so.

3.4 RI data management and sharing approaches

Not all RIs have formal policies for data archiving and sharing in place at this point, which is primarily due to the different stages of maturity of the developing RIs. However, policies are being developed. In several cases, data are currently shared on an ad hoc basis or based on long established expectations of the scientific community in question.

Biological and medical imaging data (Euro-BioImaging)

No regular archiving or sharing. There are no dedicated repositories, and data producing institutions often do not have necessary resources for data archiving. Where data are archived, this may be on poorly accessible media (e.g. tapes) and deleted after variable periods of time depending on the requirements of different funding bodies. Some archiving is in institutional databases where data is not easily discoverable/accessible.

The Euro-BioImaging principles state that data will be the property of the user who visited the node and generated them. There is no mandatory sharing of data. However, users will be encouraged to share data and those who want to share their data will be supported in the form of publicly available repositories for standardized, annotated image datasets of general relevance for the research community.

Biobank data (BBMRI)

Of the several thousand European biobanks united in BBMRI-EU, most have their own data management, archival and access infrastructure, in particular to ensure data protection and necessary consent of the participants/donors.

The data sharing policies within BBMRI are very diverse as they are closely tied to the informed consent given by the data subjects to the individual biobanks. To increase discoverability and access to the data held by the large number of individual biobanks involved in BBMRI, data catalogues have been developed at the EU-level[26] and some national nodes (BBMRI.se[27], BBMRI-NL, BBMRI.dk). In addition, coordinated actions are being taken to set up infrastructure and tools for data sharing while protecting privacy, to harmonize data and IT across biobanks, and to address ELSI issues.

Chemical screening data (EU-OPENSCREEN)

Raw data typically stay at screening sites. Provided that necessary funding is available, processed data and metadata will be transferred to the European Chemical Biology Database (ECBD)[28], which will be integrated into UniChem[29] and thus connected to ChEMBL[30]. It will also be connected to PubChem and BARD[31] (BioAssay Research Database). Data sharing will thus be obligatory.

Clinical trial data (ECRIN)

Patient data in clinical trials are usually collected centrally at the sponsor/coordinating site and archived in electronic databases. Trials metadata are archived on paper and/or electronically. Requirements for archiving (e.g. Guideline for Good Clinical Practice[32] E6; GCP) and for GCP-compliant data management in ECRIN data centres exist[33]. CDISC

[26] http://www.bbmri.eu

[27] http://bbmriregister.meb.ki.se:8080/AwareIM/logonGuest.aw?domain=BBMRIRegister

[28] Under development by WP12 of the EU-OPENSCREEN FP7 preparatory phase project: http://www.eu-openscreen.de/index.php?id=73

[29] https://www.ebi.ac.uk/unichem/

[30] https://www.ebi.ac.uk/chembl/

[31] http://bard.nih.gov/

[32] http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html

[33] http://www.trialsjournal.com/content/14/1/97

has developed a series of open standards for the exchange of clinical trial data[34].

Ecological Experimental Platforms (AnaEE)

There are currently no public repositories in Europe for the classes of ecosystem data produced by AnaEE users. Most platform sites manage their own data and share it on request. Some countries or funding agencies support central data archives and observatories.

AnaEE will develop a distributed data management framework based on shared standards for data formats, data quality and metadata accompanied by harmonised policies for access to the experimental platforms and re-use of data by third parties.

Environmental data (LifeWatch)

Data collection needs to be fast and effective, and supported by up-to-date data archiving. LifeWatch promotes the deployment of embedded hardware and (remotely programmable) software in the data producing equipment - such as environmental sensors - to fully ensure adherence to standards, to provide persistent identifiers (crucial for data provenance), and to capture all necessary metadata. Processed (secondary) data resulting from research deploying the LifeWatch research infrastructure must be archived since these may be input data for other scientists.

The (ECIC compliant) Statutes of LifeWatch state that data are open and can be shared in the community. Data sharing is organized through the data providers cooperating with LifeWatch on a legal basis.

Data on highly pathogenic agents (ERINHA)

At the present time, each European BSL4 laboratory has its own processes of data management, archiving and access. No formal

[34] e.g. SDTM, ODM; see: www.cdisc.org

common policies for data sharing are implemented at this point but policies are under development.

For the purposes of data sharing, several different types of data must be distinguished in the case of ERINHA. Environmental and patient data (diagnosis) underlie similar data protection obligations as described for the biobanking data above. Data on pathogens (genomic data, sequences, phenotypes) overall can be easily shared. However, some data must be protected because of their sensitivity (due to the dangerousness of the microorganisms; e.g. virulence/pathogenicity factors). Intellectual property generated from research activities such as the development or optimisation of diagnosis, therapy or prophylaxis have to be protected. Finally, data on individual infrastructure organisation and management such as biosafety (procedures, rules, instructions) and biosecurity (BSL4 personnel, security of the site) information is sensitive and is shared only within ERINHA but protected and withheld from external parties. Within ERINHA, no formal common policies for data sharing are implemented at this point but policies are under development.

Life science data (ELIXIR)

ELIXIR's mission is to manage and ensure long-term sustainability of life-science data resources. ELIXIR also develops and maintains *interoperability* and core technical services for data access and will work with other RIs and researchers to facilitate and encourage deposition of datasets (e.g. from genomics research) into public archives and maintain services to enable access and analysis.

The general policy for the major archives provided by ELIXIR partners and within ELIXIR is to make data openly and freely accessible with few restrictions on downstream use, including industry usage. The notable exception is biomolecular and omics data from clinical cohorts managed in secure archives (e.g. the EGA), which are governed by data access committees to ensure compliance with regulations and patient consent.

Marine environmental data (EMBRC)

Data is typically deposited in public archives (e.g. Pangaea[35], European Nucleotide Archive[36]). Data for which there are no public repositories are held by the data generators and shared on request.

EMBRC will develop a more formal data policy during the construction phase in 2014 and 2015 and it is very likely that this policy will strongly advocate deposition of virtually all data into accessible public repositories (exceptions would be for commercially-funded and medically sensitive data).

Microbial Resource Data (MIRRI)

For Culture Collections and Biological Resource Collections (CCs/BRCs), no centralized approach for data acquisition, storage, archiving and accessibility exists. Data quantity and quality as well as storage systems are highly diverse and not connected. Raw experimental data are archived permanently with resource accession forms. Genome data are deposited in public archives.[5] No policy exists for archiving environmental data. Depositors are responsible for assuring the quality of data associated with the biological material. MIRRI aims to improve the quantity, quality, interoperability and usage of data associated with biological material in CCs and BRCs for stakeholders in academia and industry. MIRRI will develop concepts and standards for the acquisition, evaluation, curation, integration and interoperability of existing and future data across CCs/BRCs and other national and international projects and initiatives.

Mouse disease model information and data (INFRAFRONTIER)

Curated summary genetic descriptions of mutant mouse lines archived and distributed by INFRAFRONTIER are provided by INFRAFRONTIER

[35] http://www.pangaea.de/
[36] http://www.ebi.ac.uk/ena/

itself. Original data is held by the mouse providers. Depending on the nature of the projects, raw phenotyping data produced by the INFRAFRONTIER mouse clinics are either stored, curated and annotated and made accessible locally, or it is uploaded to centralised publicly available databases such as www.mousephenotype.org, where annotation and curation takes place.

The general policy in INFRAFRONTIER is to provide Open Access to annotated data and metadata, either immediately and centrally (e.g. www.mousephenotype.org) or locally after an embargo period to allow prior publication. Exemptions are collaborations with the private sector. Interfaces for programmatic access to raw data are under development.

Structural biology (INSTRUCT)

Raw experimental data are not archived permanently except at the user's discretion and care. Atomic coordinates and the (derived) data/restraints used to generate them as well as Electron Microscopy 3D maps are deposited in the worldwide Protein Databank (wwPDB[37]).

The data sharing policies of the INSTRUCT user community are very well established, with global coordination via wwPDB. In this way, atomic coordinates are shared via the PDB, which also requires deposition of the experimental data underlying the coordinates, while other types of data are shared via the BioMagResBank[38] (NMR) and EMDB[39] (EM 3D maps).

Translational research (EATRIS)

There is no centralised approach to data archiving. Data is held on a large variety of media at the academic institutions' laboratories and sites involved in projects. EATRIS is developing data handling policies (analysis, standardisation, harmonisation and storage) tailored to the

[37] http://www.wwpdb.org/
[38] http://www.bmrb.wisc.edu/
[39] http://www.emdatabank.org/

technology platforms to streamline medicinal product development processes and e.g. enable multicentre trial activities compliant with current GxP regulations and ethical guidelines.

The general principle for data sharing within EATRIS will be that participating institutions are responsible for their own data and privacy levels and policies in compliance with national legislation and EU directives. Within the context of projects, data will be exchanged between project parties subject to the conditions set in the project agreement. EATRIS itself is never a party in these agreements, but is planning to define best practices for data and privacy policies.