*Database*, Vol. 2011, Article ID baq027, doi:10.1093/database/baq027

# Editorial

# Towards BioDBcore: a community-defined information specification for biological databases

**Pascale Gaudet[1,2,*], Amos Bairoch[1], Dawn Field[3], Susanna-Assunta Sansone[4], Chris Taylor[5], Teresa K. Attwood[6,7], Alex Bateman[8], Judith A. Blake[9], Carol J. Bult[9], J. Michael Cherry[10], Rex L. Chisholm[2], Guy Cochrane[5], Charles E. Cook[4], Janan T. Eppig[9], Michael Y. Galperin[11], Robert Gentleman[12], Carole A. Goble[7], Takashi Gojobori[13,14], John M. Hancock[15], Douglas G. Howe[16], Tadashi Imanishi[13], Janet Kelso[17], David Landsman[18], Suzanna E. Lewis[19], Ilene Karsch Mizrachi[11], Sandra Orchard[5], B.F. Francis Ouellette[20], Shoba Ranganathan[21,22], Lorna Richardson[23], Philippe Rocca-Serra[4], Paul N. Schofield[24], Damian Smedley[5], Christopher Southan[25], Tin W. Tan[22], Tatiana Tatusova[11], Patricia L. Whetzel[26], Owen White[27], Chisato Yamasaki[14] and on behalf of the BioDBCore working group**

[1]Swiss Institute of Bioinformatics, CMU, 1 Rue Michel Servet, 1211 Geneva 4, Switzerland, [2]Feinberg School of Medicine, Northwestern University, Chicago, IL, 60611, USA, [3]NERC Center for Ecology and Hydrology, Oxford, OX1 3SR UK, [4]University of Oxford, Oxford e-Research Centre, Oxford, OX1 3QG, UK, [5]European Molecular Biology Laboratory (EMBL) Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [6]Faculty of Life Sciences, The University of Manchester, Manchester M13 9PT, UK, [7]School of Computer Science, The University of Manchester, Manchester M13 9PT, UK, [8]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK, [9]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609 USA, [10]Department of Genetics, Stanford University, Stanford, CA 94305-5120, USA, [11]NCBI, NLM, National Institutes of Health, Bethesda, MD 20894, USA, [12]Genentech, 1 DNA Way, South San Francisco, CA 94080, USA, [13]Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi Koto-ku, Tokyo 135-0064, Japan, [14]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan, [15]MRC Harwell, Mammalian Genetics Unit, Harwell Science and Innovation Campus, Oxfordshire, OX11 0RD, UK, [16]The Zebrafish Model Organism Database, 5291 University of Oregon, Eugene, OR 97401-5291, USA, [17]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [18]DATABASE, The Journal of Biological Databases and Curation, Oxford University Press, Oxford OX2 6DP, UK, [19]Genomics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road 64R0121 Berkeley, CA 94720, USA, [20]Ontario Institute for Cancer Research, Suite 800, 101 College Street, Toronto, Ontario, M5G 0A3, Canada, [21]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney NSW 2109, Australia, [22]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, [23]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh, EH4 2XU, UK, [24]Dept of Physiology, Development and Neuroscience, University of Cambridge, Doning Street, Cambridge CB2 3EG, UK, [25]ChrisDS Consulting, Göteborg, Sweden, [26]Stanford Center for Biomedical Informatics Research, National Center for Biomedical Ontology, Stanford University, Stanford, CA 94305, USA and [27]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

***Corresponding author:** Tel: +41-22-379-5050; Fax: +41-22-379-5858; Email: pascale.gaudet@isb-sib.ch

Submitted 4 November 2010; Accepted 5 November 2010

This paper is also being published in *Nucleic Acids Research*, http://www.nar.oxfordjournals.org/cgi/doi/10.1093/nar/gkq1173

The present article proposes the adoption of a community-defined, uniform, generic description of the core attributes of biological databases, BioDBCore. The goals of these attributes are to provide a general overview of the database landscape, to encourage consistency and interoperability between resources; and to promote the use of semantic and syntactic standards. BioDBCore will make it easier for users to evaluate the scope and relevance of available resources. This new resource will increase the collective impact of the information present in biological databases.

*Database*, Vol. 2011, Article ID baq027, doi:10.1093/database/baq027

## Introduction

The world of public biological databases is constantly evolving, as attested by the ever-growing size of the *Nucleic Acids Research* (NAR) annual database issue and online Molecular Biology Database Collection, as well as by the creation of a new journal dedicated to databases and biocuration, *DATABASE* (1,2). A wealth of new technologies is responsible for the exponential increase in the quantity, complexity and diversity of data generated in the life sciences. The need to store and share this data helps explain the explosion in the number and variety of resources that cater to the needs of biological research. Many researchers have commented that this increased volume of data has not yet yielded proportional improvements in biological knowledge (3–5). To a great extent this is owing to the widespread and unconnected distribution of data through databases scattered around the world. Clearly, adherence to open standards, as well as powerful and reliable tools, have become a necessity to support data sharing, integration and analysis (6). The available databases can be broadly placed into three categories: (i) archival repositories; (ii) curated resources, hence the rise of biocuration described in (7), and (iii) data integration warehouses. All three offer a range of querying and mining tools to explore the data and enable knowledge discovery. In addition, databases range from well-established repositories to burgeoning, innovative resources that cover emerging scientific areas or use novel technologies. While some databases are intended as long-term, consistently maintained community resources, others are intentionally temporary in nature, their existence being limited to the lifetime of the underlying grant or research project.

As in any emerging field, standardization across the biological databases is still inadequate at many levels. Consequently, there is still unnecessary and costly duplication of efforts, poor interoperability between resources and loss of valuable data and annotations when a resource is no longer supported. Most critically, the large number and variety of resources available are major hurdles for users, who are often unable to locate the resource(s) that best fits their specific needs. Even when appropriate resources are located, combining data from different resources can be a very difficult task. Having a uniform system for describing biological databases available in a single, centralized location would benefit both users and database providers: it would be much easier for users to find appropriate resources, while publicizing specialized resources and lesser known functionality of established databases more widely.

To address some of these issues we propose the adoption of a community-defined, uniform, generic description of the 'core attributes of biological databases', which we will name BioDBCore. Such minimum information checklists are now being developed for a wide range of data types. For example, the MIBBI (Minimum Information for Biological and Biomedical Investigations) portal [http://mibbi.org; (8)] contains over 30 MI checklists. BioDBCore will contain essential descriptors common to all databases.

## Goals of the BioDBCore attributes

The goals of the proposed BioDBCore checklist are given below:

- Gather the necessary information to provide a general overview of the database landscape, and compare and contrast the various resources.
- Encourage consistency and interoperability between resources.
- Promote the uptake and use of semantic and syntactic standards.
- Provide guidance for users when evaluating the scope and relevance of a resource, as well as details of the data access methods supported.
- Ensure that the collective impact of these resources is maximized.

This working group is open to all interested parties, and has started to collect a list of attributes of the BioDBCore checklist. Proposed core attributes are presented in Table 1.

**Table 1.** Proposed core descriptors for inclusion in the BioDBCore specification

| Proposed core descriptors for a biological database |
|---|
| 1. Database name |
| 2. Main resource URL |
| 3. Contact information (E-mail; postal mail) |
| 4. Date resource established (year) |
| 5. Conditions of use (Free, or type of license) |
| 6. Scope: data types captured, curation policy, standards used |
| 7. Standards: MIs, Data formats, Terminologies |
| 8. Taxonomic coverage |
| 9. Data accessibility/output options |
| 10. Data release frequency |
| 11. Versioning policy and access to historical files |
| 12. Documentation available |
| 13. User support options |
| 14. Data submission policy |
| 15. Relevant publications |
| 16. Resource's Wikipedia URL |
| 17. Tools available |

The BioDBCore will be used to collect information about databases for use in online browsing, searching and classification. The current specification can be found as an online survey and users are encouraged to join the project and leave feedback (http://biocurator.org/biodbcore.shtml; Figure 1). Examples can be found in Table 2 and at the BioDBCore web site.

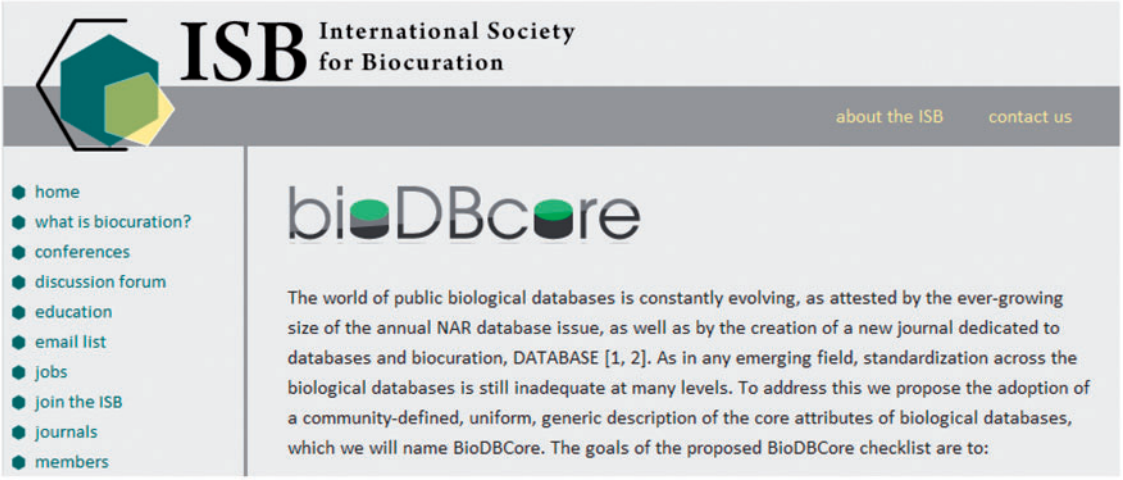


**Figure 1.** A screenshot of the BioDBCore discussion page on the ISB web site (http://biocurator.org/biodbcore.shtml).

BioDBCore is registered with MIBBI, the umbrella organization that works to promote minimal information reporting in biomedical and biological research (8).

## The BioDBCore working group

To achieve widespread uptake and adoption of the BioBDCore guidelines, these recommendations must be developed as a community effort. To get the initiative started, we have formed a working group encompassing representatives from a wide range of existing life sciences resources. This includes representatives from MIBBI, editors from key journals publishing database descriptions, staff from model organism, sequences and protein databases, members of the Asia-Pacific Bioinformatics network (APBioNet, http://www.apbionet.org/), the Bioinformatics Links Directory (http://www.bioinformatics.ca/links_directory/) (9), developers from the ELIXIR survey of European databases and leaders of the Database Description Framework (DDF) from the CASIMIR project (10). One of the working group participants, APBioNet, has developed a framework for Minimum Information about a Bioinformatics Investigation (MIABi) (11) that aims to cover all aspects of bioinformatics studies. We plan to coalesce the BioDBCore with the relevant aspects of MIABi. This is an important opportunity to build a combined framework for advancing bioinformatics standards in a coordinated manner.

The BioDBCore checklist is overseen by the International Society for Biocuration (ISB) (http://biocurator.org/), in collaboration with the BioSharing forum [http://www.biosharing.org/, (12)]. The ISB was created in 2009 to promote and support the work of biocurators and bio-programmers. One of its goals is to foster interactions between these professionals to maximize the usefulness of all resources by encouraging the interoperability of databases and supporting data sharing. The BioSharing forum works at the global level to build stable linkages between funders, implementing data-sharing policies and well-constituted standardization efforts in the biosciences domain to expedite communication and achieve harmonization and mutual support. A 'one-stop shop' portal is under development for those seeking data-sharing policy documents and information about the standards (checklists, ontologies and file-formats), linking to exiting resources, such as MIBBI.

## Participation of the biocuration community in the BioDBCore initiative

With this editorial, we announce the launch of this initiative and present for discussion an initial draft version of the specification of information to be captured. We welcome and encourage representatives of resources, included those listed in this NAR database issue, NAR Molecular Biology Database Collection (1) and the DATABASE journal to actively participate in the development of BioDBCore.

## Long-term vision and potential impact

The BioDBCore implementation will take place in three phases: (i) consultation with interested parties; (ii) collaborative development of the minimal information list. To help establish requirements, some examples can be found on the BioDBCore page of the ISB and moreover the APBioNet's BioDB100 initiative will be used to develop

*Database*, Vol. 2011, Article ID baq027, doi:10.1093/database/baq027

**Table 2.**

| | | | | | | |
|---|---|---|---|---|---|---|
| 1. Database name | dictyBase | EMAGE | Gene Ontology Database | IntAct | SGD, Saccharomyces Genome Database | MGI, Mouse Genome Informatics |
| 2. Main resource URL | http://dictybase.org | http://www.emouseatlas.org/emage | http://geneontology.org/ | http://www.wbi.ac.uk/intact | http://www.yeastgenome.org/ | http://informatics.jax.org |
| 3. Contact information | dictybase@northwestern.edu | ma-edit@hgu.mrc.ac.uk | gohelp@geneontology.org | intact-help@ebi.ac.uk | yeast-curator@yeastgenome.orog | mgi-help@informatics.jax.org |
| 4. Date resource established (year) | 2003 | 2002 | 1998 | 2003 | 1992 | 1989 |
| 5. Conditions of use | Free | Creative commons | Free | Free | Free | Free |
| 6. Scope: | | | | | | |
| Data types captured | Genome sequence; gene models including CDS and predicted proteins; phenotypes, Gene Ontology annotations, functional annotation (gene product names), gene nomenclature; strains; plasmids; free text descriptions, domains (via InterPro), orthologs (via OrthoMCL and inParanoid), protein subcellular location (via Swiss-Prot); protein existence (via Swiss-Prot), citations, researchers database | Spatially integrated *in situ* gene expression patterns in the developing mouse embryo (*in situ* hybridization, immunohistochemisty, *in situ* reporter data). Ontology based text descriptions of expression patterns. Metadata relating to the experiments. | Gene Ontology (Biological Process, Molecular Function, Cellular Component), GO annotations for proteins, functional RNAs and stable complexes. | Molecular interactions | Genome sequence; gene models including CDS and predicted proteins and non-coding RNAs; chromosomal features including telomeres, centromeres and ARS elements; mutant phenotypes; Gene Ontology annotations; gene product names; gene nomenclature; strains; plasmids; free text descriptions and literature summaries; protein domains (from InterPro); orthologs; literature citations; database of yeast researchers; functional genomics (gene expression, synthetic genetic arrays); biochemical pathways; genetic and physical interactions (from BioGRID); images of protein subcellular location (via YeastGFP); links to other tools and databases including post-translational modification databases | Genes, pseudogenes, and gene models including CDS and predicted proteins and non-coding RNAs; cytogenetic markers; genomic and genetic maps; nucleotide and protein sequence associations; spontaneous, induced, and genetically engineered alleles; transgenes; QTL; mutant and conditional phenotypes; mouse models of human disease annotations; Gene Ontology annotations; mouse anatomy, mouse phenotype ontology, gene product names; gene nomenclature; strains; SNPs; protein domains (from InterPro); mammalian orthologs; literature citations; experimental molecular reagents; functional genomics (gene expression); biochemical pathways; images of phenotypic mutants and gene expression; links to other tools and other database resources |
| Curation policy | Manual curation | Manual curation | Manual curation | Manual curation | Manual curation | Manual curation |
| Standards: MIs, Data formats, Terminologies | Gene Ontology, Dicty Anatomy Ontology, Dicty Gene Nomeclature | EMAP Mouse Anatomy Ontology, MISFISHIE, MGI (MGNC) Gene/Protein ID, MGI Mouse Strain Information, MGI Mouse Allele ID, INSDC versioned sequence ID, EMBL/PIR versioned ID, MGI probe ID. | Development of the Gene Ontology standard. | MIMIX, IMEx, Gene Ontology, MOD gene nomenclature, PSI-MI CV, PSI-MOD CV | Gene Ontology, Saccharomyces Gene Nomenclature, GenBank feature table, Sequence Ontology, ChEBI, Yeast Phenotype Ontology (YPO) | Mouse gene nomenclature, Gene Ontology, Mammalian Phenotype Ontology, Mouse Adult Anatomy |
| 7. Data formats | FASTA, OBO, GAF, GFF3 (standard) | 2D Images: jpg,gif, tiff, png, etc. (standard)—3D images: OPT (standard)—Data Domains: wlz: Probe sequence: FASTA, versioned INSDC ID (standard) | OBO v1.2, Gene Association Format (GAFs obtained via Model organism databases, UniProt-KB and other collaborators), MySQL and SQL database dumps, RDF-XML, OBO-XML, OWL | PSI-MI XML2.5, MITAB2.5 (standard) | FASTA, GenBank, GAF, GFF3 (standard) | HTMLl, tab-delimited, GFF3, images, GAF files, FASTA, XML/webservices |
| 8. Taxonomic coverage (use NCBI Taxid) | D. discoideum (44689) including all strains [PRIMARY], also some genome/EST/gene model info for D. purpureum (5786), and gene model sequences for P. pallidum (13642) and D. fasiculatum (261658) | Mus musculus (10090) | All | All | *Saccharomyces cerevisiae* (4932) | Laboratory mouse (10090) |

(Continued)

**Table 2.** Continued.

| | | | | | | |
|---|---|---|---|---|---|---|
| 9. Data accessibility/ output options | HTML, text, database reports | HTML, xml, csv, webservices, SQL, Java API, DAS | HTMLltextlXMLldatabase reportsl database dumpsl web services | PSI-MI XML2.5, MITAB2.5 | HTML, text, TAB, ASN.1, FTP, Intermine | HTML, tab-delimited, GFF3, images, GAF files, FASTA, XML/webservices, FTP, BioMart |
| 10. Data release frequency | Curators work on the 'live' database, data dumps are done weekly (sequences) or monthly (other data) | As and when available, in principle daily | Daily | Weekly | Daily | Daily |
| 11. Versioning policy/ access to historical files | No versioning but access to historical files is possible | | Versioning by date. Access to monthly releases of the full GO database going back to 2002. | Versioning by date, access to historic files available | Versioning frequency specified by datatype, database updated in real time | |
| 12. Documentation available | http://dictybase.org/FAQ/ HelpFilesIndex.html | Documentation, FAQ's, etc. found here http://genex.hgu.mrc.ac.uk/emage/help/all_help.html. Also, an information link is available on all search pages leading to a full description of the process. | http://geneontology.org/GO.contents.doc.shtml. Also, an information link is available on all AmiGO search pages leading to a full description of the interface. | www.ebi.ac.uk/intact, http://code.google.com/p/intact/ | http://www.yeastgenome.org/ aboutsgd.shtml | http://www.informatics.jax.org/ mgihome/homepages/help.shtml |
| 13. User support options | Documents, Email, web form | Documentation, FAQ's, demo movies, glossary, email, live demo at meeting exhibits, ad hoc workshops. | Written documentation on web pages, FAQ's, email helpdesk, webform, training camps. | Documents, email, webform, training | http://www.yeastgenome.org/ HelpContents.shtml http://www.openhelix.com/sgd http://www.yeastgenome.org/ help/glossary.html | Dedicated user support staff available via email, phone, customized SQL, training, tutorials, FAQs |
| 14. Data submission policy | Data from published literature. Some HTP data corresponding to published analyses is incorporated | http://www.emouseatlas.org/ emage/data_submission/all_submission_options.html | Daily updates to GAF repository from verified submitting groups (approximately 30 at present time). Submissions from other groups accepted after quality assurance agreements. | Data accepted as part of publication process, released on article publication by Journal | Data from published literature. Some HTP data corresponding to published analyses is incorporated | Data from published literature, contributed data sets. http:// www.informatics.jax.org/ submit.shtml |
| 15. Relevant publications | PMID: 18974179, PMID: 14681427 | PMID:19767607, PMID:18077470, PMID:16381949. | PMID: 10802651, PMID: 14681407, PMID: 19920128 | PMID: 19850723 | PMID:10592186, PMID:11125055, PMID:11752257, PMID:12073322, PMID:14681421, PMID:15153302, PMID:15608219, PMID:16381907, PMID:17001629, PMID:17142221, PMID:17982175, PMID:19906697, PMID:20157474, PMID:9169866, PMID:9297238, PMID:9399804, PMID:9847146, PMID:9885151 | PMID:19864252 PMID:18981050 PMID:18158299 PMID:17135206 PMID:16381933 PMID:15608240 |
| 16. Resource's Wikipedia URL | http://en.wikipedia.org/wiki/ DictyBase | | http://en.wikipedia.org/wiki/Gene_Ontology | | http://en.wikipedia.org/wiki/ Saccharomyces_Genome_ Database | |
| 17. Tools available | BLAST, BioMart, Generic Genome Browser, TextPresso, MetaCyc (dictyCyc) | LOSSST (Spatial Query Tool), Gene Query Tool, Anatomy Query Tool, GO Query Tool, 'Find Similar' Spatial Query Tool, MAPaint, Spatial Clustering Tool, Webservices, Java API, DAS Query Tool, Formatted URL Query Tool | Ontology Browseer (AmiGO), BLAST, GOTerm Finder, GOOSE (SQL query tool), GO Slimmer, Visualization, Web Services, Galaxy | | BLAST (variety of fungal genome data sets), GO Query Tools (GO Slim Mapper, GO Term Finder), GBrowse for chromosomal sequence and features, GBrowse for protein sequence features, short sequence pattern matching tool (PATMATCH), oligonucleotide primer design (webprimer), genome restriction enzyme cutting site analysis, Synteny Viewer between *S. cerevisiae* and Saccharomyces senso strico, links between P-POD (Princeton Protein Ontology Database), microarray tools (SPELL, expression connection), YeastMine (Intermine fast database searching for *S. cerevisiae* data), full-text search (Textpresso) | mouseBLAST (mouse, human, rat), Ontology Browsers, VLAD, Batch Quesy, BioMart, Gbrowse, MGI GO_slim |

*Database*, Vol. 2011, Article ID baq027, doi:10.1093/database/baq027

further working examples (11); and (iii) in the longer term, completion of stable guidelines and their implementation as a public submission web site that will allow data entry and easy update by database providers, in collaboration with the existing database collections and the BioSharing standards portal to reduce duplication of effort. Many of the members of the BioDBCore working group have experience and expertise in establishing such services.

We are aware that the adoption of this specification requires significant effort from all participating groups. However, the long-term benefits, both for the specific adopters and for the community as a whole provides considerable compensation for this effort. The complete, uniform and centralized descriptions of databases should benefit both users and data providers by providing easy access to the scope of each resource. This will be particularly valuable for specialized resources that are only used within with a restricted research community. We envisage that having such rich information readily available may facilitate collaboration between resources currently outside each other's immediate networks. We expect the BioDBCore guideline to be useful not only to users of life sciences resources, but also to drive the evolution of databases themselves. For example, the initial version of BioDBCore includes a field to describe data-submission policies. Currently, many databases do not provide such documents. We hope that by including such a field in BioDBCore, they will be encouraged to develop them. A longer term application of the information captured by BioDBCore is to allow bird's eye views of the database world to emerge by drawing connections between them into a resource network, showing the flow of data between different sites and how each complements the other.

*Conflict of interest*. None declared.

## References


1. Cochrane,G.R. and Galperin,M.Y. (2010) The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acids Res.*, **38**, D1–D4.
2. Landsman,D., Gentleman,R., Kelso,J. and Ouellette,B.F.F. (2009) DATABASE: a new forum for biological databases and curation. *DATABASE,* doi:10.1093/*bap002* (Advance access published 26 March 2009).
3. Attwood,T.K., Kell,D.B., McDermott,P. *et al*. (2009) Calling International Rescue: knowledge lost in literature and data landslide! *Biochem. J.*, **424**, 317–333.
4. Seringhaus,M.R. and Gerstein,M.B. (2007) Publishing perishing? Towards tomorrow's information architecture. *BMC Bioinform.*, **8**, 17.
5. Philippi,S. and Kohler,J. (2006) Addressing the problems with life-science databases for traditional uses and systems biology. *Nat. Rev. Genet.*, **7**, 482–488.
6. Goble,C. and Stevens,R. (2008) State of the nation in data integration for bioinformatics. *J. Biomed. Inform.*, **41**, 687–693.
7. Howe,D., Costanzo,M., Fey,P. *et al*. (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.
8. Taylor,C.F., Field,D., Sansone,S.A. *et al*. omoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.*, **26**, 889–896.
9. Brazas,M.D., Yamada,J.T., Ouellette,B.F.F. Evolution in bioinformatic resources: 2009 update on the Bioinformatics Links Directory. *Nucleic Acids Res.*, **37**, W3–W5.
10. Smedley,D., Schofield,P., Chen,C.K., *et al*. (2010) Finding and sharing: new approaches to registries of databases and services for the biomedical sciences. *DATABASE*, doi:10.1093/bap014 (Advance access published 2 July 2010).
11. Tan,T.W., Tong,J.C., De Silva,M. *et al*. (2010) Advancing standards for bioinformatics activities: persistence, reproducibility, disambiguation and Minimum Information about a Bioinformatics Investigation (MIABi). *BMC Genomics*, **11**(Suppl. 4), S27.
12. Field,D., Sansone,S.A., Collis,A. *et al*. (2009) Omics Data Sharing. *Science*, **326**, 234–236.