Article

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

GAUDET, Pascale, BAIROCH, Amos Marc

Abstract

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.

Reference

GAUDET, Pascale, BAIROCH, Amos Marc. Towards BioDBcore: a community-defined information specification for biological databases. *Nucleic acids research*, 2011, vol. 39, no. Database, p. D7-D10

PMID : 21097465
DOI : 10.1093/nar/gkq1173

Published online 18 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D7–D10***
doi:10.1093/nar/gkq1173

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[12,13], Carole A. Goble[7], Takashi Gojobori[14,15], John M. Hancock[16], Douglas G. Howe[17], Tadashi Imanishi[14], Janet Kelso[13,18], David Landsman[13], Suzanna E. Lewis[19], Ilene Karsch-Mizrachi[11], Sandra Orchard[5], B. F. Francis Ouellette[13,20], Shoba Ranganathan[21,22], Lorna Richardson[23], Philippe Rocca-Serra[4], Paul N. Schofield[24], Damian Smedley[5], Christopher Southan[25], Tin Wee Tan[22], Tatiana Tatusova[11], Patricia L. Whetzel[26], Owen White[27] and Chisato Yamasaki[14] 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, [4]Oxford e-Research Centre, University of Oxford, Oxford, OX1 3QG, [5]European Molecular Biology Laboratory (EMBL) Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, [6]Faculty of Life Sciences, The University of Manchester, Manchester M13 9PT, [7]School of Computer Science, The University of Manchester, Manchester M13 9PT, [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, [10]Department of Genetics, Stanford University, Stanford, California 94305-5120, [11]NCBI, NLM, National Institutes of Health, Bethesda, MD 20894, [12]Genentech, 1 DNA Way, South San Francisco, CA 94080, USA, [13]DATABASE, The Journal of Biological Databases and Curation, Oxford University Press, Oxford OX2 6DP, UK, [14]Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi Koto-ku, Tokyo 135-0064, [15]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan, [16]MRC Harwell, Mammalian Genetics Unit, Harwell Science and Innovation Campus, Oxfordshire, OX11 0RD, UK, [17]The Zebrafish Model Organism Database, 5291 University of Oregon, Eugene, Oregon 97401-5291, USA, [18]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [19]Genomics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road 64R0121 Berkeley, California 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, [24]Department 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 and [27]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

Received and Accepted November 1, 2010

*To whom correspondence should be addressed. Tel: +41 22 379 5050; Fax: +41 22 379 5858; Email: pascale.gaudet@isb-sib.ch

This paper is also being published in the DATABASE journal, http://www.database.oxfordjournals.org/cgi/doi/10.1093/database/baq027

ABSTRACT

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.

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 to:

(i) Gather the necessary information to provide a general overview of the database landscape, and compare and contrast the various resources.
(ii) Encourage consistency and interoperability between resources.
(iii) Promote the uptake and use of semantic and syntactic standards.
(iv) Provide guidance for users when evaluating the scope and relevance of a resource, as well as details of the data access methods supported.
(v) 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. 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.

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 the Supplementary Data and at the BioDBCore web site.

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 further working examples (11) and (iii) in the longer term, completion of stable guidelines and their implementation as a public submission website that will allow data entry and easy update by database providers, in collaboration

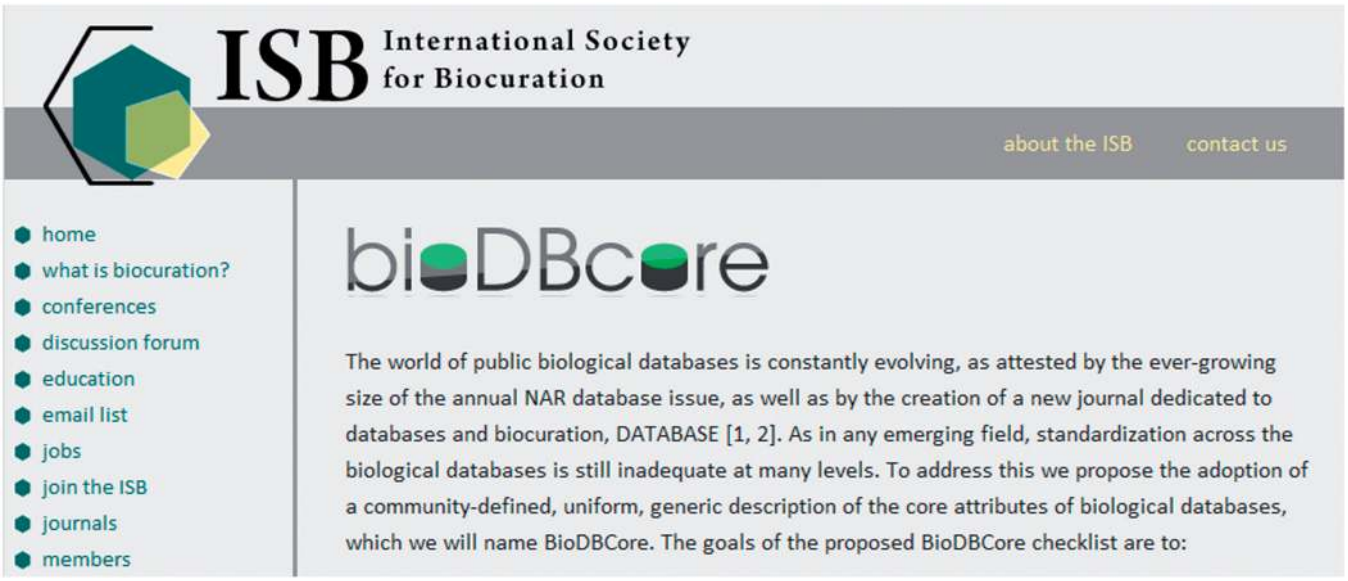

Figure 1. A screenshot of the BioDBCore discussion page on the ISB web site (http://biocurator.org/biodbcore.shtml).

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.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: Invited paper.

Conflict of interest statement. 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 acess published online, 26 March 2009).
3. Attwood,T.K., Kell,D.B., McDermott,P., Marsh,J., Pettifer,S.R. and Thorne,D. (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., Gojobori,T., Hannick,L., Hide,W., Hill,D.P., Kania,R., Schaeffer,M., St Pierre,S. *et al.* (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.
8. Taylor,C.F., Field,D., Sansone,S.A., Aerts,J., Apweiler,R., Ashburner,M., Ball,C.A., Binz,P.A., Bogue,M., Booth,T. *et al.* (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.*, **26**, 889–896.
9. Brazas,M.D., Yamada,J.T. and Ouellette,B.F.F. (2009) 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., Aidinis,V., Ainali,C., Bard,J., Balling,R., Birney,E., Blake,A., Bongcam-Rudloff,E. *et al.* (2010) Finding and sharing: new approaches to registries of databases and services for the biomedical sciences. *DATABASE*, doi:10.1093/baq014 (Advance acess published online, 2 July 2010).
11. Tan,T.W., Tong,J.C., De Silva,M., Lim,K.S. and Ranganathan,S. (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., Booth,T., Dukes,P., Gregurick,S.K., Kennedy,K., Kolar,P., Kolker,E., Maxon,M. *et al.* (2009) 'Omics Data Sharing. *Science*, **326**, 234–236.