

THE DIGITAL PUGLIA PROJECT: AN ACTIVE DIGITAL LIBRARY OF REMOTE SENSING DATA

Giovanni Aloisio¹, Massimo Cafaro¹, Roy Williams²

¹ Facoltà di Ingegneria, Università degli Studi di Lecce
Via per Monteroni, 73100 Lecce, Italy
{aloisio,cafaro}@sara.unile.it

² Center for Advanced Computing Research
California Institute of Technology, Pasadena, California, USA
roy@cacr.caltech.edu

Abstract. The growing need of software infrastructure able to create, maintain and ease the evolution of scientific data, promotes the development of digital libraries in order to provide the user with fast and reliable access to data. In a world that is rapidly changing, the standard view of a digital library as a data repository specialized to a community of users and provided with some search tools is no longer tenable. To be effective, a digital library should be an active digital library, meaning that users can process available data not just to retrieve a particular piece of information, but to infer new knowledge about the data at hand. Digital Puglia is a new project, conceived to emphasize not only retrieval of data to the client's workstation, but also customized processing of the data. Such processing tasks may include data mining, filtering and knowledge discovery in huge databases, compute-intensive image processing (such as principal component analysis, supervised classification, or pattern matching) and on demand computing sessions. We describe the issues, the requirements and the underlying technologies of the Digital Puglia Project, whose final goal is to build a high performance distributed and active digital library of remote sensing data.

1 Introduction

The last three years have seen the transition from isolated personal computers to universal Internet connectivity; now in Europe and the USA we see computing centers joining into ever-larger geographically distributed collaborations. Taking these facts together, it is clear that supercomputers and data archives will soon become part of a global data and computing fabric. Some users will not be aware of the architecture or the location of the machines executing their jobs, and yet other users may construct and schedule a complex, distributed, metacomputer with heterogeneous computers and data resources providing services to some central objective.

In the paper we show how such a fabric can be exploited to create an “active digital library” of remote-sensing data. The term “library” implies that data is organized for easy access by a community of users. The term “active” implies that the library provides computing services in addition to data-retrieval services, so that users can

initiate computing jobs on remote supercomputers for processing, mining, and filtering of the data in the library.

In many cases, data must be processed by a supercomputer before a human can extract any knowledge from it; only with filtering, mining, and visualization algorithms does it expose its knowledge content. A good example of such data is Synthetic Aperture Radar (SAR) images of the surface of the Earth. In this case, sometimes the user does not need just the delivering of the ground image selected from the library, but she needs to start on this image compute-intensive image post-processing (such as mosaicking, registration, interpolation, rotation, GIS integration, or tasks such as Principal Component Analysis, Singular Value Decomposition, Maximum Likelihood Classification or fusion with other data, such as a digital elevation model). Furthermore, “on-demand computing” sessions could be also required for raw-data that have not yet been processed.

In the following, we present the *Digital Puglia* project we conceived to create an “Active Digital Library” (ADL) of remote-sensing data.

2 The Digital Puglia Project

Our ADL prototype refers to the Puglia region of southern Italy and is built using SAR raw data provided by the Italian Space Agency. The Digital Puglia ADL is based on three joined projects:

- the SARA Digital Library,
- the Globus Metacomputing Toolkit,
- the SAR processing on Wyglaf.

SARA (Synthetic Aperture Radar Atlas) [1,2] is a web-based digital library that has been running at Lecce, Caltech, and the San Diego Supercomputer Center for over a year. Data is replicated on multiple servers to provide fault-tolerance and also to minimize the distance between client and server. SARA already allows clients to download SAR images from the public domain SIRC dataset. A client navigates web pages containing Java applets that implement a GUI (Graphical User Interface) showing a map of the world. Clicking on the map zooms in on a part of the world until the user can see the coverage of the atlas in terms of the SAR images, which are perhaps 50km in size. Chosen subsets of the image can then be downloaded in any of a variety of formats.

The **Globus** project [3] is developing the basic infrastructure required to support computations that integrate geographically distributed computational and information resources. Such computations may link tens or hundreds of resources located in multiple administrative domains and connected using networks of widely varying capabilities. Existing systems have only limited abilities for identifying and integrating new resources, and lack scalable mechanisms for authentication and privacy. The Globus project is building a parallel programming environment that supports the dynamic identification and composition of resources available on

national-scale internets, and mechanisms for authentication, authorization, and delegation of trust within environments of this scale.

The **SAR processing on Wyglaf** project [4] is developing a parallel SAR processor on a Beowulf cluster of PCs (Wyglaf) available at the HPC Laboratory of the University of Lecce. Goal of the project is also to provide on-demand SAR processing capabilities.

A primary objective of the Digital Puglia project is to use heterogeneous distributed computing resources for an on-demand SAR processing service. The client can order processing of a dataset through the Internet, then retrieve the resulting multichannel images when processing is complete.

Joining SARA and Globus allows servicing multiple user requests for on-demand SAR processing. A collection of computing resources, such as the Wyglaf cluster at Lecce or the resources of the Center for Advanced Computing Research at Caltech in California, can be exploited to fulfill the user requirements in terms of time constraints and geographic location. Thus, depending on where the query originates and how stringent are the temporal constraints, a suitable ensemble of computing facilities will be selected by a simple click in the SARA GUI exploiting the Globus Toolkit.

In the rest of the paper we will concentrate on the SARA project and its new features. In fact, the SARA architecture (the kernel of the Digital Puglia Project) was changed with respect to the first version to meet new requirements (such as on-demand SAR processing) and to adapt its structure to the new emerging technologies (such as XML). In the following we refer to Digital Puglia as DPSARA.

3 Technologies Underlying the ADL Implementation

In this section, we briefly make explicit our digital library requirements. In our ADL:

- contents should be accessible on the web;
- users should be allowed to do different operations on the basis of their authorizations;
- a structured data base of SAR images must be provided and data should be searched for by means of complex queries;
- operations include, but are not limited to, retrieval of image data, on demand processing of raw data for authorized users, on demand postprocessing such as Principal Component Analysis, supervised or unsupervised classification, multitemporal image production and so on.

To meet these requirements we advise the use of the following technologies:

- a Java enabled web server which allows servlet execution (such as Sun's Java Web Server or Netscape's Enterprise Web Server);
- Java applets on the client side and Java servlets on the server side. The applets will provide the necessary interaction between the user and the library, and the servlets will be used to provide the same functionality as CGI together with the new possibility to establish interactive sessions no more stateless;

- a database supporting SQL and Sun's JDBC (Java DataBase Connectivity) which will be used as the communication protocol between servlets and the database;
- a GIS software to plot and annotate geographic maps;
- a Beowulf cluster for parallel computing on commodity hardware following the recent trend in this area;
- the MPI standard for message passing;
- HTML and XML.

The former technologies have proven themselves to be reliable and useful in the construction of distributed web-based applications. XML has gained in the last few months considerable attention, and we briefly describe its usefulness in the context of Digital Puglia.

4 XML, an emerging technology on the web

In the past the only way to exchange data between different software was the use of available import/export filters. XML has the potential to become a universal platform for data exchange and we adopt it as a standard to communicate between the different services provided by Digital Puglia.

XML [5] is a proper subset of SGML, intended to port on the web the essential features and the inherent power of SGML avoiding its well known complexity.

XML allows the creation of customized markup languages (i.e. an XML document resembles an HTML document, but the set of tags is not fixed and can be locally defined and extended). An XML file is just a plain ASCII file, thus can be generated using a common text editor. Moreover, it can be checked for well formedness and validity against its Document Type Definition (DTD), which is a subsidiary document that specifies allowed syntax. The check can be performed by a parser which can be written in C or Java.

XML thus makes interfaces transparent and easy to use. Collaborative working on a project becomes a simple matter, since once an agreement is reached in a group about what element types can be used, and the relevant DTD is written, then the people involved in a project can develop its part of the project referring to the defined DTD, aided by a validating parser to check for mistakes.

These features allow:

- intelligent search of a particular piece of information;
- intelligent check and validation of data structure;
- performing complex queries on the data;
- linking different types of information in a richer way with respect to HTML;
- creation of standards of XML element types for industry or specialized communities.

5 DPSARA metadata and XML representation

In the following we present the metadata describing our remote sensing images with reference to SARA, an XML DTD (Document Type Definition) for them, and an example image track described in XML using the DTD to check and validate the XML. The entity relationship model for the metadata is shown in Fig.1.

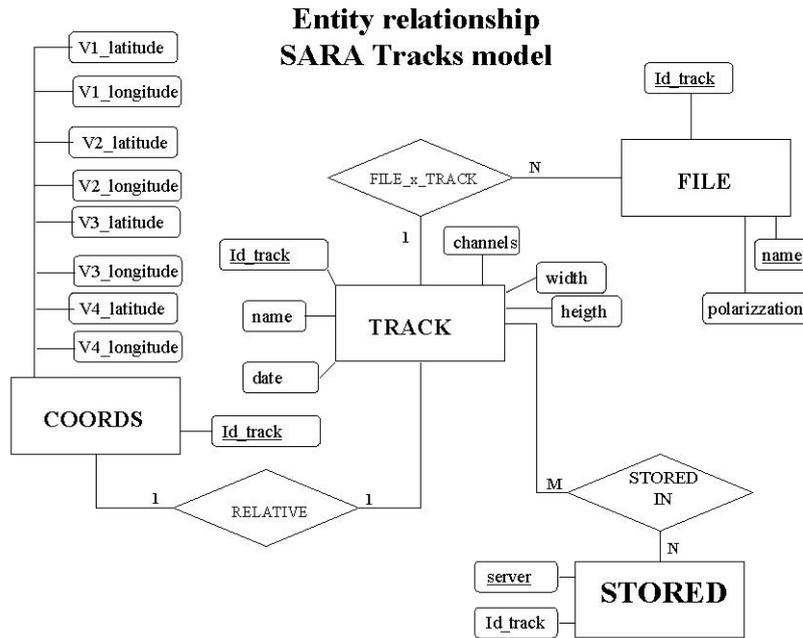


Fig. 1. Entity Relationship model for track metadata

The database consists of four tables. The Track table houses information about the track image such as its name, date of acquisition, unique id, width, height, and number of channels.

The Coords table contains the coordinates of latitude and longitude of the four vertex surrounding the image. In the File table the filenames of the files constituent the image are reported, and finally the Stored table contains the information about where the image is actually stored, that is, one of the web server which compose this distributed digital library.

A DTD for this metadata is presented in Fig.2, while an example of XML file produced according to the DTD is reported in Fig.3.

```

<?XML version='1.0'?>
<!-- DTD for SARA tracks -->
<!--Digital Puglia Project -->

<ELEMENT SARAMETADATA
(SARAQUERYRESULT*)>
<ELEMENT SARAQUERYRESULT
(SARATRACK,SARACOORDS,SARAFILES,
SARASTORED+)*>

<ELEMENT SARATRACK
(NAME,TRACKDATE,WIDTH,HEIGHT,
CHANNELS)>

<ELEMENT NAME (#PCDATA)>
<ELEMENT TRACKDATE (#PCDATA)>
<ELEMENT WIDTH (#PCDATA)>
<ELEMENT HEIGHT (#PCDATA)>
<ELEMENT CHANNELS (#PCDATA)>
<ATTLIST SARATRACK IDTRACK ID
#REQUIRED>
1

```

```

<ELEMENT SARACOORDS
(SARACOORD+)>
<ELEMENT SARACOORD (LAT,LON)>
<ELEMENT LAT (#PCDATA)>
<ELEMENT LON (#PCDATA)>
<ATTLIST SARACOORD IDTRACK IDREF
#REQUIRED>

<ELEMENT SARAFILES (SARAFILE+)>
<ELEMENT SARAFILE (POLARIZATION)>
<ELEMENT POLARIZATION (#PCDATA)>
<ATTLIST SARAFILE NAME ID
#REQUIRED>
<ATTLIST SARAFILE IDTRACK IDREF
#REQUIRED>

<ELEMENT SARASTORED EMPTY>
<ATTLIST SARASTORED SERVER ID
#REQUIRED>
<ATTLIST SARASTORED IDTRACK IDREF
#REQUIRED>
2

```

Fig. 2. Sara Track DTD

```

<?XML VERSION='1.0'?>
<!DOCTYPE SARAMETADATA SYSTEM
"SaraTrack.dtd">
<SARAMETADATA><SARAQUERYRESULT>
<SARATRACK IDTRACK = "11829">
<NAME>Sena Madureira, Brazil</NAME>
<TRACKDATE>04-16-1994</TRACKDATE>
<WIDTH>3624</WIDTH>
<HEIGHT>7995</HEIGHT>
<CHANNELS>4</CHANNELS>
</SARATRACK>
<SARACOORDS>
<SARACOORD>
<LAT>297.575</LAT>
<LON>-18.588</LON>
</SARACOORD>
<SARACOORD>
<LAT>297.206</LAT>
<LON>-18.798</LON>
</SARACOORD>
<SARACOORD>
<LAT>297.696</LAT>
<LON>-19.574</LON>
</SARACOORD>
</SARACOORDS>
</SARAMETADATA>
1

```

```

<LAT>298.066</LAT>
<LON>-19.363</LON>
</SARACOORD>
</SARACOORDS>
<SARAFILES><SARAFILE NAME =
"pr11829_byt_hh" IDTRACK = "11829">
<POLARIZATION>LHH</POLARIZATION>
</SARAFILE>
<SARAFILE NAME = "pr11829_byt_hv"
IDTRACK = "11829">
<POLARIZATION>LHV</POLARIZATION>
</SARAFILE>
<SARAFILE NAME = "pr11830_byt_hh"
IDTRACK = "11829">
<POLARIZATION>CHH</POLARIZATION>
</SARAFILE>
<SARAFILE NAME = "pr11830_byt_hv"
IDTRACK = "11829">
<POLARIZATION>CHV</POLARIZATION>
</SARAFILE>
</SARAFILES>
<SARASTORED SERVER = "CACR_HPSS"
IDTRACK = "11829">
</SARAQUERYRESULT>
</SARAMETADATA>
2

```

Fig. 3. An example XML file

6 Integration of XML-based services in SARA

XML was adopted to exchange metadata and control information between the various services composing SARA.

We use a three-tier architecture as in fig. 4 to clearly define what service does what, allowing for a more flexible separation of scope and to prepare for future integration with other services. The client tier can be a customized software or a web browser which interact with the middle tier generating queries and commands.

The middle tier is responsible of the management and execution of queries and commands received from the client layer. This can be done by using a number of different services, which can interact to provide the answer requested or execute the command issued.

The backend tier comprises the database management system, the data, a GIS and the computing resources.

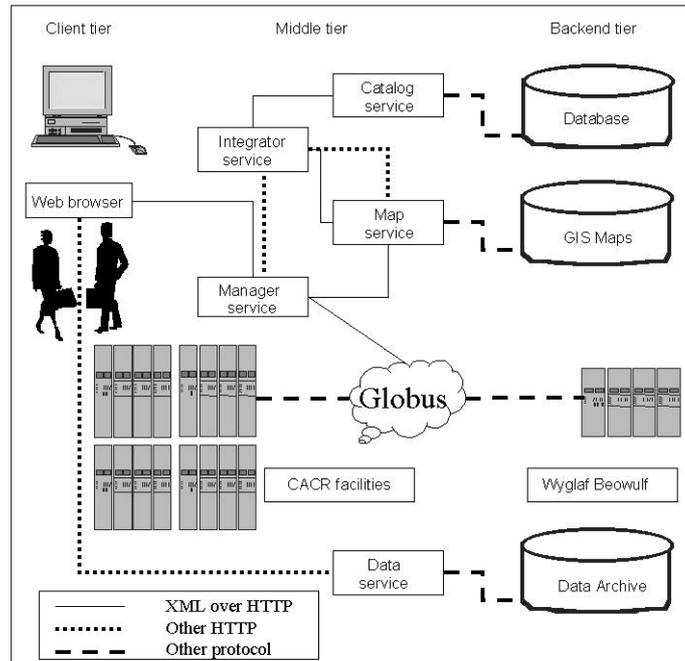


Fig. 4. Digital Puglia Three-Tier Architecture

More details on the use of XML for describing scientific data can be found in [6], and some insights about the new XML-Architecture of SARA are reported in [7].

7 On-Demand SAR processing

The “Active Digital Library” of remote-sensing data provided by the *Digital Puglia* project was conceived to allow not only retrieval of data to the client's workstation, but also customized processing of the data. Besides the usual compute-intensive image post-processing, “on-demand computing” sessions could be required for raw-data that have not yet been processed. In the case of on-demand SAR processing, the data archive must be connected to a powerful compute server at high bandwidth (controlled by a client who may be connected at low bandwidth). This means that a completely automated SAR processor must be designed and the user interface should allow control of some parameters of the SAR processing software.

The architectural solution we adopted to implement on-demand SAR processing to be integrated in the SARA interface was designed to be platform-independent.

The user (with basic password authentication) can choose on a clickable map of the world the region of interest where the SAR raw-data are available, then with a click on a simple button she can start the SAR processor. The user is also allowed to set (from a HTML form or a Java interface) some processing parameters (such as the needed image resolution).

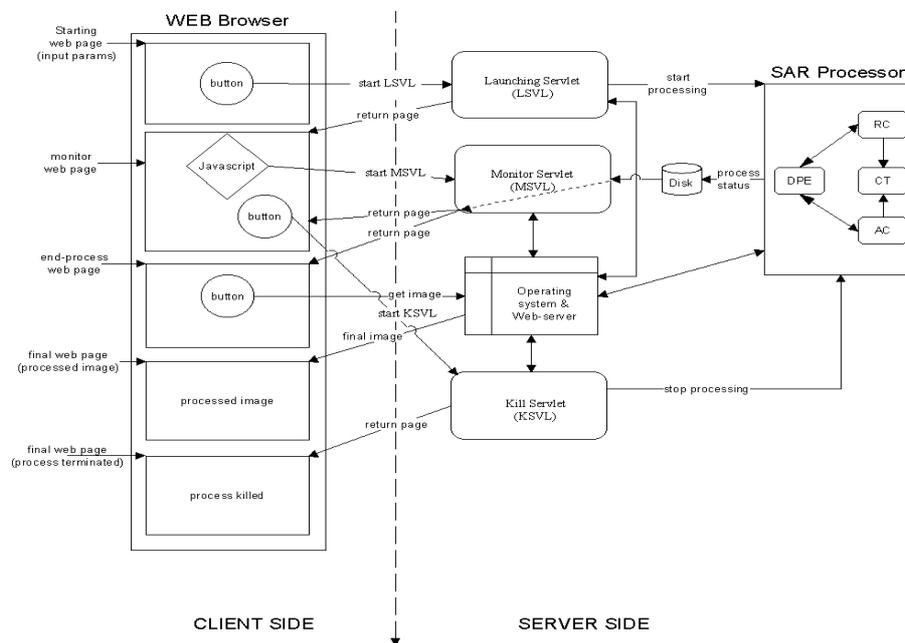


Fig. 5. On-Demand architecture

At this point a new window (monitor) is sent back to the user, showing how the processing is done. When the whole processing terminates, the final SAR image can be obtained by clicking on a button.

For the implementation we use three servlets and two JavaScript Routines. The processing is achieved by a sequential SAR processor until we complete the porting on Wyglaf.

In Fig.5, the architecture of the on-demand SAR processing system is showed. After choosing the track to be processed and setting some processing parameters, the SAR processing can be activated by a click on a button which starts the launching servlet (LSVL) on the server side.

The LSVL:

- controls the availability of computational resource;
- starts the SAR processor;
- immediately returns back to the user a web page, which reports the information on the state of the running process.

To allow the process monitoring a second servlet, the monitor servlet (MSVL) is used. The MSVL is called every 5 seconds by a Javascript routine, started inside of the first web page sent back from the LSVL.

The MSVL:

- controls that processing is going on;
- returns to the user another web page which contains the new output state of SAR processor (this page will active another monitor in a cyclic way).

It should be noted that the running SAR processor writes on file the intermediate messages about the state of the process, and this file is used by the MSVL to read the process information to be sent toward the client side.

The process ends when SAR processing is terminated unless the user interrupts it clicking a button which calls a kill servlet (KSVL): in this case the MSVL informs the user that the processing is terminated, giving the possibility to visualize the final processed image through a simple click on a button on the returned web page.

At this point the image is sent back to the user in a new browser window.

Conclusions

The Digital Puglia Project is a high performance distributed Active Digital Library of remote sensing data. The library allows not only retrieval of data to the client's workstation, but also customized processing of data. The kernel of the library is the SARA architecture that was modified with respect to its first implementation to meet new requirements and to exploit new emerging technologies, like XML.

The issue and the underlying technologies of the Digital Puglia Project were presented.

References

1. R. Williams, G. Aloisio, M. Cafaro, G. Kremenek, P. Messina, J. Patton, M. Wan, "SARA: The Synthetic Aperture Radar Atlas", <http://www.cacr.caltech.edu/sara/>
2. G. Aloisio, M. Cafaro, P. Messina, R. Williams, "A Distributed Web-Based Metacomputing Environment", Proc. HPCN Europe 1997, Vienna, Austria, Lecture Notes In Computer Science, Springer-Verlag, n.1225, 480-486, 1997.
3. I. Foster and C. Kesselman, "The Globus Project", <http://www.globus.org>
4. G. Aloisio, M. Cafaro, "The Wylaf Beowulf machine", <http://sara.unile.it/~cafaro/wylaf.html>
5. "www.xml.com"
6. R. Williams, J. Bunn, R. Moore, and J.C.T. Pool, "Interfaces to Scientific Data Archives", Report of a Workshop sponsored by the National Science Foundation, May 1998, <http://www.cacr.caltech.edu/isda>
7. G. Aloisio, G. Milillo, R.D. Williams, "An XML Architecture for High Performance Web-Based Analysis of Remote Sensing Archives", to appear on FGCS Int. Journal, North Holland

Acknowledgments

The work has been supported by the Italian Space Agency grant ASI ARS-96-118 and by the Center for Advanced Computing Research (CACR) at Caltech.