

Data Sharing in Bioinformatics: Using
the MDB's Web Services
225th ACS National Meeting

March 25, 2003 . New Orleans, LA

Jesus M. Castagnetto, Ph.D.
<jesusmc@scripps.edu>

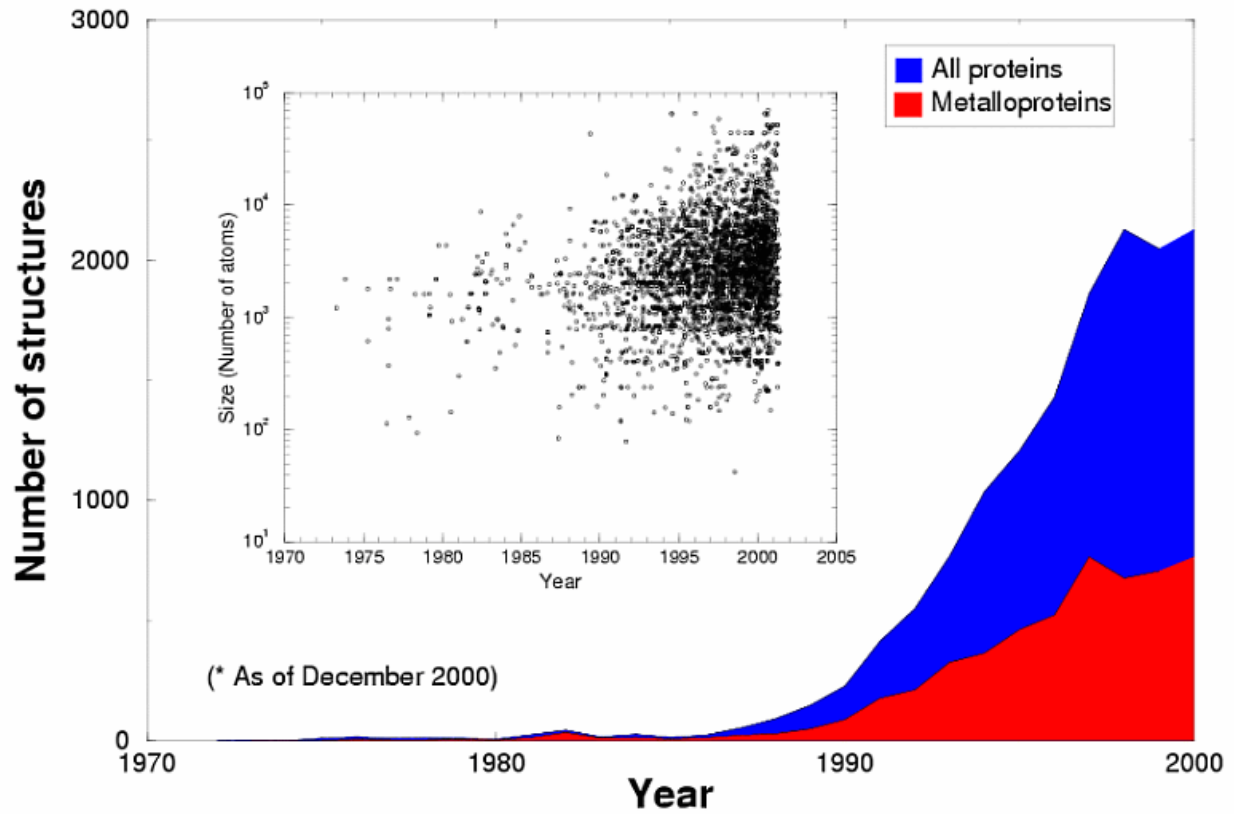
<http://metallo.scripps.edu/talks>

The MDB Web Application

- o Background
- o Its Architecture
- o Web Services: Overview, Implementation, and Examples of Use
- o Access modes: synchronous, asynchronous

Summary and Acknowledgements

Increase in the number and size of new structures



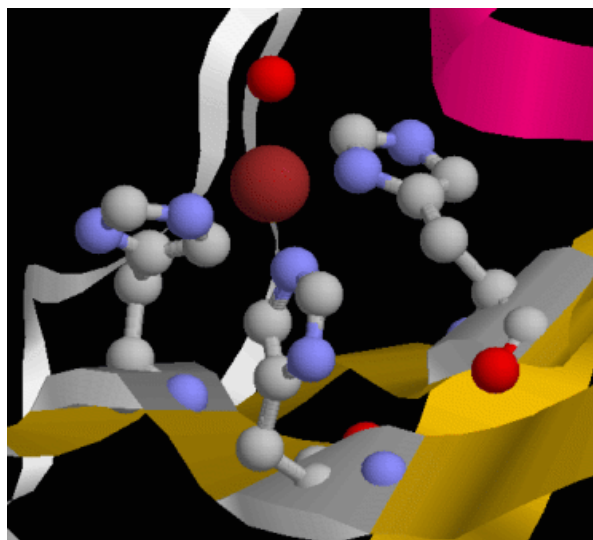
A quick definition

A protein that contains one or more metal ions which play a functional or structural role (sometimes is both).

Human Carbonic Anhydrase II (1hca), ribbon representation



A detail of the Zinc binding site



Our research program

The MDB is part of a bigger research program, which focuses on metalloprotein research and design:
The Metalloprotein Bioinformatics, Structure, and Design Program

The Objectives

- o Figure out what makes a metalloprotein tick
- o Rational design and construction of new metalloproteins

What do we need to do

1. Understand the geometrical requirements to bind a metal
2. Figure out the effect of the environmental constrains
3. Devise methods to design the sites
4. Analyze the possible candidates
5. Make the sites
6. Find out where we succeeded and where we failed
7. Go back to step 1

These are the reasons why we built the Metalloprotein-site Database and Browser ...

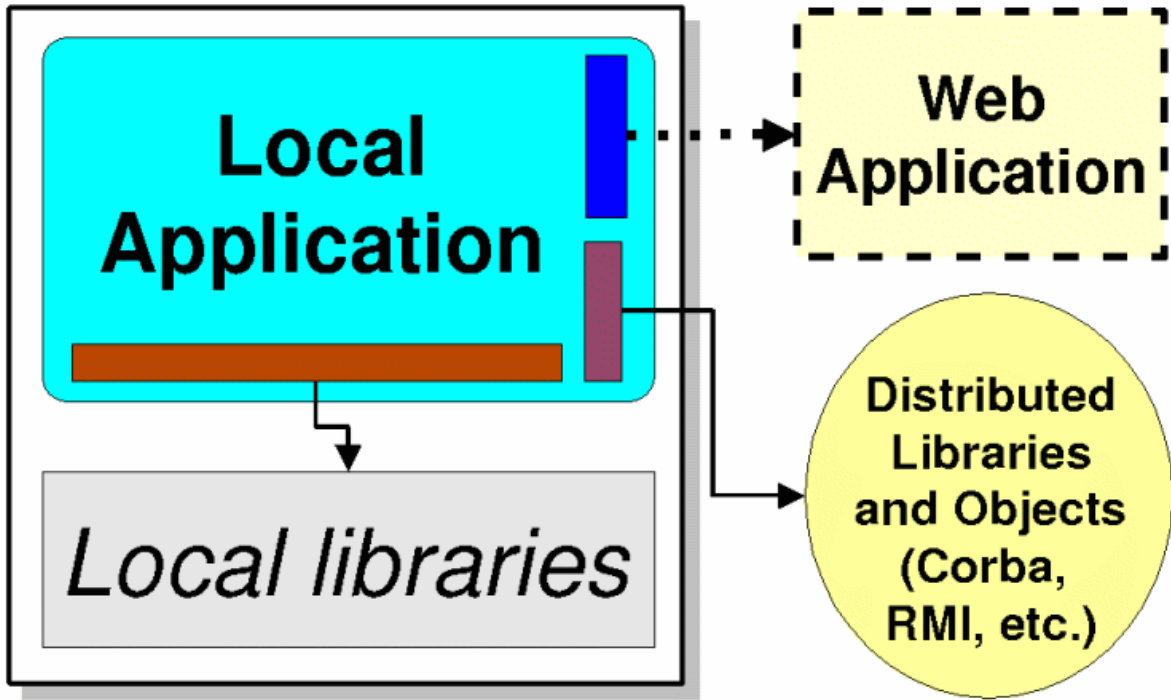
What is the MDB?

A database of quantitative data about metal-binding sites in proteins, created using automated tools and algorithms for metal site recognition and extraction.

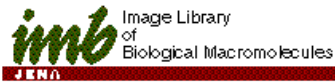
And a Web Application that allows interactive searching and analysis of said quantitative data.

... and what is becoming

All of the above, and a web based library of methods for other applications and web sites to use.



<http://www.imb-jena.de/IMAGE.html>



Access to Metal Containing Biopolymer Structures by Molecule Type

- [Proteins](#)
- [Protein–Nucleic Acid Complexes](#)
- [Nucleic Acids](#)
- [Carbohydrates](#)

via the Periodic Table of Elements

The layout of the Periodic Table of Elements is from [WebElements \[http://www.shef.ac.uk/chemistry/web-elements/\]](http://www.shef.ac.uk/chemistry/web-elements/)

Jump start: select a metal from the Periodic Table and you will get a compilation of [PDB](#) structure entries, processed by the [MDB](#), with this particular metal listed by molecule type. Only metals given in bold do occur in biopolymer 3D structures currently included in the [PDB](#)/[MDB](#). For a more complete compilation of elements occurring in hetero components of either the [PDB](#) or the [NDB](#) use the [Periodic Table Hetero Components Database Element Browser](#).

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Period																		
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54

Listing Cu containing proteins

Copper (Cu) Containing Entries

This list was generated from the Metalloprotein Database at TSRI (MDB):
http://metallo.scripps.edu/
Contact: Jesus M. Castagnetto, metallodb@scripps.edu

Nucleic Acid

3 entries found

Code	Resolution	Description
<u>1d39</u>	1.20	/DNA\$ (Z, 5'-D\$(*CP*GP*CP*GP*CP*G)-3') COPPER(II) CHLORIDE SOAKED
<u>1d40</u>	1.30	/DNA\$ (Z, 5'-D(\$M==5==*CP*GP*UP*AP\$M==5==*CP*G)-3') COPPER(II) CHLORIDE
<u>231d</u>	2.40	DNA (5'-D(*CP*GP*AP*TP*CP*G)-3') DNA

Protein

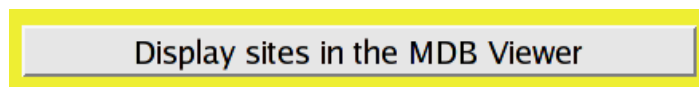
239 entries found

Code	Resolution	Description
<u>1a2v</u>	2.40	COPPER AMINE OXIDASE FROM HANSENULA POLYMORPHA METHYLAMINE OXIDASE
<u>1a3z</u>	1.90	REDUCED RUSTICYANIN AT 1.9 Å RUSTICYANIN ELECTRON TRANSPORT
<u>1a4a</u>	1.89	AZURIN MUTANT WITH MET 121 REPLACED BY HIS, PH 6.5 CRYSTAL FORM, DATA (
<u>1a4b</u>	1.91	AZURIN MUTANT WITH MET 121 REPLACED BY HIS, PH 6.5 CRYSTAL FORM, DATA (
<u>1a4c</u>	2.45	AZURIN MUTANT WITH MET 121 REPLACED BY HIS, PH 3.5 CRYSTAL FORM, DATA (

We also have a lightweight molecular viewer (written in Java, approx. 70Kb in size) that is embeddable by using simple HTML code in a page, e.g.:

```
<!-- Display sites in lhca and lnpc -->
<div align='center'>
<form method="POST"
action="http://metallo.scripps.edu/services/remote/viewer.php"
name="Remote MDB Viewer" target="_blank">
<input type="hidden" name="source_id[]" value="lhca">
<input type="hidden" name="source_id[]" value="lnpc">
<input type="hidden" name="caption" value="lhca and lnpc - All sites">
<input type="submit" name="submit" value="Display sites in the MDB Viewer">
</form>
</div>
```

... generates a button like this one ...



... which when pressed calls the remote viewer

MDB (Remote viewer) - Metalloprotein Site Database and Browser - Konqueror

Macromolecular Model Manager

Warning: Applet Window

Recent Query Cumulative Query

Show Hide Delete

Shown	Description	Metal
	pdb-1hca-_hg-1 Header: [lyase(oxo-acid) 02-apr-92 1hca] Title: [] ...	hg
	pdb-1hca-_zn-1 Header: [lyase(oxo-acid) 02-apr-92 1hca] Title: [] ...	zn
	pdb-1npc-_ca-1 Header: [hydrolase(metalloproteinase) 08-jan-92 ...	ca
	pdb-1npc-_ca-2 Header: [hydrolase(metalloproteinase) 08-jan-92 ...	ca
	pdb-1npc-_ca-3 Header: [hydrolase(metalloproteinase) 08-jan-92 ...	ca
	pdb-1npc-_ca-4 Header: [hydrolase(metalloproteinase) 08-jan-92 ...	ca
	pdb-1npc-_zn-1 Header: [hydrolase(metalloproteinase) 08-jan-92 ...	zn

1hca and 1npc - All sites

View Picking Labels

Transform Box

Stereo display

Previous View

Window on model

Window on all

Magnify Region

Reinitialize view

Float viewer

>>Dist<< >>Ang<< >>Dihed<<

The remote viewer is being used by several web sites, among them:

CaBP database: http://structbio.vanderbilt.edu/cabp_database/

IMB at Jena: <http://www.imb-jena.de/>

PROMISE: <http://bioinf.leeds.ac.uk/promise/>

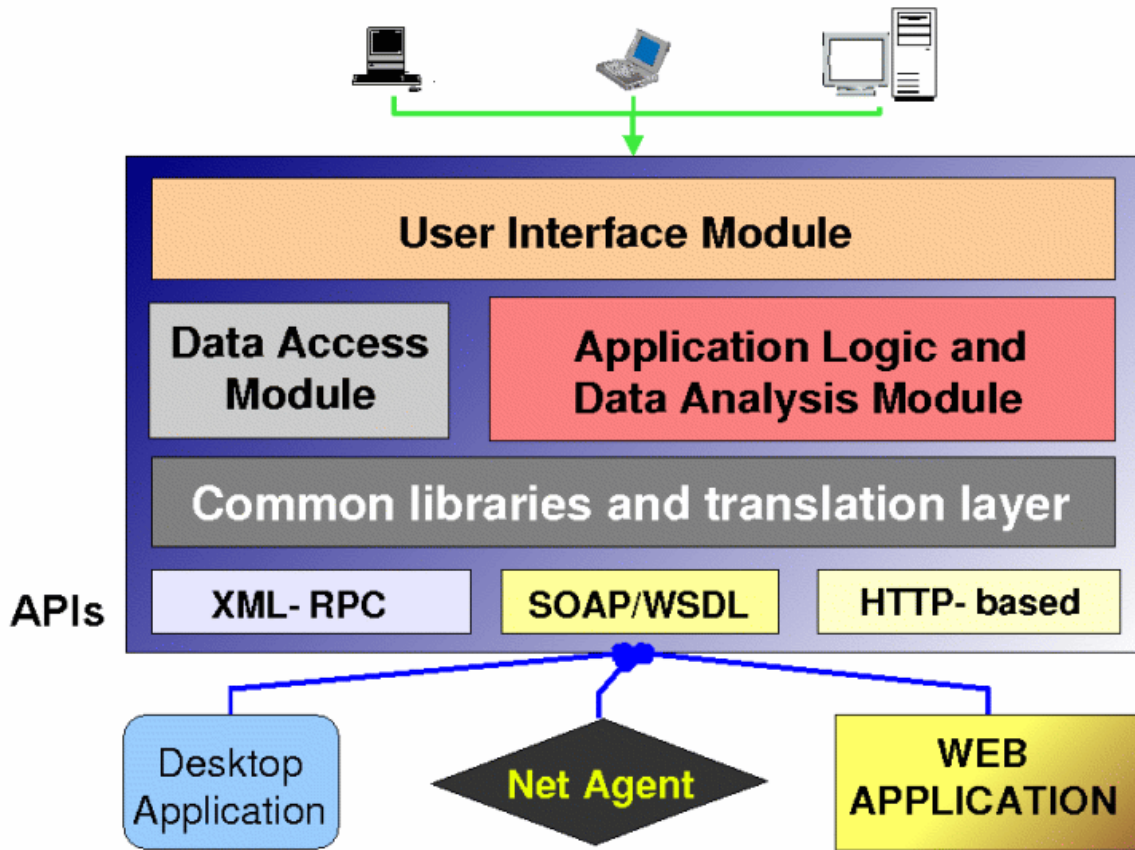
My old 'Web API' point of view ...

A web Application Program Interface, is a set of callable functions or methods that a web application exposes as a public interface for communication with external application.

... and the modern Web Services definition

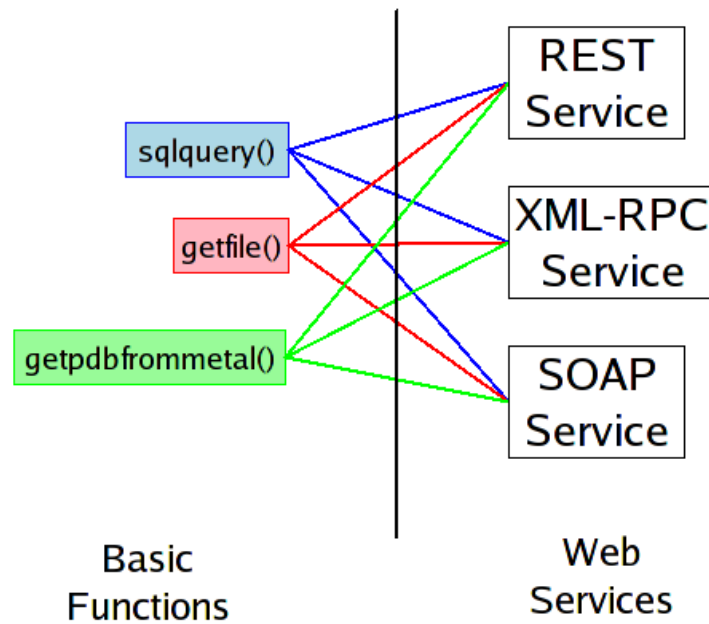
"A Web service is a collection of functions that are packaged as a single entity and published to the network for use by other programs. Web services are building blocks for creating open distributed systems..." [1]

[1] from "The Web services (r)evolution, Part 1",
<http://www-106.ibm.com/developerworks/webservices/library/ws-peerl.html>



The Base Library

A set of simple modular functions that can be wrapped in each of the services. Better than my original design of kludging everything in parallel code sets.



REST

- o URL: <http://metallo.scripps.edu/services/rest.php>
- o Functions: sql and metallopdb
- o Example: <http://metallo.scripps.edu/services/rest.php?func=metallopdb&metal=zn&mode=new&count=5&format=rss>

XML-RPC

- o URL: <http://metallo.scripps.edu/services/xmlrpc.php>
- o struct method.sql(string query)
- o struct method.metallopdb(string metal, string mode, int count, string format)

SOAP

- o URL: <http://metallo.scripps.edu/services/soap.php>
- o Methods: sql, metallopdb, rssmetallopdb

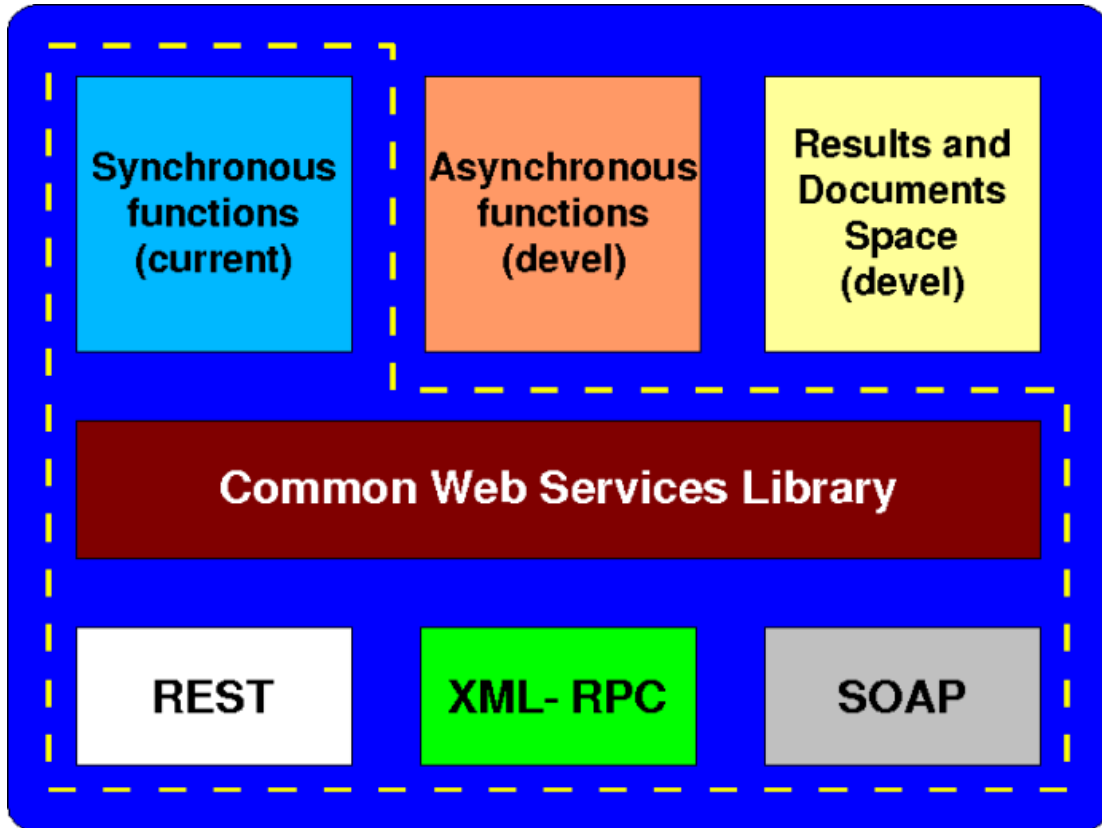
Accessing web services

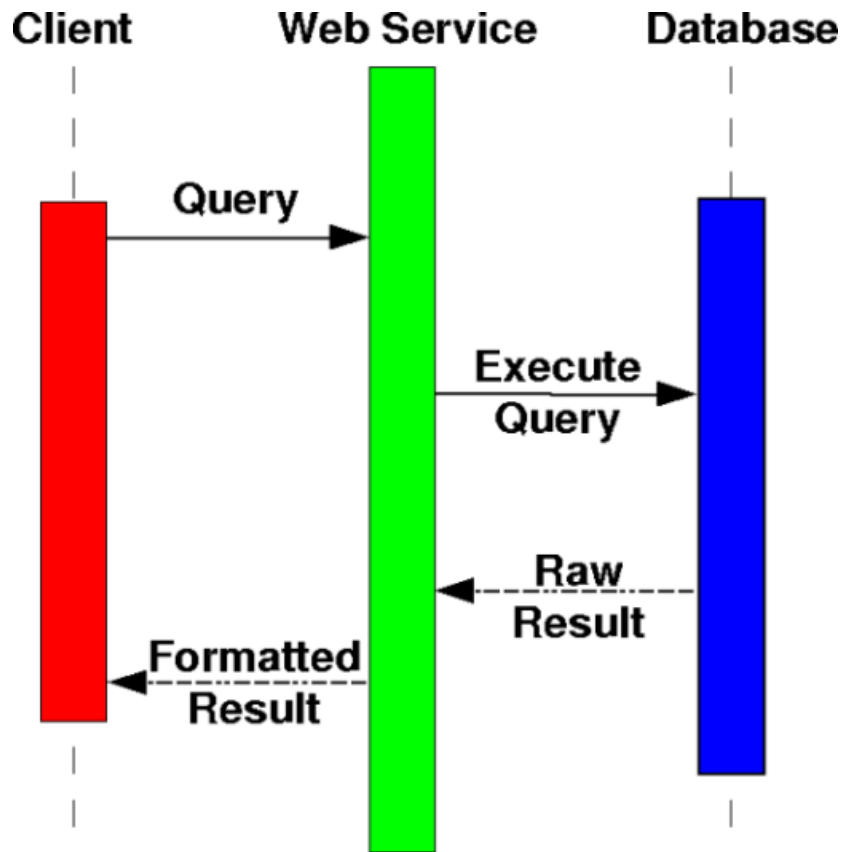
A possible problem

The client accessing the MDB web services has to wait for as long as it takes to get the results. This can be very inefficient if complex queries or analyses are requested.

... and a solution: asynchronous access

- o Decouple the request and the retrieval of the results, to allow scheduling of query execution to periods of decreased server load.
- o Query result caching: if the same query is requested by different clients, it will be executed only once.
- o Flexible result/documentation delivery. For example, an external web application can register interest in knowing the latest metalloproteins indexed in the MDB, for that it will register a callback service, a method/protocol, and a result format.





What is a tuplespace? [1]

"... (A) tuplespace is a shared datastore (the space) for simple list data structures (tuples). A very simple model is used to access the tuplespace, usually consisting of the operations write (out), take (in), read (rd) ..."

XML Spaces [2]

"... An XML-space is a tuple-space made up only of XML-tuples--that is, a public repository or buffer that can contain XML-tuples. An XML-space has a name, by which it is referenced, and it is hosted on a server, along with any number of other (and other-named) XML-spaces ..."

A Sample XML Tuple

```
<pdb_id>lhca</pdb_id>
<mdb_id>lhca_sl</mdb_id>
<metal_site>
  <metal_center>
    <metal>Zn</metal>
    <geometry>Tetrahedral</geometry>
  </metal_center>
</metal_site>
```

[1] <http://earl.strain.at/space/Tuple%20Spaces>

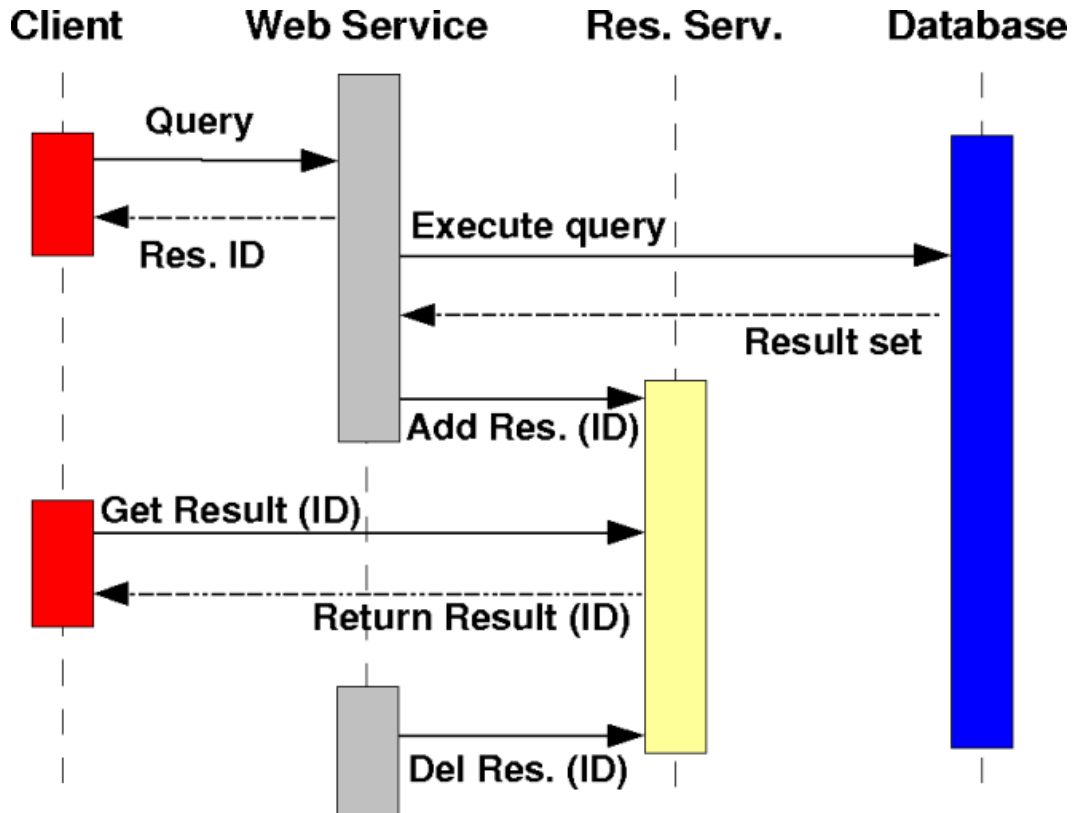
[2] "XML-Tuples and XML-Spaces, V0.7" (1999) <http://uncled.oit.unc.edu/XML/XMLSpaces.html>

Async. access

Example of asynchronous access

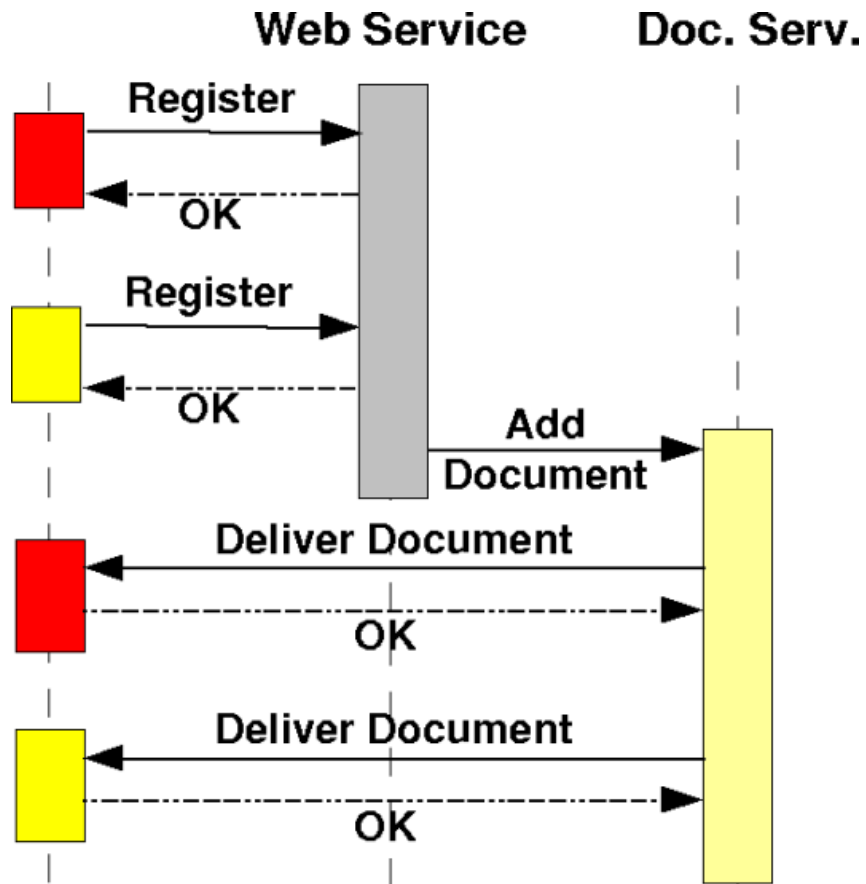
We will assume that the client connecting to the MDB web services is capable of using different protocols.

1. Web client asks for all Zn-containing proteins using the `sql()` method of the XML-RPC server.
2. XML-RPC server returns a result ID to client, then schedules query to be executed and results stored for later retrieval.
3. Query is executed and result set stored using the assigned ID.
4. Client uses SOAP to retrieve result. Server retrieves stored result set, and formats it as a complex structure using the appropriate XML Schema.



Example of callback registration

1. Web client tells the SOAP server that it wants to be informed whenever new Cu-containing proteins are indexed in the MDB. Gives as callback the URL: <http://www.example.com/newcu.jsp>, and wants to receive data in RSS format, using the HTTP POST method.
2. MDB Web services know how to use POST and format data in RSS, so a registration ID is returned. ID can be used to unregister the callback, or modify it.
3. The following week, new protein structures are indexed. Five new Cu-containing proteins are found.
4. MDB uses the registered URL and POSTs a RSS formatted document with information on the new Cu-containing proteins.
5. Callback is repeated next week unless unregistered, modified, or if an error condition was encountered the previous week (e.g. a critical HTTP error).



- o MSE: The Metal-binding Site Evaluator. Performs a full analysis on a user submitted PDB structure, generating a complete report of the metal-binding sites, their geometry, first and second shell ligands, hydrogen-bond interactions, etc. Uses our MSIT java application.

Analysis of 2sod.pdb

```

END OF PROCESS: Thu Jan 30 16:53:41 PST 2003

2sod_s1.desc

Site: 2sod_s1

[general]
number_of_metal_centers;2
number_of_first_shell_ligands;8
number_of_metal_ligand-atom_bonds;9
number_of_ligand-atom_metal_ligand-atom_angles;16
number_of_first_shell_h-bonds;7
number_of_second_shell_ligands;34
number_of_second_shell_h-bonds;19
number_of_intershell_contacts;0
number_of_intershell_h-bonds;16
number_of_disulfide-bonds;0

[metal_centers]
metal_center_1;CU.1.O;pentacoordinated;CU binding site
  metal_center_1_atom_1;CU.CU.1.O
metal_center_2;ZN.2.O;distorted tetrahedral;ZN binding site
  metal_center_2_atom_1;ZN.ZN.2.O

[first_shell_ligands]
# full residue name; denticity
HIS.44.O;1
HIS.46.O;1
HIS.61.O;1
HIS.118.O;1
HOH.3.O;1
HIS.69.O;1
HIS.78.O;1
ASP.81.O;1

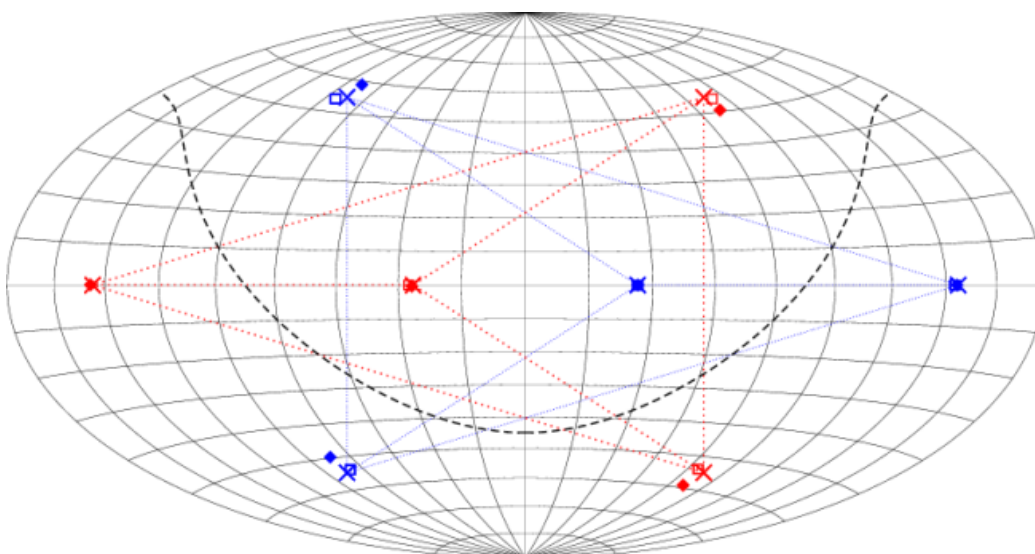
[metal_ligand-atom_bonds]
# metal atom; ligand atom; distance; type
CU.CU.1.O,ND1.HIS.44.O,2.0074546,M-L bond
CU.CU.1.O,NE2.HIS.46.O,2.1103191,M-L bond
CU.CU.1.O,NE2.HIS.61.O,2.2131639,M-L bond
CU.CU.1.O,NE2.HIS.118.O,2.0960212,M-L bond
CU.CU.1.O,O.HOH.3.O,3.1981292,M-Water bond
ZN.ZN.2.O,ND1.HIS.61.O,2.0946991,M-L bond
ZN.ZN.2.O,ND1.HIS.69.O,2.140547,M-L bond
ZN.ZN.2.O,ND1.HIS.78.O,2.0375605,M-L bond
ZN.ZN.2.O,OD1.ASP.81.O,1.9098946,M-O bond

[ligand-atom_metal_ligand-atom_angles]
# atom1; atom2; atom3; angle; dist(1-2); dist(2-3); dist(1-3);type
ND1.HIS.44.O, CU.CU.1.O, NE2.HIS.46.O, 130.22617, 2.0074546, 2.1103191, 3.7356489, L-M-L PROPER
ND1.HIS.44.O, CU.CU.1.O, NE2.HIS.61.O, 74.24425, 2.0074546, 2.2131639, 2.5630897, L-M-L PROPER

```

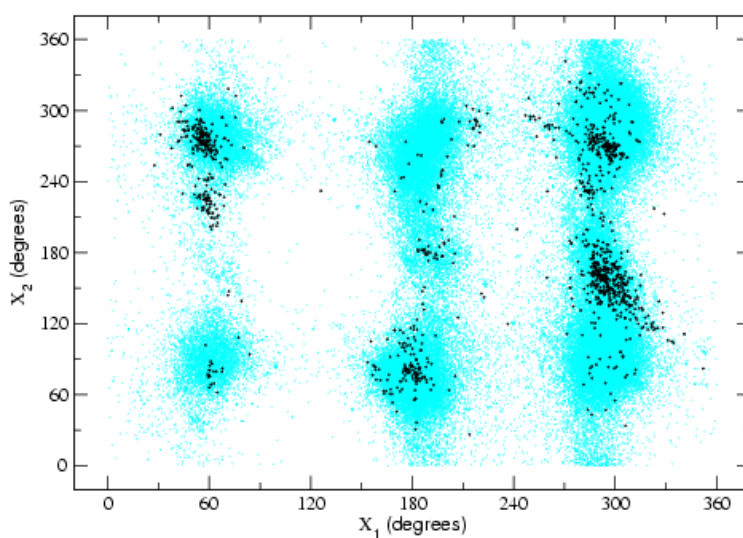
- o ClusterGeom: Java application that performs a geometrical analysis of iron-sulfur clusters ([4Fe4S] and [3Fe4S]). Generates parameters and projections that can be correlated with cluster distortion from an ideal geometry.

[4Fe4S] 2D projection



- o Expansion of the analytical tools, to include: Chi angle distributions and 2D comparisons, Bond angle distributions, Liganding pattern and parameter distribution correlations, etc.

Chi@1@-Chi@2@ distribution. All Histidines vs Zn(His)@3@X



Summary

- o Share data using a standardized file format (e.g. RSS, CML), better yet, via a standard web services protocol: REST, XML-RPC, SOAP.
- o REST is the simplest way to share. We've been doing it since the CGI spec appeared.
- o XML-RPC is simple, but data with complex structure might not be easy to represent.
- o SOAP is more complex, but it gives you more flexibility (XML Schemas, RPC and Document Literal, WSDL, Service redirection, etc.)
- o Plan on synchronous and asynchronous access to improve performance.
- o Use the MDB Web Services. Use the MDB Web Services. Use the MDB Web Services ...

Acknowledgments

People at the Metalloprotein Bioinformatics, Structure and Design Program



The "vast" MDB team (circa 1999)



\$\$ from NIH.

Users of the MDB.

This and other MDB related talks at:

<http://metallo.scripps.edu/talks>

Index

Agenda	2
PDB Stats	3
Metalloprotein	4
MDB: Background	5
MDB: Description	6
MDB as Library	7
IMB @ Jena	8
MDB Remote Viewer	10
Web Services	12
MDB: Architecture	13
The Base Library	14
Available services	15
Accessing web services	16
Access modes	17
Sync. access	18
Tuplespaces	19
Async. access	20
Async. access	21
Register interest	22
Register interest	23
Future services	24
Summary	26
Acknowledgments	27
Where to find this talk	28