

CGIAR Generation Challenge Program

Web Service Working Group

Justification for use of the BioMOBY system

Mark Wilkinson

Martin Senger

Ben Good

Alex Garcia

Jon Mendoza

Mathieu Rouard

July 6-9, 2004

preliminary recommendations

CONTENTS

Background Knowledge

What is a web service?

How do Web Services compare to existing web interfaces?

How do web service systems compare to other data search & retrieval systems such as SQL?

Why BioMOBY? Executive Summary

The BioMOBY Project Overview

Funding

License

End-user support

Project scope

Underlying technology

The registry

The messaging format

The ontologies

Pervasiveness of BioMOBY

The future of BioMOBY

The PlaNet use case

Background Knowledge

What is a web service?

A web service is a WWW-based interface to a program running on a host computer that allows that program to be run, and the results returned, from anywhere on the Internet. Web services typically use a messaging protocol called “Simple Object Access Protocol” (SOAP), which defines a message structure that contains both data, and the information the host computer needs to map the incoming data to the appropriate program. SOAP interfaces are described by a machine-readable document written in a form of XML called Web Service Description Language (WSDL). In the web service paradigm, a third party “registry” holds metadata about web services, and acts as a broker between a client who is looking for a data service, and the service provider that can accomplish this request. Human intervention is required at several steps of the process: Creation of the metadata query to discover services of interest, selection of that service, and (often) manipulation of data prior to service execution such that it conforms to the service providers interface.

How do Web Services compare to existing web interfaces?

Web Services are very similar to the more familiar web forms that are common on the Internet. In a web form, a human is presented with a set of fields into which they type whatever information is required before clicking a “submit” button and sending the information to the service provider. The main difference between these interfaces, and Web Services, is in the discovery of the interface and the ability of a machine to “understand” the interface definition and thus be able to interact with the service provider without human intervention. On the Web, humans will often use a tool such as Google to find an interface that can accomplish their needs. Google has made an index of the words on the web page, and it matches those words against your request to give you the best list of matching pages. While Google stores human readable text from web pages, a Web Service registry stores machine-readable definitions of the interface itself. For example, that the interface defines a piece of text called “sequence”, and a floating point number called “cutoff value”. This allows the programmatic discovery, and (semi-)automated execution of the discovered service. Most importantly, you find the service based on queries about the service definition itself, rather than by matching words on a web page.

How do web service systems compare to other data search & retrieval systems such as SQL?

Web Services are designed, like the Web itself, to operate in a distributed environment where data may be resident on a large number of disparate hosts anywhere in the world. They are more limiting than SQL in a number of ways:

- 1) They do not support ad hoc query construction. You can only execute queries

that the service provider has written services to support.

- 2) They are incapable of processing multi-table SQL joins, so any conditionals in the query that set conditions on data in two locations must be processed on the users machine.

However, there is currently no technology that supports SQL queries over a distributed network of databases, and as such, web services are the most flexible option available for collaborative relationships between service providers.

Why BioMOBY – Executive Summary

- The more simplistic of two extant Web Services systems (the other being myGrid - <http://www.mygrid.org.uk>) capable of mapping biological data-types to their applicable services
- Implementation within minutes!
- Strong support from the Plant genomics community (for historical reasons)
- Community-driven, needs-based development philosophy
- Open source, freely available
- Strong, responsive distributed on-line support community
- Uses existing standards as much as possible
- Arbitrarily extensible to new areas of biological and/or non-biological knowledge and data
- Increasing support in existing browsers and client-side tools and libraries.

The BioMOBY Project Overview

BioMOBY (<http://www.biomoby.org>) was founded by members of the Model Organism Database community in response to the critical need for a simple, extensible data-sharing platform. The design of the BioMOBY system was undertaken with practicality as its primary directive, and more theoretical aspects of “correct” web design were only implemented if they were of benefit to the goal of deploying a useful system that could be implemented rapidly, easily, with minimal disruption to existing systems, by service providers who had limited time and human resources. Typically, BioMOBY can be implemented by a novice user within the first day. The project is led by Dr. Mark Wilkinson from the University of British Columbia, Vancouver, BC, Canada.

BioMOBY currently has two main branches of development – MOBY Services (MOBY-S) is designed on a Web Service paradigm, while Semantic MOBY (S-MOBY) is designed with a Semantic Web paradigm. Of these two, MOBY-S is the more mature, and is the system that is being recommended for use in the Generation Challenge Program. MOBY-S defines a platform through which data and analysis services in a distributed network can be discovered and executed in a fully or semi-automated manner. Mapping of data onto relevant services is accomplished by a web service registry

(“broker”), such that only services capable of operating on, or providing, the desired data types are discovered. For historical reasons, the main users of BioMOBY thus far have come from the plant research community, however the representation of non-plant model systems is increasing rapidly.

Funding

The BioMOBY project has a good pedigree of supporting agencies. BioMOBY is directly funded by Genome Canada through the Genome Canada Bioinformatics Platform grant (Dr. Christoph Sensen, Head PI), and by the National Science Foundation of the United States through a grant to Dr. Lincoln Stein and Dr. Damian Gessler. A significant amount of indirect funding flows into the project through re-allocation of existing resources towards BioMOBY research, development, and deployment by organizations such as the European Bioinformatics Institute (Dr. Peter Rice), The San Diego Supercomputing Centre (Dr. Michael Gribskov), The Rat Genome Database (Dr. Simon Twigger), the European PlaNet consortium (Dr. Hans Werner Mewes, Dr. Heiko Schoof, et al.), and IBM (Dr. Michael Niemi) to name just a few.

License

The core of the BioMOBY codebase is developed under the Open Source Perl Artistic License (<http://www.opensource.org/licenses/artistic-license.php>). As such, all code will indefinitely be available at no charge and with no significant restrictions on end-use, modification, extension, or commercial use. Supporting code for the project (e.g. client/server side “convenience” libraries) are distributed with their own licenses, and these are also, for the most part, open source.

End-user support

The project enjoys a strong and responsive user-support community through its users mailing list. Since there are MOBY developers on every continent, the coverage of time-zones is fairly comprehensive, and questions posted to the list are generally answered within hours of posting.

Project Scope

The BioMOBY project is largely aimed at solving problems of interoperability in the biological domain; however the semantic registry system it defines is not limited to biological data, and could be applied to a much wider range of scientific or commercial

activities. At the present time it is being used primarily as a platform for achieving interoperability and integration between biological data resources, either on an ad hoc basis (client-side on-demand definition of an integrated network), or as a planned network of collaborative sites (e.g. the PlaNet consortium: <http://mips.gsf.de/proj/planet/>). BioMOBY places no restrictions on the types of data that can be integrated, and is compatible with all existing data formats, many of which have already been registered in the BioMOBY ontologies. The extension of BioMOBY to define new data-types is open and end-user driven, thus the system can scale and be configured to represent most common biological databases.

Underlying Technology

The BioMOBY system uses existing standards wherever it is practical to do so, with the caveat that existing standards that interfere with interoperability have been avoided, even if new standards had to be defined as a result.

BioMOBY consists of three core technologies:

1. A web service registry (MOBY Central) with a SOAP/HTTP-based API
2. An XML-based messaging format that is compatible with a variety of common web transport protocols such as HTTP, FTP, and SMTP
3. A set of ontologies describing biological data types, bioinformatic analysis types, and biological data domains.

The registry

The MOBY Central registry is a “yellow pages” of biological data retrieval and analysis interfaces, and is capable of returning a human-readable list of services, and their machine-readable interface definitions, in response to queries based on input data types, output data types, and/or analysis types. It is currently backed by a relational database (currently MySQL), and code restructuring is currently underway to allow a wider variety of database back-ends, including Postgres, and MSAccess through ODBC. The registry exposes a variety of API procedures for registering and discovering services, as well as retrieving the service interface definitions. MOBY Central uses a SOAP (Simple Object Access Protocol) interface and well defined XML (eXtensible Markup Language) message structure. Multiple instances of the MOBY Central registry can (and do) exist, and the creation of “boutique”, or task-specific registries can be easily accomplished; this is particularly relevant in the context of a collaborative network such as the Generation Challenge Program. It is important to note that *only one site* in the collaborative network needs to implement the registry and underlying database, thus the administrative overhead is very low.

The messaging format

The messaging format defines the scaffold of the interfaces between a data provider and a data consumer. This was derived after an extensive Use Case analysis (<http://www.biomoby.org/twiki/bin/view/General/WebHome>), and is capable of supporting most common bioinformatic web-based data retrieval and analysis processes, including batch-executions and note-booking. At this time, the BioMOBY system restricts client/server interactions to use only this messaging format, however project developers have begun to explore the possibility of allowing a wider range of messaging formats in the advent of the emergence of LSIDs as a simple and predictable identification system. Preliminary code is currently being tested.

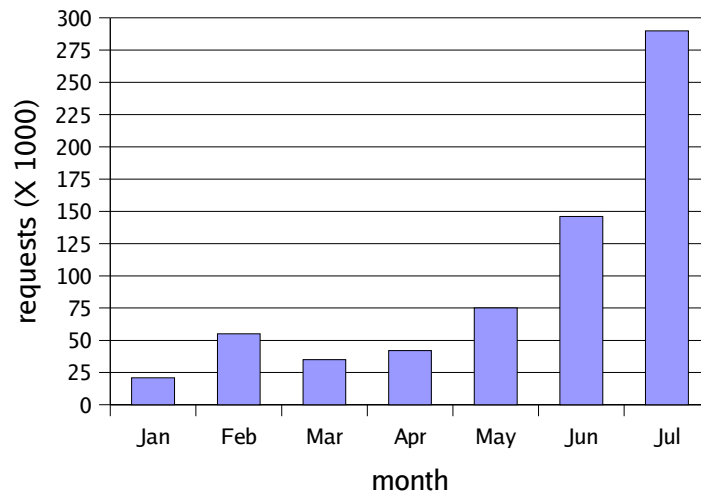
The ontologies

Three simple ontologies have been defined – the Object ontology describes the structure and relationships between data-types that can be passed in the messaging system; the Namespace ontology describes a set of semantic data types, i.e. the underlying data that is being represented in the structured Object; and the Service ontology describes the types of manipulations that can be performed on data Objects (e.g. “R etrieve”, “B LAST”, “P arse”, e tc.). The ontologies are easily extended to include new data and service types in a process that is end-user driven, and the MOBY Central API exposes most operations that are commonly performed on these ontologies such as the registration of new data types, or querying the relationship between data types. To simplify the use of these ontologies by service providers and consumers, they have been built using a standard set of paradigms established by the Gene Ontology (GO: <http://www.geneontology.org>) consortium – an ontology that is particularly familiar to biologists. In particular, the ontologies consist only of “part of”, and “is a” relationships.

Pervasiveness of BioMOBY

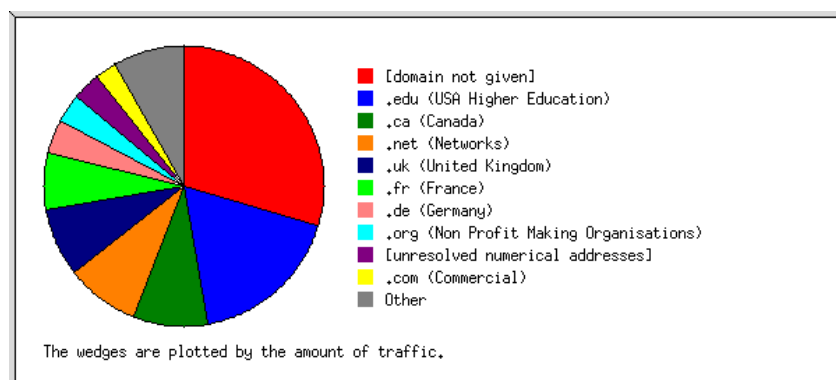
The final public version of the BioMOBY API and core code was released in late 2003. The chart below describes the usage statistics for the MOBY Central API in the first 7 months of 2004. It shows an exponential increase in use of the MOBY Central registry, with the projected numbers for July reaching almost 300,000 requests (99051 requests as of July 11th) on the public registry alone. In addition, a number of institutions have implemented site-specific registries on their own servers, so these numbers underestimate the actual usage worldwide.

Use of the MOBY Central Registry January - July, 2004.



Despite the increasing usage, the registry has considerable room for expansion before traffic becomes a problem. At its peak usage, the registry currently serves less than 8 Megabytes of data per day. This is because the registry does not itself serve biological data, and the registry messages are extremely lightweight, primarily acting as pointers to the real service providers.

Approximately 100 services are currently registered in the public registry, and these provide a wide variety of retrieval and analysis services such as Genbank and EMBL sequence records, NCBI Blast, PDB records and 3-D information, SNP and HapMap data, Gene Ontology browsing and NCBI taxonomy traversal. Currently, the main users of BioMOBY are US and Canadian academic institutions, and a variety of plant databases throughout Europe (see discussion of the PlaNet consortium below). The breakdown of BioMOBY usage by web domain is shown here:



The future of BioMOBY

BioMOBY is an extremely young project, yet it is enjoying remarkable success largely due to the simplicity of implementation and the critical need that it fulfills. The existing codebase is almost fully functional, and all of the critical functions have been written and well tested. Extensive code refactoring is underway to take advantage of emergent technologies such as the Semantic Web (e.g. Resource Description Framework), and to provide for a more flexible underlying database storage layer. We anticipate this work will be complete by the end of the year, and code updates will be extensively tested before being released. As such, these code updates should have minimal, if any, impact on deployed systems. The number of tools that use BioMOBY is currently limited, but is increasing. A simplistic MOBY Browser system is available for people to do basic data-surfing and to test their interfaces (<http://mobycentral.cbr.nrc.ca>). BioMOBY support will soon be integrated into the Rat Genome Database, and is currently being built into standalone applications such as the workflow management systems Taverna (<http://taverna.sourceforge.net/>) from the European Bioinformatics Institute, and Pegasys (<http://www.bioinformatics.ubc.ca/pegasys/>) from the Bioinformatics Centre of the University of British Columbia. With these powerful interfaces on the horizon, the expansion of BioMOBY will no doubt accelerate.

The PlaNet use case

PlaNet (<http://mips.gsf.de/proj/planet/>) is a collaborative network of plant databases throughout Europe. The various member databases specialize in genomic, proteomic, seed-stock, and phenotypic information. To achieve their collaborative network, the PlaNet partners have implemented an independent copy of MOBY Central, hosted at MIPS in Munich, Germany. Partner institutions set up MOBY-compliant services on an ad hoc basis, and register these in the PlaNet MOBY Central (PMC). Incoming queries (e.g. “what can you tell me about the locus At116473”) are first posed to the registry to determine which partner(s) are capable of providing the requested information, and then all discovered services are automatically executed and the results are presented to the user as a comprehensive view. From the perspective of the user, there is a single interface, however in the “guts” of the system the query is being posed to an indeterminate number of institutions throughout Europe and the responses are collated. Most importantly, as member institutions decide that they wish to add new services, or provide new data-types, they may do so without coordinating with the other member institutions. They simply register the existence of their service with PMC, and their data immediately begins to appear on the query portal. Thus the administrative overhead of running the collaborative network is negligible with respect to coordination of data-provision activities.