An interoperable architecture using the CDPP/AMDA service and the SPASE Model

Michel Gangloff [1], Richard Hitier [2], Florian Topf [3], E. Budnik[4], M. Bouchemit[1], E. Pallier[1]

[1] CDPP/CESR, CNRS/Université Paul Sabatier
9, avenue du Colonel Roche, BP44346, 31028 Toulouse cedex04, France
EMail: Michel.Gangloff@cesr.fr

[2] Co-Libri
Cremefer 11290 Montréal, France

[3] Space Research Institute Austrian Academy of Sciences
Schmiedlstrasse 6, 8042 Graz, Austria

[4]Noveltis
2 Avenue de l' Europe, 31520 Ramonville Saint Agne, France

ABSTRACT

AMDA (AMDA, Automated Multi Dataset Analysis) is a web based service developed at the CDPP, the French national centre for space physics data. This service allows the users to perform integrated analysis of solar system plasma data, offering usual functionalities like data visualisation but also innovative tools like automated search on the content of the data. A prototype of interoperable architecture has been built around AMDA by using the Data Model from the SPASE consortium. This prototype allows the users to access from AMDA , not only local data, but also data from remote databases in a standard way. Remote data are described with metadata compliant with the SPASE Model. The metadata are written in XML and stored in distributed registries. This architecture is based on the web-services technology and the SOAP standard. This prototype will be enhanced and used in the forthcoming HELIO project (for a European Virtual Observatory of Heliophysics) and EuroPLANET RI/IDIS project (for a future Virtual Observatory for Planetology).

Keywords: spase, soap, web-services, xquery, xml,virtual observatory

INTRODUCTION

We have developped a prototype which provides the users with access to remote data from a service, AMDA, according to the Virtual Observatory for space sciences paradigm[1].AMDA is a service, maintained by the CDPP, the french national centre for space physics data, in the framework of EuroPlaNet. EuroPlaNet is a European project, which is composed of different activities. One of these is IDIS (Integrated and Distributed Information Service). IDIS is divided into thematic nodes. Every node is responsible for the access to data and tools relevant to its thematic field. One of these thematic nodes is dedicated to plasma physics. But the study of a planetary object needs a pluridisciplinary approach (plasmas, fields, radio, dust, moon surfaces, rings , atmospheres,...). This requires to access and exploit resources of many origins, of various natures, i.e from different nodes. This prototype was developed as a demonstrator of Virtual Observatory functionalities for the Plasma Node only, and should be considered as the first step towards a Virtual Observatory for planetary sciences. AMDA already provides access to remote databases like NSSDC CdaWeb, but it is necessary to implement a specific "reader" for every data base. The goal of this prototype is to use a generic reader. For this generic reader, we need a common way of describing data, a standard metadata model. With this metadata model , a lot of data providers can publish their data, which become accessible by a Virtual Observatory service like AMDA in a standard way. The architecture of this prototype is composed of a tool, AMDA, which provides several services (download data, plot data, parameter editor, conditional or visual search, time table manager) on plasma physics data from a lot of space missions (ACE, Cluster,Geotail,ISEE,THEMIS,WIND), a metadata model (SPASE), a registry, implemented in a native XML database, a query language (Xquery), a web-services technology (SOAP) and remote data servers (Cassini MAPS Key Parameters, Venus Express Magnetic Field).

AMDA

The plasma objects studied in space physics consist of gigantic systems characterised by multiscale dynamics and fast long range couplings between key regions. For studying such systems, it is necessary to perform integrated multi-point/multi-instrument analysis in case studies and statistical studies as well.

As a generic Webtool for Space Physics data, AMDA provides the user with :

- automated event search and characterisation
- catalogue generation and exploitation
- automated database conditional extraction
- access to remote Data Centers

AMDA is divided into three main components:

- AMDA-Client is the graphical interface
- AMDA-Server is the core of AMDA and handles all functionnalities offered through the graphical interface
- AMDA-Database that feeds AMDA-Server with local or remote data

A more detailed presentation of AMDA is available in [2]

ACCESS TO REMOTE DATABASES WITH AMDA

Remote databases may be accessed from AMDA in two ways. The first one, not generic, needs a dedicated shim for every data provider : CDAWEB, MAPSKP, etc. The second one, generic, uses standard descriptions of the data to access them.

Source dependant access to remote data

In this case, a specific interpretor is necessary to access remote databases. Every interpretor must extract and organize the information necessary to transform the data flux from the remote database into an object which can be used by AMDA-Server. Data are accessed via web-services or ftp servers. See figure 1.

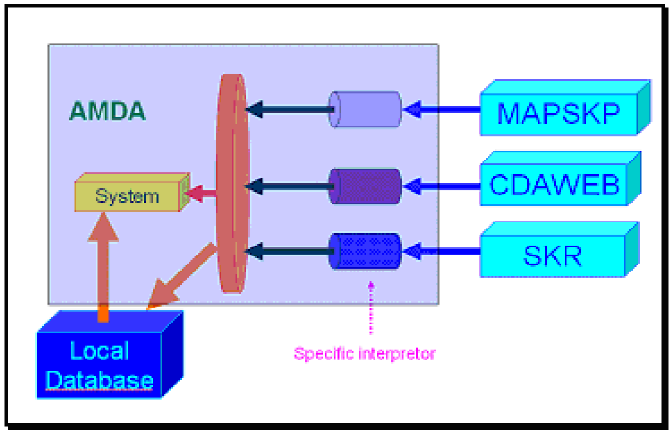


Figure 1 Remote Data Access without SPASE

Generic access to remote data

Here we use a generic approach: instead od developping a specific shim for every source of data, we use a generic metadata translator. All information necessary to handle data (metadata) is available in a descriptor,written in XML, and compliant with the SPASE metadata model.

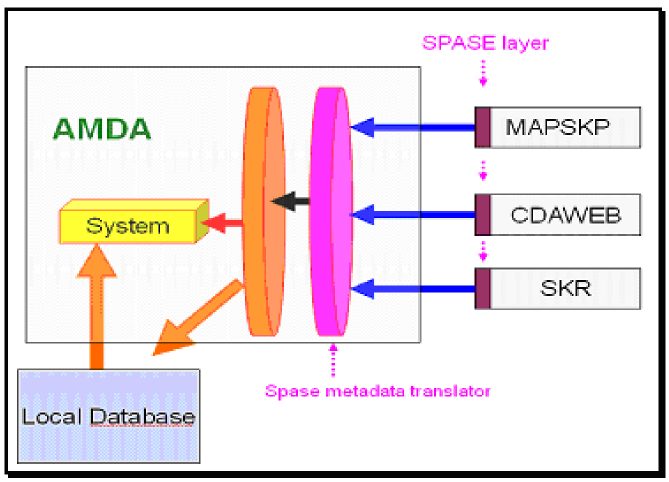


Figure 2 Remote Data Access with a SPASE Layer

ARCHITECTURE

In order to implement this generic access to remote data, we have developed the following architecture:

- A registry of metadata in an eXist database. All metadata are compliant with the SPASE model
- An XQUERY interface to handle queries
- A set of web-services with a SOAP interface

What is SPASE ?

Space Physics Archive Search and Extract [2] is an international organisation. Its main goal is to facilitate data search and retrieval across the Space and Solar Physics data environment. Spase has concentrated on defining and maintaining a standard data model and a dictionary. The main component of the data model is a “resource”. Each resource description contains information which is likely to be refrenced by more than one data product. The main resource types are: Numerical Data, Catalog, Display Data, Granule, Observatory, Instrument.

In this way, an observatory , which can be a satellite or a ground facility may house several instruments. Every instrument provides a lot of numerical data sets, each of which is composed of several physical parameters (field , particle). SPASE is widely used in the heliophysics community and as such is the best choice to describe space physics or solar data. The current version of the data model is 2.0.0.

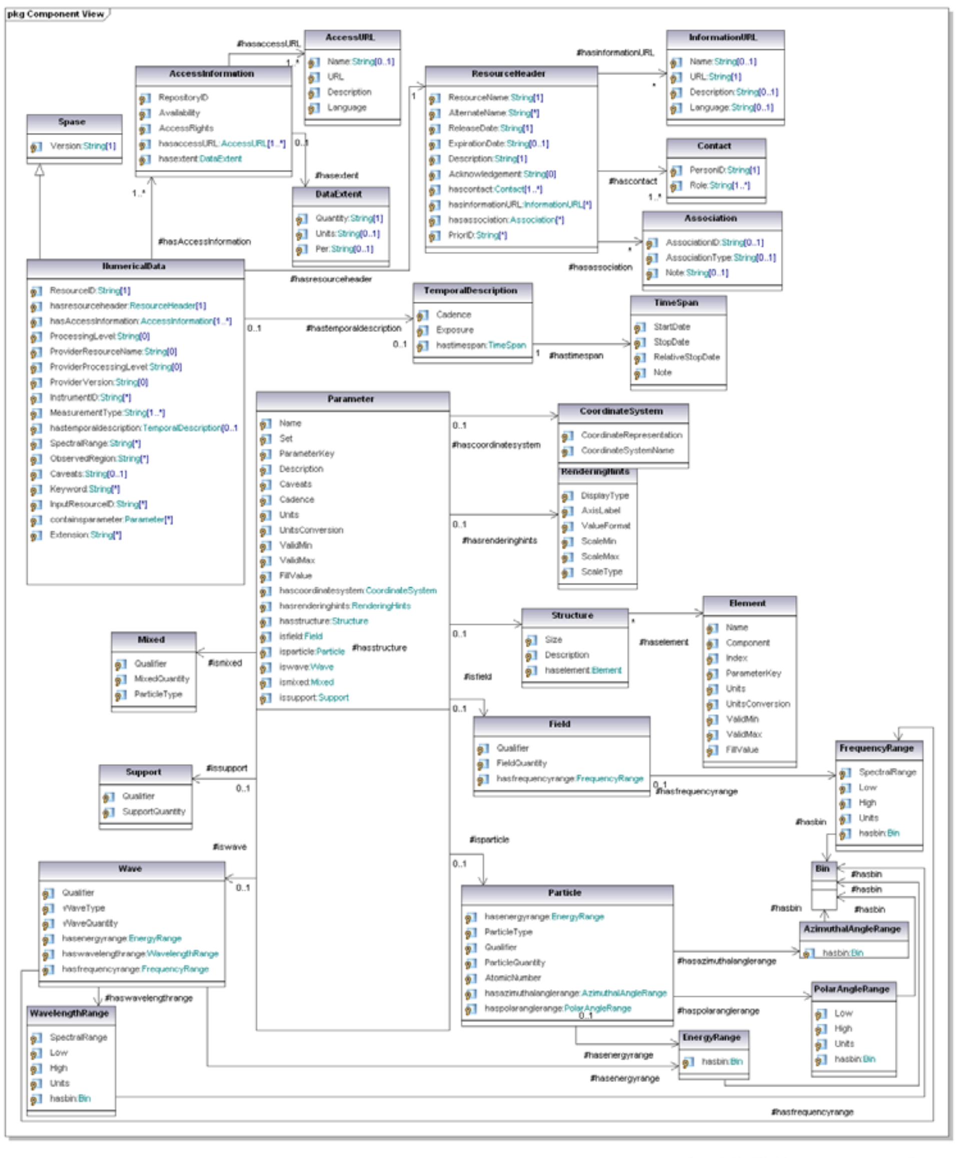


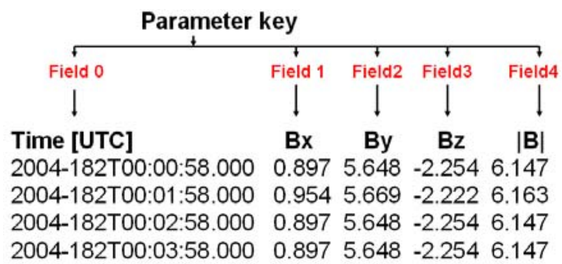
Figure 3 UML Class Diagram of SPASE Numerical Data Resource

But the SPASE data model does not contain enough information to enable the extraction of the actual data. There is no indication about the data type. Therefore, it is necessary to use an additional descriptor which contains the data type for each parameter to read. In the SPASE model, each physical parameter is assigned a “ParameterKey” to access it.

This is an example of auxiliary XML descriptor, for the components of a magnetic field vector :

```
<dataset xml:id=”MAG_KG”>
        <parameter xml:id=”VECTOR”>
                <data_type>FLOAT</data_type>
        </parameter>
        <parameter xml:id=”MAGNITUDE”>
                <data_type>FLOAT</data_type>
        </parameter>
</dataset>
```

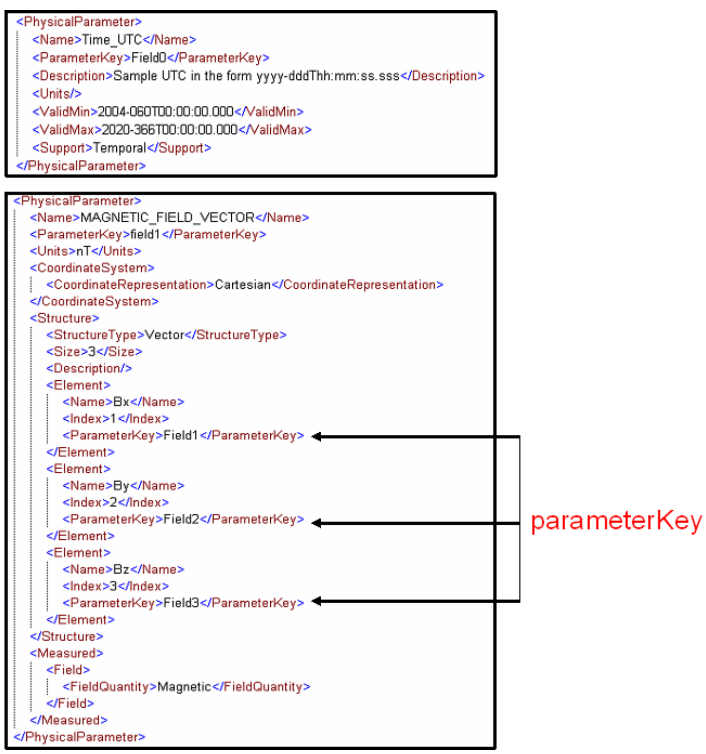
Figure 4 shows an example of tabular data, in which we can see the correspondance between the name of variables (Time, Bx, ...) and the parameter key.



Time [UTC]	Bx	By	Bz	|B|
2004-182T00:00:58.000	0.897	5.648	-2.254	6.147
2004-182T00:01:58.000	0.954	5.669	-2.222	6.163
2004-182T00:02:58.000	0.897	5.648	-2.254	6.147
2004-182T00:03:58.000	0.897	5.648	-2.254	6.147

Figure 4 Tabular data: Date Time and magnetic field vector

Figure 5 shows the corresponding extract of the SPASE descriptor, giving information for every parameter.




```
<PhysicalParameter>
  <Name>Time_UTC</Name>
  <ParameterKey>Field0</ParameterKey>
  <Description>Sample UTC in the form yyyy-dddThh:mm:ss.sss</Description>
  <Units/>
  <ValidMin>2004-060T00:00:00.000</ValidMin>
  <ValidMax>2020-366T00:00:00.000</ValidMax>
  <Support>Temporal</Support>
</PhysicalParameter>
```

```
<PhysicalParameter>
  <Name>MAGNETIC_FIELD_VECTOR</Name>
  <ParameterKey>field1</ParameterKey>
  <Units>nT</Units>
  <CoordinateSystem>
    <CoordinateRepresentation>Cartesian</CoordinateRepresentation>
  </CoordinateSystem>
  <Structure>
    <StructureType>Vector</StructureType>
    <Size>3</Size>
    <Description/>
    <Element>
      <Name>Bx</Name>
      <Index>1</Index>
      <ParameterKey>Field1</ParameterKey>
    </Element>
    <Element>
      <Name>By</Name>
      <Index>2</Index>
      <ParameterKey>Field2</ParameterKey>
    </Element>
    <Element>
      <Name>Bz</Name>
      <Index>3</Index>
      <ParameterKey>Field3</ParameterKey>
    </Element>
  </Structure>
  <Measured>
    <Field>
      <FieldQuantity>Magnetic</FieldQuantity>
    </Field>
  </Measured>
</PhysicalParameter>
```

Figure 5 An extract of a Numerical dataset descriptor

Why eXist ?

eXist [4] is an Open Source effort to develop a native XML database system, which can be easily integrated into applications dealing with XML. The database is completely written in Java and may be deployed in a number of ways, either running as a stand-alone server process, inside a servlet-engine or directly embedded into an application. eXist's query engine implements efficient, index-based query processing. The database is currently best suited for applications dealing with small to large collections of XML documents which are occasionally updated. Access through HTTP, XML-RPC, SOAP and WebDAV is provided. The eXist database is already used by some Virtual Observatories projects :

- VSPO , Virtual Space Physics Observatory
- NVO , National Virtual Observatory, IVOA
- Astrogrid , IVOA

Associated with the eXist database, we use Xquery. Xquery[5] is a powerfull programming language to extract XML data from single files or data collections, using Xpath[6] expressions.

Why SOAP ?

SOAP [7] stands for Simple Object Access Protocol, and is a W3C recommendation.It is one of the best ways to communicate between applications over HTTP. Theses applications may be running on different operating systems, use different technologies, and programming languages. SOAP eliminates the security problem encountered with technologies like CORBA which use Remote Procedure Calls(RPC). Proxies and firewalls normally block this kind of traffic.

Web services

There are three web-services implemented:

- getAvailability() asks for all data available in the remote server. The response is displayed as a data tree by AMDA.
- getDatasetInfo(DataSetID) to get metadata specific to a dataset. The response contains the SPASE XML descriptor and the auxiliary decriptor of the dataset.
- getDatasetURL(DataSetID, Start Time, End Time) to get the list of granules(files) which are part of the dataset.

wget is used to retrieve data. See figure 3

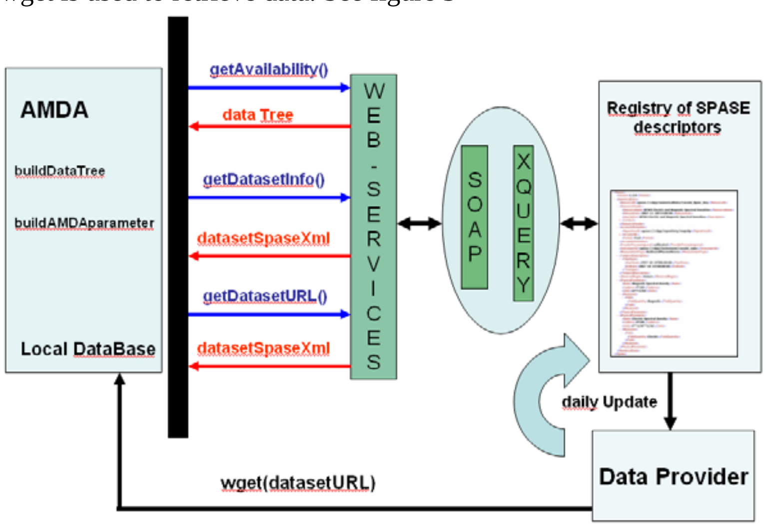


Figure 6 Web Services to search and access data

CONCLUSION

In this paper, we have presented a first prototype of Virtual Observatory using a widely known standard, the SPASE data model. We have seen that it is possible, from a scientific tool like AMDA, to access remote data from various data providers in a generic way. But the kind of query involved is very basic.With the current architecture, it is not possible to send a more precise query like: “send me the reference of all datasets including thermal plasma measurements”. This could be the purpose of a second prototype, which would use a new protocol called SPASE QL (Spase Query language)[8] to handle more complex queries.

REFERENCES

[1] – Virtual observatories for space and solar physics research Christopher C. Harvey, Michel Gangloff , Todd King , Christopher H. Perry , D. Aaron Roberts , James R. Thieman in Earth Sciences Informatics Volume1/Number1/April 2008 (http://www.springerlink.com/content/4163553805612t20/fulltext.pdf)

[2] - AMDA, Automated Multi-Dataset Analysis: A web-based service provided by the CDPP C. Jacquey, V. Génot , E. Budnik , R. Hitier , M. Bouchemit , M. Gangloff , A. Fedorov , B. Cecconi, N. André , B. Lavraud, C. Harvey , F. Dériot, D. Heulet, E. Pallier , E. Penou and J.L. Pinçon, in press in the proceedings of the 15th Cluster Workshop

[3] - Harvey, C.C., Thieman, J.R., King, T. and Roberts, D. A.,2004. SPASE – Space Physics Archive Search and Extract,In *Proceedings PV-2004 Ensuring the Long TermPreservation and Adding Value to Scientific and TechnicalData,* Frascati, Italy, ESA/ESRIN WPP-232.

[4] - http://www.exist-db.org/

[4] - http://www.w3.org/TR/xquery/

[5] - http://www.w3.org/TR/xpath/

[6] - http://www.w3.org/TR/SOAP/

[7] - http://spaseql.gsfc.nasa.gov/