Web Services for the NLM

By

Michael D. Jensen
NLM Biomedical Informatics Fellow
Summer 2004

Preceptor: Olivier Bodenreider, PhD, MD

Abstract

The surge of new life sciences tools and resources gives researchers an increasingly large pool of applications to use, query, and analyze data with. The task of actually finding the right tool can be difficult, and methods for finding these tools is currently not comprehensive. Once a tool is found, using the tool in an automated manner can be difficult and expensive in time and resources. Additionally, combining tools in a pipeline takes considerable work and is susceptible to being inoperable based on simple changes of a website by service providers. Web Services is an architecture based on open standards that allow services to communicate with other services and clients in an automated way, and allow for self-registration of services with centralized registries, giving users automated service discovery. This project analyzed and assessed various resources of the NLM and compared characteristics of these resources with those of an ideal Web Service in order to categorize resources as strong, fair, or non-candidates for Web Services implementation. The PubMed resource has three interfaces that were assessed for their characteristics as related to Web Services. Finally, the MetaMap program was implemented as a Web Service, registered with a centralized registry, and is serving as a template for other NLM resources to use for Web Services implementation. The results of this project show that most NLM resources are strong candidates for Web Services, that the PubMed interfaces offer a wide array of options and versatility but lack the registry component of true Web Services, and the implementation of a Web Service can be done with minimal time and resources.

Introduction

Within the biomedical domain, tools and resources are increasing at an unprecedented level. Dozens of new applications or databases debut on a weekly basis, creating a vast pool of resources for researchers. Several problems have risen from this surge of resources, namely a lack of automated discovery and use of these tools. Many researchers do not know where to begin looking for tools they need, beyond a standard search engine search. Very few tools actually offer APIs, or any interface other than online browser access. The input and output (I/O) of resources is heterogeneous, with standards typically only being employed internally within large organizations. With a lack of standardization of I/O, linking tools in pipelines for shuttling data through multiple analysis tools is difficult and expensive to maintain.

Web Services offer a comprehensive solution to these problems. Web Services is an architecture for automated communication and discovery of services. Within this report, Web Services will be used in a strict definition, web services are those services that utilize standardized protocols for interoperability and a central registry for automated discovery. With service providers registering the services they offer, by enabling users to query the registry based on service type, I/O, and other factors, and having reusable output of services, Web Services can provide the biomedical domain with an architecture for true interoperability and automated discovery of services.

Background

Attempted Solutions

Many solutions have been attempted to solve these problems, but none of them have proved efficacious. Screen-scraping is a popular technique involving a program that automatically scrapes through the data from an HTML web page and extracts pertinent data. This technique relies on the layout of the presentation-specific elements of the document, and will often break if a service provider modifies even the look of the web site. BioPIPE (1) is a project that combined some of the programming tools from BioPerl (2) to link tools together for performing multiple types of analysis in one run. This method also tends to be very fragile; if one program within the pipeline makes a modification, the entire pipeline becomes useless. Closed technologies like CORBA (Common Object Request Broker Architecture; 3) limit programmers to a particular language. Websites like BioWareDB.org (4) and Journals like the Nucleic Acids Research (5) Database or Web Server Issue seek to maintain lists of tools available, but matters of comprehensiveness, automated discovery, and reliability are not thoroughly addressed.

Web Services

Web Services are not necessarily a new architecture, but with the influx of XML into the data representation realm, new standards based on XML have been developed and adopted throughout entire industries. The business and financial industries have integrated Web Services into applications for communication with vendors, clients, and on intranets. Web Services are beneficial due to a collection of beneficial characteristics:

1) Standardized protocols for I/O
2) Cross-platform, language independent
3) Automated architecture for high throughput analyses, querying
4) Automated service discovery
5) Integration with Semantic Web technologies

A Web Service architecture is made up of three main players (see Figure 1): service providers, a registry, and users. A service provider registers its service with the registry, and is now listed on the registry. A user can then query the registry with a request for services typically based on the input and/or output of the service. The registry then relays to the user the location and communication information to access the service, and the user sends requests directly to the service provider. The service provider analyzes the requests or processes the query, and responds with the appropriate output. The output is in a standardized format and could, without reformatting or parsing, be reused as input for another service the user may wish to utilize.

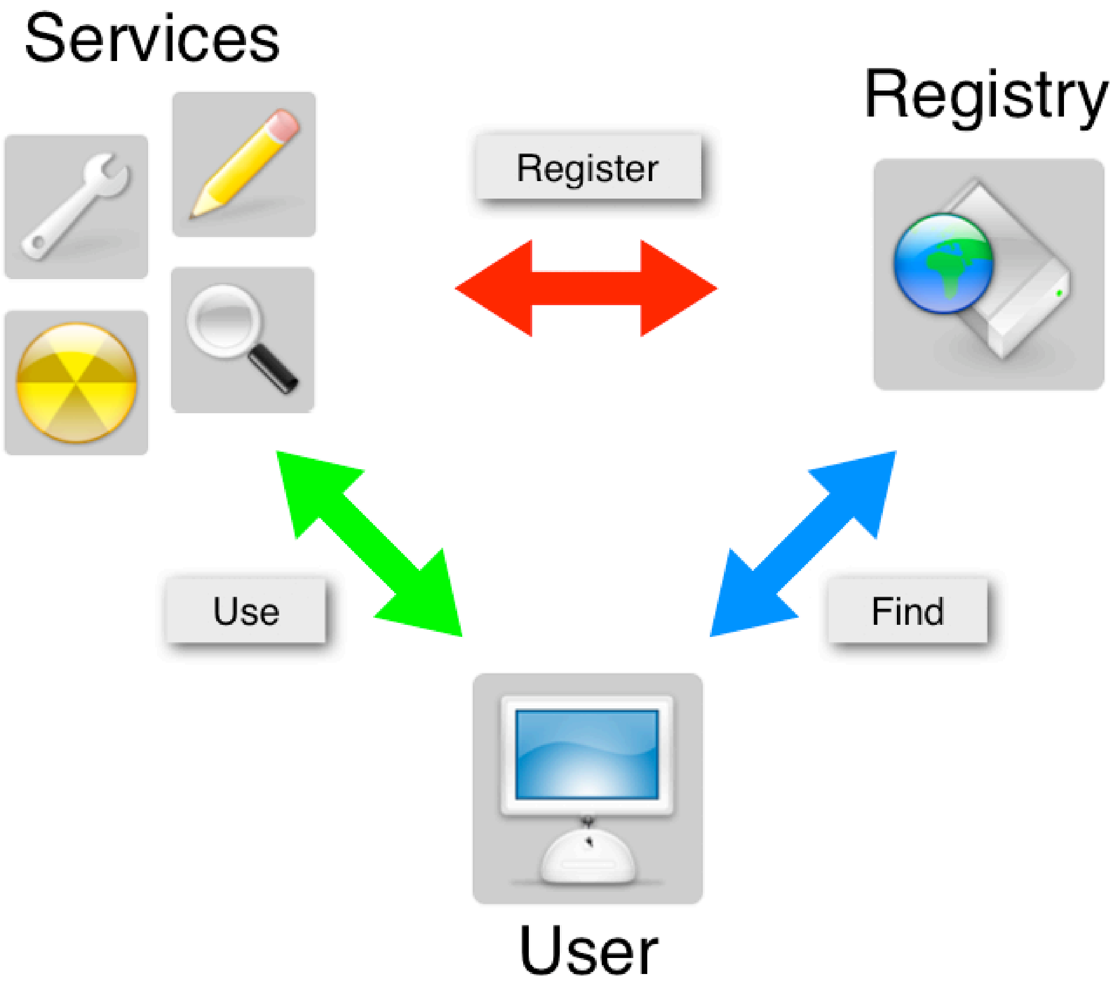

Figure 1. Web Services Architecture and Players

In the Life Sciences there are two main Web Services projects, BioMOBY (6) and myGRID (7). BioMOBY appears more advanced in the actual deployment of the architecture, and currently has more than a hundred services publicly available, along with an operational registry. Examples of services currently available by BioMOBY participants include PubMed, BLAST, EMBL databases, sequence formatting, image databases, natural language processing, gene ontology, taxonomy data, protein interactions, homology databases, and more.

WS and the NLM

The NLM has made great efforts to interoperate between its own resources and offers portals like the Gateway and Entrez to query nearly all of the resources in one sweep. The vast majority of NLM resources also provide APIs or source data for its resources, enabling researchers to embed programs within their own tools for service reuse and automated use. PubMed just recently released a SOAP (Simple Object Access Protocol) interface, the XML messaging protocol used heavily in Web Services. Despite these efforts, it is still difficult to link NLM resources with other tools.

MetaMap

MetaMap (8) discovers Metathesaurus concepts from the UMLS in submitted text. Text is parsed into components, including sentences, phrases, lexical elements, and tokens. Variants are generated from resulting phrases, candidate concepts are retrieved and evaluated, and the best candidates are organized into a final mapping. All output is in text and various flags can affect the output components and if it is in machine format or not. MetaMap has a Java API and online access through a browser. Currently the only way to use MetaMap in an automated fashion is to download the freely available code and install it on a server.

Specific Aims

This project's overall goal is to better understand how the NLM resources fit with the Web Services architecture. This project has three specific aims: (1) identify candidates of NLM resources for Web Services, (2) assess specific characteristics of PubMed's three interfaces, and (3) implement MetaMap as a web service.

Methods

Identify Web Services Candidates of NLM Resources

In order to better understand the NLM resources and their candidacy as Web Services, eight objective questions were formulated pertaining to important aspects of a Web Services implementation. The questions revolve around usage, navigation, output, and accessibility. The questions were designed to have minimal bias or subjectivity. The more commonly known resources of the NLM were analyzed, including PubMed, BLAST, Entrez Gene, MetaMap, Metathesaurus, Specialist Lexicon, Semantic Network, OMIM, MapViewer, Gateway, and GEO (9). The answers to each question were determined for an ideal Web Service, as defined previously within this report. The answers were also determined for each of the NLM resources, and then the results compared to the ideal Web Service (see Table 1 (A-C) for questions and answers). All resources were then categorized into a group based on the amount of differences between the resource and the ideal. The three groups are (1) Strong Candidates, (2) Fair Candidates, and (3) Non-Candidates.

Assess Pubmed Interfaces

PubMed (10) has three main interfaces: Browser (HTML), URL, and SOAP (11). The HTML interface is the most common to users and involves use of a browser to search the PubMed web site. The user can enter queries through a main search box, or use a query guide to help design queries. The URL interface permits a client program to access the results of a query or individual articles by retrieving the results of a URL query in an automated fashion. The output format and query parameters can all be designated in the URL. The SOAP interface is similar to the URL interface but uses XML to contain queries and form a response from the service. The SOAP interface is the newest, and was made public during the eight-week rotation. Seven objective characteristics of interfaces were selected for qualitative assessment. Information was gathered for each interface based on the author's familiarity with the interfaces and generally accepted characteristics of each interface.

Web Service Implementation of MetaMap

MetaMap was installed on a Solaris server. The MMTx Java API providing access to the various methods within MetaMap was utilized for implementation of the Web Service. Apache web server is being used to handle the web server tasks, and Perl and Java are being used for processing requests and sending out responses. The BioMOBY API v8.00 was used for implementing the Web Service. XML-Schema was utilized for producing a schema of the output data (replaces a DTD). MetaMap was registered with the BioMOBY Central Registry as "metaMapBasic". Three namespaces were also registered, UMLS_CUI, UMLS_TUI, and UMLS_SUI for the output. The output object was also registered with the registry.

Results

Most NLM Resources are Strong Candidates

The questions and respective answers for all of the resources analyzed and for the ideal Web Service are found in Table 1 (A-C). Of the eleven resources, nine were strong candidates, one was a fair candidate, and one was a non-candidate. The nine strong candidates all matched the ideal web services potential answers to the specific questions. The fair candidate, OMIM, was categorized as such due to its moderate need for automation. OMIM is composed of long paragraphs of text useful primarily for human reading, although the text can be useful for natural language processing. MapViewer, a tool used for viewing chromosomes and exploring the genes within chromosomes, was labeled as a non-candidate. This is due to its complete dependency on images for navigation. Although Web Services can handle non-text data such as images, MapViewer is most beneficial in a browsing mode.

Interestingly, individuals have created Web Service interfaces out of three of the strong candidates, PubMed, BLAST, and parts of Entrez Gene, using the available APIs and/or downloaded source code. All resources but MapViewer has either source code available or an API. This allows for integration within other software and tools, although maintenance is high with mirroring data sources if a service desires to keep well up-to-date.

The Gateway is a resource of resources. One query generates results from multiple resources and the results can be accessed directly from within the Gateway. As the effort has already been made to create a "one-stop" interface for multiple resources, this may be an easier point of entry for Web Services into some NLM resources, sparing each individual project to implement a new interface.

PubMed Interfaces

Each interface is unique, though there are many similarities between the URL and SOAP interface (Table 2). The HTML interface is geared towards browsing by a user, while the URL and SOAP interface is for automating the task of querying and fetching information about the articles. The HTML interface has the advantage of helping guide the user to design a query. The results of the HTML interface require more processing, including selecting the output format and manually saving the data. Pagination is also an issue with the HTML interface, while the other interfaces can retrieve all articles in one process. The only discovery for any of the interfaces is the HTML interface through a search engine, or linking from another resource.

Web Services Interface to MetaMap

The MetaMap program is now available for use as a Web Service. The service can be invoked by using a SOAP call with text input. The output of the service is a standardized, registered object (see Figure 1 for sample output). The service and components have been registered with the BioMOBY Central registry.

Conclusions

With all but two of the NLM resources designated as strong candidates, the implementation of Web Services for these resources is recommended. The current APIs and downloadable sources are helpful and provide a large coverage of access. The SOAP interface to PubMed is a step toward Web Services, but is missing the registration component. However, without a registry component the services are not easy to find or to communicate with. The registry is vital to solving the problem of finding services in an automated fashion.

Discussion

Analysis of the NLM resources with respect to the implementation of Web Services has resulted in understanding the types of resources that fit into the Web Services architecture. Resources with a demand for high throughput, compartmentalized and reusable output, and absence of image-based navigation are candidates for Web Services. The Gateway could be a point of entry for Web Services, as multiple resources have already been connected to a unified interface. The development of MetaMap as a Web Service can also serve as a framework for future applications. The MetaMap Web Service code can be used as a template for other services, as well as the registration scripts and XML object design.

Future work may involve a full implementation of MetaMap, allowing full access to UMLS authorized users. The authentication procedure will need developed, as well as a means for a request to submit an authorization code.

Web Services offers a powerful yet simple to implement solution to the service finding and automated usage problems in the life sciences. As demonstrated in the last few weeks of the summer rotation, the implementation of Web Services is not overly difficult or time consuming.

References

1. BioPipe, http://biopipe.org [URL: Accessed 05 Aug 2004].
2. BioPerl, http://bioperl.org [URL: Accessed 05 Aug 2004].
3. CORBA, http://www.corba.org [URL: Accessed 05 Aug 2004].
4. Biowaredb.org, http://www.biowaredb.org [URL: 05 Aug 2004].
5. Nucleic Acids Research, http://www.nar.org [URL: Accessed 05 Aug 2004].
6. BioMOBY, http://www.biomoby.org [URL: Accessed 05 Aug 2004].
7. myGrid, http://www.mygrid.org.uk [URL: Accessed 05 Aug 2004].
8. MetaMap, http://mmtx.nlm.nih.gov [URL: Accessed 05 Aug 2004].
9. NLM Website, http://www.nlm.nih.gov [URL: Accessed 05 Aug 2004].
10. PubMed Interfaces, http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html [URL: Accessed 05 Aug 2004].
11. SOAP (Simple Object Access Protocol), http://www.w3.org/TR/soap/ [URL: Accessed 05 Aug 2004].

Table 1.A. Analysis of NLM Projects for Web Service Implementation

	Ideal WS	**BLAST**	**MapViewer**	**Entrez Gene**
How important is having high throughput, automated submission for the service? (Not, Low, Moderate, High)	High Moderate	High	**Not**	High
Are links important for navigating through results? (No, Somewhat, Yes)	No Somewhat	No	**Yes**	Somewhat (other sources)
Are images a part of the output? How are they important?	No Yes (can output non-text objects, but not very reusable)	Yes Dispensable, but help to visualize sequences	Yes (integral in browsing)	Yes Dispensable, but help to visualize molecule
Reusability of output for other services (other than NLP)	Yes	Yes	**No**	Yes
Is the output textual, visual, or data oriented?	Data Textual Endpoint Visual	Data Minor Visual	Visual	Data Minor Visual
Does the output have components, such as text divided in categories, field-based data, etc.?	Yes Somewhat	Yes	**No**	Yes
Is the source code or source files available for download?	Yes No	Yes	No	Yes
Is an API available?	Yes No	Yes	No	No
What formats are the API available in?	Any (XML)	URL (HTML)	None	None
Categorization	**(Ideal)**	**Strong Candidate**	**Non-Candidate**	**Strong Candidate**

Table 1.B. Analysis of NLM Projects for Web Service Implementation

	OMIM	**GEO**	**PubMed**	**Gateway**
How important is having high throughput, automated submission for the service? (Not, Low, Moderate, High)	**Moderate**	High	High	High
Are links important for navigating through results? (No, Somewhat, Yes)	Somewhat (citations)	Somewhat (more details)	Somewhat (related resources)	Somewhat (some require link to original source)
Are images a part of the output? How are they important?	No	No	No	No
Reusability of output for other services (other than NLP)	Somewhat	Yes	Somewhat	Somewhat
Is the output textual, visual, or data oriented?	Textual	Data	Textual	Textual
Does the output have components, such as text divided in categories, field-based data, etc.?	Yes	Yes	Yes	Yes
Is the source code or source files available for download?	Yes	Yes	No	No
Is an API available?	No	No	Yes	Coming
What formats are the API available in?	None	None	URL (XML, text, HTML)	Unknown
Categorizaiton	**Fair Candidate**	**Strong Candidate**	**Strong Candidate**	**Strong Candidate**

Table 1.C. Analysis of NLM Projects for Web Service Implementation

	MetaThesaurus	**Semantic Network**	**Specialist Lexicon**	**MetaMap**
How important is having high throughput, automated submission for the service? (Not, Low, Moderate, High)	High	High	High	High
Are links important for navigating through results? (No, Somewhat, Yes)	No	Somewhat (details for data)	No	No
Are images a part of the output? How are they important?	No	No	No	No
Reusability of output for other services (other than NLP)	Yes	Yes	Yes	Yes
Is the output textual, visual, or data oriented?	Textual, Data	Data	Data	Data
Does the output have components, such as text divided in categories, field-based data, etc.?	Yes	Yes	Yes	Yes
Is the source code or source files available for download?	No	No	No	Yes
Is an API available?	Yes	Yes	Yes	Yes
What formats are the API available in?	Java, XML	Java, XML	Java, XML	Java
Categorization	**Strong Candidate**	**Strong Candidate**	**Strong Candidate**	**Strong Candidate**

Table 2. PubMed Interfaces Analysis

	HTML Interface	URL Interface	SOAP Interface
Automated vs Manual	Manual	Automated	Automated
Accessibility	Browser	Automated Program	Automated Program
Output	HTML, XML, Text, others	HTML, XML, Text, others	Standardized XML with DTD
Query design	Interface Guide	Manually entered	Manually entered
Time Involved to retrieve results	At user's pace, paginated results require multiple interactions	Immediate once set up	Immediate once set up
Quality of Results	Hand-picked results based on query and individual evaluation	Not filtered, all articles retrieved	Not filtered, all articles retrieved
Discovery	Search Engine	None	None