*doi: 10.1093/nar/gkx359*

# The EBI search engine: EBI search as a service—making biological data accessible for all

**Young M. Park, Silvano Squizzato, Nicola Buso, Tamer Gur and Rodrigo Lopez***

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

Received January 27, 2017; Revised April 11, 2017; Editorial Decision April 18, 2017; Accepted April 20, 2017

## ABSTRACT

**We present an update of the EBI Search engine, an easy-to-use fast text search and indexing system with powerful data navigation and retrieval capabilities. 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 that include nucleotide and protein sequences, genes, gene products, proteins, protein domains, protein families, enzymes and macromolecular structures, as well as the life science literature. EBI Search provides a powerful RESTful API that enables its integration into third-party portals, thus providing 'Search as a Service' capabilities, which are the main topic of this article.**

## INTRODUCTION

During 2016, more than 362 000 unique IPs accessed EBI Search from across the globe. These give rise to more than 300 million searches, comprising the web as well as usage of the RESTful API, and representing more than 1 TB of metadata downloaded that links search results with more than 1.3 billion records held in the databases at the EMBL–EBI. At present, the system can re-index all these data in less than 24 h.

There is a continuous collaboration between the developers of EBI Search and its users. Working together, the search functionality can now be integrated into third-party portals in novel, intuitive and useful ways. The concept of 'Search as a Service' (https://en.wikipedia.org/wiki/Search_as_a_service) is not new but the implementation presented here is novel and unique, and opens up further opportunities for collaboration, removing the need for developing and maintaining different search applications. These collaborations often result in the development of new features to EBI Search, which are potentially useful for other projects and ultimately benefit scientific findability, reproducibility and discoverability.

## DATA COVERAGE

EBI Search (1) provides comprehensive access to all data maintained in the public repositories hosted at the EMBL–EBI, European Molecular Biology Laboratory, European Bioinformatics Institute (2). These are divided into biological themes, as shown in Table 1. Since last reported, new data resources include the Enzyme Portal (3), providing protein- and enzyme-centric views; Rfam (4) RNA families; Ensembl Genome Variants (5); Elixir Tools Registry (6); Omics Discovery Index (OmicsDI) (https://doi.org/10.1101/049205); the EBI Metagenomics portal (7) and annotations from the InterPro protein domain and families consortium (8). Data resources have also been deprecated that include ENA (9), Whole Genome Shotgun sequences and Transcriptome Shotgun Assemblies, however, master projects remain in order to provide access to these data via the ENA (European Nucleotide Archive) portal.

EBI Search retrieves and indexes data in native XML, EBI search XML schema and various types of text files, including flat files. In addition to these, JSON format is now supported using a new schema (http://www.ebi.ac.uk/ebisearch/schemas/data_schema.json) to give data providers more flexibility.

## EBI SEARCH AS A SERVICE

The concept of EBI Search as a Service defines a novel paradigm in the use of the EBI Search API (Application Programming Interface) to create complex views of inter-related data, thus allowing users to combine and create different views. Examples of this can be seen on the OmicsDI (http://www.omicsdi.org/) portal that integrates 11 data sources relating to transcriptomics, genomics, proteomics and metabolomics; the EBI Metagenomics portal is another example that provides fast and uniform access to project, samples and individual runs, and relates these to taxonomic assignments, allowing users to explore some 80 000 metagenomic datasets from more than 140 biomes. Finally, EBI Search as a Service is used to provide fast lookup of annotations, as in the EBI HMMER service (https://www.ebi.ac.uk/Tools/hmmer).

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

**Table 1.** Data resources available through EBI Search in 2016–2017

| Category | Data resources |
|---|---|
| Genomes and metagenomes | Ensembl Genomes, Ensembl, HGNC, PomBase, DGVa, EGA, LRG, WormBase ParaSite, Metagenomics |
| 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, MetabolomeExpress, Metabolomics Workbench |
| Protein families | InterPro, TreeFam, Pfam, TreeFam, MEROPS, GPCRDB |
| Protein expression data | PRIDE, GNPS, GPMdb, MassIVE, PeptideAtlas |
| Enzymes | IntEnz, Enzyme Portal |
| Literature | MEDLINE, Patent families, Patents |
| Samples and ontologies | Taxonomy, GO, EFO, SBO, MESH, BioSamples, Elixir registry |

Yet another example is the EBI Search web interface itself: the traditional graphical user interface (GUI) has now been replaced by a RESTful client developed using Angular (https://angular.io), a popular JavaScript web development platform. The new GUI adopts the EMBL–EBI Visual Framework (https://github.com/ebiwd/EBI-Framework), this helps with keeping a uniform user experience across the institute's data resources and implements Responsive Design (http://www.w3schools.com/html/html_responsive.asp), thus providing for a better experience across different devices.

In order to document the RESTful API of EBI Search and help developers design and build web applications that manage complex biological concepts, a Swagger (http://swagger.io), an OpenAPI specification-compliant (https://www.openapis.org/specification/repo) interface has been provided that is available from: https://www.ebi.ac.uk/ebisearch/swagger.ebi. An advantage of using this technology is that it markedly improves the APIs accessibility.

As mentioned above, there are good, functional examples of portals that use the EBI Search as a Service to search, retrieve and display complex data. Table 2 lists 16 portals consuming the service API at the time of writing.

## IMPROVED SEARCH FINDABILITY AND USABILITY

Newly added features to the web interface improve findability and discoverability: Query Builder; saving results for later re-use; launching of bioinformatic tools and using RSS feeds for alerts. Also, user experience is improved by the new GUI framework.

Query Builder, which is a replacement of the advanced search page, helps users to build complex queries with an intuitive GUI. It provides a list of available data resources that users can start searching from, allowing them to combine several fields, using Boolean criteria, to build a complex query. Users can trigger searching and save the query for re-use or generate an RSS feed from it, as will be discussed later.

The EBI Search GUI now provides the ability to save search results on the client side, which can later be used for further analysis. There is now a 'Save result' button, which downloads a subset or all of the current search results in machine-readable formats such as XML (eXtensible Markup Language), JSON (JavaScript Object Notation), TSV |(Tab Separated Values) and CSV (Comma Separated Values) . The save function is built using the RESTful API, thus the generated URLs can be used in scripts or analysis pipelines. The number of entries that can be retrieved in a single operation is 100; more can be retrieved by using 'pagination' (i.e. specifying ranges of results with a maximum number of 100 between operations).

Launching bioinformatic tools with a selected subset of nucleotide or protein sequence query results is now possible. The list of tools varies depending on a selected database. For example, BLAST+ (10), Clustal Omega (11), UniSave lookup (12) and Dbfetch (13) are available for results pertaining to UniprotKB (14). When launching a tool, users will be taken to that tool's web page, with the relevant identifiers pre-filled.

EBI Search provides an alerting mechanism to help users create queries that can be used to find new or updated data. Creating RSS (http://www.rssboard.org/rss-specification) query alerts is possible from the Query Builder page and all data resource results pages. Alert examples can be found in the documentation page 'https://www.ebi.ac.uk/ebisearch/documentation.ebi'.

The new GUI loads data in an asynchronous manner, a behavior that has improved user experience during testing. Because JavaScript libraries are required by the client, the browser memory footprint has increased slightly but this is offset by faster rendering of the search results. Furthermore, a cache server has been implemented into the framework to provide high levels of query responsiveness. The move toward a JavaScript framework will make it easier to adapt to changing web practices.

## ENHANCED RESTFUL API

In the previous paper (2) the basic RESTful API was described. Feedback from users has prompted enhancements that include hierarchical facets and 'more like this' results.

Faceting is an efficient way for users to filter search results. However, some data, e.g. taxonomy (15), are difficult to fit in a single-level structure facet. The Metagenomics portal taxonomy data are now searchable and filterable as a hierarchical facet through the RESTful API, meaning it is

**Table 2.** List of resources currently consuming the EBI Search through its RESTful API

| Resource | URL/reference | Format used |
|---|---|---|
| ENA | https://www.ebi.ac.uk/ena/ | XML |
| Ensembl Genomes | https://www.ensemblgenomes.org/ | JSON |
| InterPro | https://www.ebi.ac.uk/interpro/ | XML |
| Expression Atlas | https://www.ebi.ac.uk/gxa/ | XML |
| LRG | http://www.lrg-sequence.org/ | XML |
| Job Dispatcher | http://europepmc.org/abstract/MED/25845596 | JSON/XML |
| RNAcentral | http://rnacentral.org/ | XML |
| MetaboLights | https://www.ebi.ac.uk/metabolights/ | XML |
| Enzyme Portal | https://www.ebi.ac.uk/enzymeportal/ | JSON |
| OmicsDI | https://www.omicsdi.org | JSON |
| PomBase | http://www.pombase.org/ | XML |
| Gene & Protein Summaries | https://www.ebi.ac.uk/s4/ | XML |
| WormBase Parasite | http://www.wormbase.org/ | XML |
| HMMER | https://www.ebi.ac.uk/Tools/hmmer/ | JSON/XML |
| Metagenomics | https://www.ebi.ac.uk/metagenomics/ | JSON/XML |
| identifiers.org | http://www.identifiers.org/ | XML |
| EBI Search web interface | https://www.ebi.ac.uk/ebisearch/ | JSON |

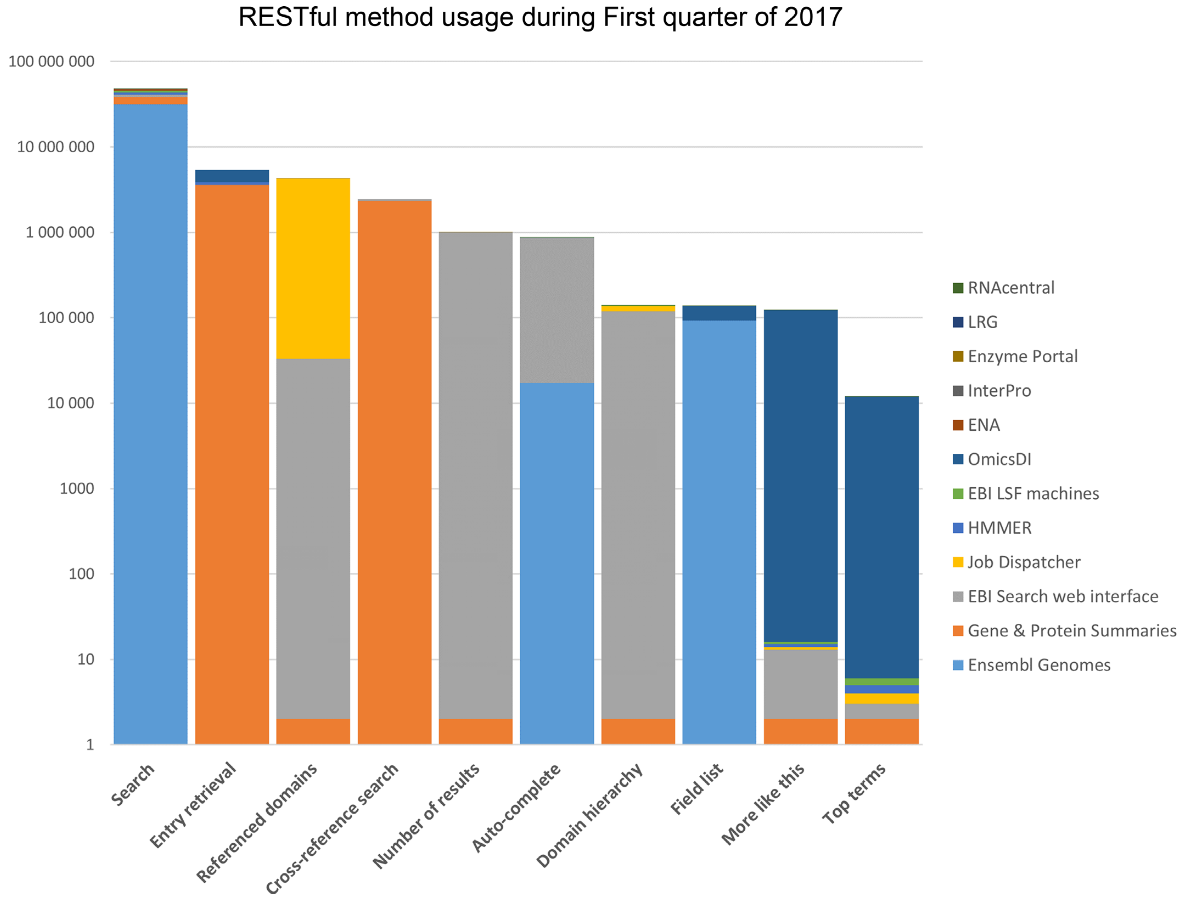

**Figure 1.** Relative use of the RESTful API methods during first quarter 2017.

now possible to navigate results across this taxonomic classification.

Another new feature, called 'more like this,' developed in collaboration with the OmicsDI team, has been introduced in the RESTful interface. Starting from an entry, the new API returns a list of similar entries. These are related by terms they have in common and which are extracted from descriptive fields according to specific criteria, such as the expected frequency of a term. The selected terms form a new query that is used to search against the same database as the original entry or against other database(s). In the OmicsDI website these types of result appear under the heading 'Similar Datasets.'

## DESCRIBING CONTENTS AND MONITORING (SEARCH ANALYTICS)

As the importance of EBI Search as a Service is growing, the users need to know about the status of databases. The main information page of the system (https://www.ebi.ac.uk/ebisearch) provides graphical overviews of the most popular terms that originate through web search boxes and the relative size of each data resource. Below these, there is a list of collaborators followed by a table of data resources that shows the current domain classification, its name, category, query on, number of entries and dates of indexing, last updated and release. This information is also available over the RESTful Web Services.

To better understand how the search engine is used by web and Web Services users, there are dedicated systems that monitor and analyse requests using Elastic Stack (http://www.elastic.co). These measure the volume of incoming and outgoing traffic, the relative use of the RESTful API methods and the source (Figure 1). Usage patterns can be captured and later used for diagnosing problems, generating usage statistics and finally, making improvements and enhancements.

## FUTURE DIRECTIONS

There is no doubt that the continuous growth of data in EBI Search will present challenges. From the data perspective, there will be more collaboration with data providers, which should bring more content to index and produce enriched views on the data. Also, future collaboration with ChEMBL (16) will bring improvement in search and results in the cheminformatics resources. With user-centered design techniques, the web interface will be reviewed in order to expose more information to the end users. Similarly, better error handling is required. In the RESTful API, more result formats, such as lists of database identifiers, which can work directly as input into biological analysis workflows will be developed.

## DISCUSSION

The implementation of industry standards, such as RESTful Web Services, has permitted the development of a scalable search infrastructure that provides fast and efficient access to a diverse and complex set of biological data. Keeping up-to-date with proteomics data generation, the output of high-throughput sequencing technologies, growing literature resources, ontologies and specialized taxonomies, is a challenge. Providing direct access to specialist data portals as well as presenting views on combined complex data is high on the development agenda and must happen in close collaboration with the data providers. In this context, the concept of EBI Search as a Service, has allowed third-parties to integrate and develop powerful search functionality that can be used by all. This has several advantages: sharing of search and result components; avoiding duplication of effort; establishing common syntax for searching, and finally, improving scientific findability, reproducibility and discoverability.

## ACKNOWLEDGEMENTS

The authors wish to acknowledge the following: Simone Badoer, Andrea Cristofori and Philip Lewis for web administration support; Yasset Perez Riverol, for the development of ideas and algorithms in the context of the OmicsDI portal; Anton Petrov, Matloob Qurechi, Simon Potter and Gift Nuka for input in relation to RNAcentral, InterPro and HMMER; Joseph Onwubiko for input to the Enzyme Portal and Peter Walter and Ken Hawkins for UX on the new GUI. Finally, we would like to thank all EMBL–EBI teams involved in providing data to the EBI Search.

## FUNDING

European Molecular Biology Laboratory (EMBL); National Institutes of Health [R01-GM-094231]; European Commission [676559]; Funding for open access charge: EMBL.

*Conflict of interest statement.* None declared.

## REFERENCES

1. Squizzato,S., Park,Y.M., Buso,N., Gur,T., Cowley,A., Li,W., Uludag,M., Pundir,S., Cham,J.A., McWilliam,H. *et al.* (2015) The EBI Search engine: providing search and retrieval functionality for biological data from EMBL-EBI. *Nucleic Acids Res.*, **43**, W585–W588.
2. 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.
3. Alcántara,R., Onwubiko,J., Cao,H., Matos,Pd., Cham,J.A., Jacobsen,J., Holliday,G.L., Fischer,J.D., Rahman,S.A., Jassal,B. *et al.* The EBI enzyme portal. *Nucleic Acids Res.*, **41**, d773–d780.
4. Eric P,Nawrocki., Sarah W,Burge., Alex,Bateman., Jennifer,Daub, Ruth Y,Eberhardt., Sean R,Eddy., Evan W,Floden., Paul P,Gardner., Thomas A,Jones., John,Tate and Robert D,Finn. (2014) Rfam 12.0: updates to the RNA families database. Nucleic Acids Research. *Nucleic Acids Res.*, **43**, D130–D137.
5. Kersey,P.J., Allen,J.E., Armean,I., Boddu,S., Bolt,B.J., Carvalho-Silva,D., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C. *et al.* (2016) Ensembl Genomes: more genomes, more complexity. *Nucleic Acids Res.*, **44**, D574–D580.
6. Ison,J., Rapacki,K., Ménager,H., Kalaš,M., Rydza,E., Chmura,P., Anthon,C., Beard,N., Berka,K., Bolser,D. *et al.* Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Res.*, **44**, D38–D47.
7. Mitchell,A., Bucchini,F., Cochrane,G., Denise,H., ten Hoopen,P., Fraser,M., Pesseat,S., Potter,S., Scheremetjew,M., Sterk,P. *et al.* EBI metagenomics in 2016–an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.*, **44**, d595–d603.
8. Finn,R.D, Attwood,T.K, Babbitt,P.C, Bateman,A., Bork,P., Bridge,A.J, Chang,H.Y., Dosztányi,Z., El-Gebali,S., Fraser,M. *et al.* InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**, D190–D199.
9. Toribio,A.L., Alako,B., Amid,C., Cerdeño-Tarrága,A., Clarke,L., Cleland,I., Fairley,S., Gibson,R., Goodgame,N., Ten Hoopen,P. *et al.* (2016) European Nucleotide Archive in 2016. *Nucleic Acids Res.*, **45**, D32–D36.
10. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421–430.
11. 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.
12. Leinonen,R., Nardone,F., Zhu,W. and Apweiler,R. (2006) UniSave: the UniProtKB Sequence/Annotation Version database. *Bioinformatics*, **22**, 1284–1285.

13. Li,W., Cowley,A., Uludag,M., Gur,T., McWilliam,H., Squizzato,S., Park,Y.M., Buso,N. and Lopez,R. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.*, **43**, w580–w584.
14. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
15. Federhen,S. (2011) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
16. Gaulton,A., Hersey,A., Nowotka,M., Bento,A.P., Chambers,J., Mendez,D., Mutowo,P., Atkinson,F., Bellis,L.J., Cibrián-Uhalte,E. *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954.