BRIEFINGS IN BIOINFORMATICS. VOL 15. NO 1. 65–78
Advance Access published on 9 October 2012
doi:10.1093/bib/bbs064

Data management strategies for multinational large-scale systems biology projects

Wasco Wruck, Martin Peuker and Christian R.A. Regenbrecht

Submitted: 13th July 2012; Received (in revised form): 4th September 2012

Abstract

Good accessibility of publicly funded research data is essential to secure an open scientific system and eventually becomes mandatory [Wellcome Trust will Penalise Scientists Who Don't Embrace Open Access. *The Guardian* 2012]. By the use of high-throughput methods in many research areas from physics to systems biology, large data collections are increasingly important as raw material for research. Here, we present strategies worked out by international and national institutions targeting open access to publicly funded research data via incentives or obligations to share data. Funding organizations such as the British Wellcome Trust therefore have developed data sharing policies and request commitment to data management and sharing in grant applications. Increased citation rates are a profound argument for sharing publication data. Pre-publication sharing might be rewarded by a data citation credit system via digital object identifiers (DOIs) which have initially been in use for data objects. Besides policies and incentives, good practice in data management is indispensable. However, appropriate systems for data management of large-scale projects for example in systems biology are hard to find. Here, we give an overview of a selection of open-source data management systems proved to be employed successfully in large-scale projects.

Keywords: *data management; data sharing; open access; data citation; systems biology*

INTRODUCTION

Dissemination of scientific data and knowledge catalyzes worldwide scientific progress. Enormous additional knowledge and insights could be extracted from existing projects if their data were publicly available. But often, these data are inaccessible for future research projects or data are stored in proprietary data formats that are not interoperable standard formats. Therefore, governments, funding agencies and the Organization for Economic Cooperation and Development (OECD) commissioned studies concerning the best practice to handle publicly funded research data. First, the US national council on research requested full and open access to publicly funded research data [1]. The OECD confirmed this demand in its guidelines arguing that open access to data enables 'testing of new or alternative hypotheses and methods of analysis' and 'exploration of topics not envisioned by the initial investigator' [2]. The studies are adopted by funding organizations so that in an increasing number of grant application calls researchers are asked to commit to an open access data sharing policy [3]. However, data sharing and good practice in data management require funding of an efficient infrastructure including training of researchers [4].

In smaller projects, data management is often realized via a wiki or similar content management systems—systems managing arbitrary contents of web pages. For large-scale projects, more professional

Corresponding author. Wasco Wruck, Institute of Pathology, Charite - Universitaetsmedizin Berlin, Chariteplatz 1, 10117 Berlin. Tel.: +49 30 2093 8951; Fax: +49 30 450 536 909; E-mail: wasco.wruck@charite.de

Wasco Wruck is working as bioinformatician at Charité, Berlin. He is responsible for data management in the multinational hepatocyte/iPS project livSYSiPS. Beyond that, he chairs the data management group of the transnational ERASysBio+ initiative consisting of 16 projects funded by the European Research Area.

Martin Peuker is head of the Central IT services at the Charitè Univeritätsmedizin Berlin.

Christian R.A. Regenbrecht is group leader for Cancer Stem Cell research and PI in various cancer and systems biology-related projects. He is head of data management and deputy coordinator for the ERASys+ livSYSiPS project.

solutions are needed. We define the requirements for a data management system in the area of systems biology and systems medicine, respectively, and give an overview of data management solutions which have been successfully employed in large-scale projects. Although there has been one short publication comparing the data management system MIMAS to other systems [5], this is the first comprehensive review of data management systems successfully employed in large-scale biology projects which are surveyed in an environment of multinational data sharing strategies.

STRATEGIES FOR PUBLICLY FUNDED RESEARCH DATA

Targeting at the maximal possible gain from the invested public funds, strategies are needed which have to ensure that resulting data from the funded projects will be publicly available and benefit the systems biology community. Below we describe the organizational and technical issues involved in reaching this goal.

Policies

The OECD study [3] was initiated by science and technology ministers and has been recommended for adoption by the executing entities. Thereafter, many funding organizations, journals and scientific institutions—e.g. Biotechnology and Biological Sciences Research Council [6], the Wellcome Trust and the Sanger Institute—have implemented data sharing policies. Field *et al.* [7] suggest to agree on a single data sharing policy consensus template emphasizing public and timely delivery of data in secure public databases with a long-term funding horizon. This suggestion appears reasonable because conflicting sharing policies would be a major obstacle for research in international consortia. Ministries and funding agencies of other countries could adapt these policies as a template.

For joint European research projects, harmonization is already recommended in the report of the EU-initiated and funded project 'PARSE.Insight' [8]: 'An integrated and international approach is desired, in which policies are geared to one another to ensure efficiency and rapid development of policies'. A full discussion of the large variety of countries' data sharing policies and infrastructures would go beyond the scope of this review. Here, Joint Information Systems Committee (JISC) reports provide more detail [9]. A straightforward approach which is already exercised in UK, Canada and USA (and recommended to funding partners in other countries, e.g. Germany) is to make data sharing mandatory in order to receive (full) funding. In Canada, it is consent 'that all funded research be made openly available for future use and ensure this is a condition attached to future funding decisions' at the latest 2016 [10].

Data management plans as required prerequisite for funding in Australia [9, 11] can then be employed to control the delivery of data assets planned. Even if there are sharing policies installed, the practices to control data availability vary as the US Government Accountability Office (GAO) [12] reports for the field of climate research. The GAO report also gives an example for withholding of grant payments because of reluctance to share data with other researchers [12].

Restrictive use of intellectual property (IP) rights and privacy are identified as main hurdles for open access to public research data. Shublaq *et al.* [13] discuss privacy issues arising from the expected higher frequency of sequencing whole genomes of individuals due to the continuously decreasing costs of the technology. An International Council for Science report detects additional impediments in 'trends toward the appropriation of data, such as genetic information and the protection of databases' [14]. A report of the Australian Prime Minister's Science, Engineering and Innovation Council provides an explanation for the culture of non-sharing as due to the pressure for commercialization and competition even between academic groups [15]. To counteract these tendencies, [14] recommends incorporating open access to research data into IP right (IPR) legislation. However, privacy issues can mostly be coped with by anonymization and thorough use of data in order to minimize the risk of misuse or disclosure. IPRs not necessarily have to be in conflict with open access to data. Data producers can grant open access to their data without relinquishing their authorship. Embargo periods can be employed to retain the release of data for a defined period of time—to enable the author to prepare a manuscript or file for a patent, etc. Nonetheless, these periods need to be justified as required by the Wellcome Trust [16]. Mechanisms to give credit to authors of data are an additional incentive to overcome the reluctance of sharing data.

Data sharing infrastructure

The OECD study already emphasized the need for development and maintenance of data management systems in order to build up and secure a stable data management infrastructure. The infrastructure and also training programs for researchers have to be funded in respect to a long-term horizon in order to preserve data and to make it available for possible future investigations. Data centers are the basis of this infrastructure, but networks of institutional repositories gain importance. Data federation of distributed international repositories—a reasonable complementary approach to central national repositories—is not easy to implement because it has to deal with different data formats, different languages in essential documents and jurisdictional hurdles like license agreements. Ruusalepp [9] reports that 'a significant portion of data sharing infrastructure funding is being allocated to developing technical solutions for data federation from different repositories in one research domain and across domains'.

REQUIREMENTS FOR LARGE-SCALE SYSTEM BIOLOGY DATA MANAGEMENT

The basic functionality of a data management system includes (i) data collection, (ii) integration and (iii) delivery. Data collection (i) maintains and provides storage guaranteeing data security. It might abstract from physical storage questions, e.g. via references in an assets catalogue. Superior to manual data collection are semi-automated approaches allowing batch import or even fully automated approaches via harvesting where data from distributed repositories are automatically transferred into the system—a technique also employed to crawl metadata from self-archived publications for the open archives initiative [17]. Central concepts of data integration (ii) are metadata (systematic descriptions of data), standardization and annotations in order to make data comparable. With respect to interoperability of different 'omics' data, the idealistic data management system has to comply to state-of-the-art community standards in the field as largely covered by the Minimum Information for Biological and Biomedical Investigations (MIBBI) checklists, e.g. Minimum Information About a Microarray Experiment (MIAME) for microarray transciptomics and MIAPE for proteomics. 'Omics' might also profit from other existing omics' standards, e.g. lipidomics from metabolomics. Subramaniam *et al*. [18] provide information about lipidomics data management. 'Omics'-specific analysis and visualization functionality are optional requirements. As most large-scale projects are expected to use a large diversity of 'omics' data, systems should be flexible to incorporate as much 'omics' data as possible via mechanisms to integrate the specific standard formats, e.g. the SysMO-SEEK Just Enough Results Model (JERM) templates (see 'SysMO-SEEK' section). However, specific requirements of projects restricted to dedicated 'omics' might be adequately met by systems tailored to that 'omics' data, e.g. BASE for microarray transcriptomics or XperimentR for a combination of transcriptomics, metabolomics and proteomics. Data delivery (iii) includes dissemination to a broad public—requiring researchers to grant access to their data. Mechanisms to make private data automatically publicly available after a dedicated 'embargo period' are under discussion in order to improve the accessibility to the scientific community. Support for publications, data publications and upload to public repositories are a requirement facilitating systems biologists' work. Extensibility to new data types and functionality as well as intelligently designed interfaces for integration with other systems are indispensable for a system in a continuously changing environment.

Further requirements are quality control and curation of data collections guaranteeing quality and long-term usability of data [19]. This plays an important role in many databases, e.g. BioModels [20] where curated data sets provide a higher quality than non-curated data sets.

Systems biology projects have to deal with two big challenges: high-throughput data and data diversity. High throughput in this context should be understood as a large number of experiments made in a short time mostly in a strongly parallelized and miniaturized fashion and is topped up by state-of-the-art next-generation sequencing techniques generating huge amounts of raw data which have to be managed in order to derive information from it [21]. One way to manage such data is cloud storage (and computing) [22]. The second big challenge is the diversity in large-scale systems biology projects comprising data types from all phases of the systems biology cycle (hypothesis—experiment—evaluation—model). Thus, these projects typically comprise Systems Biology Markup Language (SBML) models as well as nuclear magnetic resonance data,

proteomics data, microarrays and next-generation sequencing data generated in the experimental phase—only naming the most prominent ones. Another challenge emerges in large-scale systems biology projects like ERASysBio+ (85 research groups from 14 countries in 16 consortia) [23] in the context of the European Research Area or SystemsX in Switzerland (250 research groups in 62 approved projects) [24] or at the Superfund Research Center at Oregon State University [25]: here a data management system has to integrate existing data management solutions in consortia or dedicated research groups.

In addition to fulfilling these basic requirements, many data management systems include multifaceted features for data analysis, e.g. BASE [26] provides tools for microarray data analysis via a plug-in architecture, SysMO-DB integrates modeling via JWSOnline [27]. Non-functional requirements like reliability, scalability, performance and security obviously have to be respected in any data management system as well as basic functional requirements, e.g. storage management. Table 1 summarizes the specific functional requirements described earlier for an ideal data management system in large-scale systems biology projects.

Standardization, metadata and annotation

Standard data formats and annotations are required to make data comparable. The MIBBI [28] recommendations represent community standards for bioinformatics data in form of checklists specifying the minimal information needed to reproduce experiments. They are well established in many areas, e.g. MIAME [29] for microarrays, while other areas as novel sequencing techniques require further elaboration. MIBBI is complemented by many other standards, e.g. protein standards initiative—molecular interactions (PSI-MI) [30] in the proteomics area, mass Spectrometry Markup Language [31] in the mass spectrometry area, BioPax [32] as exchange format for pathway data, SBML [33] for the description of systems biology models in xml and similarly CellML [34] with the ability to describe arbitrary mathematical models although the focus is biology, CabosML [35] for the description of carbohydrate structures. There are also efforts to standardize the graphical notation in systems biology in SBGN [36] and to find lightweight syntax solutions to exchange observation data [37]. Here, we can only name the standards with the highest relevance to systems biology data management and refer to Brazma *et al*. [38] for a more detailed discussion of standards in systems biology. Many standard data formats are defined in xml to adequately represent metadata—systematic descriptions of data which is indispensable to compare and reproduce. Besides standard data formats 'ontologies' are frameworks to standardize the knowledge representation in a domain. The Open Biological and Biomedical Ontologies foundry [39] provides a suite of open-source ontologies in the biomedical area. Common examples in bioinformatics are the formal characterization of genes in the Gene Ontology [40] and of gene expression experiments in Microarray Genomics Data Society (MGED) ontologies [41].

Annotated information is required for data integration using matching identifiers, e.g. two microarray experiments on different platforms can be

Table 1: Functional requirements for data management systems in large-scale systems biology projects

No.	Requirement	Notes
1	Support for standard data formats	MIBBI, SBML, PSI-MI, mzML, etc.
2	Assistance in metadata annotation	E.g. suggesting predefined values from ontologies
3	Automation of data collection	E.g. via harvesting from distributed repositories
4	Support for modelling data	Primarily SBML models, CellML models
5	Support for upload to public repositories	NCBI GEO, EBI ArrayExpress, BioModels, JWS-Online, etc.
6	Extension system	To new data types and functionality (e.g. plug-ins)
7	Integration with heterogeneous DM systems	SW design, interfaces, web interfaces, servlets
a	Fine-grained access control	Keep data private, share with dedicated users, groups, world
b	Embargo periods	Retain data publishing until predefined time points
c	Support for publications, data publications	Providing data for supplementaries, data publications
d	Support for large data	Depends on the technique, e.g. next-generation sequencing
e	Analysis and modeling functionality	Optional
f	Connectivity to relevant external resources	(Optional) via integration (data warehouses) or links

compared by mapping the vendor-specific probe ids to ENSEMBL gene ids. However, annotations are changing and use of synonyms is common. Here, the connexin GJA1 might serve as an example which was formerly often referred to as CX43. Furthermore, many annotations are ambiguous—e.g. one ENSEMBL gene id might map to multiple Illumina probe ids. Thus, data management systems should keep original identifiers and update annotations on demand.

Level of integration

The integration of resources can be handled in manifold ways. The extremes poles are simple hyperlinks connecting resources versus full integration in a data warehouse structure [42]. BioMart [43] is an example of a data warehouse system. Castro *et al*. [44] and Smedley *et al*. [45] assess several approaches for integration of functional genomics data. Goble *et al*. [46] provide a description of the different levels of integration and locate mashups at the lightweight integration extreme. Mashups [47] are aggregations of multiple web services, e.g. an aggregation of microarray data with interactions from the human immunodeficiency virus type 1 (HIV-1), Human Protein Interaction Database for investigation of the HIV-1 [48]. Workflows [49, 50] can be regarded as a dedicated type of mashup. These techniques depend on the availability of web services, which are rapidly spreading and can be retrieved from collections like BioCatalogue [51]. Semantic web technologies have gained importance during the last years and have been employed for pilot projects, e.g. LabKey and SysMO-SEEK. Semantic web uses knowledge representation to achieve an improved exploitation of web resources [52, 53]. Semantic web concepts are Resource Description Framework (RDF) for simple descriptions of resources and relationships between them and Web ontology language for a language about ontologies (see 'Level of Integration' section). Applications of these concepts in systems biology include the Systems Biology Ontology [54] which adds semantics to models allowing, e.g. to give understandable names to reaction rate equations—and the tool RightField [55] employing ontologies to enhance annotation of data. RDFs emerge as an appropriate method to improve metadata. An example for integration of bioinformatics data from various databases via these semantic web technologies is Bio2RDF [56].

User commitment

The basic challenge is to convince the participants on all levels (PI to technician) of the necessity to share experimental data and reliably use data management systems. There is variation in the sharing-culture in dedicated scientific disciplines. One exceptional example of outstanding community spirit are the human genome project's Bermuda principles with the requirement to upload sequences to a public database within 24 h [57] but usually the willingness to share systems biology data is not that high (see 'Policies' section). Swan *et al*. [58] studied reasons why researchers do not want to share and give a detailed description that emphasizes lack of resources and lack of expertise as the main reasons why data are not shared easily. Thus, one major reason for the reluctance to share data before publication is that researchers want to prevent competitors from anticipating publications based upon their data. To leverage user commitment to share unpublished data, introduction of a rewarding system by giving citation credits to unpublished data in public databases in form of digital object identifiers (DOIs) has been proposed [59, 60]. DOIs are issued by the International DOI Foundation and are already in use for conventional citations. An interesting approach toward data citation is the signaling gateway molecule pages (SGMP)—a database providing structured data about signaling proteins [61]. SGMP does not implement DOI services directly for data objects but for review articles about signaling proteins. Now DOIs have been used for giving credit to data producers, e.g. [62, 63]. With the goal of establishing a DOI infrastructure for data in 2009, the international 'DataCite' organization [64] was founded. DataCite DOIs can be minted via DataCite's MetadataStore Application interface (API) after the associated data center, institutional repository or supplementary data archive has been registered with the MetadataStore service. The DOI minting procedure will be best integrated into data publishing software in order to keep references to data up-to-date and thus will be useful functionality for systems biology data management. A convincing argument for sharing published data is the increased citation rate as a consequence of sharing [65]. Another obstacle is the sometimes-large effort of converting data into a shareable format and the investment required for long-term preservation. A JISC report investigates the costs of preservation of research data in detail [66]. The reluctance to share

might be counteracted by covering the entire data management infrastructure with appropriate funding and by intelligent software systems alleviating the task of making data shareable. User interfaces on smartphones are already employed for clinical data management [67] and might help in making the use of systems biology data management systems more attractive and efficient.

SURVEY OF SYSTEMS BIOLOGY DATA MANAGEMENT SYSTEMS

Data management systems provide the core functionality to collect data, integrate it with other data and disseminate it. There is much variation in how these tasks are accomplished, e.g. automation of data collection, fine-grained access management, support for standard formats and submission to public repositories. For this review, we selected open-source systems biology data management systems proved to be employed successfully in large-scale projects.

SysMO-SEEK

SysMO-SEEK [68] was developed for a large-scale transnational research initiative on the systems biology of microorganisms (SysMO). Eleven projects contributed to the initiative, most of them having their own data management solution (see Figure 1). Main components are an assets catalogue and yellow pages which provide social network functionality leveraging the exchange of expertise. The JERM determines metadata based on the Investigation, Study, Assay (ISA) format [69] and compliant to MIBBI [28]. JERM enables useful comparability by a minimal compromise of metadata schemes differing between projects. JERM templates exist for the most common bioinformatics data types facilitating data exchange and data depositing in the relevant public repositories, e.g. ArrayExpress. The tool RightField alleviates the task of metadata generation using ontology annotations in spread sheets [55]. SEEK can be easily integrated into a heterogeneous landscape of data management systems because its software architecture with web interfaces ensures optimal extensibility. Projects registered to the SEEK are not forced to change their existing data management solutions. Harvesters can automatically collect assets held at distributed project sites and return them to the SEEK interface [68], where they are interpreted by extractors. SEEK is connected to many relevant external resources, modeling via JWS-Online [27] is directly integrated and a plug-in to PubMed enables linking of publications to supporting data in SEEK. SEEK alleviates the publication process by providing data for supplementaries. Data management of models is further supported by the capability to link models to data and vice versa simulated data to models. Experimental data can be compared to results from models via combined plots. Besides the SysMO project SEEK is employed in several other large-scale systems biology projects, e.g. ERASysBio+ and the Virtual Liver network. SysMO-SEEK can be installed quite comfortable via a virtual machine image.

DIPSBC

The Data Integration Platform for Systems Biology Cooperations (DIPSBC) [70] is based on the Solr/Lucene search server and on a wiki system that are brought together via a Solr search plug-in into the wiki system. The central idea is to construct a system around a search engine providing efficient access to project data and other integrated data sets. Search results are interpreted depending on the dedicated data types. Solr requires xml and thus ensures a systematic metadata model. The system is flexible. New data types can be integrated straightforwardly by writing the corresponding handlers. Most standard data types already use xml. Xml files have to be indexed before they can be searched efficiently using a syntax similar to most popular search engines. DIPSBC has been successfully used in many systems biology projects integrating diverse data types, e.g. several 'omics' data types and computational models. Although the main installation now contains about 35 Mio entries, response times are usually <1 s. Xml schema definitions exist for many data types, e.g. for next-generation sequencing data. Figure 2 shows a system chart of DIPSBC. Handlers for the import of the MiniML format from the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) repository already exist and might be easily rewritten for upload to public repositories. The installation of DIPSBC is straightforward for a bioinformatician and is well documented.

openBIS

openBIS [71] is an open-source distributed data management system for biologic information developed at the ETH Zürich. The central components of openBIS are an application server and a data store server. The application server stores metadata in a

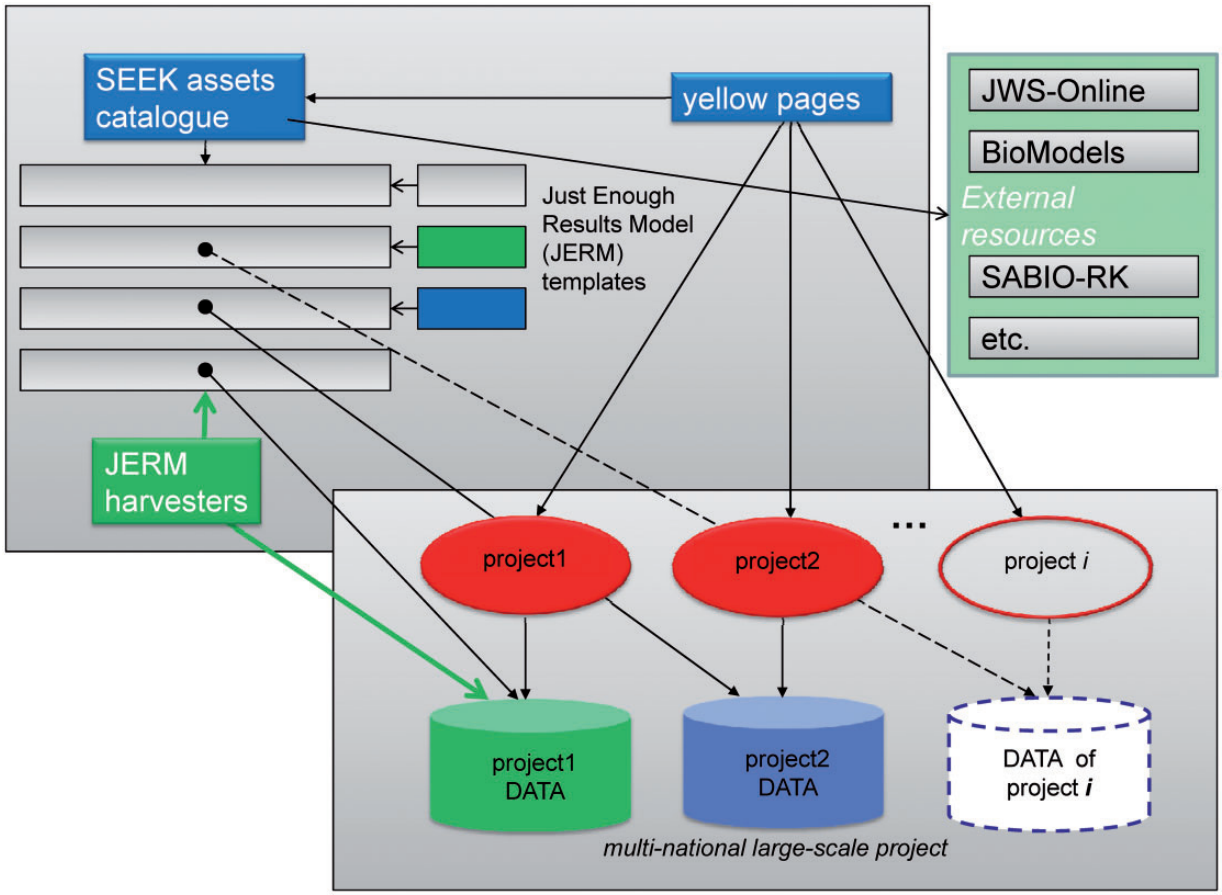

Figure 1: SysMO-DB system chart: SysMO-DB was developed for a multi-national large-scale project consisting of multiple 'sub'-projects with own data management solutions which were not changed. For that purpose the Just Enough Results Model (JERM) was introduced which aims at finding minimal information to make data comparable across project borders. JERM templates cater for compliance to MIBBI. Data of multiple projects are brought together via upload to the assets catalogue which can be performed automatically using so-called JERM 'harvesters'. The yellow pages component provides details about projects, participating people and institutions to enable exchange of expertise and association of assets to people. Many external resources are connected to SysMO-DB, e.g. integration of JWS-Online provides systems biological modelling facilities for project data.

dedicated database and interfaces to the user web browsers. openBIS employs a generic metadata model based on controlled vocabularies. The Data Store server handles high-volume data—e.g. next-generation sequencing data—in a separate database and collects new data items via drop boxes (directories monitored for incoming data) or via a remote Java API using the secure HTTPS protocol. The web interface provides visualization and retrieval functionality. openBIS is designed to allow extensions and integration of workflows with minimal effort and provides interfaces for the integration of relevant community tools. The workflow system P-Grade [72] has been integrated for proteomics data and it is planned to integrate the Galaxy workflow system [50] for analysis of next-generation sequencing data. openBIS is successfully employed in the large-scale systems biologic project SystemsX (250 research groups in 62 approved projects) and several EU projects. openBIS can be installed with some IT expertise. A demo system is online and an installation guide and comprehensive documentations are available.

XperimentR

XperimentR [73] developed at the Imperial College in London—is not distributed as software but instead provides all its features via the web with the intention to free the user from installation tasks. The system integrates three specialized data management systems: BASE for microarrays [26], OMERO [74] for microscopy imaging data and Metabolomixed—based

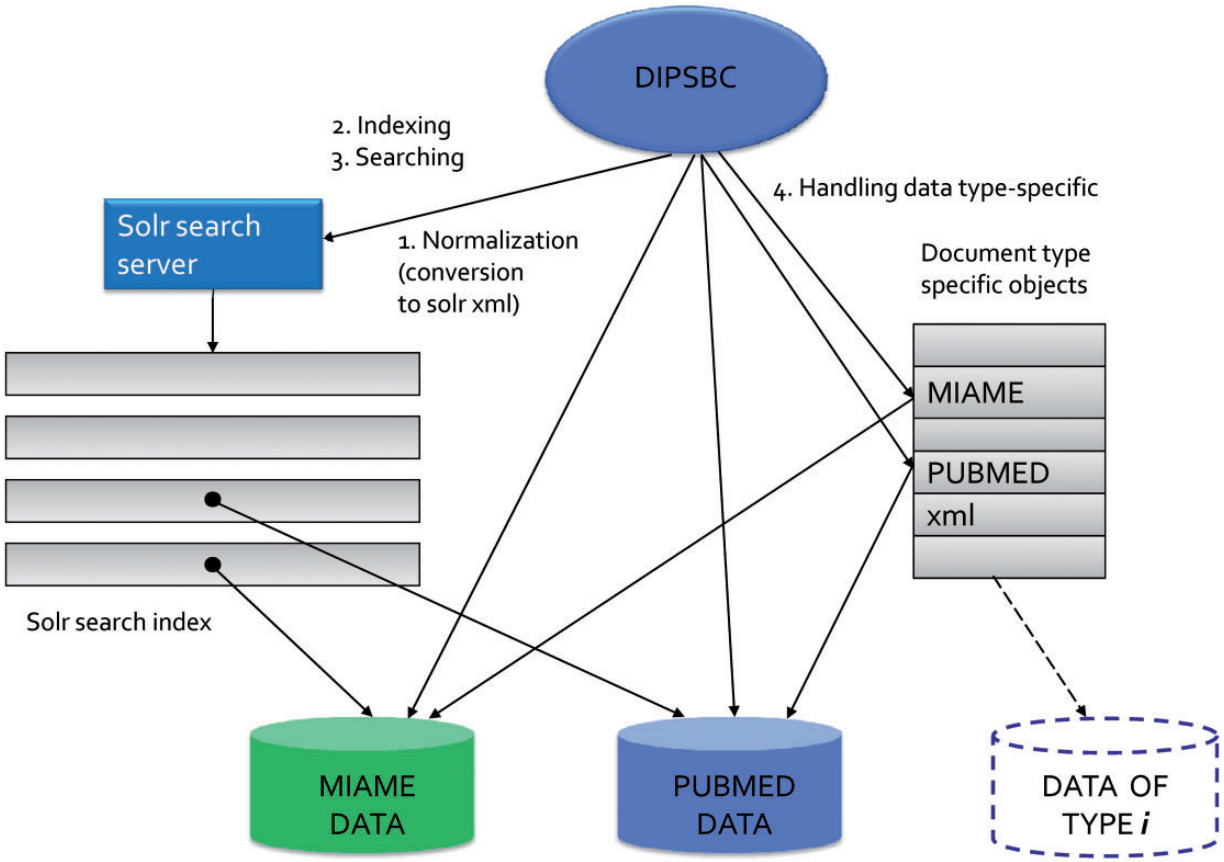

Figure 2: DIPSBC system chart. Data are first converted to the Solr xml format ('normalized') and afterward indexed by the Solr search server. Then data sets can be found efficiently and will be passed to document type specific objects which initiate processes corresponding to the dedicated data type (here MIAME data and PubMed data are shown). New data types can be introduced straightforwardly by adapting new objects derived from existing ones, e.g. the xml object.

on the open-source system omixed [75]—for metabolomics and proteomics data. XperimentR lets users describe experimental metadata via a graphical representation based on the ISA-Tab structure. As a side-effect, data can be easily exported to the ISA-Tab format. Annotation of experiments is assisted by an integrated ontology lookup service. XperimentR's web approach is only feasible for smaller projects, for large-scale projects a distribution of the software would be desirable. However, the components BASE, OMERO and omixed are distributed as stand-alone versions—but without XperimentR functionality.

Gaggle and the Bioinformatics Resource Manager

Gaggle [76] is an open-source Java software environment developed at the Institute for Systems Biology. Gaggle uses only four data types (list of names, matrices, networks and associative arrays) to integrate various data resources and software. To integrate applications, the Gaggle boss relays messages of these basic data types between 'geese' (software adapted to Gaggle). Current geese include a data matrix viewer, Cytoscape [77], the TIGR microarray experiment viewer, a R/Bioconductor goose, a simple Bioinformatics web browser and Bioinformatics Resource Manager (BRM) which adds data management functionality to Gaggle. Gaggle is easily extendible to new applications via the message interface based on the four data types. Gaggle is now employed in the integration of databases over the Internet in the US Department of Energy's effort to build up a systems biology knowledgebase [78].

The BRM [79] is a freely distributed data management system connected to Gaggle. The system is designed as client server architecture containing a PostgreSQL database storing project data, external data and metadata, a server providing access to the database and a Java front-end to the server to let the

user manage projects and analyze data. BRM is integrating user data with functional annotation and interaction data from public resources. Heterogeneous data are integrated via overlapping column values from spread sheet formats which are employed to import high-throughput data. Built-in tools allow converting data between multiple formats. BRM can export data in simple file formats and in Cytoscape. BRM can be installed straightforwardly assisted by a quick-start guide.

MIMAS

MIMAS [5] is an open-source data management system for multi-omics. Together with openBIS MIMAS is pioneer in integrating next-generation sequencing techniques. The system was originally conceived for microarrays in a MIAME compliant fashion but has been extended to sequencing techniques using MGED ontologies for annotation—in the lack of a broadly accepted standard for sequencing data. MIMAS is implemented in Perl as a web application for the Apache server and is portable between MySQL and Oracle databases. MIMAS is used in about 50 academic laboratories in Switzerland and France. Upload of data in MAGE-TAB format to European Bioinformatics Institute's (EBI) ArrayExpress repository is alleviated with a one-step upload procedure. IT expertise is needed to install MIMAS, a demo system is accessible via a guest login.

ISA software suite

The open-source ISA software suite [80] provides Java applications managing data referring to the ISA-Tab format. ISA-tab was specified in a backward compatible fashion based on MAGE-TAB [81]—the data format employed at the EBI repository Array-Express for depositing microarray data. Thus, the produced formats can be directly deposited at Array-Express. ISA is an acronym for Investigation, Study and Assay—three metadata categories constituting a hierarchical structure for the description of experiments including associated measurements (assays) and the experimental context. The BioInvestigation Index (BII) Component Manager as part of the suite administrates ISA-Tab metadata in a relational database. Other tools include the ISAconfigurator which enables customization of the editor tool ISAcreator which in turn can cater for metadata conformance to the MIBBI checklists. Access to the ISA-Tab data sets can be granted to dedicated users and distinguished as public or private. BII Manager facilitates straightforward export of data to community databases. The ISA software suite has been employed in several projects [69] and integrated into other data management systems, e.g. the Harvard stem cell discovery engine [82]. Installation of the Java applications of the software suite is simple.

BASE

BASE [26] was developed at the Lund University for microarray data management. The multi-user open-source system uses a MySQL or PostgreSQL relational database, an integrated File transfer protocol server for batch data transfer and a Tomcat servlet attached to a web server for dissemination of data. A laboratory information management system comes with BASE and thus a complete capture of all data concerning the experiment is enabled—needed for compliance with the MIAME standard. BASE supports the most popular microarray platforms and provides data analysis functionality amended by a plug-in system which also allows straightforward integration of other systems—like in 'XperimentR' section. BASE has been employed in many projects, e.g. [83]. Plug-ins provide import/export functionality for the MAGE-TAB format for data deposit in EBIs ArrayExpress repository. BASE can be installed with some IT expertise.

LabKey

LabKey server is an open-source data management system developed by LabKey Software with the focus on specimen management [84]. The LabKey instance Atlas is applied in a large-scale HIV project consisting of many consortia. LabKey provides a web server via the Tomcat servlet which is accessing a PostgreSQL or Microsoft SQLServer database. Access control is role-based and allows users to keep data private or to share with a wider community. Data are integrated using semantic web methods like RDF describing the connection of uniform resource identifiers (URIs) in combination with SQL functionality. Annotations ensure compliance of data descriptions to ontologies. Data can be integrated via view summaries based on cross-reference identifiers shared by multiple tables. Experiments can be described via general templates for assays or via specialized types which are provided for many objects including neutralizing antibody or microarrays. Extensions to new data types are facilitated by RDFs appropriateness to describe metadata. LabKey can

dynamically access external resources so that modifications in the external data sets are immediately visible.

Comparison of systems

All systems compared here have proved to be useful in large-scale projects. Figure 3 shows their performance concerning criteria corresponding to the requirements from Chapter 3. Compliance to community standards in different 'omics' areas is a central criterion which is largely covered by the MIBBI [28] and MIAME [29] recommendations.

The metadata model is often related to the standard formats used. SysMO-SEEK provides the most elaborated approach for automated data collection: harvesters are automatically looking for new data and feeding it to the system. Semi-automated approaches are dropboxes of OpenBIS and batch import facilities in most other systems.

Fine-grained access control is indispensable when multiple groups and consortia are working with the same data management system. Researchers thus can control if they want to keep their data private or with whom they want to share it. Most systems allow granting read/write permissions to users, groups and the general public. Nearly all systems' functionality can be increased via extensions which sometimes can be taken from a pool of existing software. In large-scale projects, often heterogeneous data management solutions have to be brought together. SysMO-SEEK was developed to be integrated with other data management systems via a software architecture using web interfaces for the adaptation of external systems. Also, systems like ISA software suite [69], BASE (e.g. with XperimentR), DIPSBC or Gaggle (e.g. with BRM) have been integrated with other data management solutions. However, integrating these systems or other systems with a landscape of heterogeneous data management systems in a large-scale project will not come without effort.

Support of upload to public repositories is invaluable to alleviate data sharing. SysMO-SEEK supports upload to repositories like JWS-Online, ArrayExpress, NCBI GEO. ISA software suite provides the framework for ISA-Tab also used by XperimentR and directly connected to EBI Array-Express upload. BASE, DIPSBC, Gaggle and MIMAS have implemented functionality concerning upload to NCBI GEO or Array-Express. As the DOI infrastructure for data is just in the process of being established, none of the systems compared here have special support for data DOIs, e.g. using DataCite's API for minting DOIs and notifications of URL changes. However, DOIs can be minted external to the data management systems and refer to data objects inside them, e.g. to persistent URIs as provided by the SEEK.

Modeling support distinguishes dedicated systems biology data management systems from general purpose bioinformatics systems which nevertheless might cover most features needed for systems biology data management. While DIPSBC has the BioModels [20] database indexed by its search engine and the potential of Gaggle to integrate modeling is demonstrated in preliminary versions of the Systems biology knowledgebase Kbase framework [78]—SysMO-SEEK provides the highest level of modeling support in the systems compared: integration of JWS-Online, functionality to link data and models and to plot experimental and simulated data coherently. Currently there is no system providing embargo periods—but SysMO-SEEK provides services to publish data at the end of projects. Large next-generation sequencing data are currently managed in openBIS and MIMAS. SysMO-SEEK provides an archiving solution for large data. LabKey server and Gaggle are currently offered for cloud computing but might be further improved to exploit the full potential, e.g. via MapReduce [85] pipelines. However, all systems are scalable and no great obstacles can be expected getting them on the cloud. While most systems access external resources via web services or URLs, the BRM uses data warehousing to integrate them. Additional analysis and visualization functionality with respect to different 'omics' can be a useful option and is adequately realized by BASE for transcriptomics/microarrays and XperimentR bundling BASE, Metabolomixed and OMERO for transcriptomics, metabolomics and proteomics.

CONCLUSIONS

Initiatives to foster open accessibility of publicly funded research data are now put into practice by compilation of data sharing policies and mandatory professional data management strategies in application calls. Often the willingness to share biological data is limited and it is difficult to convince researchers to share raw data even after publication. Approaches based on voluntary sharing intend to

System	1. compliance to standards/ checklists	2. meta-data model	3. automation of data collection	4. modelling support	5. upload to public repositories	6. extensibility to new functionality	7. integration with other DM systems	I. documentation	II. URL
SysMO-SEEK	MIBBI	JERM	harvesters	JWS-online for modeling Integration of data and models	JWS-online Array-Express GEO	ruby classes	good, planned in SW design	videos, presentations installation guide	www.sysmo-db.org
DIPSBC	MIAME, SBML, mzML	xml	scripting	BioModels in search engine index	MiniML datatype (GEO)	perl modules	integration in search engine index	tutorial, installation guide	dipsbc.molgen.mpg.de
openBIS	mzXml (MIBBI projected)	Generic/ controlled vocabularies	dropboxes	-	-	loosely coupled design enabling java extensions	via interfaces to the openBIS framework	user, installation, advanced guides, videos	www.cisd.ethz.ch/software/openBIS
Gaggle / BRM	NCBI GEO SOFT	EMI-ML metadata in xml	batch import	modelling integrated in gaggle-based kbase framework	GEO	java classes	via message interface	website, BRM manual	gaggle.systemsbiology.net www.sysbio.org/dataresources/brm.stm
MIMAS	MIAME	MIAME-related	batch import	-	one-step upload to Array-Express	perl modules	non-trivial	user manual	http://mimas.vital-it.ch
XperimentR	MIAME	ISA-Tab-related	batch import	-	via ISA-Tab	no, only accessible via web	via XperimentR java servlet	user guide	www3.imperial.ac.uk/bioinfsupport/resources/data_management/xperimentr
ISA tools	MIBBI	ISA-Tab-related	batch import	-	Array-Express	java classes	reference integrations	user manual	www.isa-tools.org
BASE	MIAME	MIAME-related	batch import	-	Array-Express	java plug-ins/extensions	via java servlets	user and developer manuals	base.thep.lu.se
LabKey	MIAME (only import)	semantic web (RDF)	batch import	-	-	via client API	non-trivial	installation guide, tutorials, videos	www.labkey.org

Figure 3: Benchmarking: data management systems for large-scale systems biology projects.

reward researchers with publication-like citations for publishing data or by providing attractive data management systems supporting their work. Making data citable via DOIs is promoted by the DataCite organization and now has been used in a few initial cases. Beyond voluntariness—making data management and sharing a condition of funding, using data management plans and controlling these—are efficient mechanisms to get publicly funded data shared and are on the road map in some countries. These actions will have to be accompanied by appropriate funding for data management infrastructure—including training of researchers and development of appropriate systems. All data management systems reviewed here have particular strengths coming from the dedicated contexts; they have been developed in DIPSBC, the flexibility to easily add new data types and the large collection of integrated types; ISA-Tab, the direct connection to the EBI ArrayExpress repository and the good metadata annotation facilities proved in many projects; XperimentR, the good microscopy, metabolomics, microarray and annotation facilities; Gaggle, the good extensibility and the collection of interesting extensions already existing; openBIS and MIMAS, the next-generation sequencing integration; BASE, the good microarray management and extension system proved in many projects and LabKey, the specimen management and semantic web features. However, SysMO-SEEK has clear advantages as an out-of-the box solution for large-scale systems biology projects because it can efficiently integrate

a heterogeneous landscape of multiple data management systems, is MIBBI-compliant, provides the most elaborate modeling support, fine-grained access control, automatic data harvesting, assists metadata annotation and supports relevant public repositories.

The systems reviewed here provide a useful fundament for an efficient data management infrastructure but might be further advanced by adding or optimizing these features: automatic data collection, support for conversion to standard formats, upload to public repositories, preparation of publication supplementaries and data publications and applying techniques like data warehouses and semantic web for a better exploitation of inherent knowledge and synergies. Then they will contribute to make sharing more attractive by alleviating researchers' work.

Key Points

- Incentives to share data can be given by data citation credits (datacite).
- Open access to research data can be advanced via making sharing a condition of funding.
- Data management systems might be made attractive via alleviating and improving researchers' work, e.g. support for standard formats and publications.
- The reviewed systems proved useful for systems biology projects at least in dedicated environments, SysMO-SEEK out-of-the-box provides most useful features for large-scale systems biology projects.

FUNDING

The authors acknowledge support from the German Federal Ministry of Education and Research (BMBF GRANT 0315717A), which is a partner of the ERASysBio+ initiative supported under the EU ERA-NET Plus scheme in FP7.

REFERENCES

1. *Committee on Issues in the Transborder Flow of Scientific Data, National Research Council: Bits of Power: Issues in Global Access to Scientific Data*. Washington, D.C.: The National Academies Press, 1997.
2. Organization for Economic Co-operation and Development (OECD). *Principles and Guidelines for Access to Research Data from Public Funding. OECD Publications 2007*.
3. *Wellcome Trust will Penalise Scientists Who Don't Embrace Open Access. The Guardian 2012*. http://www.guardian.co.uk/science/2012/jun/28/wellcome-trust-scientists-open-access (24 September 2012, date last accessed).
4. Lyon L. Dealing with data: roles, rights, responsibilities and relationships. Consultancy Report 2007;54.
5. Gattiker A, Hermida L, Liechti R, *et al*. MIMAS 3.0 is a multiomics information management and annotation system. *BMC Bioinformatics* 2009;**10**:151.
6. McAllister D, Collis A, McAllister D, Ball M. *BBSRC Data Sharing Policy*. *Nat Precedings* 2011; doi:10.1038/npre.2011.6015.1 (Advance Access publication 8 June 2011).
7. Field D, Sansone S-A, Collis A, *et al*. 'Omics data sharing. *Science* 2009;**326**:234–6.
8. Kuipers T, Hoeven JVD. Insight into digital preservation of research output in Europe. Survey Report FP7-2007-223758 PARSE.Insight. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf, 2009 (2 May 2012, date last accessed).
9. Ruusalepp R. Infrastructure planning and data curation: a comparative study of international approaches to enabling the sharing of research data. JISC report, 2008. http://www.jisc.ac.uk/media/documents/programmes/preservation/national_data_sharing_report_final.pdf (24 September 2012, date last accessed).
10. *2011 Canadian Research Data Summit Final Report Released - Research Data Strategy*. http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/news/data_summit_report.html (2 May 2012, date last accessed).
11. *Data Management Planning*. http://ands.org.au/guides/data-management-planning-awareness.html (25 April 2012, date last accessed).
12. *U.S. GAO - Climate Change Research: Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research*. http://www.gao.gov/products/GAO-07-1172 (26 April 2012, date last accessed).
13. Shublaq NW, Coveney PV. Merging genomic and phenomic data for research and clinical impact. *Stud Health Technol Inform* 2012;**174**:111–5.
14. International Council for Science. ICSU Report of the CSPR Assessment Panel on Scientific Data and Information. *International Council for Science* 2004; 26–7.
15. PMSEIC Working Group on Data for Science: FROM DATA TO WISDOM: Pathways to Successful Data Management for Australian Science, Report to PMSEIC, 2006.
16. Ball A. *Review of the State of the Art of the Digital Curation of Research Data,* 2010, 26. Project Report. Bath: University of Bath, (ERIM Project Document erim1rep091103ab11). http://opus.bath.ac.uk/18774/2/erim1rep091103ab11.pdf (24 September 2012, date last accessed).
17. *Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0*. http://www.openarchives.org/OAI/openarchivesprotocol.html (10 May 2012, date last accessed).
18. Subramaniam S, Fahy E, Gupta S, *et al*. Bioinformatics and systems biology of the lipidome. *Chem Rev* 2011;**111**: 6452–90.
19. Beagrie N. Digital curation for science, digital libraries, and individuals. *Int J Digit Curation* 2008;**1**:3–16.
20. Li C, Donizelli M, Rodriguez N, *et al*. BioModels Database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol* 2010;**4**:92.
21. Clarke L, Zheng-Bradley X, Smith R, *et al*. The 1000 Genomes Project: data management and community access. *Nat Methods* 2012;**9**:459–62.

22. Schadt EE, Linderman MD, Sorenson J, *et al*. Computational solutions to large-scale data management and analysis. *Nat Rev Genet* 2010;**11**:647–57.
23. ERASysBio | *Home*. http://www.erasysbio.net/ (3 May 2012, date last accessed).
24. *SystemsX.ch – The Swiss Initiative in Systems Biology - What is SystemsX.ch?* http://www.systemsx.ch/about-us/what-is-systemsxch/ (3 May 2012, date last accessed).
25. Hobbie KA, Peterson ES, Barton ML, *et al*. Integration of data systems and technology improves research and collaboration for a superfund research center. *J Lab Autom* 2012;**17**: 275–83.
26. Vallon-Christersson J, Nordborg N, Svensson M, Hakkinen J. BASE - 2nd generation software for microarray data management and analysis. *BMC Bioinformatics* 2009;**10**:330.
27. Olivier BG, Snoep JL. Web-based kinetic modelling using JWS Online. *Bioinformatics* 2004;**20**:2143–4.
28. Taylor CF, Field D, Sansone S-A, *et al*. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotech* 2008;**26**: 889–96.
29. Brazma A, Hingamp P, Quackenbush J, *et al*. Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nat Genet* 2001;**29**: 365–71.
30. Kerrien S, Orchard S, Montecchi-Palazzi L, *et al*. Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.* 2007;**5**:44.
31. Martens L, Chambers M, Sturm M, *et al*. mzML–a community standard for mass spectrometry data. *Mol Cell Proteomics* 2011;**10**: r110.000133.
32. Demir E, Cary MP, Paley S, *et al*. The BioPAX community standard for pathway data sharing. *Nat Biotechnol* 2010;**28**: 935–42.
33. Hucka M, Finney A, Sauro HM, *et al*. The Systems Biology Markup Language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003;**19**:524–31.
34. Lloyd CM, Halstead MDB, Nielsen PF. CellML: its future, present and past. *Prog Biophys Mol Biol* 2004;**85**:433–50.
35. Kikuchi N, Kameyama A, Nakaya S, *et al*. The carbohydrate sequence markup language (CabosML): an XML description of carbohydrate structures. *Bioinformatics* 2005; **21**:1717–8.
36. Novère NL, Hucka M, Mi H, *et al*. The systems biology graphical notation. *Nat Biotechnol* 2009;**27**:735–41.
37. Adamusiak T, Parkinson H, Muilu J, *et al*. Observ-OM and Observ-TAB: universal syntax solutions for the integration, search, and exchange of phenotype and genotype information. *Hum Mutat* 2012;**33**:867–73.
38. Brazma A, Krestyaninova M, Sarkans U. Standards for systems biology. *Nat Rev Genet* 2006;**7**:593–605.
39. Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**: 1251–5.
40. Ashburner M, Ball CA, Blake JA, *et al*. Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.
41. *MGED NETWORK: Ontology Working Group (OWG)*. http://mged.sourceforge.net/ontologies/index.php (27 April 2012, date last accessed).
42. Triplet T, Butler G. Systems biology warehousing: challenges and strategies toward effective data integration. In: *DBKDA 2011, The Third International Conference on Advances in Databases, Knowledge, and Data Applications* 2011, pp. 34–40.
43. Kasprzyk A. BioMart: driving a paradigm change in biological data management. *Database* 2011;**2011**.
44. Garcia Castro A, Chen Y-PP, Ragan MA. Information integration in molecular bioscience. *Appl Bioinformatics* 2005; **4**:157–73.
45. Smedley D, Swertz MA, Wolstencroft K, *et al*. Solutions for data integration in functional genomics: a critical assessment and case study. *Brief Bioinformatics* 2008;**9**:532–44.
46. Goble C, Stevens R. State of the nation in data integration for bioinformatics. *J Biomed Inform* 2008;**41**:687–93.
47. Butler D. Mashups mix data into global service. *Nature* 2006;**439**:6–7.
48. Nolin M-A, Dumontier M, Belleau F, Corbeil J. Building an HIV data mashup using Bio2RDF. *Brief Bioinform* 2012; **13**:98–106.
49. Hull D, Wolstencroft K, Stevens R, *et al*. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* 2006;**34**:W729–32.
50. Giardine B, Riemer C, Hardison RC, *et al*. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 2005;**15**:1451–5.
51. Bhagat J, Tanoh F, Nzuobontane E, *et al*. BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res* 2010;**38**:W689–94.
52. Antezana E, Kuiper M, Mironov V. Biological knowledge management: the emerging role of the Semantic Web technologies. *Brief Bioinformatics* 2009;**10**:392–407.
53. Stephens S, LaVigna D, DiLascio M, *et al*. Aggregation of bioinformatics data using Semantic Web technology. *Web Semant*. 2006;**4**:216–21.
54. Courtot M, Juty N, Knüpfer C, *et al*. Controlled vocabularies and semantics in systems biology. *Mol Syst Biol* 2011;**7**:543.
55. Wolstencroft K, Owen S, Horridge M, *et al*. RightField: embedding ontology annotation in spreadsheets. *Bioinformatics* 2011;**27**:2021–2.
56. Belleau F, Nolin M-A, Tourigny N, *et al*. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008;**41**:706–16.
57. Marshall E. Bermuda rules: community spirit, with teeth. *Science* 2001;**291**:1192.
58. Swan A, Brown S. To share or not to share: publication and quality assurance of research data outputs. Report commissioned by the Research Information Network (RIN). Annex: detailed findings for the eight research areas (June 2008), 2008;26–9.
59. Data producers deserve citation credit. *Nat Genet* 2009;**41**: 1045.
60. Editors. Credit where credit is overdue. *Nat Biotech* 2009;**27**: 579.
61. Dinasarapu AR, Saunders B, Ozerlat I, *et al*. Signaling gateway molecule pages–a data model perspective. *Bioinformatics* 2011;**27**:1736–38.
62. Yan G, Zhang G, Fang X, *et al*. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nat Biotechnol*. 2011;**29**:1019–1023.

63. Derradji-Aouat A. *Numerical Work for Oceanide Project*. A Progress Report.
64. Brase J. DataCite - a global registration agency for research data. In: *Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 2009.* COINFO '09. IEEE, 257–61.
65. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2007;**2**:e308.
66. *Keeping Research Data Safe (Phase 1)*. http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx (3 May 2012, date last accessed).
67. Xie F, Zhang D, Wu J, *et al*. Design and implementation of the first nationwide, web-based Chinese Renal Data System (CNRDS). *BMC Med Inform Decis Mak* 2012;**12**:11.
68. Wolstencroft K, Owen S, du Preez F, *et al*. The SEEK: a platform for sharing data and models in systems biology. *Meth Enzymol* 2011;**500**:629–55.
69. Sansone S-A, Rocca-Serra P, Field D, *et al*. Toward interoperable bioscience data. *Nat Genet* 2012;**44**:121–6.
70. Dreher F, Kreitler T, Hardt C, *et al*. DIPSBC - data integration platform for systems biology collaborations. *BMC Bioinformatics* 2012;**13**:85.
71. Bauch A, Adamczyk I, Buczek P, *et al*. openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics* 2011;**12**:468.
72. Kacsuk P, Sipos G. Multi-grid, multi-user workflows in the P-GRADE grid portal. *J Grid Comput* 2005;**3**:221–38.
73. *XperimentR*. http://www3.imperial.ac.uk/bioinfsupport/resources/data_management/xperimentr (14 October 2011, date last accessed).
74. Allan C, Burel J-M, Moore J, *et al*. OMERO: flexible, model-driven data management for experimental biology. *Nat Methods* 2012;**9**:245–53.
75. *Omixed (Scientific Data Organised Via Web Services) Home*. http://www.omixed.org/ (4 May 2012, date last accessed).
76. Shannon P, Reiss D, Bonneau R, Baliga N. The Gaggle: an open-source software system for integrating bioinformatics software and data sources. *BMC Bioinformatics* 2006; **7**:176.
77. Shannon P, Markiel A, Ozier O, *et al*. A software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
78. *Genomic Science Program Systems Biology Knowledgebase Projects*. http://genomicscience.energy.gov/compbio/kbaseprojects.shtml (3 May 2012, date last accessed).
79. Shah AR, Singhal M, Klicker KR, *et al*. Enabling high-throughput data management for systems biology: The Bioinformatics Resource Manager. *Bioinformatics* 2007;**23**: 906–9.
80. Rocca-Serra P, Brandizi M, Maguire E, *et al*. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* 2010;**26**:2354–6.
81. Rayner T, Rocca-Serra P, Spellman P, *et al*. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics* 2006;**7**:489.
82. Ho Sui SJ, Begley K, Reilly D, *et al*. The Stem Cell Discovery Engine: an integrated repository and analysis system for cancer stem cell comparisons. *Nucleic Acids Res* 2011;**40**:D984–91.
83. Olsson E, Honeth G, Bendahl P-O, *et al*. CD44 isoforms are heterogeneously expressed in breast cancer and correlate with tumor subtypes and cancer stem cell markers. *BMC Cancer* 2011;**11**:418.
84. Nelson E, Piehler B, Eckels J, *et al*. LabKey Server: an open source platform for scientific data integration, analysis and collaboration. *BMC Bioinformatics* 2011;**12**:71.
85. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM* 2008;**51**: 107–113.