



HEALTH-F4-2007-200754

www.GEN2PHEN.org

D3.8 DAS Implementation for GEN2PHEN Data

WP3 – Standard Data Models And Terminologies

**V2.0
Final**

Lead beneficiary: EMBL
Date: 11/02/2010
Nature: Prototype
Dissemination level: PU (Public)



 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	2/27

TABLE OF CONTENTS

DOCUMENT INFORMATION	3
DOCUMENT HISTORY	3
DEFINITIONS	4
1. INTRODUCTION	5
1.1. INTEROPERATION OF DATA RESOURCES	5
1.1.1. <i>The Distributed Annotation System</i>	5
1.1.2. <i>Adoption of DAS within bioinformatics domain</i>	6
2. TECHNICAL OVERVIEW OF DAS	6
2.1. EXAMPLE DAS USE CASES.....	7
3. DAS PROPOSAL FOR SNP AND GEN2PHEN DATA	9
4. DAS IMPLEMENTATION FOR GEN2PHEN DATA	10
4.1. SUPPORT FOR DAS WITHIN ENSEMBL	10
4.1.1. <i>Integrating external DAS data</i>	10
4.1.2. <i>Supported coordinate systems</i>	11
4.1.3. <i>Supported Ensembl web pages</i>	11
4.2. THE ENSEMBL DAS REFERENCE SERVER	12
4.2.1. <i>Example requests</i>	12
4.3. INTEGRATION OF GEN2PHEN DATA SOURCES INTO ENSEMBL	13
4.3.1. <i>The SNPedia resource</i>	13
4.3.2. <i>SNPedia as a DAS source</i>	14
4.3.3. <i>GEN2PHEN data displayed via DAS in Ensembl</i>	15
APPENDICES	16
A. THE DISTRIBUTED ANNOTATION SYSTEM (DAS) PROTOCOL AND IMPLEMENTATION FROM DOWELL, ET AL [1]	16
B. DAS 1.53E SPECIFICATION FROM JENKINSON, ET AL [7]	21
REFERENCES	27

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies		Security: PU
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)		Version: v2.0 Final

Document Information

Grant Agreement Number	HEALTH-F4-2007-200754	Acronym	GEN2PHEN
Full title	Genotype-To-Phenotype Databases: A Holistic Solution		
Project URL	http://www.gen2phen.org		
EU Project officer	Frederick Marcus (Frederick.Marcus@ec.europa.eu)		


Deliverable	Number	3.8	Title	DAS Implementation for GEN2PHEN Data
Work package	Number	3	Title	Standard Data Models And Terminologies

Delivery date	Contractual	Month 24	Actual	11/02/2010
Status	Version 2.0		final <input checked="" type="checkbox"/>	
Nature	Report <input type="checkbox"/> Prototype <input checked="" type="checkbox"/> Other <input type="checkbox"/>			
Dissemination Level	Public <input checked="" type="checkbox"/> Confidential <input type="checkbox"/>			

Authors (Partner)	Paul Flicek (EMBL)		
Responsible Author	Paul Flicek		Email flicek@ebi.ac.uk
	Partner	EMBL	Phone +44 (0)1223 492581

Document History

Name	Date	Version	Description
P. Flicek	10/02/2010	1.0	Initial version
P. Flicek	11/02/2010	2.0	Revisions after review

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	4/27

Definitions

- Partners of the GEN2PHEN Consortium are referred to herein according to the following codes:

ULEIC – University of Leicester (UK) – Coordinator

EMBL – European Molecular Biology Laboratory (Germany) – Beneficiary

FIMIM – Fundació IMIM (Spain) – Beneficiary

LUMC – Leiden University Medical Center (Netherlands) – Beneficiary

INSERM – Institut National de la Santé et de la Recherche Médicale (France) – Beneficiary

KI – Karolinska Institutet (Sweden) – Beneficiary

FORTH – Foundation for Research and Technology Hellas (Greece) – Beneficiary

CEA – Commissariat à l’Energie Atomique (France) – Beneficiary

EMC – Erasmus Universitair Medisch Centrum Rotterdam (Netherlands) – Beneficiary

UH.FGC – Helsingin Yliopisto (Finland) – Beneficiary

UAVR – Universidade de Aveiro (Portugal) – Beneficiary

UWC – University of the Western Cape (South Africa) – Beneficiary

CSIR – Council of Scientific and Industrial Research (India) – Beneficiary

SIB – Swiss Institute of Bioinformatics (Switzerland) – Beneficiary

UNIMAN – The University of Manchester (UK) – Beneficiary


BIOBASE – BioBase GmbH. (Germany) – Beneficiary

deCODE – Islensk Erfoagreining EH (Iceland) – Beneficiary

PHENO – Phenosystems S.A. (Belgium) – Beneficiary

BCP – Biocomputing Platforms Ltd. Oy (Finland) – Beneficiary

- Grant Agreement:** The agreement signed between the beneficiaries and the European Commission for the undertaking of the GEN2PHEN project (HEALTH-200754).
- Project:** The sum of all activities carried out in the framework of the Grant Agreement by the Consortium.
- Work plan:** Schedule of tasks, deliverables, efforts, dates and responsibilities corresponding to the work to be carried out for the GEN2PHEN project, as specified in Annex I to the Grant Agreement.
- Consortium:** The GEN2PHEN Consortium, conformed by the above-mentioned legal entities.
- Consortium agreement:** agreement concluded amongst GEN2PHEN participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties’ obligations to the Community and/or to one another arising from the Grant Agreement.

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	5/27

1. INTRODUCTION

Biological information management nearly always occurs in the context of federated databases. The reasons for this are multiple, but the most important include (1) the importance of staying connected to a wet laboratory of innovation and its daily needs; (2) the sheer diversity of biological information leading to the need of many differently structured datasets unique to different sub-domains in biology; and (3) the more mundane aspects of separate funding streams to individual scientists. Although there are a number of clear “aggregation” databases, such as UniProt, Ensembl, and ArrayExpress, which collect data (in some cases from submissions, in some cases by literature review and in some cases by direct interaction with key collaborators) these large aggregation datasets can never encompass the complete diversity of biological information. It is therefore important that maximal inter-operation between databases be engineered with an overall goal to allow seamless traversal of the biological space as needed.


1.1. Interoperation of data resources

Interoperation in any IT domain has a long and usually chequered history, with both successes and quite dramatic failures. Interoperation techniques can be classified into two ways: the type of technology used for the interoperation, and the semantic richness of the interoperation. Experience has shown that the technical aspect—although important—is usually less critical for success than the semantic level for the interoperation. The deeper the semantic interoperation, the more complex it is to achieve systems working systems with semantic agreement. In contrast, lighter interoperation, although conceptually less complete is far more likely to succeed. A final consideration is the level of experience in the field of the interoperation technique; here a technique known to a broad community is more valuable because it is easier to achieve interoperation and there are more robust toolkits for this interoperation.

1.1.1. *The Distributed Annotation System*

A successful interoperation technique is the Distributed Annotation System (DAS), originally specified by Robin Dowell, Sean Eddy and Lincoln Stein [1]. The DAS system, which was originally designed on the basis of genomic coordinates has been gradually extended to work with other coordinate systems such as gene or protein information. DAS is a REST¹ style service, meaning that it is built directly upon existing http technology. The “semantic level” of DAS is extremely simple – it in effect allows the transmission of a small amount of text or images to another site attached to an identifier. The transmitted information may optionally be restricted to a particular range or subset of the data (obviously this is critical for genomic coordinates). The only smart aspect of DAS compared to other web technologies (eg, RSS) is the support of multiple “coordinate systems” (or perhaps better called overlapping namespaces), a

¹ REST (representational state transfer) describes an architecture style of networked systems including the World Wide Web. REST services have a number of characteristics including a memoryless client-server interaction system in which client requests contain all of the information required for the server to respond.

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	6/27

common feature in biological resources. DAS can be described as a “smart client, dumb server” protocol, with most of the complexity needing to be handled in the client; this is ideal for the common situation in biology where one has a large number of medium to small databases needing to be integrated with only a few large clients, usually web based (such as the Ensembl Browser).

1.1.2. Adoption of DAS within bioinformatics domain

DAS has now been adopted by both Ensembl and UniProt as key federation technologies, and there are over 401 different DAS servers from 53 institutions in 17 countries registered at www.dasregistry.org with a variety of outputs and clients (see Figure 1).

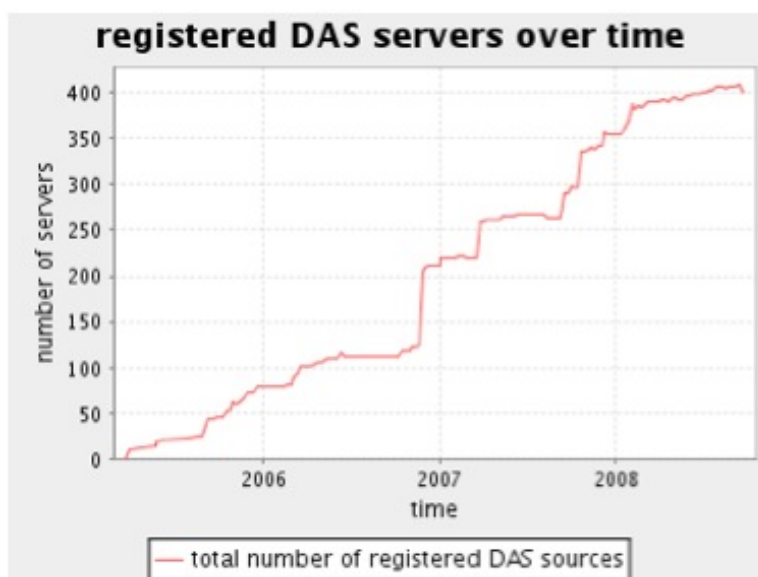



Figure 1: The growth of registered DAS sources over time as measured by DASregistry.org.

Overall DAS is a successful, stable technology able to be used for a variety of aspects. This deliverable describes the DAS protocol and the extension of DAS to GEN2PHEN data initially in the form of providing annotations on SNP data. We demonstrate this functionality within Ensembl by providing annotations from SNPedia.

2. Technical Overview of DAS

Figure 2 provides an overview of DAS as it is used currently. True clients interact with large web based systems such as Ensembl or Uniprot. These large systems provide a useful view on a series of biological items – such as genomic features or genes. DAS is presented to the user as series of

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	7/27

additional “tracks” or information that the user can switch on. To facilitate the discovery and choosing of this information, a centralised DAS registry provides a structured repository of submitted and active DAS servers (the DAS registry also provides statistics about up-time and validation, allowing the application to present more information about the reliability of the DAS server to the client). The client then activates a series of DAS servers, and the information is rendered “in page” for each active DAS server.

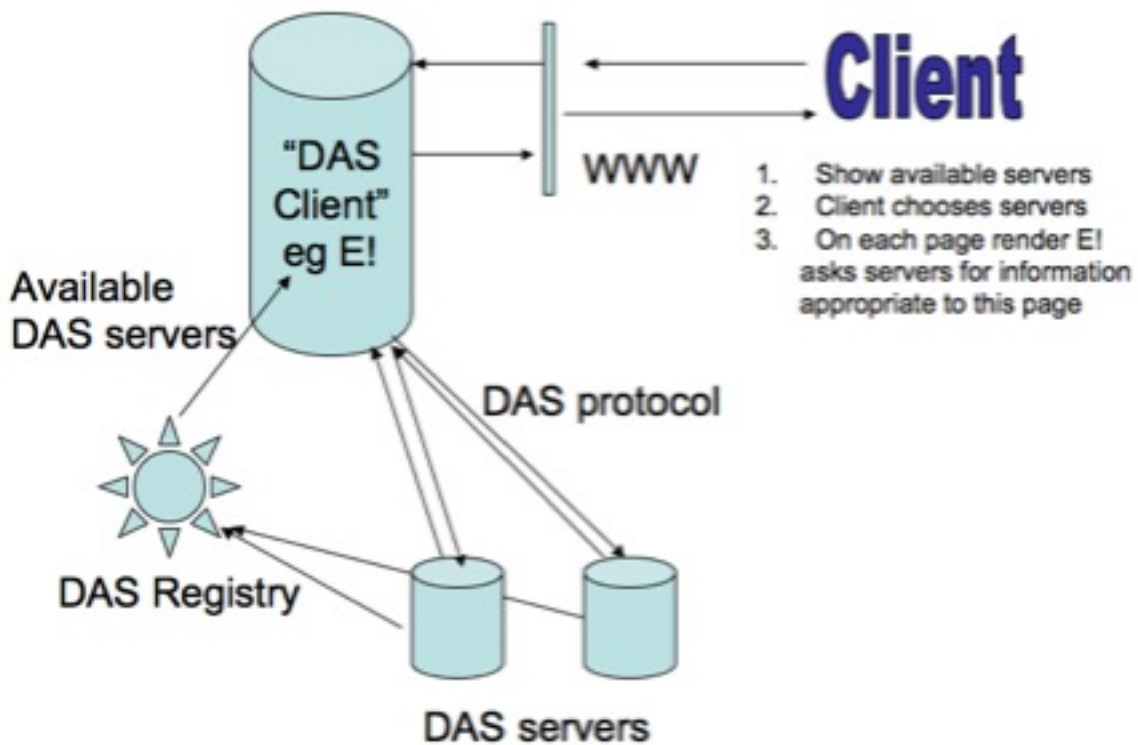



Figure 2: A schematic overview of a DAS request within the context of the Ensembl DAS client (denoted as E!). Client users (upper right) request information on available resources and this request is relayed by the client to the DAS Registry. Once the requested resource is chosen, the DAS client within Ensembl requests information from the relevant DAS servers. Once provided, the external information is integrated into the appropriate Ensembl visualisation and this data is returned to the user.

2.1. Example DAS use cases

Figures 3 and 4 show some use typical use cases in Ensembl. In figure 3, a new “track” of mutations from the EUCOMM project is rendered “in-line” to the Location page visualisation in Ensembl on genomic coordinates. Clicking on the EUCOMM picture then leads to a project-

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	8/27

specific web page hosted at the Sanger Institute. Figure 4 shows a Gene DAS source, in this case from Array Express on gene expression. Here more textual information is rendered into the main gene page of Ensembl including pictures. These are rendered “in page” but are entirely generated by the ArrayExpress DAS server. Clicking on these images lead the user back to the ArrayExpress site.

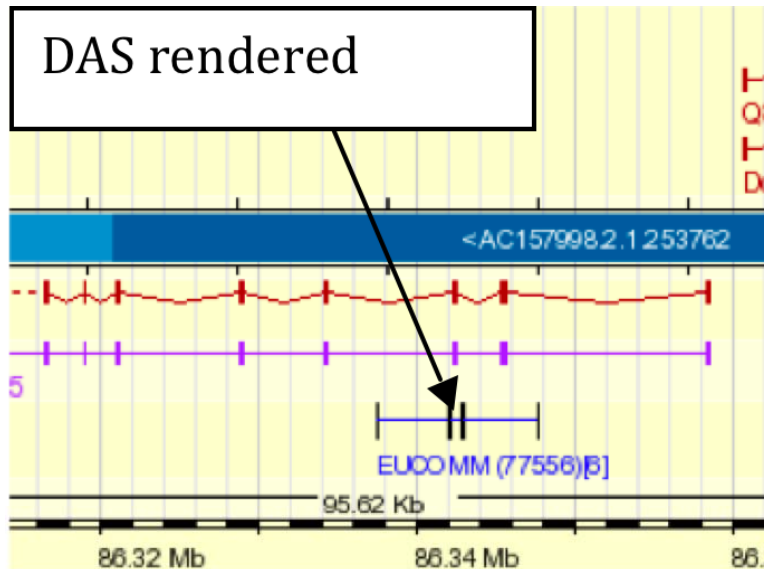



Figure 3: EUCOMM mutations from a DAS server at the Sanger Institute displayed on the Ensembl Location page.

In both cases the effect for the browsing user is that this information has been “integrated” or “federated” into the main Ensembl system, meaning that they can easily keep track of what they want to see both in a genomic context and a gene context. Although the EUCOMM data, representing mouse mutations, is in some sense already an example of GEN2PHEN data served and integrated through the DAS protocol, the significant feature addressed by this deliverable is the extension of the DAS system to also handle SNP data.

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	9/27

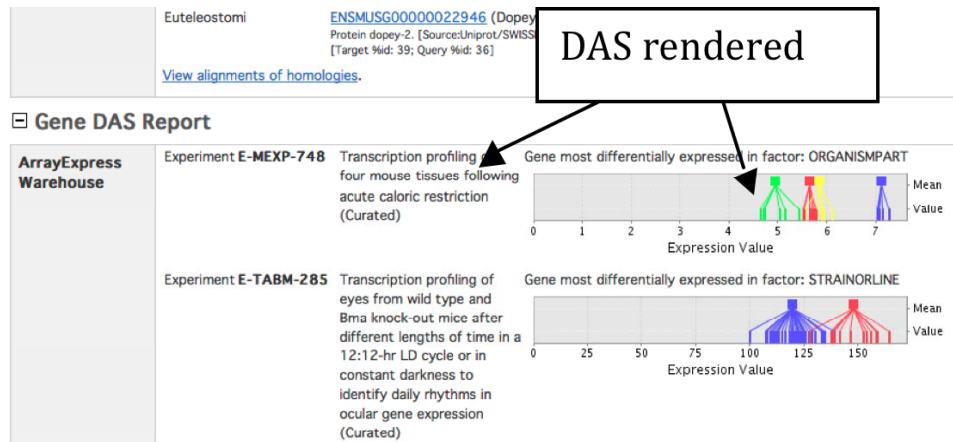



Figure 4: A Gene DAS source providing information from the ArrayExpress DAS server at the EBI. Information about expression patterns for the given gene are provided and incorporated into a custom DAS display.

3. DAS Proposal for SNP and GEN2PHEN data

Analogous to the Gene-DAS proposal (http://www.dasregistry.org/extension_genedas.jsp) “extending” the DAS protocol to SNPs requires acknowledging that this protocol can be used on variation objects but that the start and end is redundant (though see below for a proposed aspect to this). What is needed is an agreement on potential “authority/type” for identifiers are allowed. The proposal is as follows:

Authority	Type	Meaning
NCBI	Rs-id	The RS id for a variation
Ensembl	Ensembl-snp-id	For species where Ensembl called SNPs before RS ids are issued
LRG	LRG id	If an LRG served annotation corresponds perfectly to a SNP position, display this annotation to the SNP page

	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	10/27

The first two are relatively straightforward using the “natural” identifier space of variations from NCBI or Ensembl². DAS is ideal for this case of potentially differing namespaces in that the client’s are tasked to work out the right identifier to send to the right server. In the specification each server registers itself as capable of responding to specific authority/types and it is the client’s task to send the right information to the right servers. The final case with LRG information is again focused on appropriate client implementation. Here the “genomic DAS” behaviour of a server in LRG coordinates would be honoured, and in particular if an LRG annotation was recorded with start and end coordinates in the LRG space that corresponded to the SNP, this annotation would be displayed in the SNP page. This allows all of the advantages of the stable LRG coordinate system to be used to integrate data between clients and servers.

4. DAS implementation for GEN2PHEN Data

4.1. Support for DAS within Ensembl

Ensembl makes use of DAS in two ways. (1) External data may be integrated into the website. (2) Ensembl data may be integrated into other applications via the Ensembl DAS server. This deliverable focuses on the requirements for the first of these DAS applications. Future developments may include providing G2P data directly from Ensembl by making use of the Ensembl DAS server and, therefore, information about this functionality is included in this section.


4.1.1. Integrating external DAS data

As described above, DAS provides a standardised method to serve custom annotation information and to integrate data sets for display in other resources. Ensembl allows attachment and configuration of external DAS sources to several Ensembl genome browser displays (see section 4.1.3 for a full list)

Several DAS resources provided by the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI) are already available and pre-configured by default. The instructions for accessing these tracks are as follows:

- Navigate to the appropriate tab and page section. For example the Location tab and Region in detail section.

² Ensembl variations are always submitted to NCBI, but there can be up to a 6 month delay for the roundtrip of submission, processing to display, so handling supporting DAS visualisation during this delay is potentially important. However, this is far less important for Human where dbSNP is both more comprehensive and also processing occurs faster

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	11/27

- Click the "Configure this page" link.
- Choose a track to enable from the available categories.

Further annotation information resources can be integrated into Ensembl by attaching a valid data source provided by a DAS annotation server. These DAS servers may be located anywhere on the Internet. To attach a DAS source in this category, a user would do the following:

- Navigate to the appropriate tab and page section. For example the Location tab and Region in detail section.
- Click on the "Add custom data to page" link at the left.
- Click on "Attach DAS" at the left, and follow the instructions.
- Save and close the window.

After closing the window, the source will be automatically enabled.

4.1.2. Supported coordinate systems

Whilst DAS was originally designed to exchange annotations of reference sequences in chromosome or clone coordinate systems, GeneDAS and ProteinDAS are extensions to the DAS protocol used to exchange gene and protein annotations independent of genomic location information. Currently, Ensembl supports annotations based on several different coordinate systems. Some coordinate systems allow annotations which are positional in nature (i.e. refer to a location within a sequence), whereas others are concerned with text-based non-positional annotations.


The views that a particular DAS source may be displayed on depends on the data being served and the coordinate system of its annotations:

In genomic coordinate systems (e.g. chromosome, clone, contig), annotations are positional. In gene coordinate systems (Ensembl, Entrez, HUGO or MGI), annotations are non-positional. In protein coordinate systems (Ensembl, UniProt or IPI), annotations may be positional OR non-positional.

4.1.3. Supported Ensembl web pages

Ensembl supports DAS on the following sections of the website:

Region overview (positional annotations)
Region in detail (positional annotations)
Gene -> External data (non-positional annotations)

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	12/27

Transcript -> External data (non-positional annotations)

Variation -> Incorporation of phenotype annotations³

4.2. The Ensembl DAS reference server

Ensembl provides a DAS reference server which gives access to genomic sequences, the latest Ensembl gene predictions, and for some species, karyotypes and ditags. A list of the sources currently served from the Ensembl DAS reference server may be found as XML documents at:

<http://www.ensembl.org/das/dsn> (listing generated from the standard DAS data source name: DSN request)

<http://www.ensembl.org/das/sources> (a listing with extended information)

4.2.1. Example requests

DAS request URLs have a specific format (see Appendix for full specification):

protocol://site-prefix/das/data-source/command?arguments

For example:

Command:

http://www.ensembl.org/das/Homo_sapiens.NCBI36.transcript/features?segment=13:31787617,31871806

Result:

Request all transcripts (exons really) in the region [31787617,31871806] on human chromosome 13 (this is where the gene BRCA2 is located in the NCBI 36 assembly).


Command:

http://www.ensembl.org/das/Gallus_gallus.WASHUC1.reference/sequence?segment=1:1,1000

Result:

Request the first 1000 bp of the first chicken chromosome.

³ New feature reported first in this document

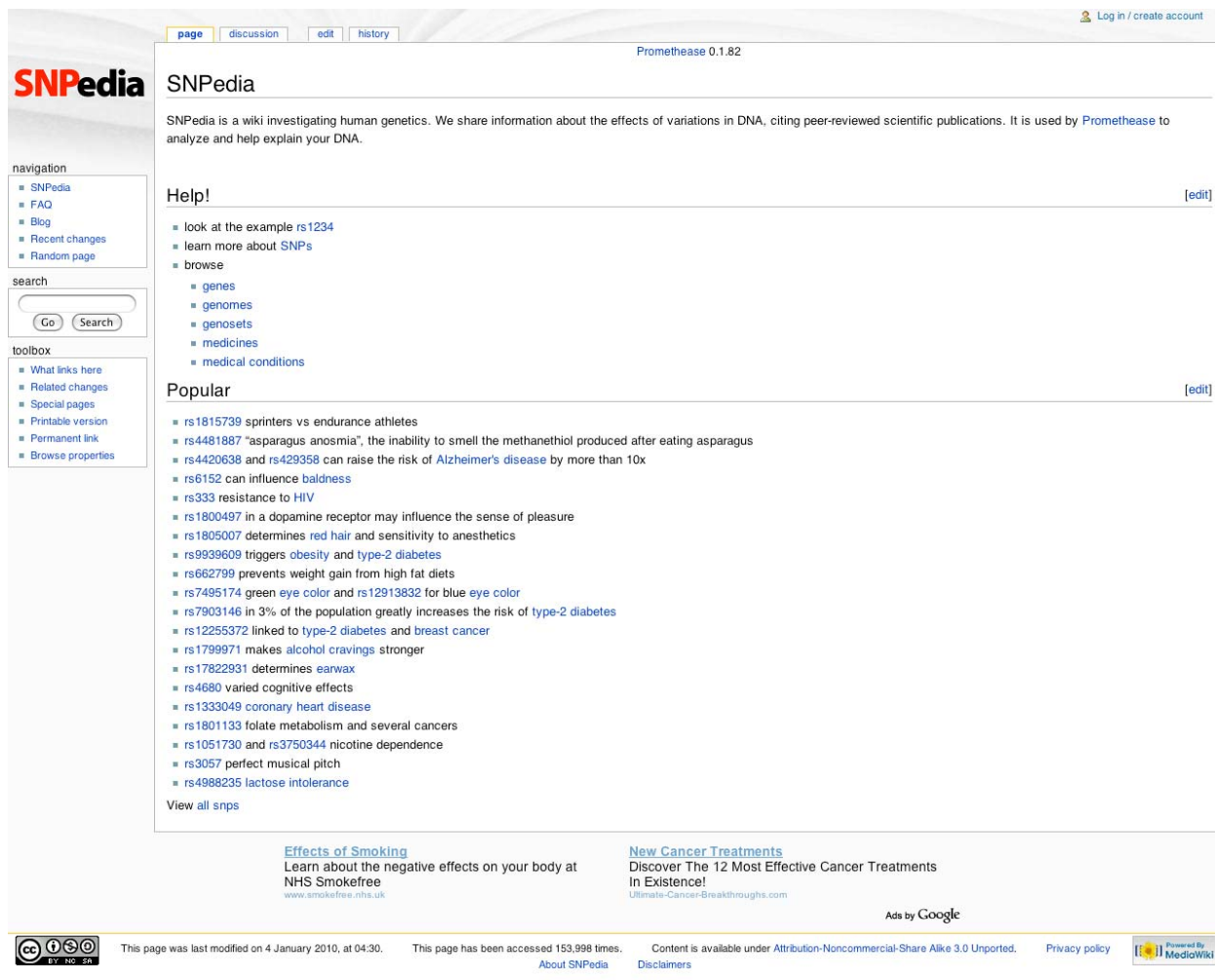
 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	13/27

4.3. Integration of GEN2PHEN data sources into Ensembl

4.3.1. The SNPedia resource


SNPedia is an Internet-based wiki site (<http://www.snpedia.com>) dedicated to the collection of information about SNPs in the human genome and to support the interpretation of these SNPs (see Figure 5). As such, it is designed to be a central resource that collects information about the connections between genotype and phenotype. As a wiki, it is possible to update in essentially real time and have accurate information reflecting the most current knowledge available in the published literature or other sources.

As of the date of this report, SNPedia has information on more than 10,000 SNPs making it one of the single most comprehensive SNP annotation resources available.



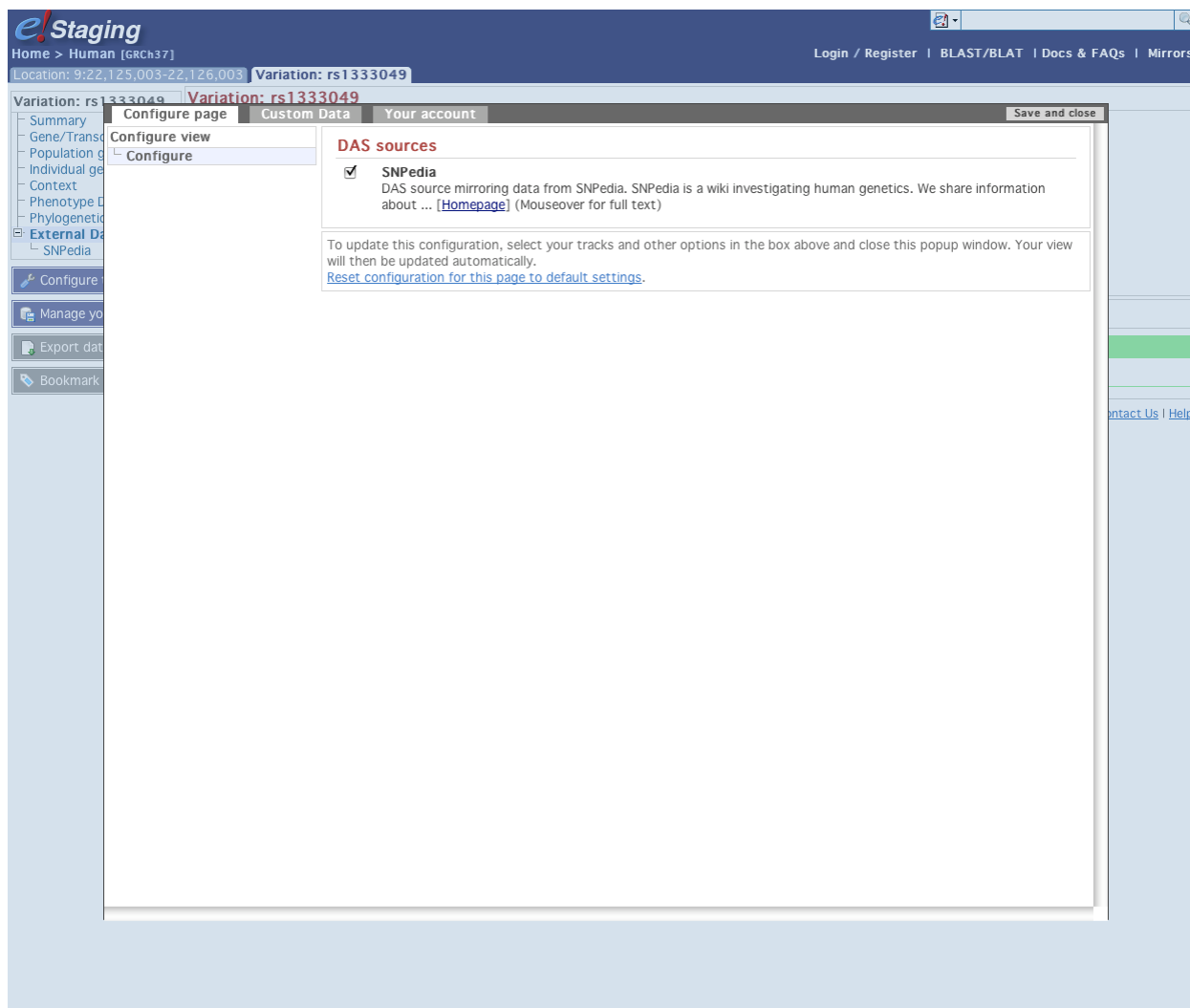
The screenshot shows the SNPedia homepage. At the top, there are navigation tabs for 'page', 'discussion', 'edit', and 'history', and a 'Log in / create account' link. The main heading is 'SNPedia' with a sub-heading 'Promethease 0.1.82'. Below this is a brief description: 'SNPedia is a wiki investigating human genetics. We share information about the effects of variations in DNA, citing peer-reviewed scientific publications. It is used by Promethease to analyze and help explain your DNA.' There are two main sections: 'Help!' and 'Popular'. The 'Help!' section includes a list of links: 'look at the example rs1234', 'learn more about SNPs', and 'browse' (with sub-links for genes, genomes, genosets, medicines, and medical conditions). The 'Popular' section lists several SNPs with their associated effects, such as 'rs1815739 sprinters vs endurance athletes', 'rs4481887 asparagus anosmia', and 'rs4420638 and rs429358 can raise the risk of Alzheimer's disease by more than 10x'. The footer contains promotional banners for 'Effects of Smoking' and 'New Cancer Treatments', along with a Google AdSense logo and a Creative Commons license.

Figure 5: The SNPedia homepage.

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	14/27


4.3.2. SNPedia as a DAS source

We converted SNPedia into a DAS source following the proposal in section 3 as a proof of principle for the general integration of GEN2PHEN data using DAS. This DAS source has been added to the default configuration of Ensembl thus making it selectable via the Ensembl page configuration interface (see Figure 6).



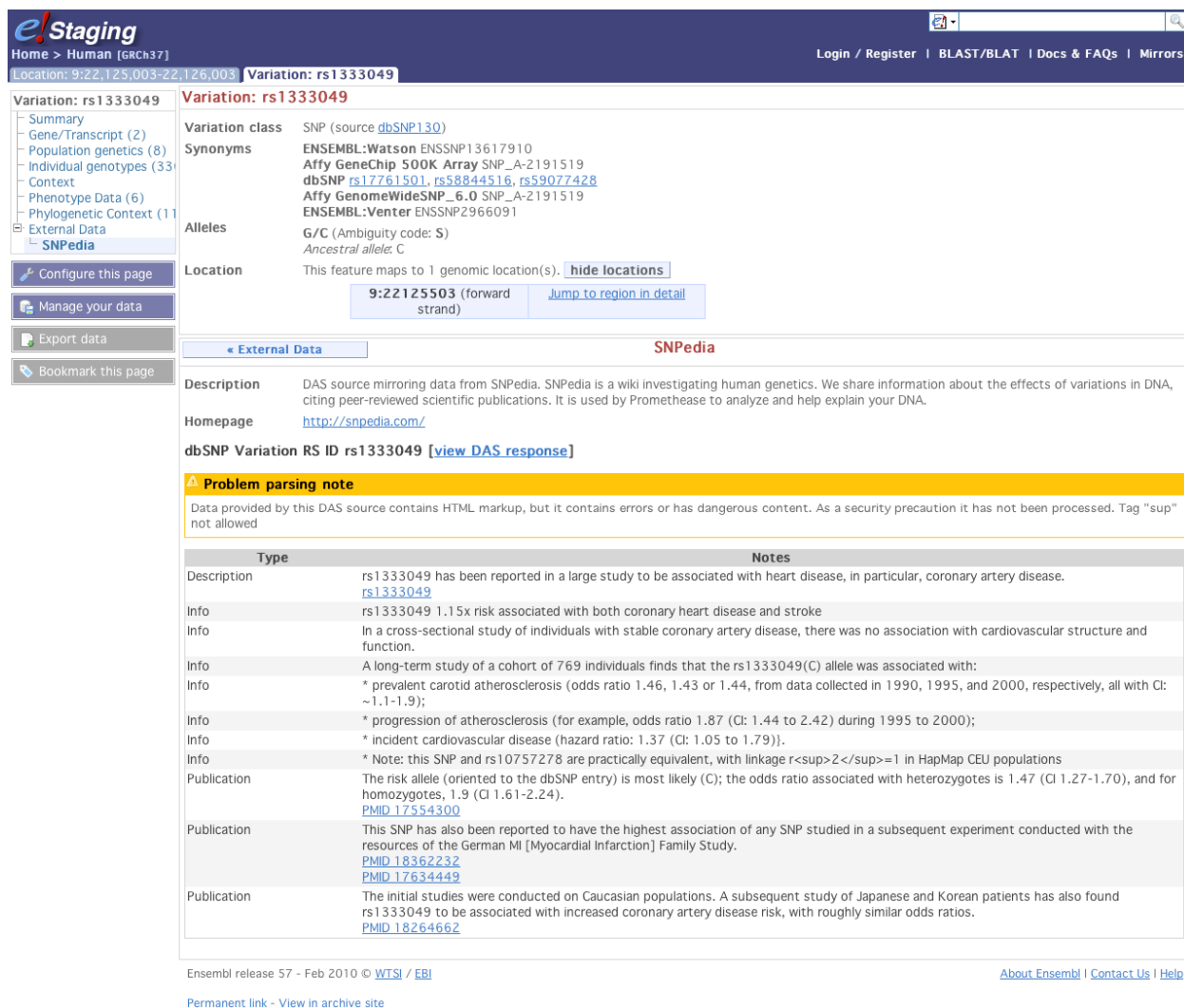
The screenshot shows the Ensembl configuration page for a specific variation (rs1333049). A 'Configure' popup window is open, displaying the 'DAS sources' section. In this section, the 'SNPedia' source is checked, indicating it is selected. Below the list, there is a note: 'To update this configuration, select your tracks and other options in the box above and close this popup window. Your view will then be updated automatically. [Reset configuration for this page to default settings.](#)' The background interface includes a navigation menu on the left with options like 'Summary', 'Gene/Transcript', 'Population genetics', 'Context', 'Phenotype Data', 'Phylogenetic', 'External Data', and 'SNPedia'. The top of the page shows the 'e/ Staging' logo and navigation links like 'Login / Register', 'BLAST/BLAT', 'Docs & FAQs', and 'Mirrors'.

Figure 6: The Ensembl configuration page showing the SNPedia DAS source which has been selected. As additional GEN2PHEN data is made available for SNPs via DAS, they would appear on this page if incorporated into the Ensembl default sources or through the DAS registry for all DAS sources that have provided information to the DAS registry.

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	15/27


4.3.3. GEN2PHEN data displayed via DAS in Ensembl

Figure 7 shows the final integration of the SNPedia data within Ensembl. This prototype provides a general method to incorporate any DAS sources into clients supporting the proposal in section 3. Ensembl supports this capability as of Release 57 (expected March 2010) and will publicise this new feature as a way to encourage investigators both within and outside the GEN2PHEN project to make data available via DAS and to build additional DAS clients capable of displaying GEN2PHEN data.



The screenshot shows the Ensembl Variation page for rs1333049. The page is titled "Variation: rs1333049" and includes a navigation menu on the left with options like "Summary", "Gene/Transcript (2)", "Population genetics (8)", "Individual genotypes (33)", "Context", "Phenotype Data (6)", "Phylogenetic Context (1)", "External Data", and "SNPedia". The main content area shows the variation class as "SNP (source dbSNP130)", synonyms including "ENSEMBL:Watson ENSSNP13617910" and "Affy GeneChip 500K Array SNP_A-2191519", and alleles as "G/C (Ambiguity code: S)" with the ancestral allele being "C". The location is "9:22125503 (forward strand)". Below this, there is a section for "External Data" from "SNPedia". The description states: "DAS source mirroring data from SNPedia. SNPedia is a wiki investigating human genetics. We share information about the effects of variations in DNA, citing peer-reviewed scientific publications. It is used by Promethease to analyze and help explain your DNA." The homepage is "http://snpedia.com/". The dbSNP Variation RS ID is "rs1333049 [view DAS response]". A yellow warning box indicates a "Problem parsing note" stating that the data contains HTML markup but has errors or dangerous content. Below this is a table of "Notes" with columns "Type" and "Notes". The notes include: "rs1333049 has been reported in a large study to be associated with heart disease, in particular, coronary artery disease.", "rs1333049 1.15x risk associated with both coronary heart disease and stroke", "In a cross-sectional study of individuals with stable coronary artery disease, there was no association with cardiovascular structure and function.", "A long-term study of a cohort of 769 individuals finds that the rs1333049(C) allele was associated with: * prevalent carotid atherosclerosis (odds ratio 1.46, 1.43 or 1.44, from data collected in 1990, 1995, and 2000, respectively, all with CI: ~1.1-1.9); * progression of atherosclerosis (for example, odds ratio 1.87 (CI: 1.44 to 2.42) during 1995 to 2000); * incident cardiovascular disease (hazard ratio: 1.37 (CI: 1.05 to 1.79)).", "* Note: this SNP and rs10757278 are practically equivalent, with linkage r²=1 in HapMap CEU populations", "The risk allele (oriented to the dbSNP entry) is most likely (C); the odds ratio associated with heterozygotes is 1.47 (CI 1.27-1.70), and for homozygotes, 1.9 (CI 1.61-2.24).", "This SNP has also been reported to have the highest association of any SNP studied in a subsequent experiment conducted with the resources of the German Mi [Myocardial Infarction] Family Study.", "The initial studies were conducted on Caucasian populations. A subsequent study of Japanese and Korean patients has also found rs1333049 to be associated with increased coronary artery disease risk, with roughly similar odds ratios." The footer of the page includes "Ensembl release 57 - Feb 2010 © WTSI / EBI" and "About Ensembl | Contact Us | Help".

Figure 7: The Ensembl Variation tab showing data directly imported from the SNPedia DAS source. Information provided by the DAS source includes information about the publications citing the SNP, information on associations and phenotype information and other details. As SNPedia is updated, this information is incorporated in real time into Ensembl.

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	16/27

APPENDICES

The DAS protocol was originally specified approximately 10 years ago [1] and has been extended several times since [2-7]. In order to make this deliverable internally consistent, the following sections provide details of the specification, which are then refined below to represent the implementation available for G2P data.

A. The Distributed Annotation System (DAS) Protocol and Implementation from Dowell, et al [1]

DAS implementation

The basic system is composed of a genome server, one or more annotation servers, and an annotation viewer. The genome server is responsible for serving genome maps, sequences, and information related to the sequencing process. Annotation servers are responsible for responding to requests on a region and delivering annotations. The client, an annotation viewer, is a lightweight application whose behavior is analogous to a web browser. The viewer communicates with the genome and annotation servers using a well defined language specification.


At a fundamental level, all annotations can be reduced to their coordinates relative to a particular sequence landmark. The DAS viewer retrieves annotations from the various annotation servers and uses the sequence coordinates to generate an integrated index of what is on the genome. This integration is then presented to the user in tabular or graphical form. Annotation providers can provide a suggestion of how their annotations should be rendered in a graphical display, and can provide links back to their databases and web sites to allow the researcher to retrieve further information about the annotation.

Because it relies entirely on sequence coordinates to achieve integration, DAS does not attempt to resolve semantic contradictions between different data sources. The goal of the system is to provide indexing and visualization, thereby making contradictions between annotations visible.

Reference sequence

The distributed annotation system relies on there being a common "reference sequence" on which to base annotations. The reference server consists of a set of "entry points" into the sequence, and the lengths of each entry point. Entry points will vary from genome to genome. For some genome projects, entry points correspond to entire chromosomes. For others, entry points may be a series of contigs.

The entry points describe the top level items on the reference sequence map. It is possible for each entry point to have substructure, basically a series of subsequences (components) and their start and end points. This structure is recursive. Annotations take the form of a statement about a

	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	17/27

region of the reference sequence. Each annotation is unambiguously located by providing its position as the start and stop positions relative to a "reference sequence."

To give a concrete example, the *C. elegans* reference map consists of six top level entry points, one per chromosome. Each chromosome is formed from several contigs called "superlinks," and each superlink contains one or more smaller contigs called "links." Links in turn are composed of one or more fully-sequenced clones [8]. One could refer to an annotation by specifying its start or stop positions in clone, link, superlink, or chromosome coordinates.


The reference sequence server is responsible for providing the reference sequence map and the underlying DNA. The server can provide a list of sequence entry points or given a component of the map it can return its parent and children components. The reference server can provide arbitrarily long stretches of raw DNA sequence given a reference subsequence, start position, and stop position. Needless to say, bandwidth becomes a limiting factor for retrieving multi-megabase segments of DNA. However, in practice it is rare for users to retrieve more than a gene's worth of raw DNA at a time.

Annotation servers

Annotation servers are specialized for returning lists of annotations across defined regions of the genome. Each annotation is anchored to the genome map by way of a start and stop position relative to one of the entry points. Annotations have an identifier that is unique to the providing server and a structured description of its nature and attributes. The general description of an annotation follows loosely the general feature format (GFF) which intentionally aims for a basic lowest common denominator description <http://www.sanger.ac.uk/Software/formats/GFF/>. Annotations may also be associated with URLs where additional human or machine readable information about the annotation can be found.

The annotator is free to describe his annotations using any terms which he feels are appropriate, as DAS does not impose a controlled vocabulary. Annotations have categories, types, and methods defined by the annotator. The annotation type corresponds to a biological significance description. In the Eddy Lab RNA track of the HGP three types are defined, "tRNA", "snoRNA", and "miscRNA". The annotation method is intended to describe how the annotated feature was discovered, and may include a reference to a software program. The annotation category is a broad functional category. "Homology", "variation" and "transcribed" are example categories. This structure allows researchers to add new annotation types if the existing list is inadequate without entirely losing all semantic value. It is intended that larger annotation servers provide URLs to human-readable information that describes its types, methods and categories in more detail.

Another optional feature of annotation servers is the ability to provide hints to clients on how the annotations should be rendered visually. This is done by returning a DAS "stylesheet."

	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies		Security: PU
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)		Version: v2.0 Final

Stylesheets use the type and category information to associate each annotation with a particular graphical representation, a glyph.

Although the servers are conceptually divided between reference servers and annotation servers, there is in fact no key difference between them. A single server can provide both reference sequence information and annotation information. The main functional difference is that the reference sequence server is required to serve the coordinate map and the raw DNA, while annotation servers have no such requirement.

Specification

The main component of DAS is the XML specification, which defines all valid DAS communication. As with HTML, our goal is a language which is human readable, easily parsed, and extensible. 1 provides a summary of version 1.01 of the DAS specification.

While a client can query multiple servers simultaneously, the communication between the client and any single server follows a simple client server model. Clients query the reference and annotation servers by sending a formatted URL request to each server. Each URL has a site-specific prefix, followed by a standardized path and query string. The standardized path begins with the string /das. This is followed by URL components containing the data source name and a command. For example:


<http://stein.cshl.org/das/elegans/features?segment=ZK154:1000,2000>

In this case, the site-specific prefix is <http://stein.cshl.org/>. The request begins with the standardized path /das, and the data source, in this case /elegans. This is followed by the command /features, which requests a list of features relative to a given set of named arguments (?segment=ZK154:1000,2000). The data source component allows a single server to provide information on several genomes.

Servers process the request and return a response as defined by the DAS specification, typically a formatted XML document. The response from the server to the client consists of a standard HTTP header with DAS status information within that header followed optionally by an XML file that contains the answer to the query. The DAS status portion of the header consists of two lines. The first is X-DAS-Version and gives the current protocol version number, currently DAS/1.0. The second line is X-DAS-Status and contains a three digit status code which indicates the outcome of the request. The defined status codes are listed in Table A1.

Table A1. Server Status Codes Server status codes are modeled after the familiar status codes of the HTTP 1.0 protocol.

Code	Meaning
200	OK, data follows
400	Bad command (command not recognised)

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	19/27

401	Bad data source (data source unknown)
402	Band command arguments (arguments invalid)
403	Bad reference object (reference sequence unknown)
404	Bad stylesheet (request stylesheet unknown)
405	Coordinate error (out of bounds/invalid)
500	Server error, not otherwise specified
501	Unimplemented feature

An example HTTP header: (*provided by server*)

HTTP/1.1 200 OK

Date: Sun, 12 Mar 2000 16:13:51 GMT

Server: Apache/1.3.6 (Unix) mod_perl/1.19

Last-Modified: Fri, 18 Feb 2000 20:57:52 GMT

Connection: close

Content-Type: text/plain

X-DAS-Version: DAS/1.0

X-DAS-Status: 200

DATA FOLLOWS ...

The specification outlines seven basic queries which a client can use to interrogate a DAS server. The valid queries are briefly summarized in Table A2. Two queries, "dsn" and "entry points", essentially provide information to the client about the structure of the server and the reference sequence. The "dna" query can be used to fetch a segment of DNA from a reference server. A client can request annotations, "features", or a summary of the annotations available, "types", from any DAS server. The main annotation content query, "features", basically follows the general feature format (GFF). The servers provide a "stylesheet" to suggest representations to the client's graphical display. When more information is desired about a particular annotation, the client makes a "link" request. The "link" request, the only query which does not return a structured XML document, returns HTML. It is anticipated that DAS clients will hand off the link requests to the local web browser or other web-accessible genome database.


 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	20/27


Table A2. Queries Summary The basic seven queries of the DAS 1.01 specification.

Command	Basic Format	Scope
dsn	PREFIX/das/dsn	both
entry-points	PREFIX/das/DSN/entry points	reference
dna	PREFIX/das/DSN/dna?segment=SEG	reference
types	PREFIX/das/DSN/types?segment=SEG	both
features	PREFIX/das/DSN/features?segment=SEG	both
stylesheet	PREFIX/das/DSN/stylesheet	both
link	PREFIX/das/DSN/link?field=TAG;id=ID	both

Servers

A server is expected to respond to the DAS specification's defined queries with the appropriate content, usually XML. The details of server implementation are left to the various annotation source providers. We provide a sample Perl script for converting ACeDB-based databases into DAS servers, and the Dazzle Java library does the same thing for annotation databases based on the Ensembl code base (T. Down, personal communication, 2001).

The first reference DAS server was written for WormBase [8] and piggybacks on the WormBase software architecture: an Apache/mod_perl web server communicating with an ACeDB database via the AcePerl database access library. The Perl DAS server accepts incoming DAS requests, translates them into the ACeDB query language, reformats the results as XML, and returns them.

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	21/27

B. DAS 1.53E specification from Jenkinson, et al [7]

Coordinate systems

DAS may be used to annotate different types of data. In order to distinguish these, coordinate systems describe the various reference data types DAS supports. Each coordinate system may be thought of as a model that bioinformaticians commonly use to denote biological entities and locations of features within them. A coordinate system has four parts:


1. The category or type of annotatable entity. For example a chromosome, gene, protein sequence or protein structure.
2. The authority or project responsible for defining the coordinate system. For example NCBI, UniProt or Ensembl.
3. The version, used where entities themselves are not versioned (as in genomic assemblies).
4. The species, for coordinate systems containing only entities from a single organism.

Though coordinate systems are normally used to describe the location of a feature within a reference entity (for example residue 26 of UniProt sequence P15056), some annotations are not always associated with a sequence location but rather the entity itself (for example database cross-references). Such features are commonly called non-positional features and are used most when annotating genes, which themselves are often thought of as abstract entities. The difference between annotating an entity versus a region of an entity's sequence is conceptual and requires no special implementation for a data source, but does have implications for a client's display.

DAS commands

A DAS source may offer one or more different services to clients, determined by the commands it implements. A DAS command is a request issued by a client for a certain class of data, such as a sequence or annotations of a sequence. The server responds with an XML document representing the requested data. DAS defines a model for constructing the query (a specific URL format), a model for representing the data (an XML document type) and its means of transport (HTTP). Each command has similar but distinct query and data models. Version 1.53 of the DAS specification (<http://www.biodas.org/documents/spec.html>) has five main commands:

1. entry points – fetches a list of entities a source can annotate
2. sequence – fetches the sequence of a segment of DNA, protein et cetera
3. features – the most commonly implemented command; fetches annotations located within a segment
4. types – fetches a list of the types of feature a source or segment has

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	22/27

5. stylesheet – fetches instructions for displaying features

DAS sources that offer sequences are often referred to as reference sources because they provide the reference entry points for other commands on the same or different servers. Sources implementing the features command are by contrast referred to as annotation sources because they provide annotations based on a reference sequence. This distinction is largely historical since some DAS sources are conceptually both reference and annotation sources, and DAS has since expanded to cover non-sequence data.

The DAS specification has also been extended with several other commands, such as those offering 3D structures and alignments. These are discussed in the Results section.

DAS registry

The steady growth in both the number and diversity of publicly available DAS sources necessitated the development of a method for the discovery of DAS services. Previously reported is the implementation of such a mechanism in the form of the DAS Registry [5]. This service allows data providers to publish their DAS sources, allowing their automatic discovery by compatible clients. This discovery feature has been incorporated into most client implementations and libraries. The registry also performs service validation on registered sources to check that they are both functioning and conforming to the DAS specification. The number of registered sources has steadily increased since the DAS registry was created (see Figure 1).


In recent years the DAS protocol has been expanded beyond the core specification to cater for the data integration needs of additional areas of biological research. However these extensions have yet to be incorporated into the specification itself, the latest version of which is 1.53. Instead, collectively they form an extended version of the DAS protocol, version 1.53E. This protocol, documented at http://www.dasregistry.org/spec_1.53E.jsp, comprises five additional commands, an ontology for protein features, a server-side data preparation option (binning) and additional options for stylesheets. The extensions it offers are all optional for both servers and clients.

New commands

The DAS 1.53E specification defines five new commands.

Structure

Similar to the "sequence" command, this command allows DAS sources to act as reference sources for 3D structures. Clients may request the structure of a given entity, and the source responds with an XML representation of the atomic structure. PDB structures are currently served by a data source maintained by the Wellcome Trust Sanger Institute.

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	23/27

Alignment

This command provides a flexible mechanism for exposing pairwise and multiple alignments of entities. As well as full alignments, clients can request partial alignments containing entities within a given range of a query entity. This is particularly useful for clients wishing to display alignments containing large numbers of entities, such as the protein family alignments displayed on the Pfam website (<http://pfam.sanger.ac.uk>).

DAS alignments may additionally be used by clients as a means of converting between coordinate systems. For example, the Sanger Institute maintains an alignment DAS source that offers mappings between the UniProt and PDB databases. Using an alignment as an intermediary, it is possible for clients such as SPICE [2] to project features from one coordinate system to the other.

Interaction

The interaction command is used for unifying and integrating different sources of molecular interaction data. A DAS source implementing this command supplies XML representations of molecular interactions.


The DAS representation of an interaction is flexible enough to allow many types of interactions, including those for which the interacting region is known and those for which it is not. The XML document contains a list of interactions and a list of the interacting entities (termed "interactors"), with each interaction referencing two or more interactors. In addition to standard attributes such as name and database source, both interactions and interactors may be further described with additional custom properties.

An interaction DAS source can be queried using one or more interactor identifiers, whereupon the DAS source returns interactions involving them all. The client can also request that interactions be filtered by their custom properties, specifying either interactions for which a given property is defined or those for which the property matches a given value.

Volmap

The volmap command is used for syndicating 3D structure volume map data from electron microscopy. It accepts a single "query" ID, and the simple XML response contains metadata for the volume map and a link to the raw data. Unlike other DAS commands, the data itself is not encapsulated in XML due to its large size. The 3DEM group at the Spanish National Center for Biotechnology offers DAS reference and annotation servers for volume map data, and have developed the PeppeR client to facilitate its display [9].

Sources

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	24/27

The sources command is different from other DAS commands in that it is not implemented by individual DAS sources. Instead it is typically implemented by the servers on which DAS sources are hosted, and provides metadata describing their DAS sources. This allows clients and end users to discover the services a server offers. The command details for each source:

1. The capabilities (commands) the source responds to.
2. The coordinate systems the source offers data for.
3. A contact email address.
4. Custom properties that describe the source further (such as the project the source belongs to).

Through the sources command, the DAS Registry can automatically 'mirror' individual servers, significantly augmenting the federation capabilities of the DAS protocol.

Protein feature ontology


The DAS protocol is intended to facilitate user-driven data integration such as graphical interfaces, and to enable data providers to quickly and easily expose their data. For these reasons, although the data transport mechanism has a defined structure, unlike other data integration technologies DAS does not impose strict semantic constraints on the data itself. Whilst this has resulted in widespread adoption, data shared via DAS are typically not amenable to automated analysis because the relationships between data types cannot be reliably inferred and it is difficult to assess their relative significance. To address this shortcoming, the DAS/1.53E specification defines an ontology for sharing protein feature annotations within a controlled vocabulary, developed jointly by the BioSapiens, UniProt and Gene Ontology projects. Currently, 34 BioSapiens DAS sources are committed to implementing the ontology in their annotations, though any source may choose to do so.

The ontology is an optional extension to the features DAS command, and because it is implemented by convention rather than by modifying the XML schema it is fully backwards compatible. The ontology itself is actually a composite of three ontologies:

1. Sequence Ontology [10], an established ontology describing features of biological sequences.
2. PSI-MOD, an ontology for post-translational modification terms.
3. A new ontology for BioSapiens-specific terms not covered elsewhere, such as literature references and other non-positional annotations.

Command extensions

The DAS 1.53E specifications defines two new optional extensions to existing commands.

	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	25/27

Binning

A core principle of DAS is the notion of servers being relatively simple, which lowers the requirements for data providers to expose their data. However, some DAS sources can potentially serve very large numbers of annotation features for a given segment of sequence. This creates problems for user-driven clients that rely on fast response times. Often, the client is not capable of rendering all these features because the user interface has insufficient resolution. For example, a DAS source might annotate every base in a megabase region of the genome, but the user of a graphical client will not be able to see every annotation.

To solve the speed issue, the Ensembl DAS client takes advantage of this fact. It adds a maxbins parameter to a "features" command request. This parameter informs the DAS source of the client's maximum available rendering space by means of the number of 'bins' that features may be placed into. The DAS source may then choose to optimise its response by only returning features that are renderable by the client (i.e. maximum one per bin). It is up to the DAS source to decide which features it should filter.


Advanced stylesheets

Some DAS sources opt to provide stylesheets – generic blueprints that allow a client, if it so wishes, to render features according to the intention of the DAS source provider. The core specification defines several glyphs that a feature can be rendered as such as boxes, lines and arrows. Stylesheets, as in other DAS commands, are provided in XML format and work by specifying the size, colour and type of glyph to be rendered for each type of annotation provided by the features command.

Though stylesheets work well in representing sequence annotations such as exons, it is often desirable for some feature annotations to be rendered in more elegant formats. The 1.53E specification contains new glyph types for the "stylesheet" command that allow a server to define new ways of rendering data. The most notable of these are instructions for rendering plots according to a feature's score property. Different plot types include histograms, colour gradients, line plots and tiling arrays (wiggle plots).


Implementation

The DAS specification has several client implementations. The Ensembl genome browser [11] incorporates a DAS client for several of its "views", and is able to display data from a wide variety of genomic, gene and protein sequence coordinate systems. It also integrates with SPICE [2], a Java Web-Start application that uses DAS alignments to combine protein sequence and structural annotations. Using SPICE, protein sequence annotations can be projected onto and visualised within a 3D structure. The DASMIweb portal (<http://dasmi.de>) integrates protein-protein and domain-domain interaction datasets. The iPfam website also integrates interaction data, comparing the interaction topologies of different sources by overlaying them in a node

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	26/27

graph (<http://ipfam.sanger.ac.uk/graph>). Other clients include Dasty [3], a web-based standalone DAS client implemented in Javascript, and the Pfam website (<http://pfam.sanger.ac.uk>).

Several solid client implementations are based on open source libraries, which are available for the Perl and Java programming languages. These include Bio::Das::Lite (<http://search.cpan.org/~rpettt/Bio-Das-Lite/>) and the Dasobert component of BioJava (<http://www.spice-3d.orgldasobert/>). DAS server implementations are also provided for both languages: ProServer [12] and LDAS (<http://www.biodas.org/servers/LDAS.html>) for Perl; Dazzle (<http://www.biojava.org/wiki/Dazzle>) and MyDas (<http://code.google.com/p/mydas>) for Java.

 HEALTH-200754	D3.8 DAS Implementation for GEN2PHEN Data		
	WP3: Standard Data Models And Terminologies	Security: PU	
	Author(s): Paul Flicek (EMBL – European Bioinformatics Institute)	Version: v2.0 Final	27/27

References

1. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L: **The distributed annotation system.** *BMC Bioinformatics* 2001, **2**:7.
2. Prlić A, Down TA, Hubbard TJ: **Adding some SPICE to DAS.** *Bioinformatics* 2005, **21 Suppl 2**:ii40-ii41.
3. 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* 2005, **21**:3198-3199.
4. Olason PI: **Integrating protein annotation resources through the Distributed Annotation System.** *Nucleic Acids Res* 2005, **33**:W468-W470.
5. Prlic A, Down TA, Kulesha E, Finn RD, Kahari A, Hubbard TJ: **Integrating sequence and structural biology with DAS.** *BMC Bioinformatics* 2007, **8**:333.
6. Andreeva A, Prlić A, Hubbard TJ, Murzin AG: **SISYPHUS--structural alignments for proteins with non-trivial relationships.** *Nucleic Acids Res* 2007, **35**:D253-D259.
7. Jenkinson AM, Albrecht M, Birney E, Blankenburg H, Down T, Finn RD, Hermjakob H, Hubbard TJ, Jimenez RC, Jones P, Kähäri A, Kulesha E, Macías JR, Reeves GA, Prlic A: **Integrating biological data - the Distributed Annotation System.** *BMC Bioinformatics* 2008, **9 Suppl 8**:S3.
8. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J: **WormBase: network access to the genome and biology of *Caenorhabditis elegans*.** *Nucleic Acids Res* 2001, **29**:82-86.
9. Macías JR, Jiménez-Lozano N, Carazo JM: **Integrating electron microscopy information into existing Distributed Annotation Systems.** *J Struct Biol* 2007, **158**:205-213.
10. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M: **The Sequence Ontology: a tool for the unification of genome annotations.** *Genome Biol* 2005, **6**:R44.
11. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Gräf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kähäri A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A, Searle S: **Ensembl 2008.** *Nucleic Acids Res* 2008, **36**:D707-D714.
12. Finn RD, Stalker JW, Jackson DK, Kulesha E, Clements J, Pettett R: **ProServer: a simple, extensible Perl DAS server.** *Bioinformatics* 2007, **23**:1568-1570.