

The BioDAS project - <http://www.biodas.org/>

Andrew Jenkinson^{1,A}, Ewan Birney¹, Hagen Blankenburg³, Robert Finn², Henning Hermjakob¹, Tim Hubbard², Felix Kokocinski², Eugene Kulesha¹, Gabrielle Reeves¹ and Andreas Prlic²

¹ European Bioinformatics Institute, Cambridge, CB10 1SD, UK

² Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, UK

³ Max Planck Institute for Informatics, Saarbrücken, 66123, Germany

^A andy.jenkinson@ebi.ac.uk

The Distributed Annotation System (DAS) is an open protocol outlining a mechanism for the dynamic integration of biological data from geographically diverse sources. It was designed as a framework defining the communication protocol, data formats and visualisation standards for distributed genomic sequence annotations. The BioDAS community maintains client and server libraries implementing the DAS protocol in both Java and Perl programming languages. All are distributed with open source licenses (GPL/L-GPL). Various full client implementations, including Ensembl, Dasty and SPICE, are also available on an open source basis.

The core premise of DAS is that rather than aggregate data from multiple sites into a centralised database, data remains distributed. Within this basic strategy, data providers retain control over access and there is no requirement for synchronisation. However, proprietary interfaces restrict the usability of the data. DAS attempts to alleviate this concern by providing a common interface and format for obtaining and displaying annotations. The principal strength of DAS in this endeavour is its “dumb server, clever client” architecture. Since data providers are more numerous than visualisation clients, DAS aims to be simple for data providers to implement and has been widely adopted for this reason. DAS sees use both by public databases wishing to expose their data, and by groups of researchers wishing to share data between themselves.

In recent years, DAS has been extended to facilitate the communication of additional data types beyond genomic sequences and their annotations. There is now extensive use of DAS in communicating protein sequences and annotations, as well as 3D structures. Similarly, pairwise and multiple alignments are supported, with the inclusion of a paging option that is crucial for very large alignments such as Pfam protein family alignments. Latest developments include the addition of flexible support for molecular interactions, DASMI, as well as the introduction of the Protein Feature Ontology for standardisation of protein annotation types. Both have been adopted by the BioSapiens network, a collaboration of 25 institutions based in 14 countries throughout Europe.

DAS also remains actively supported in the genomic community, with the number of sources steadily increasing. Ensembl is committed to the use of DAS for federating data in its genome browser, and a number of high profile projects such as the collaborative ENCODE project have adopted DAS. In addition, the nature of contemporary functional genomics data has also driven further development of DAS. The high volume and density of data generated by sequencing or microarray experiments creates a technical challenge, in that researchers often wish to visualise huge volumes of data from a remote source simultaneously. DAS has introduced measures to improve efficiency in response to this requirement, and also additional visualisation options such as histograms and colour gradients. A new extended version of the DAS specification, 1.53E, encompasses all extensions made to the protocol to date.

The continuous growth in use of DAS has necessitated the development of a registry of public servers. The DAS Registry (<http://www.dasregistry.org/>), a discovery broker between servers and clients, is now an integral part of client applications - allowing client software to filter applicable servers and end users to distinguish and discover the various data available. As a result of this inclusion, DAS can be considered in principle a Service Oriented Architecture (SOA), having the three required components of service providers (servers), service requestors (clients) and service broker (registry). As evidence of the continued growth of DAS, registered DAS sources (i.e. only those publicly advertised) number around 400. The total number in use worldwide is far in excess of this.

DAS/2 emerged several years ago following a bifurcation of the BioDAS project with the aim of adding and improving features. At present and following considerable development to the specification, DAS/2 has some server and client implementations and offers features not available in DAS (e.g. writeback, for read-write operations). However, other features of DAS are not supported in DAS/2 (e.g. interactions), and no provision is made for backwards compatibility. Since future funding for DAS/2 has not been secured, we intend to explore possibilities for reconciling the two projects.