Published online 11 May 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W155–W160***
doi:10.1093/nar/gkq331

The Ontology Lookup Service: bigger and better

Richard Côté[1,*], Florian Reisinger[1], Lennart Martens[2,3], Harald Barsnes[4], Juan Antonio Vizcaino[1] and Henning Hermjakob[1]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK, [2]Department of Medical Protein Research, VIB, Ghent, Belgium, [3]Department of Biochemistry, Ghent University, Ghent, Belgium and [4]Department of Informatics, University of Bergen, Norway

Received January 29, 2010; Revised April 4, 2010; Accepted April 17, 2010

ABSTRACT

The Ontology Lookup Service (OLS; http://www.ebi.ac.uk/ols) has been providing several means to query, browse and navigate biomedical ontologies and controlled vocabularies since it first went into production 4 years ago, and usage statistics indicate that it has become a heavily accessed service with millions of hits monthly. The volume of data available for querying has increased 7-fold since its inception. OLS functionality has been integrated into several high-usage databases and data entry tools. Improvements in the data model and loaders, as well as interface enhancements have made the OLS easier to use and capture more annotations from the source data. In addition, newly released software packages now provide easy means to fully integrate OLS functionality in external applications.

INTRODUCTION

Ontologies and controlled vocabularies (CVs) have more than demonstrated their essential function when dealing with large volumes of complex data currently being generated by high-throughput multi-domain analysis techniques (1). They provide a framework around which large data sets can be systematically annotated and queried. For this framework to function efficiently, however, the ontologies and CVs must be made available to the user community.

The Ontology Lookup Service (OLS) has been in production since mid-2005 and has quickly become one of the most accessed services in the Proteomics Services team at the EBI, with monthly usage figures in the millions of hits. This includes both the programmatic as well as the interactive interfaces that the service offers. The OLS has been previously described and readers are invited to refer to the original publication for in-depth information on the technical architecture and data models (2,3).

The core functionality of the OLS has remained largely unchanged since its inception, allowing users to query ontologies and CVs by name or identifier as well as obtaining metadata, such as synonyms, definitions, cross references and other annotations, for a given term. Users can also traverse the relationships between terms. The usability and volume of data captured, however, has been enhanced and this will be expanded below.

The OLS has always been designed to be used in other projects as a means to integrate ontology and CV annotation and query functionality. A SOAP web service has been available since the OLS went into production. A full description of the web service has already been published (2,3) and users who wish to make use of it are encouraged to go to the OLS web service developer section for the most up-to-date documentation and code samples (http://www.ebi.ac.uk/ontology-lookup/WSDLDocumentation.do).

AVAILABLE DATA

The first OLS publication described it as containing 42 ontologies, accounting for roughly 135 000 terms. Over a 4-year period, the data loaded into the OLS has been expanded to 79 ontologies, representing over 971 000 unique terms (Figure 1). These cover far-ranging topics such as model organism anatomy and development, physiology and disease, instrumentation and methods and many others. In the 2 years since the OLS was previously published in NAR, 25 new ontologies have been added (Table 1). Users are encouraged to go online at http://www.ebi.ac.uk/ontology-lookup/ontologyList.do to access a full listing of currently available ontologies and CVs.

The ontologies and CVs loaded in the OLS are maintained by various external groups that are domain experts in their fields. To maintain the OLS as up-to-date as possible with the current state of knowledge, the ontology providers are polled on a daily basis and updated files are downloaded and parsed to update the core OLS database. Currently, the OLS loaders poll six

*To whom correspondence should be addressed. Tel: +44 1223 492 610; Fax: +44 1223 494 468; Email: rcote@ebi.ac.uk

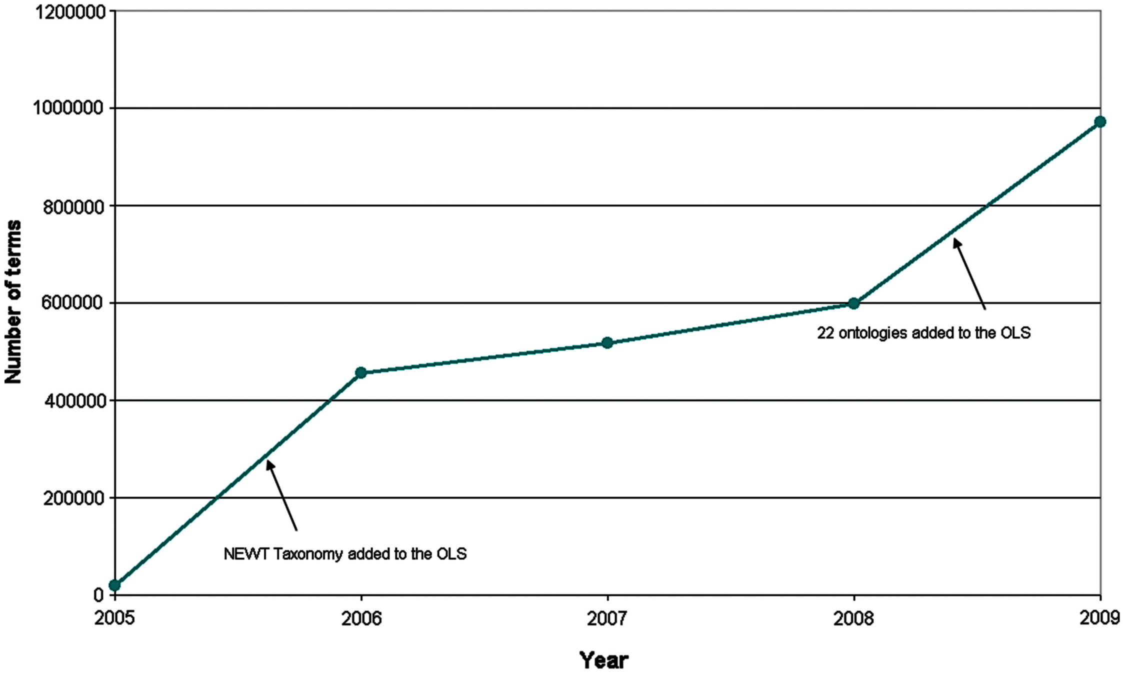


Figure 1. Growth chart of the OLS data content. The amount of data loaded into the OLS, based on unique terms, has shown a 7-fold increase since the service went online. The first large increase is due to the incorporation of the NEWT taxonomy and the second large increase is due to the addition of a large number of ontologies at once.

Table 1. A list of ontologies that have been added to the OLS in the last 2 years

Ontology Prefix	Ontology name
AAO	Amphibian Gross Anatomy Ontology
APO	Yeast Phenotype Ontology
ATO	Amphibian Taxonomy
CCO	Cell Cycle Ontology
EFO	ArrayExpress Experimental Factor Ontology
ENA	European Nucleotide Archive Submission Ontology
FBsp	Flybase Taxonomy
FMA	Foundational Model of Anatomy Ontology
HAO	Hymenoptera Anatomy Ontology
HOM	Homology Ontology
HP	Human Phenotype Ontology
IDO	Infectious Disease Ontology
LSM	Leukocyte Surface Marker Ontology
MIAA	Minimal Information about Anatomy Ontology
MIRO	Mosquito Insecticide Resistance Ontology
MPATH	Mouse Pathology Ontology
MS	Mass Spectrometry Ontology
PAR	Protein Affinity Reagents Ontology
PRO	Protein Ontology
TADS	Tick Gross Anatomy Ontology
TTO	Teleost Taxonomy
WBbt	*C. elegans* Gross Anatomy Ontology
WBls	*C. elegans* Development Ontology
WBPhenotype	*C. elegans* Phenotype Ontology
ZFA	Zebrafish Anatomy and Development Ontology

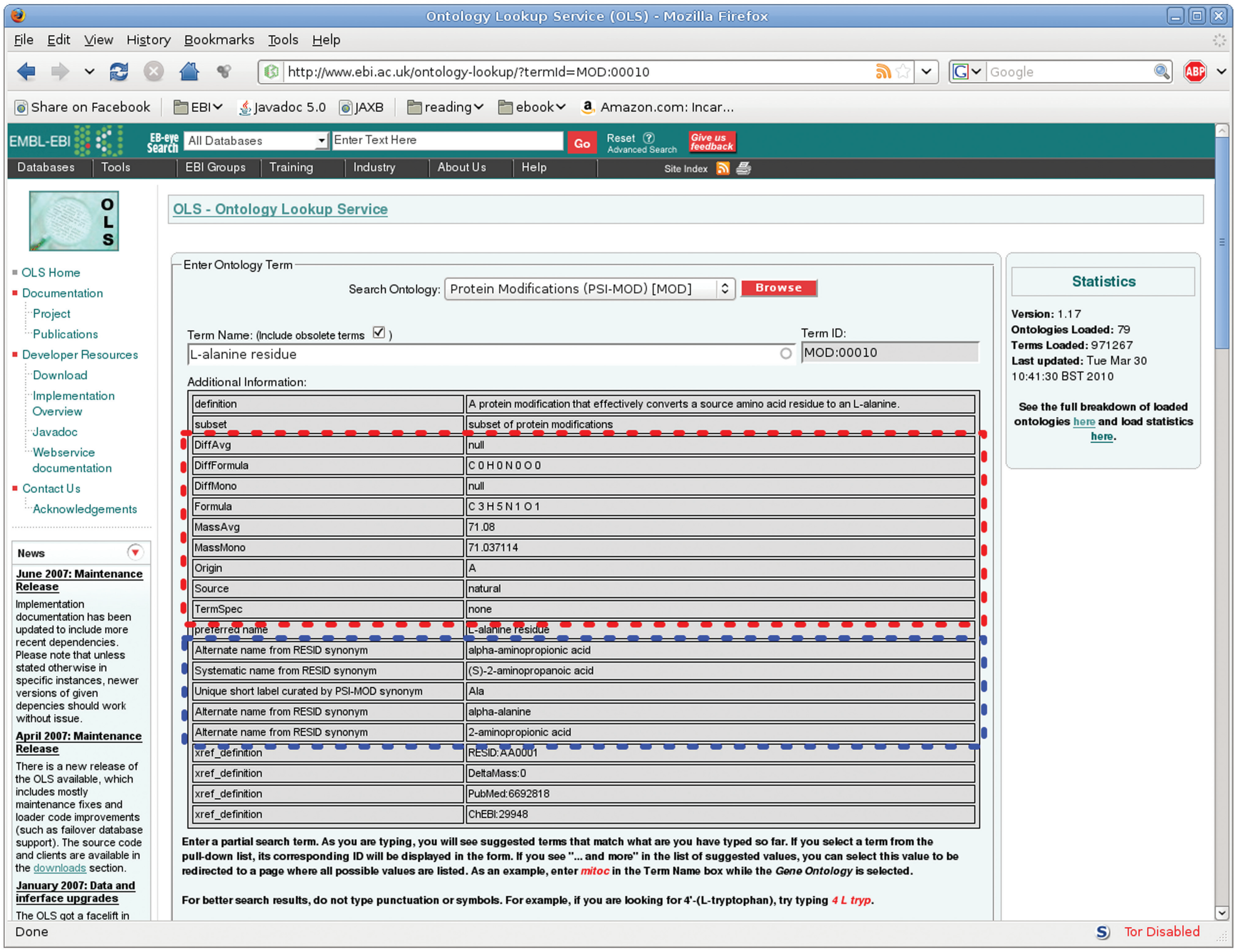
different Concurrent Versioning System (CVS) repositories, complemented with three Subversion (SVN) Version Control repositories thanks to the recently added SVN support. A mechanism to download individual files available by HTTP or FTP has also been implemented, which allows the loaders to track changes in files that are not in CVS or SVN

The OLS codebase is made available under the permissive Apache 2.0 Open Source License and is freely available from the Google Code project repository (http://code.google.com/p/ols-ebi/). A weekly updated MySQL database dump is also made available from the EBI FTP server (ftp://ftp.ebi.ac.uk/pub/databases/ols).

DATA MODEL IMPROVEMENTS

The OLS data loaders have been upgraded to be able to parse ontologies produced according to the Open Biomedical Ontology (OBO) 1.2 specification (http://www.geneontology.org/GO.format.obo-1_2.shtml) and can now capture previously unavailable information, such as custom name–value pairs and new synonym types (Figure 2). Another important feature of the OBO 1.2 specification is the ability to 'import' other ontologies and create relationships between local and imported terms.

In order to avoid loading multiple copies of imported ontologies, the loaders and database back end has been refactored such that each ontology is only loaded once. The OLS loaders are configured so that ontologies and CVs define one or more term prefixes that are local to itself (e.g. GO for the Gene Ontology). If the loaders encounter term identifiers that begin with a non-local prefix, they will query the OLS database and retrieve the latest version of the term in question and then proceed as normal. In this way, relationships across linked ontologies always refer to the most up-to-date data. These cross-ontology links can now also be queried and browsed, as shown in Figure 3.



OLS - Ontology Lookup Service

Enter Ontology Term

Search Ontology: Protein Modifications (PSI-MOD) [MOD]

Term Name: (Include obsolete terms ☑)

L-alanine residue

Term ID:

MOD:00010

Additional Information:

definition	A protein modification that effectively converts a source amino acid residue to an L-alanine.
subset	subset of protein modifications
DiffAvg	null
DiffFormula	C 0 H 0 N 0 O 0
DiffMono	null
Formula	C 3 H 5 N 1 O 1
MassAvg	71.08
MassMono	71.037114
Origin	A
Source	natural
TermSpec	none
preferred name	L-alanine residue
Alternate name from RESID synonym	alpha-aminopropionic acid
Systematic name from RESID synonym	(S)-2-aminopropanoic acid
Unique short label curated by PSI-MOD synonym	Ala
Alternate name from RESID synonym	alpha-alanine
Alternate name from RESID synonym	2-aminopropionic acid
xref_definition	RESID:AA0001
xref_definition	DeltaMass:0
xref_definition	PubMed:6692818
xref_definition	ChEBI:29948

Enter a partial search term. As you are typing, you will see suggested terms that match what are you have typed so far. If you select a term from the pull-down list, its corresponding ID will be displayed in the form. If you see "... and more" in the list of suggested values, you can select this value to be redirected to a page where all possible values are listed. As an example, enter *mitoc* in the Term Name box while the *Gene Ontology* is selected.

For better search results, do not type punctuation or symbols. For example, if you are looking for 4'-(L-tryptophan), try typing *4 L tryp*.

Statistics

Version: 1.17
Ontologies Loaded: 79
Terms Loaded: 971267
Last updated: Tue Mar 30 10:41:30 BST 2010

See the full breakdown of loaded ontologies here and load statistics here.

News

June 2007: Maintenance Release

Implementation documentation has been updated to include more recent dependencies. Please note that unless stated otherwise in specific instances, newer versions of given depencies should work without issue.

April 2007: Maintenance Release

There is a new release of the OLS available, which includes mostly maintenance fixes and loader code improvements (such as failover database support). The source code and clients are available in the downloads section.

January 2007: Data and inferface upgrades

The OLS got a facelift in

Figure 2. An example of custom synonyms and annotations. The synonyms in the blue box are uniquely defined in the PSI-MOD ontology. The annotations in the red box are examples of how the OLS can capture user-defined name–value attribute pairs.

INTERACTIVE USER INTERFACE IMPROVEMENTS

Users of the OLS website typically do one of two things: query the database using the auto-suggestion search box or browse an ontology (or a subset thereof). Once a term has been highlighted, either from the search suggestions or from the ontology browser, the user will be shown a table containing all the metadata associated with this term (synonyms, definitions, comments, cross-references and any other annotations that were captured during the loading process). When using the ontology browser, a graph showing either all the possible paths from the selected term to the root term(s) of the ontology and the relationships between all involved terms (Figure 3) or a local relationship graph with only the direct parent terms and children terms will also be shown. The type of graph to be displayed is configurable from the ontology browser interface. These graphs are clickable image maps that will zoom and re-root the ontology browser to the selected term.

REUSABLE CODE COMPONENTS

As mentioned previously, the OLS has always provided a SOAP web service. This service has been used by several large projects, such as PRIDE (4), IntAct (5), CheBI (6) and the Proteomics Standards Initiative (PSI) (7). However, the main drawback to its wider acceptance and uptake has been the lack of a simple GUI component that could easily be plugged in to existing code projects.

This has now been solved through the release of the open source OLS Dialog GUI component (8) (Figure 4). The OLS Dialog can easily be integrated into existing Java applications and gives access to the full range of query types supported by the OLS. Users can search for terms by name or by identifier, as well as use a graphical ontology browser to navigate an ontology and select a term. It is also possible to query terms from the PSI protein modification (PSI-MOD) (9) based on captured annotations from the source ontology. Users can select the type of annotation to query and enter a mass in Daltons and a desired precision and obtain all of the PSI-MOD entries that fit those parameters (e.g. find all PSI-MOD entries whose annotated monoisotopic mass is 120 D ± 1 D)

The OLS Dialog has been developed as part of the PRIDE Converter toolkit (10), which allows users to convert multiple mass spectroscopy file formats into

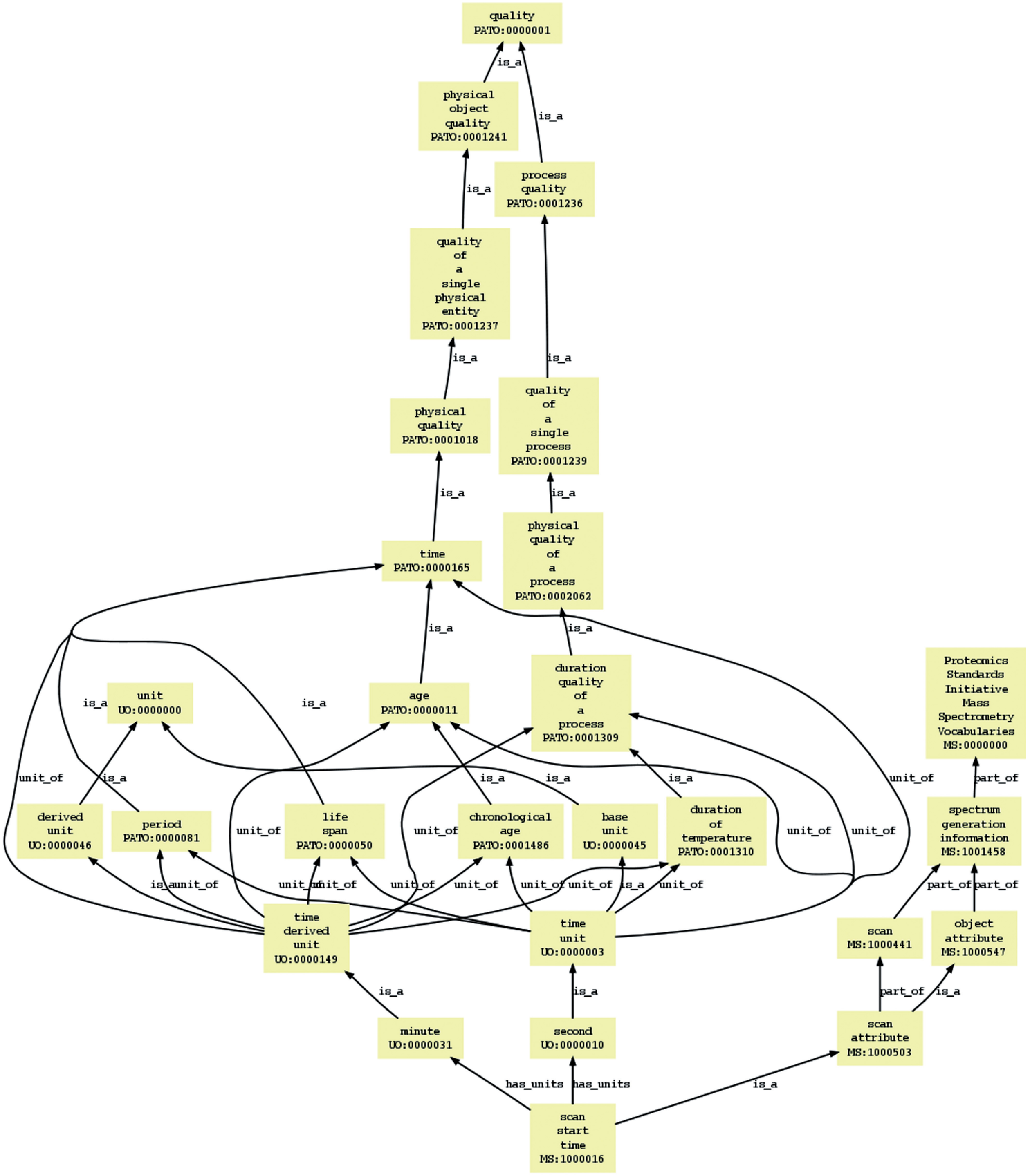
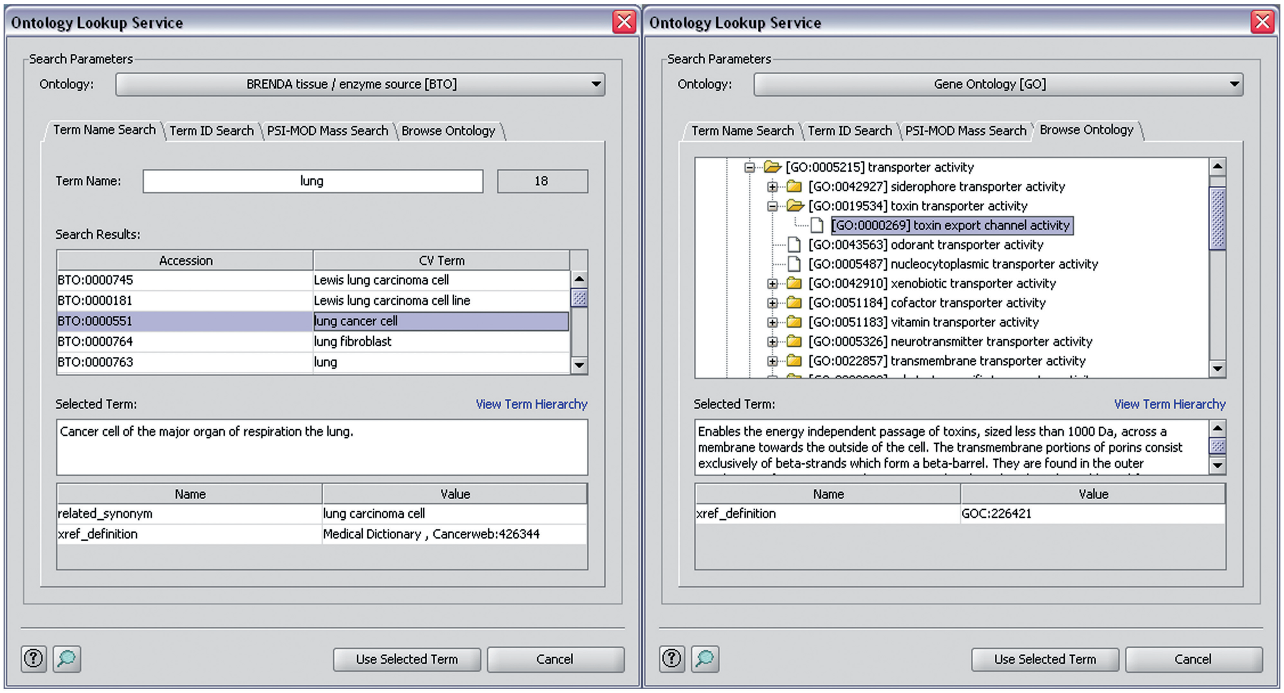

quality PATO:0000001
physical object quality PATO:0001241
process quality PATO:0001236
quality of a single physical entity PATO:0001237
quality of a single process PATO:0001239
physical quality PATO:0001018
physical quality of a process PATO:0002062
time PATO:0000165
Proteomics Standards Initiative Mass Spectrometry Vocabularies MS:0000000
unit UO:0000000
age PATO:0000011
duration quality of a process PATO:0001309
derived unit UO:0000046
period PATO:0000081
life span PATO:0000050
chronological age PATO:0001486
base unit UO:0000045
duration of temperature PATO:0001310
spectrum generation information MS:1001458
time derived unit UO:0000149
time unit UO:0000003
scan MS:1000441
object attribute MS:1000547
minute UO:0000031
second UO:0000010
scan attribute MS:1000503
scan start time MS:1000016
is_a
unit_of
part_of
has_units

Figure 3. An example term hierarchy graph from the Ontology browser of the OLS. When using the ontology browser, selecting a term will provide a graphical display of all paths from that term to the ontology root term(s). Users can click on the terms to zoom the ontology browser to a particular term. Note the cross-ontology links in this example. The term *scan start time* from the MS (mass spectrometry) ontology has a relation to the *second* and *minute* terms of the unit ontology (UO), which in turn has relations to the PATO (phenotypic quality ontology).

PRIDE XML in preparation for submission to the PRIDE database and requires users to annotate their submission files with terms from specific ontologies. User feedback has indicated that the PRIDE Converter and OLS Dialog have made submissions to PRIDE much easier and this has been made apparent in the submission figures to PRIDE (11).

DISCUSSION

The OLS has matured into a stable system and has proven to be popular beyond our initial expectations. Besides being used as a stand-alone system, its functionality has been incorporated into several independent tools and large-scale projects and is also being used by several



Ontology: BRENDA tissue / enzyme source [BTO]

Term Name: lung 18

Search Results:

Accession	CV Term
BTO:0000745	Lewis lung carcinoma cell
BTO:0000181	Lewis lung carcinoma cell line
BTO:0000551	lung cancer cell
BTO:0000764	lung fibroblast
BTO:0000763	lung

Selected Term:

Cancer cell of the major organ of respiration the lung.

Name	Value
related_synonym	lung carcinoma cell
xref_definition	Medical Dictionary , Cancerweb:426344

Ontology: Gene Ontology [GO]

- [GO:0005215] transporter activity
  - [GO:0042927] siderophore transporter activity
  - [GO:0019534] toxin transporter activity
    - [GO:0000269] toxin export channel activity
  - [GO:0043563] odorant transporter activity
  - [GO:0005487] nucleocytoplasmic transporter activity
  - [GO:0042910] xenobiotic transporter activity
  - [GO:0051184] cofactor transporter activity
  - [GO:0051183] vitamin transporter activity
  - [GO:0005326] neurotransmitter transporter activity
  - [GO:0022857] transmembrane transporter activity

Selected Term:

Enables the energy independent passage of toxins, sized less than 1000 Da, across a membrane towards the outside of the cell. The transmembrane portions of porins consist exclusively of beta-strands which form a beta-barrel. They are found in the outer

Name	Value
xref_definition	GOC:226421

Figure 4. Two screenshots of the OLS Dialog GUI component. The OLS Dialog allows Java application developers to seamlessly integrate OLS functionality in existing tools. Users can query the OLS by term name or ID. They can also locate terms by browsing an ontology and search the PSI-MOD ontology entries by term annotations specific to the ontology. In the left panel, a search on term names will also include partial matches and synonyms. In both cases, when a term is selected, the relevant associated metadata will be displayed and a graph similar to Figure 3 can be shown (not shown in these examples).

ontology developers as the primary ontology browser (12, 13 as examples).

When it went into production in mid-2005, the OLS was without peer. While it was true that each major ontology provider (GO, TAIR, FlyBase, Wormbase, etc.) generally provided its own website to browse their individual ontology, there was no unified resource to interactively and programmatically query multiple ontologies using a single, constant interface. Other services quickly followed suit and current systems that perform a similar function now include the National Center for Biomedical Ontology (NCBO) BioPortal (14) and the National Cancer Institute BioPortal (http://bioportal.nci.nih.gov/ncbo/faces/index.xhtml), which uses a scaled-down version of the NCBO BioPortal codebase.

A continuous increase in the number and scope of ontologies and CVs made available, coupled with an enhanced data model and better cross-ontology support will ensure that the OLS keeps its place as a valuable tool for a broad segment of the scientific community. The development team and its collaborators are always trying to make it easier to integrate OLS functionality into other projects, and the release of the OLS Dialog will go a long way towards achieving this goal. Ontology developers who wish to make their ontology available to the OLS can do so easily and through a variety of means, thanks to a versatile and automated loading process.

The OLS team is always looking for feedback to improve the project. Users are encouraged to contact pride-support@ebi.ac.uk for comments, problems and suggestions for new functionality.

ACKNOWLEDGEMENTS

We thank all users who have contributed recommendations and requests. Numerous improvements, already executed and some ongoing, come directly from this feedback.

FUNDING

The OLS is funded by the European Commission 'Serving Life-science Information for the Next Generation' (SLING), grant agreement number 226073 (Integrating Activity) within Research Infrastructures of the FP7. Formerly OLS was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) ISPIDER grant and European Union (EU) FP6 'Free European Life-science Information and Computational Services' (FELICS) [contract number 021902 (RII3)] grants. Funding for open access charge: EU FP6 'Felics' [contract number 021902 (RII3)].

Conflict of interest statement. None declared.

REFERENCES

1. Rubin,D.L., Shah,N.H. and Noy,N.F. (2007) Biomedical ontologies: a functional perspective. *Brief. Bioinformatics*, **9**, 75–90.

2. Côté,R.G., Jones,P., Apweiler,R. and Hermjakob,H. (2006) The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, **7**, 97.
3. Côté,R.G., Jones,P., Martens,L., Apweiler,R. and Hermjakob,H. (2008) The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res.*, **36**, W372–W376.
4. Vizcaíno,J.A., Côté,R.G., Reisinger,F., Foster,J.M., Mueller,M., Rameseder,J., Hermjakob,H. and Martens,L. (2009) A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics*, **9**, 4276–4283.
5. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) Intact–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
6. De Matos,P., Alcantara,R., Dekker,A., Ennis,M., Hastings,J., Haug,K., Spiteri,I., Turner,S. and Steinbeck,C. (2010) Chemical Entities of Biological Interest: an update. *Nucleic Acids Res.*, **38**, D249–D254.
7. Hermjakob,H. (2006) The HUPO proteomics standards initiative—overcoming the fragmentation of proteomics data. *Proteomics*, **6(Suppl 2)**, 34–38.
8. Barsnes,H., Côté,R.G., Eidhammer,I. and Martens,L. (2010) OLS Dialog: an open-source front end to the Ontology Lookup Service. *BMC Bioinformatics*, **11**, 34.
9. Montecchi-Palazzi,L., Beavis,R., Binz,P.A., Chalkley,R.J., Cottrell,J., Creasy,D., Shofstahl,J., Seymour,S.L. and Garavelli,J.S. (2008) The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.*, **26**, 864–866.
10. Barsnes,H., Vizcaíno,J.A., Eidhammer,I. and Martens,L. (2009) PRIDE Converter: making proteomics data-sharing easy. *Nat. Biotechnol.*, **27**, 598–599.
11. Vizcaíno,J.A., Côté,R.G., Reisinger,F., Barsnes,H., Foster,J.M., Rameseder,J., Hermjakob,H. and Martens,L. (2010) The Proteomics Identifications database: 2010 update. *Nucleic Acids Res.*, **38**, D736–D742.
12. Parkinson,H., Kapushesky,M., Kolesnikov,N., Rustici,G., Shojatalab,M., Abeygunawardena,N., Berube,H., Dylag,M., Emam,I., Farne,A. *et al.* (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.
13. Gaudet,P., Williams,J.G., Fey,P. and Chisholm,R.L. (2008) An anatomy ontology to represent biological knowledge in Dictyostelium discoideum. *BMC Genomics*, **9**, 130.
14. Noy,N.F., Shah,N.H., Whetzel,P.L., Dai,B., Dorf,M., Griffith,N., Jonquet,C., Rubin,D.L., Storey,M.A., Chute,C.G. *et al.* (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.*, **37**, W170–W173.