Published online 19 May 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W689–W694***
doi:10.1093/nar/gkq394

BioCatalogue: a universal catalogue of web services for the life sciences

Jiten Bhagat[1], Franck Tanoh[1], Eric Nzuobontane[2], Thomas Laurent[2], Jerzy Orlowski[3], Marco Roos[4,5], Katy Wolstencroft[1], Sergejs Aleksejevs[1], Robert Stevens[1], Steve Pettifer[1], Rodrigo Lopez[2] and Carole A. Goble[1,*]

[1]School of Computer Science, The University of Manchester, Manchester, M13 9PL, [2]EMBL European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK [3]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, PL-02-109 Warsaw, Poland [4]Informatics Institute, University of Amsterdam, Amsterdam, 1098 SJ and [5]Human Genetics Department, Leiden University Medical Centre, NL-2333 ZA Leiden, Netherlands

Received January 31, 2010; Revised April 27, 2010; Accepted April 29, 2010

ABSTRACT

The use of Web Services to enable programmatic access to on-line bioinformatics is becoming increasingly important in the Life Sciences. However, their number, distribution and the variable quality of their documentation can make their discovery and subsequent use difficult. A Web Services registry with information on available services will help to bring together service providers and their users. The BioCatalogue (http://www.biocatalogue.org/) provides a common interface for registering, browsing and annotating Web Services to the Life Science community. Services in the BioCatalogue can be described and searched in multiple ways based upon their technical types, bioinformatics categories, user tags, service providers or data inputs and outputs. They are also subject to constant monitoring, allowing the identification of service problems and changes and the filtering-out of unavailable or unreliable resources. The system is accessible via a human-readable 'Web 2.0'-style interface and a programmatic Web Service interface. The BioCatalogue follows a community approach in which all services can be registered, browsed and incrementally documented with annotations by any member of the scientific community.

INTRODUCTION

As of 2010, there are more than 1400 publicly available bioinformatics tools and databases on the Web (1,2), with over 100 new Web servers providing interactive analysis tools reported in 2009 alone (3). These published resources are just the tip of a very large iceberg, and many others exist in relative obscurity, advertised only via project or laboratory Web Pages.

Though interactive access to these resources via Web Pages has been of enormous benefit to the community over the years, there is a growing demand for programmatic interfaces that allow these tools and databases to be linked together in automated analysis pipelines (4). Web Services are becoming an increasingly popular way of providing robust remote access (5), and this approach has been adopted by major service providers including the EMBL-EBI (6), KEGG (7), NCBI (8) and the DDBJ (9). Web Services can easily be accessed from most programming languages, or chained together as workflows using free tools [e.g. Taverna (10) or Kepler (11)], or their commercial equivalents [e.g. PipeLine Pilot (http://accelrys.com/products/pipeline-pilot/)].

The resources to which Web Services provide access are distributed across centres, projects, countries and disciplines and, for the most part, are currently likely to be discovered by word-of-mouth, Google searches, or from simple on-line lists such as http://www.xmethods.net/ or http://www.webservicelist.com/. As the number of Web Services has grown, so has the need for gathering information about them into one place. Table 1 gives a short summary of prominent public service registries that are relevant to the Life Sciences. These broadly fall into two categories: those that represent collections of services based on a specific schema and/or technology [e.g. BioMOBY Central (12,13), the DAS registry (14) and those that do not (e.g. seekda (http://www.seekda.com/) and the European Model for Bioinformatics Research and Community Education (EMBRACE) registry (15)]. Although some have major commercial or institutional backing, others have grown out of fixed term

*To whom correspondence should be addressed. Tel: +44 161 275 6195; Fax: +44 161 275 6204; Email: carole.goble@manchester.ac.uk

Table 1. A summary of existing on-line collections of Web Services

Collection	URL	Service types	Description	No of services
DAS Service Registry	www.dasregistry.org	DAS services only	Validates services as conforming to the DAS schema, and monitors their up-time. Allows searching based on free-text descriptions as well as categorization by provider according to a controlled vocabulary of types.	~750
EMBRACE Registry	www.embraceregistry.net	SOAP, REST, DAS, BioMOBY	Principally contains services developed during the EMBRACE and BioSapiens projects, with a small number of external 'guest' services. Monitors service behaviour and validates endpoint descriptions. Basic community- and provider-tagging and categorization according to controlled vocabularies.	~800
BioMOBY Central	www.biomoby.org	BioMOBY services only	Strongly typed services, categorized according to the BioMOBY ontology. Monitors up-time and behaviour, and automatically removes unresponsive services.	~1560
seekda	www.seekda.com	SOAP	Unlike the other collections reported here where content is manually added by the community, seekda 'scavenges' for Web Services in the style of a search engine. The system is not specific to the Life Sciences, and contains services relevant to numerous disciplines.	~28 500
BioCatalogue	www.biocatalogue.org	SOAP, REST, with BioMOBY and DAS in development	The BioCatalogue aggregates Life Science-specific content from other sources, classifying it according to an ontology. Entries can be annotated by the community, and verified manually by a curator. Provides monitoring of up-time and validation of service interfaces.	~1620

projects and hence their long-term future is unclear. Alongside registry building, there have also been ongoing efforts to describe Web Services with rich semantic annotations using ontologies and modern ontology languages. Examples include SSwap [16], Feta [17], SADI (http://sadiframework.org/) and BioMOBY.

Drawing together the experience of these existing initiatives, the BioCatalogue provides a universal catalogue of Web Services for the Life Sciences. Launched in June 2009 and hosted at the EMBL-EBI, it allows registration of services that are specific to the Life Sciences (such as those for protein sequence or molecular structures) as well as more generic services that are of direct utility in this domain (e.g. text mining and image analysis). The catalogue does not host these services itself; instead it provides a mechanism to discover services and annotate them. The BioCatalogue has five key properties:

(1) It provides a single up-to-date port-of-call for finding Life Science Web Services, regardless of their technology or provenance. As well as allowing new registration of services manually and via its own Web Service interface, it aggregates contributions from other registries. For example, the catalogue carries service registrations from the EMBRACE Registry and domain-specific services from seekda.
(2) It offers a long-term sustained resource for service descriptions that is also a safe haven for securing the contents of registries beyond their originating projects (e.g. EMBRACE and BioSapiens services).
(3) It adds uniform and rich annotations to the services that harmonize their descriptions regardless of source or type. The annotations explain what the service does and how to use it. The descriptions draw upon existing and emerging work in the Semantic Web Services [e.g. Semantic Annotation for WSDL (SAWSDL) (18) and Semantic Annotation for REpresentational State Transfer (SA-REST) (19)]. Annotations from the EMBRACE and Feta registries have been contributed to the catalogue. Content is monitored by a full-time curator assisted by the registered members.
(4) It provides a rich range of facilities, adopting the best components of other registries where available, e.g. the EMBRACE service monitoring framework and endpoint validation software, and the use of seekda for service scavenging.
(5) It addresses the combined needs of service providers, users, annotators and developers alike, enabling the catalogue's content to be readily extended, curated and used by the community.

Currently, the BioCatalogue has over 300 registered members. It describes 1627 Web Services (1585 SOAP services, and 42 REST services) from over 158 different providers from 25 countries. All the services of the major data centres (EMBL-EBI, DDBJ and NCBI) are present.

USING THE BIOCATALOGUE

The BioCatalogue can be accessed via two mechanisms: a human-readable 'Web 2.0'-style interface which supports browsing, searching and the manual creation and annotation of service entries; and a Web Service API for programmatic access.

The 'Web 2.0' interface

The BioCatalogue's Web interface provides faceted browsing, extensive link-based navigation and filtering on multiple criteria including service categories, keywords,

Table 2. A representative sample of BioCatalogue REST API methods, accessible via http://www.biocatalogue.org/

Endpoint	Description
/search.xml?q={query}	Search by keyword to retrieve relevant services, SOAP operations, service providers, users and registries.
/services.xml	Services index.
/services/filters.xml	Filter services based on categories, tags, countries, submitters and service providers.
/services/{id}.xml	Details about a specific service.
/services/{id}/monitoring.xml	Monitoring details for a specific service.
/services/{id}/annotations.xml	Annotations on a specific service.[a]
/service_providers.xml	Service providers index.
/service_providers/{id}.xml	Details about a specific service provider.
/users.xml	Users index.
/users/{id}.xml	Details about a specific user (aka member).
/users/{id}/annotations_by.xml	Annotations by a specific user.[a]
/registries.xml	Registries index (lists registries from which BioCatalogue has sourced data).
/annotations.xml	Annotations index.[a]
/soap_services/{id}/operations.xml	SOAP operations on a specific SOAP service.
/soap_operations.xml	SOAP operations index.
/soap_operations/filters.xml	Filters that can be applied to the SOAP operations index. These include filters for tags on inputs and outputs.
/soap_operations/{id}.xml	Details about a specific SOAP operation.
/soap_operations/{id}/annotations.xml	Annotations on a specific SOAP operation.[a]
/soap_inputs/{id}.xml	Details about a specific SOAP input.
/soap_inputs/{id}/annotations.xml	Annotations on a specific SOAP input.[a]
/soap_outputs/{id}/annotations.xml	Annotations on a specific SOAP output.[a]
/rest_services/{id}.xml	Details about a specific REST service.
/rest_services/{id}/annotations.xml	Annotations on a specific REST service.[a]
/service_tests/{id}.xml	Details about a specific service test.
/service_tests/{id}/results.xml	Monitoring test results for a specific service test.
/test_results.xml	Monitoring test results index.
/test_results/{id}.xml	Details about a specific monitoring test result.
/categories.xml	Service categories index.
/categories/{id}.xml	Details about a specific service category.

[a]Output also available in JSON format.
Parameters enclosed in braces should be replaced by the appropriate search term or identifier. For example http://www.biocatalogue.org/search.xml?q=protein returns the XML description of all services containing the keyword 'protein'. A full list of methods, with documentation is available from http://apidocs.biocatalogue.org/.

providers, location and service type. Displayed information such as service popularity based on view statistics, comments from other users and the number and quality of annotations, helps to identify suitable services and find alternative or similar services.

All available information held on a service, including its annotations, tags and provider documentation is included in the search. Searching is facilitated by term suggestion based on tags, previous user searches and terms from the myGrid ontology (20). The 'Search by Data' feature matches a sample of a user's input data against example input data provided in the service annotations, allowing the user to discover services that provide methods for analysing their data. The BioCatalogue is configured so as to be indexable by generic Web search engines (e.g. Google) as well as being explicitly indexed in the specialist EB-eye (21) search engine.

Announcements and release notes are posted on Twitter and syndicated on RSS feeds. Registry entries may be bookmarked using social bookmarking systems such as Delicious (http://delicious.com/) or Digg (http://digg.com/). Users may log in using OpenID, Google, Facebook, Twitter, Yahoo! or Verisign accounts, simplifying registration, and limiting username and password proliferation.

The Biocatalogue web service interface

The BioCatalogue provides a REST Web Service API, enabling tools such as Taverna and registry aggregation sites such as ONIX (http://www.ncri-onix.org.uk/) to access its contents. The main exchange format is XML, with JSON (http://www.json.org/) output available for the annotations. The API broadly reflects the same functionality that can be accessed via the interactive Web interface. Table 2 outlines the main XML endpoints and their functions. Full documentation, along with code examples, is available from http://apidocs.biocatalogue.org/.

SERVICE ANNOTATION

The descriptions of the services registered in the BioCatalogue are drawn from service providers, the user community and monitoring and usage analysis. Each annotation is associated with a source (automatic analysis, other registries, the providers or named curators) and can take the form of structured data, free text, tags or ontology terms. Annotations are divided into four main categories:

- Functional: outlines the task of a service, the type(s) of analyses possible, information relating to underlying

data resources used, its various operations, the function and format of any inputs and outputs, and whether parameters are mandatory or optional. Examples of input data or service usage are provided where available. Services are classified into multiple categories based on their biological category (e.g. proteomics) and their technology (e.g. text mining).

- Operational: describes the mechanisms and any conditions and assumptions necessary to execute a service (e.g. restrictions by the service provider placed on the number of invocations allowed in a given interval). We previously observed (22) that many service providers structure their services to work in idiomatic ways: (i) combining the numerous useful functions beneath a single service interface [e.g. SoapLab (23), GenePattern (24) and RapidMiner (http://rapid-i.com)]; (ii) requiring operations to be combined to deliver a task [e.g. the EMBL-EBI Web Services (6)]; or (iii) prescribing that the services' interface be mapped to a semantic signature [e.g. BioMOBY (12), SSWAP (16) and SADI].
- Profile: records objective analyses drawn from monitoring metrics automatically mined from other resources: for example, workflow management systems and subjective comments about the use and usability of services from the user community.
- Provenance: includes details of where the service is hosted, who submitted the service to the registry, and who has provided annotation. Changes to the service description (e.g. its WSDL document) or its associated annotations are also recorded in order to provide a history of the service as well as an audit trail.

The BioCatalogue currently holds more than 33 000 annotations. Approximately a third of services have all operations described. As much documentation as possible is automatically extracted from the published service interfaces, and additional annotations may be added during or after initial submission by the contributor. These semantic service descriptions can be imported and exported in formats compliant with SAWSDL (18) and SA-REST (19) standards.

MONITORING WEB SERVICES

The status, reliability and stability of a Web Service are often the deciding factors for choosing a service. The BioCatalogue has adopted the EMBRACE Registry's system for monitoring service availability, service interface changes and service functionality (15). Availability is indicated using a simple 'traffic light' mechanism, whereby green means the service is active, yellow means it has one or more unresolved issues, and red means it is currently unavailable. Service interface changes are managed by periodically re-parsing interface documents and comparing them with the existing entry. Functionality is checked by the submission of scripts that exercise specific aspects of the services, managed by a separate server. By automatically monitoring changes, a history of service versions and performance can be provided and users relying on specific services can be notified of these changes by RSS subscription or Twitter.

Usage of the BioCatalogue is monitored to build up a profile of searches and access. This reveals relationships between services, including usage patterns; for example, services that are commonly used together, and/or services that provide similar functionality, which may be used as substitutes if one of these services becomes unavailable.

COMMUNITY CONTRIBUTION TO CONTENT

Members can register a Web Service, share their views, make comments or annotations on any service and provide examples of service usage with relevant input and output data. Automatic harvesting of service annotations provides the foundation on which user-provided annotations rest. Submission of services and annotations contribute to the reputation of a member, encouraging further contributions. Content is monitored by a full-time curator who oversees content and coordinates a small pool of curators to help members improve annotations and adopt best practices.

The BioCatalogue team includes several service providers, including the EMBL-EBI. Other providers are encouraged to contribute. As well as an active 'friends' mailing list, online news feeds and a wiki (http://www.biocatalogue.org/wiki/), 'annotation jamborees'—virtual or face to face group efforts to annotate a large set of Web Services and to discuss best practices, new features, directions and general issues—are organized periodically. These jamborees serve as a resource review and a team-building forum as well as a source of new annotations.

All descriptions are attributed and open to scrutiny and all monitoring results are available. Documentation is provided at various levels of detail covering guidelines and best practices for service creation and execution. Help pages provide instructions or links on how to test and run services with different tools: GUI tools, such as soapUI (http://www.soapui.org/), SOAP Client (http://ditchnet.org/soapclient/) or workflow execution engines, such as Taverna and Kepler. Pointers to commonly used software libraries that can be used to incorporate Web Services into new programs in different programming and scripting languages, and links for creating new Web Services or writing a Web Service API to an existing tool, are also provided.

CONCLUSION

The first phase of the BioCatalogue has focused on the design and development of its Web interface and API, on establishing its core content and on the building of a contributing community. Since its launch in 2009, it has had over 14 000 visits and is successfully growing a community of contributors and users. The majority of visitors use the search and browsing features to discover services. Of the 300 or so registered members, a subgroup of around 20 actively contribute high quality manual annotations.

In cooperation with their respective developers, services generated during the EMBRACE and BioSapiens projects and relevant services found by the seekda search engine have already been included, and content from BioMOBY Central and the DAS registry will be added shortly. Thus the bulk of current services have been accumulated from registries, by scavanging, and by the major service providers. We now observe a growing number of more specialist service providers each adding a small number of domain-specific services to the catalogue.

The next phase of development concentrates on extending functionality and content, improving the quality and coverage of service curation, and integration with other systems. Support for tagging with community-curated ontologies will be extended. The myGrid ontology is already used and the EMBRACE project's EDAM ontology (http://sourceforge.net/projects/edamontology/) is under review.

Contributions will be made easier by the release of a write-API, providing members with the ability to register and update services programmatically. Consequently, profiles derived from other service-using and monitoring software, like the Taverna workflow system and its Web Service workflow library myExperiment (http://www.myexperiment.org/), and the service monitoring systems of BioMOBY, DAS and QBIOS (http://qbios.gforge.inria.fr/) will be integrated to form aggregated profiles.

The BioCatalogue aims to satisfy the needs of service providers, users and experts in the field, bringing them together in a common effort to make Web Services for biology more visible, better documented and easier to use. It is an important 'one stop shop' where users can locate Web Services that implement the analysis relevant for their experiments, learn how these services work and, most importantly, learn how to make the most of these valuable resources.

ACKNOWLEDGEMENTS

Authors would like to thank all members of the BioCatalogue focus group, Strategy Advisory Board and other people who help us to improve the registry: Duncan Hull, Benjamin Good, Chrysanthi Ainali, Olivier Sallou, Chris Rawlings, Anil Wipat, Jo Dicks, Robert Gill, Steve Kemp, Antoine H.C. van Kampen, Holger Lausen, Terry Payne, Mark Wilkinson, Janusz Bujnicki, Paul Gordon, Khalid Belhajjame, Philip McDermott, Dave De Roure and all participants of annotation jamborees. Special acknowledgments are given to our partners and all projects that cooperate with BioCatalogue to ease and popularize usage of Web Services in Life Sciences, especially to EU EMBRACE network, OMII-UK, BioMOBY Central, seekda, myExperiment, myGrid, EU BioSapiens network and NBIC.

FUNDING

Funding for open access charge: Biotechnology and Biological Sciences Research Council (BB/F01046X/1, BB/F010540/1 to BioCatalogue project); the European Commission via the EMBRACE project (LHSG-CT-2004-512092); EMBO (ASTF 338.00-2009 to Development on Search By Data).

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. 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.
3. Benson,G. (2009) Nucleic acids research annual web server issue in 2009. *Nucleic Acids Res.*, **37**, W1–W2.
4. Goble,C., Stevens,R., Hull,D., Wolstencroft,K. and Lopez,R. (2008) Data curation + process curation = data integration + science. *Brief Bioinform.*, **9**, 506–517.
5. Romano,P., Marra,D. and Milanesi,L. (2005) Web services and workflow management for biological resources. *BMC Bioinformatics*, **6(Suppl 4)**, S24.
6. McWilliam,H., Valentin,F., Goujon,M., Li,W., Narayanasamy,M., Martin,J., Miyar,T. and Lopez,R. (2009) Web services at the European Bioinformatics Institute-2009. *Nucleic Acids Res.*, **37**, W6–W10.
7. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–360.
8. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **37**, D5–D15.
9. Kaminuma,E., Mashima,J., Kodama,Y., Gojobori,T., Ogasawara,O., Okubo,K., Takagi,T. and Nakamura,Y. (2010) DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Res.*, **38**, D33–D38.
10. Hull,D., Wolstencroft,K., Stevens,R., Goble,C., Pocock,M.R., Li,P. and Oinn,T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, W729–W732.
11. Altintas,I., Berkley,C., Jaeger,E., Jones,M., Ludascher,B. and Mock,S. (2004) Kepler: an extensible system for design and execution of scientific workflows. In *16th International Conference on Scientific and Statistical Database Management, Proceedings,* pp. 423–424.
12. Wilkinson,M.D. and Links,M. (2002) BioMOBY: an open source biological web services proposal. *Brief. Bioinform.*, **3**, 331–341.
13. Wilkinson,M., Schoof,H., Ernst,R. and Haase,D. (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services the PlaNet exemplar case. *Plant Physiol.*, **138**, 5–17.
14. Prlic,A., Down,T.A., Kulesha,E., Finn,R.D., Kahari,A. and Hubbard,T.J. (2007) Integrating sequence and structural biology with DAS. *BMC Bioinformatics*, **8**, 333.
15. Pettifer,S., Thorne,D., McDermott,P., Attwood,T., Baran,J., Bryne,J.C., Hupponen,T., Mowbray,D. and Vriend,G. (2009) An active registry for bioinformatics web services. *Bioinformatics*, **25**, 2090–2091.
16. Gessler,D.D.G., Schiltz,G.S., May,G.D., Avraham,S., Town,C.D., Grant,D. and Nelson,R.T. (2009) SSWAP: a simple semantic web architecture and protocol for semantic web services. *BMC Bioinformatics*, **10**, 309.
17. Lord,P., Alper,P., Wroe,C. and Goble,C. (2005) Feta: a light-weight architecture for user oriented semantic service discovery. In *The Semantic Web: Research and Applications,* Vol 3532/2005 of *Lect. Notes Compur Sci,*. Springer, Berlin/Heidelberg, pp. 17–31.
18. Vitvar,T., Bournez,C., Farrell,J. and Kopeck,J. (2007) SAWSDL: semantic annotations for WSDL and XML schema. *IEEE Internet Comput.*, **11**, 60–67.

19. Sheth,A.P., Gomadam,K. and Lathem,J. (2007) SA-REST: semantically interoperable and easier-to-use services and mashups. *IEEE Internet Computing*, **11**, 91–94.
20. Wolstencroft,K., Alper,P., Hull,D., Wroe,C., Lord,P.W., Stevens,R.D. and Goble,C.A. (2007) The mygrid ontology: bioinformatics service discovery. *Int. J. Bioinform. Res. Appl.*, **3**, 303–325.
21. Valentin,F., Squizzato,S., Goujon,M., McWilliam,H., Paern,J. and Lopez,R. (2010) Fast and efficient searching of biological data resources–using EB-eye. *Brief. Bioinform.*, February 11, 2010 [Epub ahead of print;doi:10.1093/bib/bbp065].
22. Lord,P., Bechhofer,S., Wilkinson,M.D., Schiltz,G., Gessler,D., Hull,D., Goble,C. and Lincoln,S. (2004) Applying semantic Web Services to bioinformatics: experiences gained, lessons learnt. In *International Semantic Web Conference,* Vol. 3298/2004 of *Lecture Notes in Computer Science*. Springer, Berlin/Heidelberg, pp. 350–364.
23. Senger,M., Rice,P. and Oinn,T. (2003) Soaplab-A Unified Sesame Door to Analysis Tools. In Cox,S.J. (ed.), *Proceedings, UK e-Science, All Hands Meeting, 2–4 September*. Nottingham, UK, pp. 509–513.
24. Reich,M., Liefeld,T., Gould,J., Lerner,J., Tamayo,P. and Mesirov,J.P. (2006) Genepattern 2.0. *Nat. Genet.*, **38**, 500–501.