Published online 3 December 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D1–D4***
doi:10.1093/nar/gkp1077

The 2010 *Nucleic Acids Research* Database Issue and online Database Collection: a community of data resources

Guy R. Cochrane[1,*] and Michael Y. Galperin[2]

[1]EMBL–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received October 16, 2009; Revised November 2, 2009; Accepted November 3, 2009

ABSTRACT

The current issue of *Nucleic Acids Research* includes descriptions of 58 new and 73 updated data resources. The accompanying online Database Collection, available at http://www.oxfordjournals.org/nar/database/a/, now lists 1230 carefully selected databases covering various aspects of molecular and cell biology. While most data resource descriptions remain very brief, the issue includes several longer papers that highlight recent significant developments in such databases as Pfam, MetaCyc, UniProt, ELM and PDBe. The databases described in the Database Issue and Database Collection, however, are far more than a distinct set of resources; they form a network of connected data, concepts and shared technology. The full content of the Database Issue is available online at the *Nucleic Acids Research* web site (http://nar.oxfordjournals.org/).

COMMENTARY

This Database Issue of *Nucleic Acids Research* (NAR) includes descriptions of 58 new data resources and updates to 73 previously published data resources. The online Database Collection that accompanies the issue holds 1230 data resources, a growth of 5% over last year (http://www.oxfordjournals.org/nar/database/a/).

Continuing a decade-long tradition, the Database Issue and Database Collection serve two functions: (i) to introduce molecular and cell biologists that make up the regular readership of NAR to the databases that might be useful to them and (ii) to provide database developers a venue to publish articles to promote their resources and introduce their work to the community that might benefit from it. Based on a number of measures (such as the numbers of downloads, literature citations and web links from outside sources), the NAR Database Issue and Database Collection have been extremely successful. Despite rather strict acceptance criteria (1), the number of submitted articles greatly exceeds the capacity of a single annual issue. In order to accommodate this, Oxford University Press, the publisher of NAR, has recently launched the new journal Database: The Journal of Biological Databases and Curation (http://database.oxfordjournals.org/). We hope that the availability of this new journal, as well as that of our other sister journal, Bioinformatics, will provide a publication venue for databases that could not be accepted in the NAR Database Issue because of their limited scope, absence of manual curation or orientation to a limited readership.

The data resources of the Database Issue and the Database Collection make up an invaluable infrastructure upon which much of life science has come to rely. Far more than a collection of distinct information sources, the resources form an extensive and evolving network of connected data, common concepts and shared technologies, driven forward by the collective efforts of developers, curators and database managers. While there is no moderator of this network and no overall controller of its growth, through peer review and editorial processes, the Database Issue and Database Collection provide a valuable quality assurance service to the reader.

In this editorial, we first outline some of the new and updated databases that will be of interest to readers of the Database Issue. While individually these databases offer great utility, it is perhaps as part of the community of people, data and technology that the resources offer up some of their richest uses; we complete our introduction to the Database Issue with a commentary on this community.

NEW AND UPDATED DATABASES

In addition to the usual updates on the database services at the US National Center for Biotechnology Information

*To whom correspondence should be addressed. Tel: +44 1223 492 564; Fax: +44 1223 494 468; Email: cochrane@ebi.ac.uk

(NCBI) and the European Bioinformatics Institute (EBI), this issue includes a comprehensive listing of Japanese databases provided by the Japanese National BioResource Project [http://www.nbrp.jp (2)]. The geographic distribution of the featured databases continues to grow; the phiSITE (http://www.phisite.org), a database of gene regulation in bacteriophages (3), is the first database in the list from Slovakia.

Several articles in this issue feature updates on the status of databases that have been included in the Database Collection after being described first in other journals. These include the Eukaryotic Linear Motif database [ELM, http://elm.eu.org/ (4)], the Catalogue Of Somatic Mutations In Cancer [COSMIC, http://www.sanger.ac.uk/genetics/CGP/cosmic/ (5)], MicrobesOnline [http://www.MicrobesOnline.org (6)], the Immune Epitope Database [http://www.immuneepitope.org/ (7)] and PDBselect [http://bioinfo.tg.fh-giessen.de/pdbselect/ (8)]. Several other articles describe updated features of such popular resources as the Comprehensive Microbial Resource [CMR, http://cmr.jcvi.org/ (9)], PrimerBank [http://pga.mgh.harvard.edu/primerbank/ (10)] and the Therapeutic Target Database [http://xin.cz3.nus.edu.sg/group/cjttd/ttd.asp (11)], which have been last described in NAR several years ago.

In previous issues, update articles were permitted only brief descriptions of the latest changes in the respective resource. We felt that this limitation was unnecessary and that readers might benefit from more extensive and detailed descriptions of key database resources. For this issue, we have invited the authors responsible for several popular data resources to submit extended papers to provide a deeper insight into the organization and goals of their respective resources and would put the recent changes in these resources into a broader context. We are very happy with several excellent papers that resulted from this initiative, including comprehensive descriptions of the recent changes in Pfam [http://pfam.sanger.ac.uk/ (12)], MetaCyc [http://metacyc.org/ (13)], UniProt [http://www.uniprot.org/ (14)], IntAct [http://www.ebi.ac.uk/intact/ (15)], the Eukaryotic Linear Motif database [ELM, http://elm.eu.org/ (4)], the Comprehensive Microbial Resource [CMR, http://cmr.jcvi.org/ (9)] and the Integrated Microbial Genomes system [IMG, http://img.jgi.doe.gov/ (16)].

In addition, we have included extensive descriptions of three key databases recently unveiled by the EBI: the Gene Expression Atlas [http://www.ebi.ac.uk/gxa, (17)], Ensembl Genomes [http://www.ensemblgenomes.org, (18)] and the Protein Data Bank in Europe [PDBe, http://www.ebi.ac.uk/pdbe/ (19)]. We expect to continue this approach with longer articles next year; database authors who would like to submit such descriptions of their resources are encouraged to contact Michael Galperin at nardatabase@gmail.com in advance.

A COMMUNITY OF DATA RESOURCES

Some of the greatest efforts that have brought connectivity between the records of distinct resources were conceived not to serve pre-defined sets of users, but rather as open-ended initiatives that would provide broad utility across multiple domains. Perhaps the best known of these is the Gene Ontology (GO) project [http://www.geneontology.org/ (20)], which finds itself at the heart of the community of resources described in this Database Issue and Database Collection, with many of them providing GO annotations. Annotation of a common GO term to data objects in distinct resources allows the user to infer a conceptual relationship between the objects that is described by the term; by extension, more distant relationships can be inferred by using ontological relationships within GO to reach terms common to distinct data objects.

A shared approach to the development of data models is a theme in the Database Collection. For example, many model organism databases, such as Flybase [http://flybase.org/ (21)], Beetlebase [http://beetlebase.org/ (22)], ParameciumDB [http://paramecium.cgm.cnrs-gif.fr/db/index (23)] and wFleaBase [http://wfleabase.org/ (24)], have adopted the Chado database schema as underlying data structure for their resources. Chado, delivered and maintained by the GMOD community, provides database and interface technology for a broad range of information typically required by users of a model organism database, including genomic data, expression data, phenotypic information and literature collections (25). Centred on ontologies and controlled vocabularies, the schema is extensible through its system of domain-specific modules. Model organism databases that adopt Chado benefit from reduced development costs, as they are immediately able to use the substantial body of technologies (such as genome browsers, gene pages, search tools) that are freely available. In addition, users of these databases can apply their knowledge and experience of Chado-based interfaces to all model organism databases that have adopted the schema.

A key strategy for many resources in the Database Collection is partnering to exchange data, as a means of achieving comprehensive coverage and to reduce overall effort in data management. Successful data exchange relies not least on agreement to structure information in compatible ways. In 1982, the databases of the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org) established the Feature Table Definitions and continue to maintain the definitions to this day. The document defines a formalised text format in which biological features and their sequence coordinates can be described. INSDC feature table format, as defined in the Feature Table Definitions, remains the file format under which INSDC annotation data are kept in daily global synchrony.

An unsung hero, perhaps, in this community is the taxonomic backbone upon which almost all of its resources hang. The NCBI Taxonomy project (http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook&part=ch4) was established in 1991 and has been adopted as the *de facto* standard taxonomic classification of biomolecular data. While there are minor deviations and while for many resources with limited taxonomic range (such as the model organism databases), a model of taxonomy is

not implicitly required, those resources that deal with multiple taxa all share this one system. The heaviest users are the large generalist resources [INSDC, PDBe (19), UniProt (14)], but even specialist resources adopt the system. The benefit to the user of the resource, of course, is the ability to approach, filter and link biomolecular information for any given taxon or set of taxa.

Finally, there are the direct relationships between items of information in the resources. These take many forms: perhaps molecular sequences are similar between objects, the same genes are described, a functional correspondence between objects exists or common technology for presenting data binds two resources together. Some of these relationships are derived computationally, some manually curated from the literature and some exchanged between databases in reciprocal exchanges. The strengths of these relationships also vary; some resources share data directly leading to exact mappings between their objects, others report common attributes of their objects. Relationships can be asserted explicitly in cross-references, implied through common references to objects in tertiary resources or the user can be left to inject his/her own knowledge and creativity to bring a specific area of the network into sharper focus.

For those generating and interpreting data to be fed into the resources of the Database Collection, shared technologies and concepts need to be used. For this to happen, they need to be understood and readily available. Community standardisation initiatives, with their repertoire of minimal reporting standards and technology development initiatives to support the use of their standards, make their contribution here. Seminal work in the microarray field under the Microarray Gene Expression Data (MGED) consortium led to the development and adoption of MIAME—Minimal Information about a Microarray Experiment—a checklist of items of information required to render microarray data reusable beyond the initial analysis of those who generated the data (26). Since this time, a whole host of minimal reporting standards have been developed to better the usability of data, across genomics and environmental sequencing [the MIGS, MIMS and MIENS standards of the Genomics Standards Consortium and the emerging MINSEQE standard from MGED (27), http://gensc.org/, http://www.mged.org/minseqe/], proteomics [the MIAPE standard (28)], cell-based assays (MIACA; http://miaca.sourceforge.net/), phylogenetics [MIAPA (29)], systems biology [MIRIAM (30)] and many more.

Plenty of additional cases of such collective investment exist, but from the examples of vocabulary, taxonomic backbone, shared data models, common file formats and cross-references development initiatives alone, the value to the user is already clear. A continued attention to collective effort in such areas remains key to optimising the utility of our community of resources.

As then, we prepare grant proposals to generate, interpret and present our latest and greatest data, we make reference to interoperability, shared effort to develop technologies and the need for cooperation and collaboration. The community of resources described in this Database Issue and Database Collection provides a compelling example of the many successes that can be won with investment in interoperability and shared effort. Curators, developers, managers and users of life science databases know well the ongoing importance of developing and maintaining connectivity between our resources and cooperation between those involved.

ACKNOWLEDGEMENTS

The authors thank Scott Federhen (NCBI) and Emily Dimmer and Cath Brooksbank (EBI) for providing background information that helped in the preparation of this editorial; Sir Richard Roberts and Alex Bateman for many helpful comments; Patricia Anderson, Martine Bernardes-Silva, Karen Otto and Gail Welsh for excellent editorial assistance; and the Oxford University Press teams lead by Claire Bird and Radha Dutia for their help in compiling this issue.

FUNDING

European Molecular Biology Laboratory (to G.R.C.); Intramural Research Program of the US National Institutes of Health (to M.Y.G.). Funding for open access charge: Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Galperin,M.Y. and Cochrane,G.R. (2009) Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. *Nucleic Acids Res.*, **37**, D1–D4.
2. Yamazaki,Y., Akashi,R., Banno,Y., Endo,T., Ezura,H., Fukami-Kobayashi,K., Inaba,K., Isa,T., Kamei,K., Kasai,F. *et al.* (2010) NBRP databases: databases of biological resources in Japan. *Nucleic Acids Res.*, **38**, D26–D32.
3. Klucar,L., Stano,M. and Hajduk,M. (2010) phiSITE: Database of gene regulation in bacteriophages. *Nucleic Acids Res.*, **38**, D366–D370.
4. Gould,C.M., Diella,F., Via,A., Puntervoll,P., Gemünd,C., Chabanis-Davidson,S., Michael,S., Sayadi,A., Bryne,J.C., Chica,C. *et al.* (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.*, **38**, D167–D180.
5. Forbes,S., Tang,G., Bindahl,N., Bamford,S., Dawson,E., Cole,C., Kok,C.Y., Jia,M., Ewing,R., Menzies,A. *et al.* (2010) COSMIC (the catalogue of somatic mutations in cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res.*, **38**, D652–D657.
6. Dehal,P.S., Joachimiak,M.P., Price,M.N., Bates,J.T., Baumohl,J.K., Chivian,D., Huang,K.H., Keller,K., Novichkov,P.S., Dubchak,I.L. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.
7. Vita,R., Zarebski,L., Greenbaum,J., Emami,H., Hoof,I., Salimi,N., Damle,R., Sette,A. and Peters,B. (2010) The Immune Epitope Database 2.0. *Nucleic Acids Res.*, **38**, D854–D862.
8. Griep,S. and Hobohm,U. (2010) PDBselect 1992-2009 and PDBfilter-select. *Nucleic Acids Res.*, **38**, D318–D319.
9. Davidsen,T., Beck,E., Ganapathy,A., Montgomery,R., Zafar,N., Yang,Q., Madupu,R., Goetz,P., Galinsky,K., White,O. *et al.* (2010) The Comprehensive Microbial Resource (CMR). *Nucleic Acids Res.*, **38**, D340–D345.
10. Spandidos,A., Wang,X., Wang,H. and Seed,B. (2010) PrimerBank: a resource of human and mouse PCR primer pairs

for gene expression detection and quantification. *Nucleic Acids Res*., **38**, D792–D799.
11. Zhu,F., Han,B.C., Kumar,P., Liu,X.H., Ma,X.H., Wei,X.N., Huang,L., Guo,Y.F., Han,L.Y., Zheng,C.J. *et al*. (2010) Update of TTD: Therapeutic Target Database. *Nucleic Acids Res*., **38**, D787–D791.
12. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Ceric,G., Forslund,K., Holm,L. *et al*. (2010) The Pfam protein families database. *Nucleic Acids Res*., **38**, D211–D222.
13. Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F., Kaipa,P., Karthikeyan,A.S., Kothari,A., Krummenacker,M. *et al*. (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*., **38**, D473–D479.
14. The UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*, **38**, D142–D148.
15. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al*. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res*., **38**, D525–D531.
16. Markowitz,V.M., Chen,I.A., Palaniappan,K., Chu,K., Szeto,E., Grechkin,Y., Ratner,A., Anderson,I., Lykidis,A., Mavromatis,K. *et al*. (2010) The integrated microbial genomes (IMG) system: an expanding comparative analysis resource. *Nucleic Acids Res*., **38**, D382–D390.
17. Kapushesky,M., Emam,I., Holloway,E., Kurnosov,P., Zorin,A., Malone,J., Rustici,G., Williams,E., Parkinson,H. and Brazma,A. (2010) Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Res*., **38**, D690–D698.
18. Kersey,P.J., Lawson,D., Birney,E., Derwent,P.S., Haimel,M., Herrero,J., Keenan,S., Kerhornou,A., Koscielny,G., Kähäri,A. *et al*. (2010) Ensembl Genomes - Extending Ensembl across the taxonomic space. *Nucleic Acids Res*., **38**, D563–D569.
19. Velankar,S., Best,C., Beuth,B., Boutselakis,C.H., Cobley,N., Sousa da Silva,A.W., Dimitropoulos,D., Golovin,A., Hirshberg,M., John,M. *et al*. (2010) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res*., **38**, D308–D317.
20. The Gene Ontology Consortium. (2010) The Gene Ontology enters its second decade. *Nucleic Acids Res*., **38**, D331–D335.
21. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al*. (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res*., **37**, D555–D559.
22. Kim,H.S., Murphy,T., Xia,J., Caragea,D., Park,Y., Beeman,R.W., Lorenzen,M.D., Butcher,S., Manak,J.R. and Brown,S.J. (2010) BeetleBase in 2010: revisions to provide comprehensive genomic information for *Tribolium castaneum*. *Nucleic Acids Res*., **38**, D437–D442.
23. Arnaiz,O., Cain,S., Cohen,J. and Sperling,L. (2007) ParameciumDB: a community resource that integrates the *Paramecium tetraurelia* genome sequence with genetic data. *Nucleic Acids Res*, **35**, D439–D444.
24. Colbourne,J.K., Singan,V.R. and Gilbert,D.G. (2005) wFleaBase: the Daphnia genome database. *BMC Bioinformatics*, **6**, 45.
25. Mungall,C.J. and Emmert,D.B. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
26. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al*. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet*., **29**, 365–371.
27. Field,D., Garrity,G., Gray,T., Morrison,N., Selengut,J., Sterk,P., Tatusova,T., Thomson,N., Allen,M.J., Angiuoli,S.V. *et al*. (2008) The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol*., **26**, 541–547.
28. Taylor,C.F., Paton,N.W., Lilley,K.S., Binz,P.A., Julian,R.K. Jr, Jones,A.R., Zhu,W., Apweiler,R., Aebersold,R., Deutsch,E.W. *et al*. (2007) The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol*., **25**, 887–893.
29. Leebens-Mack,J., Vision,T., Brenner,E., Bowers,J.E., Cannon,S., Clement,M.J., Cunningham,C.W., dePamphilis,C., deSalle,R., Doyle,J.J. *et al*. (2006) Taking the first steps towards a standard for reporting on phylogenies: minimum information about a phylogenetic analysis (MIAPA). *Omics*, **10**, 231–237.
30. Le Novere,N., Finney,A., Hucka,M., Bhalla,U.S., Campagne,F., Collado-Vides,J., Crampin,E.J., Halstead,M., Klipp,E., Mendes,P. *et al*. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol*., **23**, 1509–1515.