BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 12 2007, pages 1568–1570
doi:10.1093/bioinformatics/btl650

Databases and ontologies

ProServer: a simple, extensible Perl DAS server

Robert D. Finn, James W. Stalker, David K. Jackson, Eugene Kulesha,
Jody Clements and Roger Pettett*
Wellcome Trust Sanger Institute, Wellcome Trust Geome Campus, Hinxton, Cambridge, CB10 1SA, UK

Advance Access publication January 18, 2007
Associate Editor: Alfonso Valencia

ABSTRACT

Summary: The increasing size and complexity of biological databases has led to a growing trend to federate rather than duplicate them. In order to share data between federated databases, protocols for the exchange mechanism must be developed. One such data exchange protocol that is widely used is the Distributed Annotation System (DAS). For example, DAS has enabled small experimental groups to integrate their data into the Ensembl genome browser. We have developed ProServer, a simple, lightweight, Perl-based DAS server that does not depend on a separate HTTP server. The ProServer package is easily extensible, allowing data to be served from almost any underlying data model. Recent additions to the DAS protocol have enabled both structure and alignment (sequence and structural) data to be exchanged. ProServer allows both of these data types to be served.

Availability: ProServer can be downloaded from http://www.sanger.ac.uk/proserver/ or CPAN http://search.cpan.org/~rpettett/. Details on the system requirements and installation of ProServer can be found at http://www.sanger.ac.uk/proserver/.

Contact: rmp@sanger.ac.uk

Supplementary Materials: DasClientExamples.pdf

1 INTRODUCTION

High-throughput projects, such as the sequencing of the human genome, have resulted in a deluge of data. Thus, a key challenge in modern bioinformatics is to put in place mechanisms for programmatic exchange of and access to large volumes of data from disparate resources. As biological databases increase in size and complexity, the classical mechanism of data exchange, database duplication, can become impractical. For example, the underlying database for Ensembl release 41 is over 700GB, containing data on 25 different genomes, spread across hundreds of tables. An alternative to duplication of databases is to interlink the distributed resources, termed *federation*. However, for data exchange to take place between federated databases, a mechanism and standardization of format must be agreed. One such protocol for data exchange over a network is the Distributed Annotation System (DAS).

Briefly, the DAS protocol standardizes queries and responses for DNA or protein sequences, along with their annotations, regardless of the underlying data architecture (Dowell *et al.*, 2001). In common with other Web Services, client requests are made via HTTP to servers which process them and return results encapsulated in XML. Since the original DAS specification was developed, DAS has been used extensively for the annotation of genome (DNA) sequences and, more recently, of proteins. In addition, many extensions to the DAS protocol have been proposed. Two of these extensions, developed over the past year, allow the exchange of sequence alignments and protein structure data. These extensions enable annotations to be integrated across DNA, protein sequence and protein structure.

We have developed ProServer, a simple, user-friendly package for making data available using the DAS protocol.

2 PROSERVER

ProServer has been available from the Wellcome Trust Sanger Institute since 2003. We have recently added additional functionality and documentation to improve usability and robustness.

2.1 Architecture and flexibility

ProServer is a standalone, lightweight DAS server, written in Perl and designed to have low system requirements. The architecture of ProServer is represented in Figure 1. At the top level, there is a daemon executable which acts as a broker between requests and the code that will handle them. The server is configured using a '.ini'-format configuration file holding settings for data sources provided by the server (Fig. 1). The networking requirements of ProServer are handled by the POE package (http://poe.perl.org/), which implements a portable multitasking and networking framework for Perl.

The server sits upon a number of source adaptors with each one dedicated to a single underlying data store. Incoming queries are passed from the daemon to the appropriate source adaptor. The source adaptors handle access to the data store and the conversion of queries into generic DAS response data structures. Linking source adaptors to their data stores are generic transport helpers which are responsible for handling data acquisition (Fig. 1).

Depending on the format of any new data set, it may be necessary to implement a new transport helper (Fig. 1). However, ProServer comes bundled with helpers for some common data stores, for example: flatfile, GFF. MySQL, Oracle and SRS getz. These transport helpers are all simple, command-line or socket-handling modules.

*To whom correspondence should be addressed.

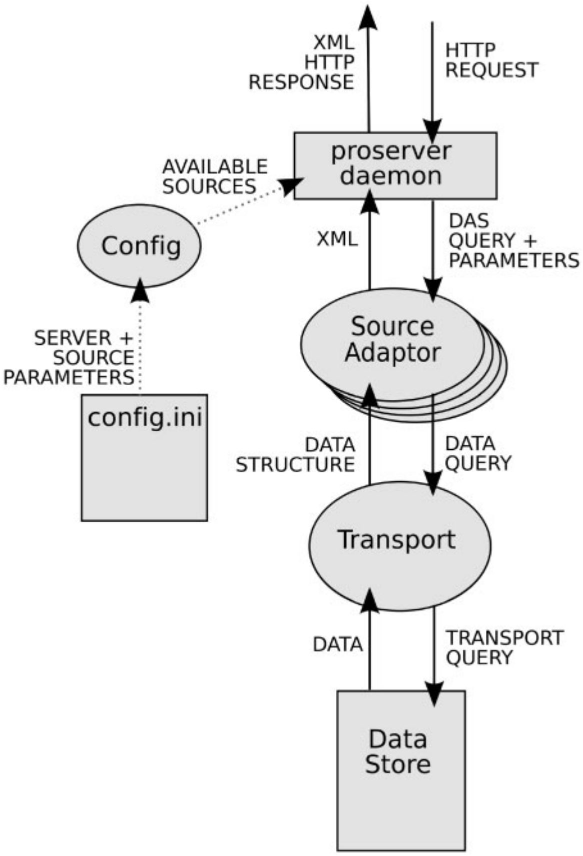

Fig. 1. A schematic representation of the ProServer architecture. (See text for details.)

How are new data sources made available using ProServer? In order to expose different sources of data, a new source adaptor class must be written. All such source adaptors inherit and extend generic functionality, namely the data retrieval methods that are applicable for the new data set (e.g. *features* or *sequence*). The superclass transparently handles the transformation of data to XML (Fig. 1). Finally, the details of the new source adaptor are entered into the ProServer configuration file.

Specific details regarding installation, security (and best practice usage) and scalability can be found in the ProServer README file.

2.2 Other DAS servers

Two alternative systems for creating DAS servers are Dazzle (http://www.derkholm.net/thomas/dazzle/) and the Lightweight Distributed Annotation Server (LDAS, http://biodas.org/servers/LDAS.html). Dazzle is written in Java, whereas LDAS is written in Perl. Dazzle has comparable functionality to ProServer, whereas LDAS does not have alignment or structure capabilities. Both of these alternatives require an available web server (e.g. Apache) and time-consuming configuration. Despite the fact that ProServer does not use a web server such as Apache, our experience indicates that the main bottleneck is input–output rather than computation. Thus, providing the underlying data store is organized optimally (as it would need to be with the alternative servers), then similar performance and scalability should be achievable regardless of the DAS server used.

The simplicity of ProServer means that it can be deployed by users with only basic bioinformatics skills. Therefore, groups can expose their data set, whether large or small, to the scientific community with little additional overhead to the storage of the data set.

2.3 Examples

ProServer was originally developed in conjunction with the Ensembl project (Birney *et al*., 2006), to display features such as gene predictions on chromosomes. Since the original release of ProServer, with only features and sequence capabilities, we have added both alignment and structure functionality. This has enabled the Pfam database (Finn *et al*., 2006) to provide access to their protein sequence alignments using ProServer to serve data directly from the underlying MySQL database (see http://das.sanger.ac.uk/das/pfamAlign).

Other projects outside of the Wellcome Trust Sanger Institute are already using ProServer for the exchange of feature annotations: examples include Gene3D (Yeast *et al*., 2006), CBS (Ólason, 2005) and AnoEST (Kriventseva *et al*., 2005).

The supplementary materials contain a list of some of the most popular clients that provide an insight to the use of DAS within the scientific community.

3 CONCLUSIONS

The federation of databases removes the effort involved with their duplication and maintenance and the problem of asynchronous versions. The relatively simple architecture of ProServer means that even small groups can make their data available via DAS, even if it is stored only as a flatfile, without the overheads of running a web server. Once available, this resource can be readily integrated into other data sets. For example, the Ensembl browser (Birney *et al*., 2006) allows new DAS sources to be added and displayed alongside existing annotation, even if they are only internally available at an institute. Rare chromosomal abnormality data from the DECIPHER project (http://decipher.sanger.ac.uk/syndromes) are being displayed in a genomic context via the Ensembl browser using ProServer, for example, the 1p36 microdeletion (http://www.sanger.ac.uk/turl/72d).

A survey of the DAS registry (Prlic *et al*., 2006) (http://www.dasregistry.org/) demonstrated that 69 of the 103 registered feature servers were using ProServer at the time of writing. Thus, ProServer is already being widely used to provide data via the DAS protocol. The recent extensions increase the range of data types available via DAS, allowing the transfer of annotations between aligned objects. These extensions are leading to the development of exciting new clients that bring together different data types (Prlic A *et al*., 2005). ProServer's minimal requirements and simplicity, together with the widespread use of DAS in bioinformatics makes ProServer a valuable tool for publishing data in a common, standard format. The accessibility and programmatic integration of distributed resources into state of the art tools and websites is accelerating novel scientific discoveries.

4 AVAILABILITY

ProServer is available from http://www.sanger.ac.uk/proserver/ or CPAN http://search.cpan.org/~rpettett/ and has been tested on Tru64, Linux and Mac OS X architectures running Perl 5.6.1 and above. ProServer has also been run under Windows using the Cygwin environment.

ACKNOWLEDGMENTS

The authors thank Andreas Kahari and Stefan Gräf for contributing bug fixes. This work was funded by The Wellcome Trust and a MRC (UK) E-science grant (G0100305). Funding to pay the Open Access publication charges was provided by The Wellcome Trust.

Conflict of Interest: none declared.

REFERENCES

Birney,E. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, D556–D561.

Dowell,R.D. *et al*. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.

Finn,R.D. *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res*., **34**, D247–D251.

Kriventseva,E.V. *et al*. (2005) AnoEST: Toward *A*. *gambiae* functional genomics. *Genome Res*., **15**, 893–899.

Prlic,A. *et al*. (2005) Adding some SPICE to DAS. *Bioinformatics*, **21**(Suppl. 2), ii40–ii41.

Prlic,A. *et al*. (2006) The Distributed Annotation System for integration of biological data. In Leser U., Naumann F., Eckman B. (eds.) *Data Integration in the Life Sciences: Third International Workshop, DILS 2006*. *Proceedings*. Springer, Berlin/Heidelberg, pp. 195–203.

Ólason,P.Í. (2005) Integrating protein annotation resources through the Distributed Annotation System. *Nucleic Acids Res*., **33**, W468–W470.

Yeats,C. *et al*. (2006) Gene3D: modelling protein structure, function and evolution. *Nucleic Acids Res*., **34**, D281–D284.