Published online 08 April 2015 *Nucleic Acids Research, 2015, Vol. 43, Web Server issue* ***W585–W588***
doi: 10.1093/nar/gkv316

The EBI Search engine: providing search and retrieval functionality for biological data from EMBL-EBI

Silvano Squizzato, Young Mi Park, Nicola Buso, Tamer Gur, Andrew Cowley, Weizhong Li, Mahmut Uludag, Sangya Pundir, Jennifer A. Cham, Hamish McWilliam and Rodrigo Lopez*

European Bioinformatics Institute, EMBL Outstation, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, Cambridge, UK

Received January 30, 2015; Revised March 18, 2015; Accepted March 28, 2015

ABSTRACT

The European Bioinformatics Institute (EMBL-EBI—https://www.ebi.ac.uk) provides free and unrestricted access to data across all major areas of biology and biomedicine. Searching and extracting knowledge across these domains requires a fast and scalable solution that addresses the requirements of domain experts as well as casual users. We present the EBI Search engine, referred to here as 'EBI Search', an easy-to-use fast text search and indexing system with powerful data navigation and retrieval capabilities. API integration provides access to analytical tools, allowing users to further investigate the results of their search. The interconnectivity that exists between data resources at EMBL-EBI provides easy, quick and precise navigation and a better understanding of the relationship between different data types including sequences, genes, gene products, proteins, protein domains, protein families, enzymes and macromolecular structures, together with relevant life science literature.

INTRODUCTION

The European Bioinformatics Institute (EBI) (1) hosts data from life science experiments comprising assembled genomes; nucleotide sequences; protein sequences; macromolecular structures; small ('drug-like') molecules; gene expression; molecular interactions; reactions, pathways and diseases; protein families; enzymes; literature; and samples and ontologies. These represent discrete categories containing one or more specialised data resources that are curated and annotated by experts from around the world. Searching for biological information within and across these resources is a challenge. In this article we discuss how EBI Search (previously known as 'EB-eye') (2) provides a solution that is fast and scalable, that allows users to query and review results using faceted navigation (3) and filters based on common fields. Users can move seamlessly from the search into specialised web portals where more detail and functionality is available. The search engine is built using the Apache Lucene library (http://lucene.apache.org) and is constantly updated with new data and is under continuous review by scientists as well as specialists in web usability and design. In 2010, EBI Search had 400 million entries. During 2014 it surpassed one billion entries, which are accessible over the web as well as programmatically using SOAP and RESTful Web Services. During 2014 EBI Search was used by more than 393 000 unique Internet Protocols (IPs) addresses that generated 290 million requests.

DATA COVERAGE

EBI Search provides a uniform and consistent search and retrieval functionality spanning many individual data resources, split into thematic categories (Table 1). For example, in the '*Nucleotide sequences*' category all ENA (4) data collections are represented together with data from RNAcentral (5); in the '*Protein sequences*' category users can find UniprotKB and UniParc (6); '*Protein families*' contains InterPro (7), '*Genomes*' includes Ensembl (8) and Ensembl Genomes (9); '*Gene expression*' includes the baseline and differential Expression Atlases (10), '*Macromolecular structures*' includes PDBe (11); '*Small molecules*' contains ChEBI (12) and ChEMBL (13); and '*Reactions, pathways and diseases*' includes OMIM (http://omim.org), Reactome (14) and Rhea (15). A complete list of data resources is available from https://www.ebi.ac.uk/ebisearch/aboutebisearch.ebi.

EBI Search is automatically updated. This is triggered by a set of scripts that monitor the production cycles of the source data resources, ensuring search results are always up-to-date.

WEB INTERFACE

The main entry points to EBI Search are bespoke search boxes, found throughout the website, into which users can type simple phrases, database identifiers, keywords, gene

*To whom correspondence should be addressed. Tel: +44 0 1223 494 423; Fax: +44 0 1223 494 468; Email: rls@ebi.ac.uk

Table 1. Data resources available through EBI Search

Category	Data resources
Genomes	Ensembl Genomes, Ensembl, HGNC, PomBase, DGVa, EGA, LRG, WormBase ParaSite
Nucleotide sequences	ENA, RNAcentral, NRNL1, NRNL2, IMGT/HLA, IPD-KIR, IPD-MHC
Protein sequences	UniProtKB, UniParc, UniRef, EPO, JPO, KIPO, USPTO, NRPL1, NRPL2
Macromolecular structures	PDBe, EMDB
Small molecules	ChEBI, ChEMBL, Ligands
Gene expression	ArrayExpress, Expression Atlases
Molecular interactions	IntAct
Reactions, pathways and diseases	Rhea, Reactome, BioModels, MetaboLights, OMIM
Protein families	InterPro, TreeFam, MEROPS, GPCRDB
Enzymes	IntEnz
Literature	MEDLINE, Patent families, Patents
Samples and ontologies	Taxonomy, GO, EFO, SBO, MESH, BioSamples

symbols, species, and molecule and disease names. This is aided by auto-complete, which suggests terms based on indexed content in the system. Search queries can be single or multiple terms combined with Boolean logic (e.g. OR, AND, NOT), and expansion of terms using wildcard characters is supported. The query syntax of the EBI Search engine follows Apache Lucene query parser syntax and its implementation is explained in detail in the help pages: https://www.ebi.ac.uk/ebisearch/documentation.ebi.

Search results pages

EBI Search executes a query against a vast amount of indexed data, so one challenge is how to present results in a coherent and intuitive way. Search results are organised into the aforementioned biological categories (e.g. '*Genomes*', '*Nucleotide sequences*', '*Gene Expression*', etc.). Within each category, the top ranking results are presented with the option for the user to expand any category of interest to see all matches. This overview also shows the number of results by category, which helps users identify data resources of interest. Typically, each entry on a query result page displays primary identifiers hyperlinked to the main data resource web portal. Additionally, titles, names and descriptions are shown.

Database cross-references are available via a '*Related data*' button. The relationships between entries in different data resources can be found by navigating through cross-references, which in EBI Search can be implicitly declared by the provider or inferred by the system. A '*Views*' button provides access to alternative formats of the data (e.g. EMBL format for ENA entries, PDB format, etc.) served via the dbfetch application (16). This button also provides access to analytical tools, such as NCBI BLAST+ (17) and InterProScan 5 (18), which apply to protein sequences in the results.

Custom fields are provided for some data resources. For example, UniProtKB query results contain primary and secondary accession numbers, IDs, names, description, species and review status. Results from the '*Literature*' category contain titles, author lists, journals and publication dates.

Facets help the user filter and narrow down results

The available facets of a search result are presented on the left-hand side when users select a data resource or a category. The text descriptions are followed by check boxes or filtering links for selecting results according to data-specific attributes such as taxonomy, keywords and controlled vocabularies. As an example, search results in UniProtKB can be filtered using the '*Organisms*' facet, and reviewed entries (i.e. UniProtKB_SwissProt) can be selected by using the '*Status*' facet (e.g. 'Reviewed' or 'Unreviewed'). Not all facets are keyword-based, for example, the '*Type'* facet in InterPro results; some facets represent ranges, for instance, the '*Publication date*' in the '*Literature*' category. Common and custom facets are shown in Table 2.

Automatic generation of human-readable 'Gene & protein summaries'

'*Gene & protein summaries*' are available at the top of the main results page when queries contain established gene names (i.e. HGNC gene nomenclature) or common database identifiers in Ensembl or UniProtKB (i.e. accession numbers). These summaries are organised into five sections presented as tabs, namely: gene, expression, protein, protein structure and literature, and apply to the following model organisms: *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana* and *Escherichia coli K-12*. These gene-centric summaries are generated by a separate application (https://www.ebi.ac.uk/s4/), which uses EBI Search's SOAP programmatic interface to identify, retrieve and display data from the main resources portals using the DAS protocol (19).

Programmatic access to search functionality via Web Services

In addition to SOAP, a new RESTful Web Service API has been available since June 2014 in response to demand for search services from developers using the latest web technologies and wanting to integrate search functionality into their portals. These Web Services provide methods, which can be grouped into three main types: '*meta-data*' (i.e. retrieving information about searchable data resources), '*search and retrieval*' functionality, '*navigation*' (i.e. explor-

Table 2. Custom and common facets available through EBI Search

Category	Common facets	Custom facets
Genomes	Organisms	
Nucleotide sequences	Organisms	Genomic mapping, Expert databases, RNA types (in RNAcentral)
Protein sequences	Organisms	Keywords and Status (in UniProtKB)
Macromolecular structures	Organisms	
Gene expression	Organisms	Organism part (in Expression Atlases)
Reactions, pathways and diseases	Organisms	Type, Compartment name and Keywords (in Reactome)
Protein families		Type (in InterPro)
Literature	Publication date	
Samples and Ontologies		Ontology (in GO)

ing cross-references) and '*filtering*' (i.e. narrowing down results using facets). Further details and sample clients can be found in the documentation pages at https://www.ebi.ac.uk/Tools/webservices/services/eb-eye_rest.

Since the launch of EBI Search in 2007, search functionality has been integrated into projects such as ENA, Ensembl Genomes, InterPro, LRG (20), Rhea, MetaboLights (21), Enzyme Portal (22) and PomBase (23). Novel pipeline processes or analytical workflows can be created by combining methods from EBI Search and other Web Services. For example, entry identifiers from UniProtKB can be sent to the dbfetch Web Service (WSDbfetch), which retrieves the corresponding sequence entries in batch fashion. These sequence entries can in turn be sent to analytic tools Web Services (16) such as Clustal Omega (24). The current version of the APIs covers the existing functionality available in the web interface of EBI Search, including facets and autocomplete. Search result formats include: XML, JSON, CSV and TSV, enabling integration of results into third-party frameworks such as AngularJS (https://angularjs.org). An example of such integration is the RNAcentral portal (http://www.rnacentral.org) launched in 2014 (5).

FUTURE DIRECTIONS

As the volume of data and the number of data resources continue to grow, providing continuous search functionality is a big challenge. This will be achieved by improving code and simplifying configuration and hardware requirements, analysing user queries and exploring novel technologies. Improving the user experience is a central focus based on user-centred design techniques. Users will be able to select results and download these in formats that can be consumed programmatically for further post-processing (e.g. for further analysis in local pipelines and workflows). In addition to the previously mentioned formats, RSS 2 (http://www.rssboard.org/rss-specification) will be available to help users pre-generate queries, which can be repeated over time to check for new or updated data. These bespoke alerting systems can be enacted using widely available RSS clients or built-in browser tools. Methods to simplify the launching of applications such as BLAST from search results are also being tested that use the tools provided by the Job Dispatcher framework (25). Lastly but importantly, the SOAP API will be phased out during 2016 in order to concentrate resources on the RESTful interface, which is easier to use and more scalable.

DISCUSSION

EBI Search is built on top of technologies that allow fast indexing and searching of vast amounts of data. The implementation of a scalable search system relies on the quick and efficient uptake of the latest technologies and also in their successful integration into existing compute infrastructures with little or negligible cost. The EBI Search engine has been successful at acting as both a 'global search' that presents results across many distinct knowledge domains and as a 'local search' thanks to the implementation of industry standard Web Services. Keeping up-to-date with changes in search technologies as well as with changes in the underlying data is challenging. However, this drives forward the development of simpler, more responsive and more efficient methods of finding and re-using biological information. By providing direct access to the primary data sources (web portals), where biological entities are fully annotated and displayed in the way expected by specialists, the EBI Search engine can provide access to a large range of data sources for a cheaper cost than multiple search engines. EBI Search must not be confused with an integration platform or a data warehouse. It enables interoperability between distinct underlying data resources and analytical tools, ultimately delivering a powerful and reproducible way to interpret biological search results.

ACKNOWLEDGEMENTS

We would like to thank the following staff and collaborators for their support: Anton Petrov, Dietmar Sturmayr, Francis Rowland, Philip Lewis, Andrea Cristofori and Simone Badoer.

FUNDING

BBSRC Award 'The RNAcentral database of non-coding RNAs'; BB/J019232/1; European Molecular Biology Laboratory (EMBL). Funding for open access charge: European Molecular Biology Laboratory (EMBL).
Conflict of interest statement. None declared.

REFERENCES

1. Brooksbank,C., Bergman,M.T., Apweiler,R., Birney,E. and Thornton,J. (2014) The European Bioinformatics Institute's data resources 2014. *Nucleic Acids Res.*, **42**, D18–D25.
2. 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.*, **11**, 375–384.

3. Russell-Rose,T. and Tate,T. (2012) *Designing the Search Experience - The Information Architecture of Discovery*, Morgan Kaufmann, Waltham, MA.
4. Silvester,N., Alako,B., Amid,C., Cerdeño-Tárraga,A., Cleland,I., Gibson,R., Goodgame,N., Ten Hoopen,P., Kay,S., Leinonen,R. *et al.* (2015) Content discovery and retrieval services at the European Nucleotide Archive. *Nucleic Acids Res.*, **43**, D23–D29.
5. The RNAcentral Consortium. (2015) RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Res.*, **43**, D123–D129.
6. UniProt Consortium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.
7. Mitchell,A., Chang,H.Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
8. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
9. Kersey,P.J., Allen,J.E., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C., Hughes,D.S., Humphrey,J., Kerhornou,A., Khobova,J. *et al.* (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.*, **42**, D546–D552.
10. Petryszak,R., Burdett,T., Fiorelli,B., Fonseca,N.A., Gonzalez-Porta,M., Hastings,E., Huber,W., Jupp,S., Keays,M., Kryvych,N. *et al.* (2014) Expression Atlas update–a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42**, D926–D932.
11. Gutmanas,A., Alhroub,Y., Battle,G.M., Berrisford,J.M., Bochet,E., Conroy,M.J., Dana,J.M., Fernandez Montecelo,M.A., van Ginkel,G., Gore,S.P. *et al.* (2014) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **42**, D285–D291.
12. Hastings,J., de Matos,P., Dekker,A., Ennis,M., Harsha,B., Kale,N., Muthukrishnan,V., Owen,G., Turner,S., Williams,M. and Steinbeck,C. (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.
13. Gaulton,A., Bellis,L.J., Bento,A.P., Chambers,J., Davies,M., Hersey,A., Light,Y., McGlinchey,S., Michalovich,D., Al-Lazikani,B. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.
14. Croft,D., Mundo,A.F., Haw,R., Milacic,M., Weiser,J., Wu,G., Caudy,M., Garapati,P., Gillespie,M., Kamdar,M.R. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, D472–D477.
15. Morgat,A., Axelsen,K.B., Lombardot,T., Alcántara,R., Aimo,L., Zerara,M., Niknejad,A., Belda,E., Hyka-Nouspikel,N., Coudert,E. *et al.* (2015) Updates in Rhea-a manually curated resource of biochemical reactions. *Nucleic Acids Res.*, **43**, D459–D464.
16. McWilliam,H., Li,W., Uludag,M., Squizzato,S., Park,Y.M., Buso,N., Cowley,A.P. and Lopez,R. (2013) Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.*, **41**, W597–W600.
17. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics.*, **10**, 421.
18. Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics.*, **30**, 1236–1240.
19. Jenkinson,A.M., Albrecht,M., Birney,E., Blankenburg,H., Down,T., Finn,R.D., Hermjakob,H., Hubbard,T.J., Jimenez,R.C., Jones,P. *et al.* (2008) Integrating biological data–the Distributed Annotation System. *BMC Bioinformatics.*, **9**, S3.
20. MacArthur,J.A., Morales,J., Tully,R.E., Astashyn,A., Gil,L., Bruford,E.A., Larsson,P., Flicek,P., Dalgleish,R., Maglott,D.R. *et al.* (2014) Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Res.*, **42**, D873–D878.
21. Haug,K., Salek,R.M., Conesa,P., Hastings,J., de Matos,P., Rijnbeek,M., Mahendraker,T., Williams,M., Neumann,S., Rocca-Serra,P. *et al.* (2013) MetaboLights–an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.*, **41**, D781–D786.
22. Alcántara,R., Onwubiko,J., Cao,H., Matos,P.D., Cham,J.A., Jacobsen,J., Holliday,G.L., Fischer,J.D., Rahman,S.A., Jassal,B. *et al.* (2013) The EBI enzyme portal. *Nucleic Acids Res.*, **41**, D773–D780.
23. Wood,V., Harris,M.A., McDowall,M.D., Rutherford,K., Vaughan,B.W., Staines,D.M., Aslett,M., Lock,A., Bähler,J., Kersey,P.J. *et al.* (2012) PomBase: a comprehensive online resource for fission yeast. *Nucleic Acids Res.*, **40**, D695–D699.
24. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539–544.
25. Goujon,M., McWilliam,H., Li,W., Valentin,F., Squizzato,S., Paern,J. and Lopez,R. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, **38**, W695–W699.