# The Distributed Annotation System for Integration of Biological Data

Andreas Prlić[1], Ewan Birney[2], Tony Cox[1], Thomas A. Down[1], Rob Finn[1], Stefan Gräf[2], David Jackson[1], Andreas Kähäri[2], Eugene Kulesha[1], Roger Pettett[1], James Smith[1], Jim Stalker[1], and Tim J.P. Hubbard[1]

[1] The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK ap3@sanger.ac.uk,
WWW home page: http://das.sanger.ac.uk/registry/
[2] EMBL - European Bioinformatics Institute, Hinxton, UK

**Abstract.** The Distributed Annotation System (DAS) is a protocol for sharing of biological data which allows for dynamical data integration. It has become widely used in both the genome and protein bioinformatics communities. Here we provide an overview of the available DAS infrastructure and present our latest developments, including a registration server that facilitates service discovery by DAS clients while automatically monitoring service availability. Currently there are 108 registered DAS servers, provided by 24 institutions in 10 countries.

## 1 Introduction

Annotation of biological data, such as genome and protein sequences, is one of the central tasks in biological research. This is done by different means, for example manually, computationally and experimentally. There are a number of centralized resources available that are working on the integration of these data. They are facing the problems of how to manage the vast amount of data that is available, the need for frequent updates and releases, and how to exchange data with other institutions and users.

The Distributed Annotation System (DAS) is a protocol that addresses these issues and facilitates the sharing of biological data [1]. It is based on the idea that annotation data is not aggregated into large centralized databases, but instead is spread over multiple sites, generally maintained by the original data creators. DAS is frequently used for

1. integration of personal data into bioinformatics resources,
2. integration of the annotations from external sources into local applications,
3. access to most recent data versions without the need for local installations,

DAS is a web service protocol built upon well established open technologies (HTTP and XML), with some similarities to SOAP-based services. Where SOAP services use XML requests and responses for the transport of information, DAS provides a data model, a query model, and a transport. The returned XML

documents contain objects like *sequence* or *feature*. All data are provided by DAS servers and it is up to a DAS client to retrieve the annotations from multiple servers and to integrate these into a visualization that is presented to the user (see Fig. 1). For a detailed description of the DAS protocol see `http://www.biodas.org/documents/spec.html`.

The DAS protocol was originally designed to serve annotation for genomes. Resources like the *Ensembl* genome browser utilize this protocol to visualize new or personal data in the context of other annotations [2]. Different web pages, *"views"*, provide access to annotation data for e.g. chromosomes, transcripts, genes, or proteins. Each of these views acts as a DAS client. A management interface allows users to configure a list of DAS servers from which annotation should be retrieved. Once a new server has been added in the configuration, Ensembl establishes the contact to the server, fetches the data, and displays it together with other annotations. In this setup the Ensembl web server acts as a data-proxy and the users can access all data via their web browsers.

Over the last few years DAS has also been used to share annotations of proteins. We recently presented *SPICE*, a browser of protein structures, sequences, and their annotations, which is built on DAS [3]. SPICE is a Java application that installs and runs locally using the Java Web-Start technology. It can be launched by simply following a link on a web page. SPICE provides an integrated view of protein sequence and structure and can project annotations from one coordinate system onto another. This, for example, allows it to display protein sequence annotations with respect to their position on the protein structure. SPICE is integrated with Ensembl (see Fig. 2).

*Dasty* is another protein DAS client [4]. It is a Java application with a Macromedia Flash front-end, and all DAS communication is done via a dedicated server. Other DAS clients that can be easily integrated into web pages are ProView (`http://www.sanger.ac.uk/proview/`) or the CBS DAS Viewer [5].

DAS has been widely adopted in the bioinformatics community, because it is simple to use and simple to set up. Both DAS servers and client software are available with implementations in multiple languages: In Perl there is support for setting up a DAS server using ProServer (`http://www.sanger.ac.uk/proserver/`) or LDAS (`http://www.biodas.org/servers/LDAS.html`), while users who prefer Java can use Dazzle (`http://www.derkholm.net/thomas/dazzle/`). Client libraries are also available in Perl, e.g. the Bio::DasLite library (`http://search.cpan.org/~rpettett/Bio-DasLite/`), and in Java (`http://www.biojava.org/`, `http://www.spice-3d.org/dasobert/`), making integration of DAS support into new and existing bioinformatics tools easy.

Several collaborations are providing support for DAS. The BioSapiens Network of Excellence (`http://www.biosapiens.info/`) is providing a large number of DAS sources, which are listed at the BioSapiens Information Resource (`http://www.biosapiens.info/page.php?page=biosapiensdir`). BioSapiens also provides a Portal that can query UniProt and provides access to several DAS clients (`http://www.biosapiens.info/page.php?page=das_portal`). Another

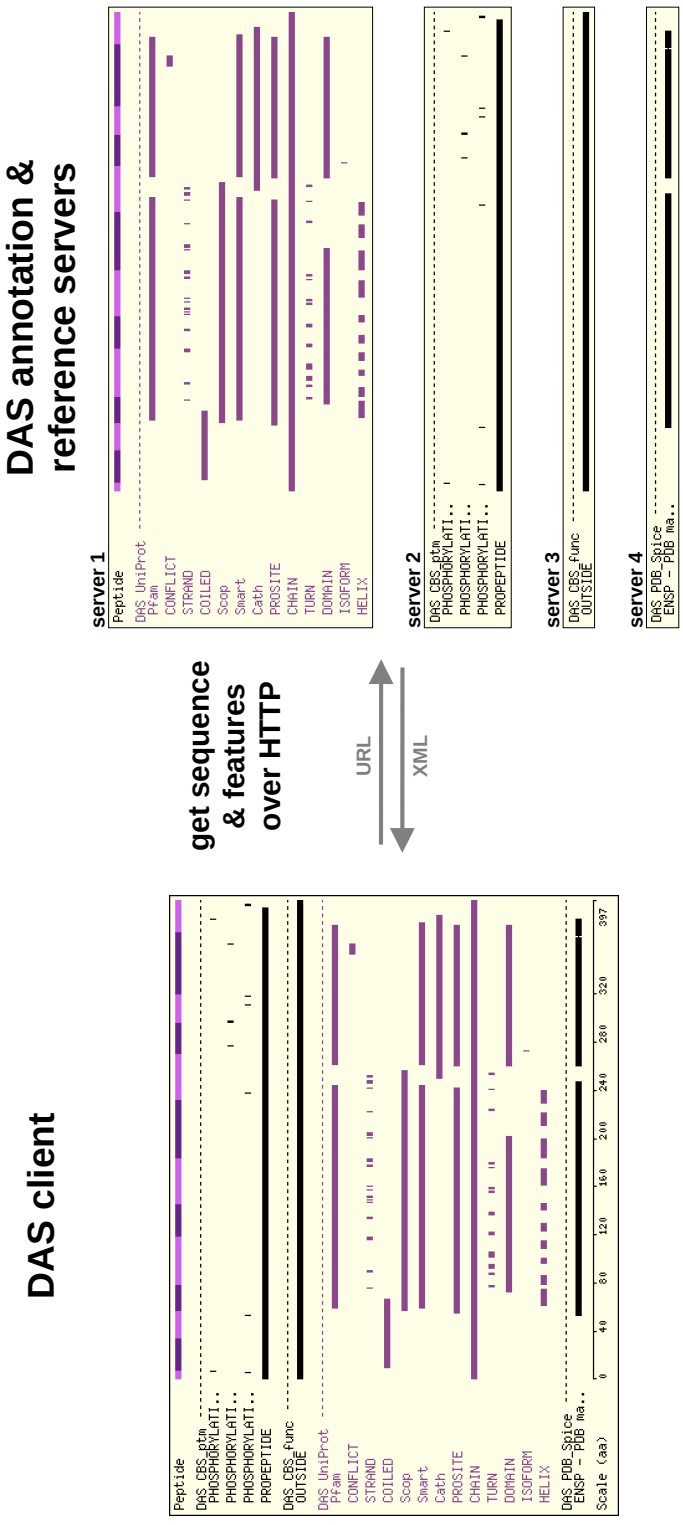


**Fig. 1.** A DAS client retrieves data from several DAS servers. In this schematic example the UniProt DAS server provides the reference sequence and some annotations. Other DAS servers are available that provide additional annotations for the same sequence. The Ensembl ProtView integrates the data into a common display.

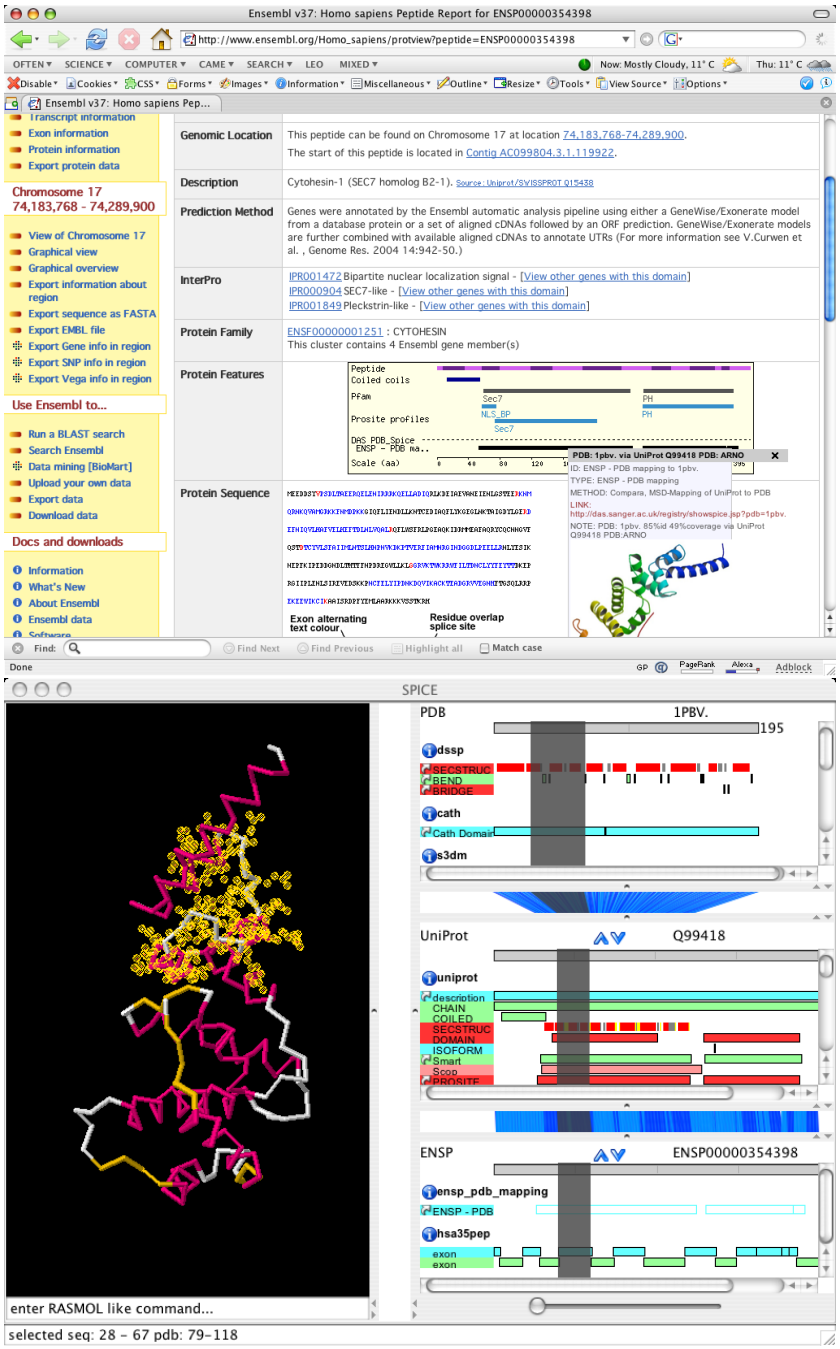


**Fig. 2.** A screenshot of the Ensembl ProtView and SPICE. A DAS track is available in Ensembl that shows which protein structures are aligned to the protein. By clicking on the DAS track the SPICE browser can be launched. SPICE shows the protein structure, the UniProt sequence, and the Ensembl predicted protein (ENSP) sequence. One of the exons is selected and its position can be projected onto the 3D structure.

project that provides support for DAS is the eFamily project (`http://www.efamily.org.uk/`).

### 1.1 Registration of DAS servers

DAS servers are divided into two categories. *Reference sources* provide the object to be annotated, e.g. a sequence or the 3D structure. *Annotation sources* provide the features of these objects. A number of different DAS sources have been released over the years providing annotations for different organisms and on different levels. The DAS protocol does not suggest how DAS clients can discover DAS sources that provide annotation. So far this has been done by hard-coding a list into a client, or requiring users to directly enter the URLs used to communicate with individual servers. Also, DAS does not define how to deal with the fact that different annotation servers provide the data for different types of biological objects. To address this we have developed a DAS registration server.

The DAS registration server fulfils several purposes:

1. It allows users (or their client software) to query and retrieve lists of available DAS sources via either a web interface or a XML web service for programmatic access.
2. It is able to direct a user to any of the most common DAS clients and attach the registered DAS server that the user is interested in seeing annotations from.
3. It automatically validates DAS sources to make sure they provide valid DAS-XML.
4. It can notify the administrator of a DAS source if the server has been down for a while.
5. It groups the registered DAS sources according to the coordinate system of the provided data.

There are three components to a Service Orientated Architecture concept: Service provider, service requestor, and a service registry. Connecting these components together are three operations: publish, find, and bind. The original DAS protocol partially implements this architecture, with DAS servers being the service providers and the clients being the requestors. By providing the new registration service that supports discovery, DAS has become a full Service Oriented Architecture.

## 2 System and methods

In this section we present various concepts and terminology that is applicable for DAS.

### 2.1 Coordinate Systems

DAS is used to annotate many different object types: genomes, gene loci, protein sequences, and structures are currently the most common cases. For each type, there are a number of meaningful 'sets' of objects — for example, the chromosome sequences in a particular assembly of the human genome. To allow data integration, clients must be able to find all the DAS servers which annotate a particular set of objects. We call the description of these sets *coordinate systems*. It can also be thought of as a "namespace". The following information is used for their description:

(1) The *authority* (or name). This is the name of the institution that defines the identifiers or accession numbers for a particular set of objects. In case of genome assemblies this field also contains the version of the assembly. For example, UniProt is an authority to assign protein sequence accession codes, while the currently used build for the human genome is NCBI 36.

(2) The *type* of object that is being annotated. This entity refers to the "physical dimension" of the data. Currently supported are *Chromosome, Clone, Contig, Gene_ID, NT_Contig, Protein Sequence, Protein Structure*, and *Scaffold*.

(3) The *organism*. The scientific name of an organism. This field is optional, since some DAS sources provide annotations for more than one organisms.

### 2.2 DAS capabilities

The DAS specification (version 1.5, `http://biodas.org/documents/spec.html`) defines a number of commands that can be sent to DAS servers. They are *sequence, features, types, entry_points, dna, stylesheet*. These are supported by the registration server together with the DAS extensions required for protein 3D structure annotations, *structure, alignment*, as described at `http://www.efamily.org.uk/xml/das/documentation/`.

### 2.3 Validation

There exist a number of different server side implementations to provide data via a DAS source. Frequently used ones include Dazzle (`http://www.derkholm.net/thomas/dazzle/`), ProServer (`http://www.sanger.ac.uk/proserver/`), and LDAS (`http://biodas.org/download/ldas/`) but sometimes individually implemented CGI scripts are used as well. In order to ensure a DAS source communicates in valid DAS-XML it can only be registered if it successfully validates by returning a correct DAS response for each of the capabilities for which it is registered. For this a *test code* is required, which is an accession code that has been annotated and for which features are provided.

Once a DAS source has been registered, the registry software contacts it periodically and attempts to validate it. Successful validation attempts are logged, and a graphical summary of the availability of a DAS source is available via the registry's web interface. If the DAS source can not be validated for more than two days, a *watchdog* can (optionally) inform the server's administrator. If a

server is down for a longer period of time, the server administrator can be contacted to inquire about the status of the server. If the server remains unavailable for an extended period, it will be removed from the listing.

### 2.4 Auto-activation

At the present stage the registry communicates with three DAS clients: Ensembl [2], SPICE [3], and Dasty [4]. Each of these can retrieve a list of available DAS sources from the repository. With appropriate client support, the registry can also communicate back in the reverse direction: DAS sources can be activated in a client by clicking on an icon in the registry web interface. In Ensembl the DAS server can be automatically added to the configuration of a particular view. The registry also provides a *send to friend* mechanism to share auto-activation links by email.

### 2.5 Implementation

The registration server at its core is a web service, backed by a MySQL database. The service can be accessed at `http://das.sanger.ac.uk/registry/services/das:das_directory?wsdl`. It can be used by different DAS clients to query and retrieve server listings or to validate DAS sources. A HTML front-end is provided which allows manual interaction with the registry (`http://das.sanger.ac.uk/registry/`). These web pages are based also on the web service. The HTML front-end is implemented as a set of JSP pages running on a Resin server.

## 3 Discussion

DAS is a data integration technology which is widely used in the genome and protein bioinformatics communities. A repository for registering and discovering DAS servers has been missing so far. Here we provide such a service. The DAS registry can interact with DAS clients and auto-activate a DAS source in the DAS client. On the management side the registry ensures that DAS sources follow the specification, and helps administrators to monitor the availability of their DAS sources. If a DAS server has been inactive for a while, we usually contact the administrators in order to query the status of the server. If it has become obsolete it can be removed from the repository.

We are participating in the development of DAS/2, a major update to the core DAS protocol (`http://biodas.org/documents/das2/das2_protocol.html`). DAS/2 will add the ability to search sets of features by identifiers and other properties (for instance, to find a gene given its name), and provides servers with extension mechanisms, allowing DAS features to be annotated with additional structured information (in XML format) as well as textual notes. DAS/2 also specifies an upload mechanism, so advanced clients can write back manually curated annotations to a DAS server.

The registration server currently contains 108 DAS sources provided by 24 institutions in 10 countries. Over the last year the number of registered DAS sources has been constantly growing (see Fig. 3). If this trend continues, at some point additional tools might be required for users to maintain an overview of the provided data. One way to achieve so could be to provide a user rating system, similar to what is known from popular online stores. Ideally such a system would be supported within the DAS clients, so user could rate a DAS source in a client, which would be communicated back to the registration server. DAS clients could sort DAS servers according to their popularity.

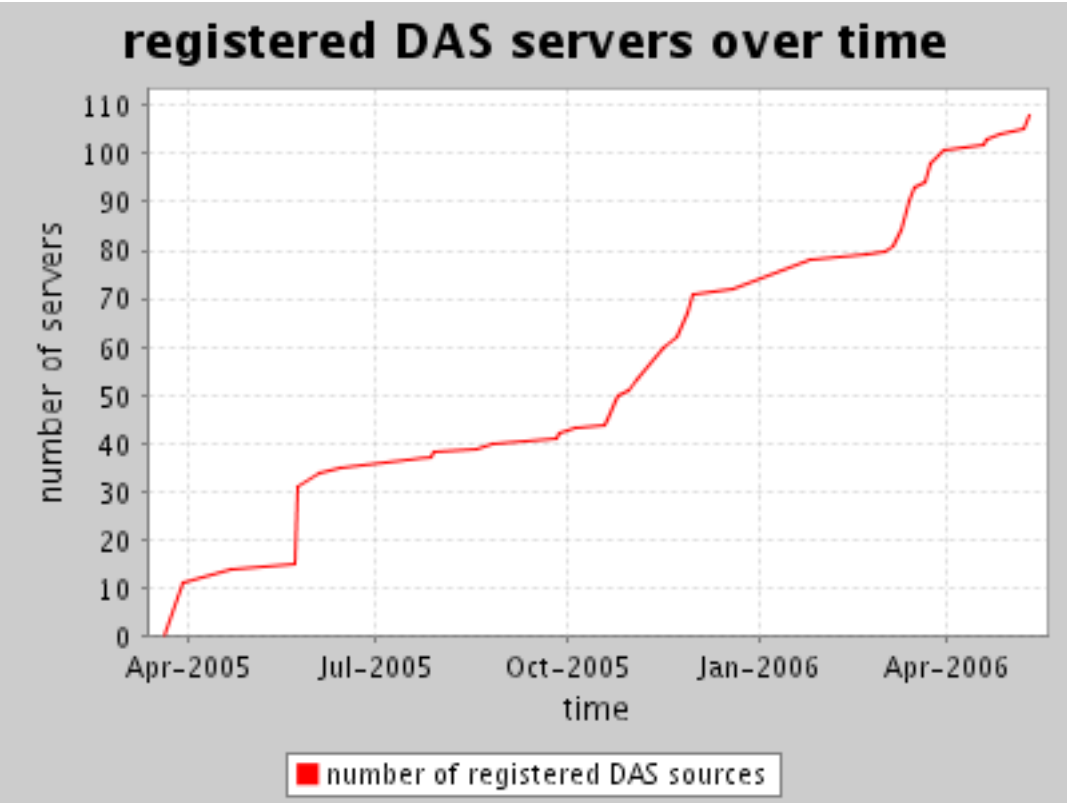


**Fig. 3.** The number of registered DAS sources over time. Currently there are 108 DAS sources available from 24 institutions in 10 countries.

## 4 Acknowledgments

We want to thank everybody who provides DAS servers and shares the data with the community. The system would not work without you. This work has been supported by the Medical Research Council, The Wellcome Trust, and the BioSapiens Network of Excellence. All source code is available under LGPL from `http://www.derkholm.net/svn/repos/dasregistry/`.

## References

1. Dowell, R.D., Jokerst, R.M., Day, A., Eddy, S.R., Stein, L.: The distributed annotation system. BMC Bioinformatics. **2** (2001) 7–7
2. Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X.M., Flicek, P., Gräf, S., Hammond, M., Herrero, J., Howe, K., Iyer, V., Jekosch, K., Kähäri, A., Kasprzyk, A., Keefe, D., Kokocinski, F., Kulesha, E., London, D., Longden, I., Melsopp, C., Meidl, P., Overduin, B., Parker, A., Proctor, G., Prlić, A., Rae, M., Rios, D., Redmond, S., Schuster, M., Sealy, I., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Stabenau, A., Stalker, J., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., Hubbard, T.J.P.: Ensembl 2006. Nucleic Acids Res **34**(Database issue) (2006) D556–61
3. Prlić, A., Down, T.A., Hubbard, T.J.P.: Adding Some SPICE to DAS. Bioinformatics **21 Suppl 2** (2005) ii40–ii41
4. Jones, P., Vinod, N., Down, T., Hackmann, A., Kahari, A., Kretschmann, E., Quinn, A., Wieser, D., Hermjakob, H., Apweiler, R.: Dasty and UniProt DAS: a perfect pair for protein feature visualization. Bioinformatics **21**(14) (2005) 3198–9
5. Olason, P.I.: Integrating protein annotation resources through the Distributed Annotation System. Nucleic Acids Res **33**(Web Server issue) (2005) W468–70