Finding and sharing: new approaches to registries of databases and services for the biomedical sciences

Smedley, Damian; Schofield, Paul; Chen, Chao-Kung; Aidinis, Vassilis; Ainali, Chrysanthi; Bard, Jonathan; Balling, Rudi; Birney, Ewan; Blake, Andrew; Bongcam-Rudloff, Erik; Gkoutos, Georgios

Published in:
Database: The Journal of Biological Databases and Curation

DOI:
10.1093/database/baq014

Publication date:
2010

Citation for published version (APA):
Smedley, D., Schofield, P., Chen, C-K., Aidinis, V., Ainali, C., Bard, J., Balling, R., Birney, E., Blake, A., Bongcam-Rudloff, E., & Gkoutos, G. (2010). Finding and sharing: new approaches to registries of databases and services for the biomedical sciences. *Database: The Journal of Biological Databases and Curation*, *2010*, [baq014]. https://doi.org/10.1093/database/baq014

Database, Vol. 2010, Article ID baq014, doi:10.1093/database/baq014

Original article

Finding and sharing: new approaches to registries of databases and services for the biomedical sciences

Damian Smedley[1,*], Paul Schofield[2], Chao-Kung Chen[1], Vassilis Aidinis[3], Chrysanthi Ainali[4], Jonathan Bard[5], Rudi Balling[6], Ewan Birney[1], Andrew Blake[7], Erik Bongcam-Rudloff[8], Anthony J. Brookes[9], Gianni Cesareni[10], Christina Chandras[3], Janan Eppig[11], Paul Flicek[1], Georgios Gkoutos[12], Simon Greenaway[7], Michael Gruenberger[2], Jean-Karim Hériché[13], Andrew Lyall[1], Ann-Marie Mallon[7], Dawn Muddyman[2], Florian Reisinger[1], Martin Ringwald[11], Nadia Rosenthal[14], Klaus Schughart[15], Morris Swertz[16], Gudmundur A. Thorisson[9], Michael Zouberakis[3] and John M. Hancock[7]

[1]European Bioinformatics Institute, Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, [2]Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge, CB2 3EG, UK, [3]Institute of Immunology, Biomedical Sciences Research Center Alexander Fleming, 34 Fleming Street, 16672 Athens, Greece, [4]Kings College London, Strand, London, WC2R 2LS, [5]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, OX1 3QX, [6]Interdisciplinary Centre for Systems Biology, University of Luxembourg, Campus, Limpertsberg, 162A, Avenue de la Faiencerie, L-1511 Luxembourg, [7]Bioinformatics Group, MRC Harwell, Harwell, Oxfordshire, OX11 0RD, UK, [8]The Linnaeus Centre for Bioinformatics, Swedish University of Agricultural Sciences, S-750 07 Uppsala, Sweden, [9]Department of Genetics, University of Leicester, Leicester, UK, [10]Department of Biology, University of Rome Tor Vergata, Rome, Italy, [11]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA, [12]Department of Genetics, University of Cambridge, Cambridge, CB2 3EG, UK, [13]European Molecular Biology Laboratory, Heidelberg, Germany, [14]EMBL-Monterotondo Outstation, Via Ramarini 32, 00015 Monterotondo-Scalo (RM), Italy, [15]Experimental Mouse Genetics, Helmholtz Centre for Infection Research & University of Veterinary Medicine, Hannover, Inhoffenstrabe 7, D-38124 Braunschweig, Germany and [16]University Medical Center Groningen, Department of Genetics, PO Box 30001, NL-9700 RB, Groningen, The Netherlands

***Corresponding author:** Tel: +44 1223 494451; Fax: +44 1223 494468; Email: damian@ebi.ac.uk

Submitted 26 April 2010; Accepted 20 June 2010

The recent explosion of biological data and the concomitant proliferation of distributed databases make it challenging for biologists and bioinformaticians to discover the best data resources for their needs, and the most efficient way to access and use them. Despite a rapid acceleration in uptake of syntactic and semantic standards for interoperability, it is still difficult for users to find which databases support the standards and interfaces that they need. To solve these problems, several groups are developing registries of databases that capture key metadata describing the biological scope, utility, accessibility, ease-of-use and existence of web services allowing interoperability between resources. Here, we describe some of these initiatives including a novel formalism, the Database Description Framework, for describing database operations and functionality and encouraging good database practise. We expect such approaches will result in improved discovery, uptake and utilization of data resources.

Database URL: http://www.casimir.org.uk/casimir_ddf

Biologists currently face a daunting challenge when trying to discover which of the multitude of computational and data resources to use in analysing their results and developing their hypotheses. The basic task of identifying appropriate online resources in a research field is non-trivial and typically involves *ad hoc* Internet trawling, recommendations from colleagues or literature searching. This is then followed by the more complex task of establishing whether the resource is relevant, reliable, well curated, and maintained. If programmatic access is required, discovering

Database, Vol. 2010, Article ID baq014, doi:10.1093/database/baq014

whether this exists and how to utilize it is another challenge. As time is short, most researchers often end up using familiar resources, which are not always the best or most relevant, while the developers and funders of under-utilized but valuable resources essentially waste time and money. What is required is a solution that helps to maximize the usefulness of each resource to the overall community. At present, approaches are being developed to construct two types of registry. One type, 'databases of databases', deal with describing the contents and other metadata about databases. The other type, web service registries, deal with the explicit description of services available at particular sites (not always databases). We present the two areas separately, but ultimately we expect solutions to arrive that merge the two approaches.

Registries of databases

Comprehensive, top-level registries of biological resources are currently provided by the Nucleic Acids Research Molecular Biology Database Collection,[1] the BioMedCentral Catalog of Databases on the Web (http://databases.biomedcentral.com) and the Bioinformatics.ca Links Directory.[2] However, they do not collect extensive metadata beyond a brief description of the resource and URL, can only be browsed by the category each registry has assigned or searched by the resource name, and lack much of the detailed information that the community requires. A number of projects [e.g. CASIMIR (Coordination and Sustainability of Mouse Informatics Resources)[3] and ENFIN[4]] have identified this problem and are producing 'database of databases' (registries) for their field of expertise.[5]

A registry of resources needs to be more than just a list of databases and textual descriptions to be useful to the biological and bioinformatics communities. To achieve its aim of helping scientists find the most relevant resource for their needs, it needs to provide at the very least browsing and searching by the type of data contained in each resource, i.e. the biological scope of the resource. A typical approach, as used by all the registries described above and the MRB (Mouse Resource Browser)[6] registry developed by a number of the authors of this article, is for a community to define a list of categories (a controlled vocabulary) that covers their scientific domain and then to tag each resource with one or more of these terms. Use of existing and newly developed ontologies for these tags would certainly facilitate future interoperability of the various registries being developed.

While developing MRB, user feedback suggested that it would be helpful if users could go beyond simple categorization of the scope of resources to discover metadata describing database operations and functionality. We therefore set out to capture the utility, accessibility and ease of use of a resource, along with its potential interoperability with other tools and databases. The types of questions that we wanted to be able to answer from this metadata included whether the resource uses automated or manual curation, how often it updates and whether there is a way to track back to different versions, does it provide good technical documentation and user support, does it use recognized standards to record and structure its data, and finally does it go beyond simple web browsing to allow programmatic access and output in standard formats?

Data are always easier to capture and search if a consistent standard is used and we therefore developed a Database Description Framework (DDF; Table 1) as part of the CASIMIR project. Although produced for the MRB, the DDF is generically applicable to any biological database and can be adapted for the requirements of any biological community. For each heading or category, there is a three-tier assessment criterion, a number chosen for simplicity and ease of use. The aim of the DDF is not to make 'value judgements' about a resource, but to summarise what it does and what functionalities it supports, with the categories simply reflecting the degree of complexity or sophistication of the database. What is useful or relevant for some databases need not be so for others, and each needs to be assessed in terms of its own remit and user community. The DDF is also intended to be helpful in disseminating and supporting good database practice, in providing backing for resources aspiring to improve the levels of their service, and in giving objective criteria that can be used by external assessors to measure a resource's progress towards their stated goals.

caBIG, the NCI Cancer Bioinformatics grid[7] has produced a similar framework for capturing resource metadata but with a stronger focus on the technical assessment of the resources that wish to participate in the project. As caBIG has a well-defined set of tasks and a user community tied to the specific vision and funding, their categories and levels are less generic than those in the DDF and more focused on assessing whether databases reach a required level of interoperability to interact with the other components of this particular project.

Registries of web services

As well as capturing the scope and database practices of resources, registries need to be explicit about the modes of programmatic access that databases provide (e.g. web services) as these are increasingly used to build database networks and cyberinfrastructure.[8–10] This technical information is often hard to find in publications or even on database web sites, but can radically change the strategy adopted by bioinformaticians needing to access the database—for example, integration into automated or

Table 1. The CASIMIR Database Description Framework (DDF)

Category	Level 1	Level 2	Level 3
Quality and Consistency	No explicit process for assuring consistency	Process for assuring consistency, automatic curation only	Process for assuring consistency with manual curation
Currency	Closed legacy database or last update more than a year ago	Updates or versions more than once a year	Updates or versions more than once a month
Accessibility	Access via browser only	Access via browser and database reports or database dumps	Access via browser and programmatic access (well defined API, SQL access or web services)
Output formats	HTML or similar to browser only	HTML or similar to browser and sparse standard file formats, e.g. FASTA	HTML or similar to browser and rich standard file formats, e.g. XML, SBML (Systems Biology Markup Language)
Technical documentation	Written text only	Written text and formal structured description, e.g. automatically generated API docs (JavaDoc), DDL (Data Description Language), DTD (Document Type Definition), UML (Unified Modelling Language), etc.	Written text and formal structured description and tutorials or demonstrations on how to use them
Data representation standards	Data coded by local formalism only	Some data coded by a recognised controlled vocabulary, ontology or use of minimal information standards (MIBBI)	General use of both recognised vocabularies or ontologies, and minimal information standards (MIBBI)
Data structure standards	Data structured with local model only	Data structured with formal model, e.g. an XML schema	Use of recognised standard model, e.g. FUGE
User support	User documentation only	User documentation and Email/web form help desk function	User documentation as well as a personal contact help desk function/training
Versioning	No provision	Previous version of database available but no tracking of entities between versions	Previous version of database available and tracking of entities between versions

semi-automated work flows using Taverna[11] such as that developed by CASIMIR.[12] Unfortunately, traditional web-service description languages such as WSDL do not provide the required detail on the biological context of the inputs and outputs of each service to allow automated data and service integration. Biocatalogue[13] and its predecessor, the EMBRACE service registry[14], address this lack of semantics by providing sites for the registration, curation, discovery and monitoring of web services for the whole biological community. Curation of information about web services is open to anyone and uses a combination of free text, tags, ontology terms and example values to describe what each service does, the type of web service (REST, SOAP, soaplab) and in particular the input and outputs in terms of what type of biological data and data formats are expected. Biocatalogue clearly addresses a vital requirement of the community and already some 1173 services have been annotated, despite the project only running for just over a year. Having a single, well-designed solution rather than multiple competing efforts is likely to improve further uptake, and we propose that all registries of databases utilize Biocatalogue to annotate the services provided by their resources rather than separately performing this task.

Dissemination issues and solutions

Capturing metadata as described for the DDF or the Biocatalogue project is not easy. Our initial DDF metadata for over 220 resources was captured as part of a detailed MRB questionnaire sent to each resource, and active manual curation had to be used to fill in the gaps in responses. This is expensive and time-consuming and, after the first pass, there is a requirement to keep the captured data up to date, and this is not easily met.

To eliminate the cost of a central curation effort, it would be much better if each resource curated their own metadata and made it accessible to the wider scientific community. As an example of this, we produced a DDF extension to the Drupal content management system (http://drupal.org), which allows curators to log-in and categorize their databases in terms of DDF categories and

Database, Vol. 2010, Article ID baq014, doi:10.1093/database/baq014

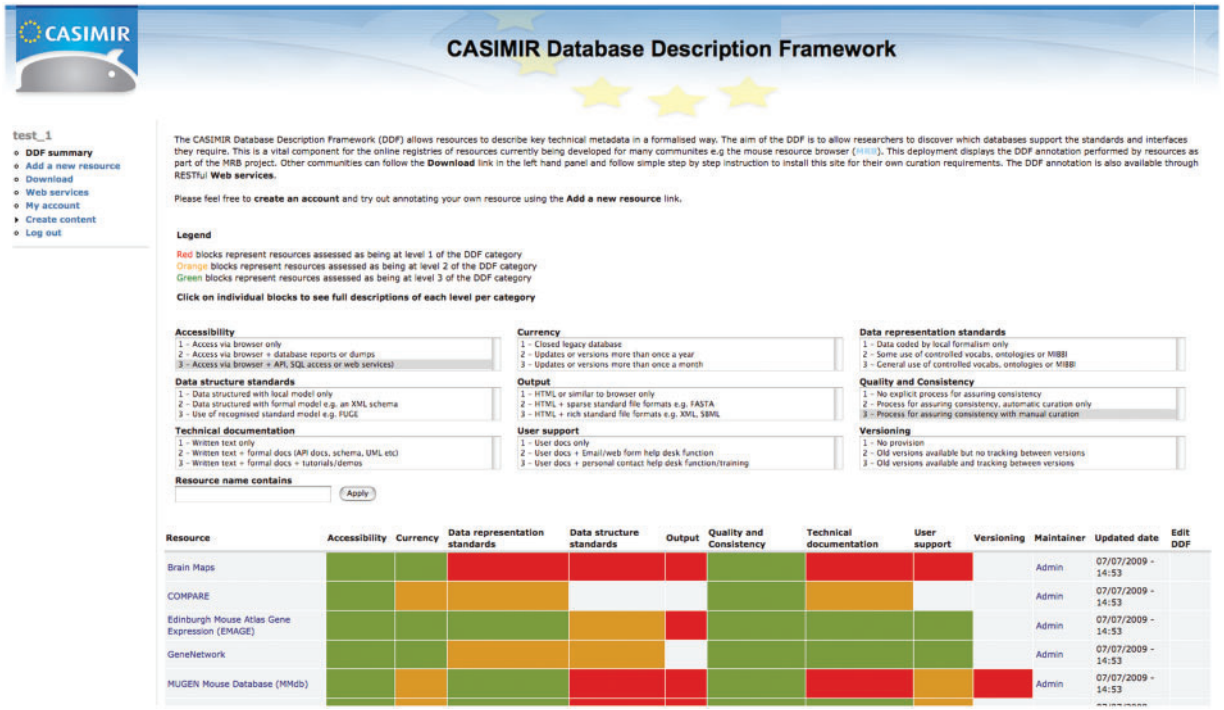

Figure 1. The DDF query and annotation tool. This tool allows any user to browse a set of resources that have been annotated using the DDF categories. Searches for resources by DDF category and level are also possible. In addition, resource maintainers can log-in and edit their existing annotations or annotate a new resource using a simple web form. This tool is freely available and easy to install for other communities that wish to create their own registry of resources.

levels using a simple web form. The resulting metadata is then browsable and searchable either through a web interface or programmatically through RESTful web services. An example deployment is viewable at www.casimir.org.uk/casimir_ddf (Figure 1) and is currently populated with the metadata for the MRB project. We encourage interested readers to visit our site and for maintainers of resources to curate their metadata using it. The Drupal framework is easily extensible to allow curation of other data associated with each resource, so allowing the production of a customisable community registry. The system is expected to be of great value to communities developing registry resources or individual informaticians wanting to establish quickly which features a database provides (the software is freely available under an open-source license). The REST web services allow a central DDF portal to be established offering the collection and sharing of data from individual database registries as well as avoiding redundancy in curation efforts.

Biocatalogue have used a combination of central and community curation from the outset to capture data on web services and the large number of services already described is testament to such an approach. Again, the provision of easy to use web tools that suggest particular tags and ontology terms to use in the annotation increases the likelihood of achieving a high level of community engagement and annotation quality.

Community curation requires pro-active participation. Communities need to acknowledge; (i) a central site where they can find relevant resources would be useful, and; (ii) the only practical means of achieving this is for each database to self-curate its entry using a clearly articulated and standardized set of benchmarks and tools such as provided by the DDF and Biocatalogue solutions. Individual resources would also benefit from this small amount of curation effort as the central registry will direct users to them, who might not previously have known about their resource. Although the creators and maintainers of a resource are best placed to describe the associated metadata, a self-curation approach can raise data quality issues, but these should be minimized if the annotation tools are well designed i.e. fast and easy to use, with clear descriptions of what is being asked for, and responses presented as a lists of terms rather than free text. However, even with a well-designed annotation tool, registries are still likely to require some central curation for validating submitted data (e.g. the DDF tool allows administrator level access to check new submissions).

In summary, there is now a clear need for registries to be built that address biological categorization of databases

and services, annotate any services provided and capture metadata on database best practises. Considerable progress has been made on standardizing the capture of each of these by such approaches as the DDF and Biocatalogue, but the community would benefit from coordination to produce full registries combining all these approaches. However, the value of a standard is dependent on its uptake by the community as can be seen, for example, in the MIBBI family of minimal information standards.[15] Uptake of a standard is, of course, as much a social issue as one of producing the right technologies for the community. Here, support from funding agencies and journals will be vital in establishing the practice of publishing database and services metadata. All curators can enhance the value of their databases by posting a minimal amount of information about their resource on a community site. The task has minimal cost, but will provide considerable value to investigators, database developers, informaticians and funding agencies.

Funding

The work described above was supported by Seventh Framework Programme of the European Commission contracts to CASIMIR (LSHG-CT-2006-037811) and ENFIN (LSHG-CT-2005-518254). Funding for open access charge: CASIMIR LSHG-CT-2006-037811.

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. Brazas,M.D., Yamada,J.T. and Ouellette,B.F. (2009) Evolution in bioinformatic resources: 2009 update on the Bioinformatics Links Directory. *Nucleic Acids Res.*, **37**, W3–W5.
3. Hancock,J.M., Schofield,P.N., Chandras,C. *et al*. (2008) CASIMIR: Coordination and Sustainability of International Mouse Informatics Resources. *Proceedings of the 8th IEEE International Conference on Bioinformatics and Bioengineering*, doi:101109/BIBE.2008.4696712.
4. Reisinger,F., Corpas,M., Hancock,J. *et al*. (2008) In: Bairoch,A., Cohen-Boulakia,S. and Froidevaux,C. (eds), *Data Integration in the Life Sciences: 5th International Workshop, DILS 2008*, Evry, France, June 25–27, 2008. Springer, Berlin, pp. 132–143.
5. Babu,P.A., Udyama,J., Kumar,R.K. *et al*. (2007) DoD2007: 1082 molecular biology databases. *Bioinformation*, **2**, 64–67.
6. Zouberakis,M., Chandras,C., Swertz,M. *et al*. (2010) Mouse Resource Browser—a database of mouse databases. *Database*, doi:10.1093/database/baq010.
7. Saltz,J., Oster,S., Hastings,S. *et al*. (2006) caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics*, **22**, 1910–1916.
8. Stein,L. (2008) Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nat. Rev. Genet.*, **9**, 678–688.
9. Foster,I. (2005) Service-oriented science. *Science*, **308**, 814–817.
10. Hey,T. and Trefethen,A.E. (2005) Cyberinfrastructure for e-Science. *Science*, **308**, 817–21.
11. Hull,D., Wolstencroft,K., Stevens,R. *et al*. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, W729–W732.
12. Smedley,D., Swertz,M.A., Wolstencroft,K. *et al*. (2008) Solutions for data integration in functional genomics: a critical assessment and case study. *Brief. Bioinformatics*, **9**, 532–544.
13. Bhagat,J., Tanoh,F., Nzuobontane,E. *et al*. (2010) BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res.*, **38**, W689–W694.
14. Pettifer,S., Thorne,D., McDermott,P. *et al*. (2009) An active registry for bioinformatics web services. *Bioinformatics*, **25**, 2090–2091.
15. Taylor,C.F., Field,D., Sansone,S.A. *et al*. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.*, **26**, 889–896.