EAST/XML, THE KEYS FOR FUTURE LONG-TERM DATA ARCHITECTURES ?

Carlos GUERREIRO, Claire-Michelle PELLIZZARI
CS Systèmes d'Information
ZAC de la Grande Plaine, Rue Brindejonc des Moulinais
BP 5872 – 31506 Toulouse Cédex 5
carlos.guerreiro@c-s.fr, michelle.pellizzari@c-s.fr

1 Abstract

1.1 Context

Long-term preservation of scientific and technical data is currently a major challenge in our world of constant technological developments that can affect solutions previously believed to be permanent. Without a global and standardised approach, stored data might quickly become unusable.

1.1 Objectives

The objective of the paper is to sketch the long-term **data architectures of the future**. The main requirements to meet are :

- a standardised architecture in the fields of data and data description formats,
- an enforced independence from technological evolutions (norms and technologies) with a great capacity to evolve,
- a new vision of friendly data services : data always accessible and exchangeable in a easy way (internal communication, external world) : man interfaces, machine interfaces, new appearing interfaces (Web, wireless technologies, PDA,…),
- an enhanced flexibility with several views or meanings of the same data (explicit separations between presentation logic, business logic and contents),
- a reducing costs philosophy always in mind.

The challenge is to leverage the data workshop concepts (set of norms and collections of tools) towards an unified view of data through a long-term data architecture centred on the users needs.

1.2 EAST/XML data architectures : ready for future ?

EAST Technology [1] is a data workshop relying on both following international recommendations : Enhanced Ada SubseT CCSDS 644.0-B-1 (also ISO N° FDIS 15889 norm), and Data Entity Dictionary Specification Language (CCSDS norm enabling to associate semantic attributes to data).

Extensible Mark-up Language (XML) [3] is a W3C recommendation for describing rules for structuring information using embedded mark-ups. It also describes a language for formally declaring the vocabularies used.
As in XML, EAST offers an hierarchical tree view of the data, but an EAST data description is external to the data itself. The power of EAST is that it can also easily describe binary data.

The idea is to combine the two technological standards to take full advantages of the power of EAST and XML.
The principle is simple. Wherever XML (or EAST) is suitable, apply XML (or EAST) : EAST is particularly suitable for describing large bulks of binary data whereas XML is particularly adequate when used as a communication protocol. Furthermore, using XML enforces the openness to the external world. The *long-term* requirement will be achieved using standardised data and data formats and using open source tools : for example, GNAT and gcc for EAST, and XML open source tools.

There are many issues to address (EAST/XML bridges, XML dialect for communication protocol, relations to other CCSDS standards, …) but thinking the two technologies as *complementary partners* instead of *incompatible adversaries* is a good starting point. The *sine qua non* condition for building the long-term data architectures of tomorrow is to federate the existing stable technologies and standards with the emerging ones into an unified view which take into account the advantages of the whole.

2 Background

2.1 EAST technology

EAST Technology is a data workshop relying on the following international recommendations:

- **EAST** (Enhanced Ada SubseT) primarily imagined by CNES and designed in the framework of CCSDS Panel II (CCSDS 644.0-B-1 and ISO 15889:2000). As the name implies (Enhanced Ada SubseT), **EAST** is based upon the ADA language (in fact, EAST is 100% compliant with the ADA syntax). EAST was designed to create **non-ambiguous descriptions of data formats** including **syntactic** (logical and physical) **information**.
- **DEDSL** (Data Entity Dictionary Specification language) designed in the framework of CCSDS Panel II. **DEDSL** allows to add **semantic information** to data by the means of semantic attributes. Two

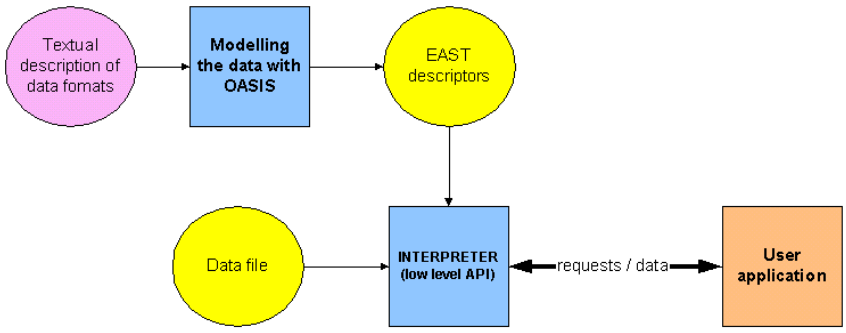
implementation are available: one in PVL (Parameter Value Language), the other one in XML (eXtensible Markup Language).

Through a co-ordinated process between CNES and CS, EAST technology has been developed (primarily in an Research & Technology Development context) with the following major objectives in mind:

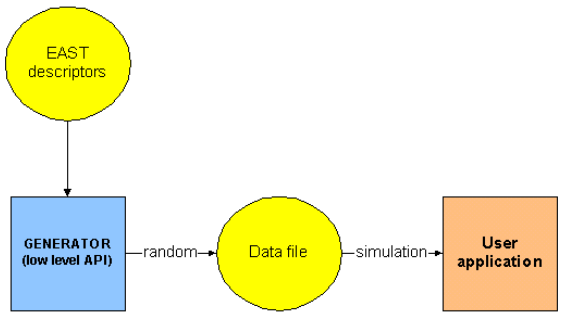
- to provide **complete**, **perennial**, easily understandable and evolutionary descriptions of data formats, including **syntactic** and **semantic** information,
- to provide engineers, scientists and end-users with **generic tools** for supporting the technology.

The first step for using EAST technology is to describe the structure and meaning of the data using OASIS tool. OASIS is a data modelling tool allowing the users to thoroughly describe their data through a user friendly graphical interface. OASIS produces as outputs a description file (often called Data Description Record – DDR) that gathers all syntactic information about the data and a DED file (Data Entity Dictionary) containing all semantic information. This DDR/DED files are then used by all other tools to process the data.

Most of the project having adopted EAST technology use the low-level APIs (INTERPRETER and data GENERATOR) to develop their customised applications that read/write data initially described with OASIS.
DUW (a tool allowing to allowing the production/modification of data files) is often used to produce test data files, ASCII_DUMP is used to have an ASCII representation of binary files (visual checking) and DATA_CHECKER is often used to check the format of data files produced by an application that does not use the EAST technology to generate the files.
In order to clarify the operational use, the following two figures depict summarily two real-use scenarios of EAST technology:



Data modelling (OASIS) and access (INTERPRETER)



Production of random data for simulation purposes

The main benefits gained from making use of EAST technologies could be summarised as follows:

- **Ensuring data preservation**: being free from non perennial file formats, and using a well defined formalism (instead of textual descriptions like Word or PDF documents) with a computer compatible format (ADA subset, PVL, XML),
- **Adding value to data**: with possibilities to define “data about data”, i.e. syntactic and semantic metadata allowing an easier understanding of the data by human-beings or computers and adding real value to the data,
- **Making data life cycle easier**: providing a suite of tools for all phases of data handling, allowing to process data with almost no developments.

2.1.1 Lessons learned

- **Advantages and benefits**:
  - *Concepts*: The practice has demonstrated that the strong concepts of the technology are really good and powerful concepts, that really fit end-user requirements and concerns (data preservation, metadata, data handling, etc.). EAST technology can produce non-ambiguous, well formalised, computer compatible descriptions of data formats. The potential benefits are easy to understand.
  - *Tools*: Globally, the data modeller OASIS tool is seen as a powerful tool, really user-friendly, and the INTERPRETOR and GENERATOR are seen as useful and generic low level tools (APIs) that can be specialised (by coding upper layers) to suit ones needs.
- **Drawbacks**:
  - *Concepts and use*: EAST concepts are really strong, but using the technology can sometimes require some knowledge of the underlying principles. This, because the modelling scheme may have serious performance impacts.
  - *Technical knowledge and understanding*: An in-depth technical knowledge and understanding of EAST tools mechanisms (notably, for the low level APIs), is sometimes required in order to make the best use of the tools. For instance, parsing a huge data file can lead to very poor performances if the user do not know (nor understand) the existing optimisation mechanisms.
  - *Performances*: EAST technology performances have been improved from since the early stages. Nevertheless, more improvements are possible (sometimes required, always expected by users).
  - *Lack of some functionalities*: Some functionalities would make the user life easier, with for instance: copy/paste, type libraries management, batch processing, project management, user oriented logbook messages, user-friendliness, support for algorithms…
  - *Lack of architectural point of view*: EAST technology is mainly a data workshop (a set of norms and tools): there is no indication (methodology or facilities) of how to integrate EAST tools within a more global architecture, or of how to connect these tools with a distributed middleware or an existing infrastructure.

2.2 XML technology

2.2.1 XML technologies

The following table gives an overview of the main XML technologies which are of particular interest.

XML technology	Possible applications in data architectures	Brief description and comments
XML	Data model	The universal format for structured documents and data. http://www.w3.org/XML/
XML Schema	Validation	This specification addresses means for defining the structure, content and semantics of XML documents by using a XML Schema (overcomes the DTD's limits). http://www.w3.org/XML/Schema
XSL & XSLT	Presentation logic, post-processing, etc.	XSL is a language for expressing style sheets for XML documents and XSLT is a language for transforming XML documents into other XML documents. http://www.w3.org/Style/XSL/
XML Query (XQuery)	Access, search, query data	XQuery is an XML query language allowing to make queries on an XML document or a set of XML documents. http://www.w3.org/TR/xquery/
XForms	Automates the generation of forms/MMI	Defines the new generation of Web forms that separates the purpose of the form (what the form does) from the presentation (how it looks). http://www.w3.org/MarkUp/Forms/
SVG	Presentation layer	SVG is a language for describing two-dimensional graphics in XML. http://www.w3.org/Graphics/SVG
SOAP	Communication protocol / interoperability	Simple Access Object Protocol: lightweight XML protocol for exchange of information. SOAP HTTP allows to embed this XML dialect within an HTTP protocol. http://www.w3.org/2000/xp/
XML-RPC	Communication protocol / interoperability	XML-Remote Procedure Call: a Remote Procedure Calling protocol that works over the Internet by encoding the exchanged message using XML in HTTP request. http://www.xmlrpc.com/

2.2.2 Lessons learned

- **Advantages and benefits**
  - W3C standard compliance.
  - Easier maintenance since XML documents are human-legible.
  - Enforced portability since XML is not platform-dependent.
  - Enforced interoperability with other languages and applications.
  - Wide range of open-source, freeware tools
  - …
- **Drawbacks**
  - XML is somewhat verbose
  - Some specifications are evolving quickly and are not really mature at the time being (e.g. SOAP, XML Query,…)
  - *Lack of architectural point of view*: there is no indication (methodology or facilities) of how to use/integrate XML within a more global architecture, how to mix together your architecture and XML, …

3 Key architectural requirements

The following key requirements are essential to prepare the data architectures of tomorrow:

- ***Open:*** the architecture shall be open to external world in order to make it communicate with existing systems of facilities (e.g. archive facility, GS, …) and to ease its integration with *real world*. The interfaces are to be based upon *de facto* technological or normative standards to allow a wide range of possible connections to external systems.
- ***Modular:*** layered architecture making explicit separations between presentation logic, business logic and back-end logic, and allowing different levels of use (developer interface, MMI interface, Services interface,…).
- ***Adaptable:*** user-friendly, and easily customisable without recompilation (MMI customisation, for instance).
- ***Extensible:*** it has to be built upon a plug and play concept (plug-in philosophy) allowing easy incorporation/replacement of external (user-specific) tools with no need for recompilation or coding.
- ***Standardised:*** a standardised architecture in the fields of data and data description formats (EAST CCSDS norm, OAIS model, XML and derivatives), with an enforced independence from technological evolutions with a great capacity to evolve,
- ***User oriented***: a new vision of friendly data services: data always accessible and exchangeable in a easy way (internal communication, external world): man interfaces, machine interfaces, new appearing interfaces (e.g. Web,…),
- ***Reliable and scalable:*** improved performances and fault tolerance.
- ***Cost-efficient:*** generic reused mechanisms, generic concepts of plug and play, reducing costs philosophy always in mind, use of freeware products.
- **Preparing the future** is to be a major principle (favouring innovation, new technologies, new concepts).

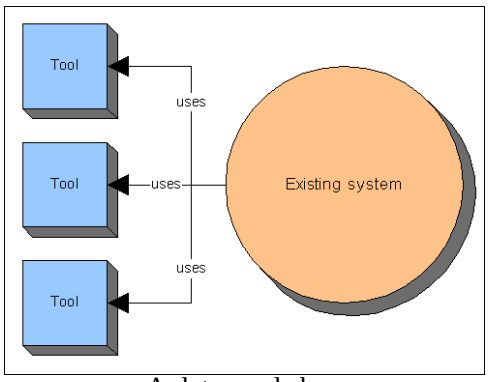
4 Key design factors

The challenge is to leverage the data workshop concepts (set of norms and collections of tools) towards an unified view of data through a **data handling architecture centred on the users needs**.

Are identified below some potential scenarios of use of the data architecture of the future. The power of this architecture relies on its capabilities to be viewed, adapted and used in a wide range of different use cases.

- *Used as a data workshop*:

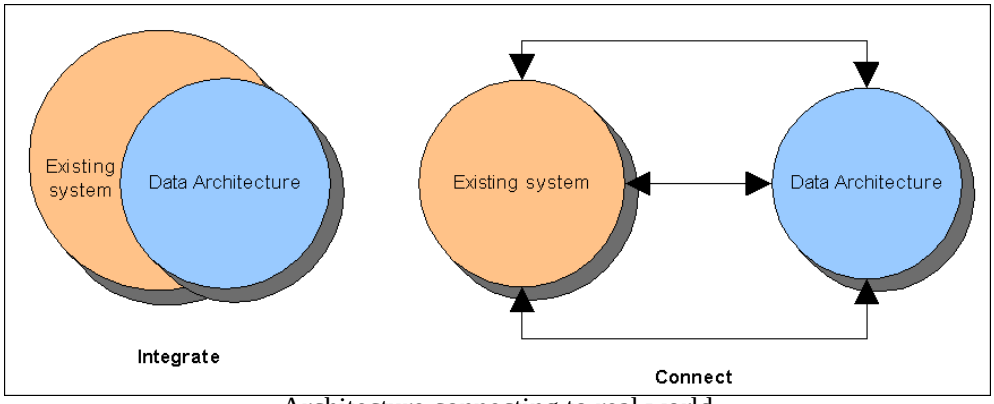
Users can directly use it as a series of enhanced tools covering the entire data life cycle (with almost no architectural concept in mind) as it is the current practice with existing EAST tools. The “data workshop” can be used in a system without any dedicated effort, as it is a combination of independent tools.



A data workshop

- *Connecting to external world:*

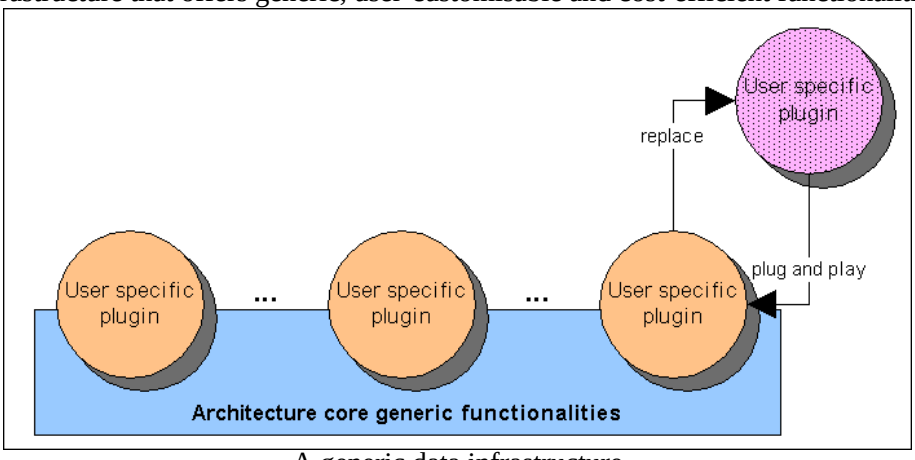
The strong idea with this use scenario is to use the great openness of this architecture to connect/integrate (in)to an existing system or with an existing middleware layer. For instance, if we imagine an archiving facility, it would be interesting to integrate theses architecture in a more global archiving facility for ensuring data preservation.



Architecture connecting to real world

- *A generic data infrastructure:*

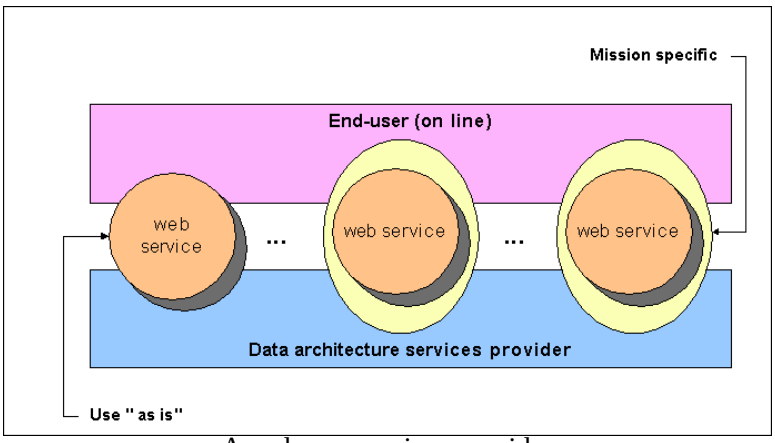
This scenario really takes advantage of this generic architecture allowing users to build their own system upon the infrastructure provided. This is possible by the enhanced modularity of the architecture and its strong concepts of dynamic customisation and plug-and-play. This way, users have capabilities to plug-in their own specific processes, their own business logic (for post-processing for example), and their own MMI in an unified infrastructure that offers generic, user-customisable and cost-efficient functionalities.



A generic data infrastructure

- *A end-user services provider:*

This scenario envisages the use of the architecture as a building-block architecture dedicated to the provision of data services for "*very* end-users", i.e. users really at the end of the chain (for example, for in-line catalogue data presentation, direct on-line delivering of data products,…). The Web Services paradigms included in our vision of the architecture should provide the basic blocks to allow further developments and customisation in this direction.



A end-user services provider

5 How EAST and XML can help ?

Some typical usage of EAST and XML are given below:

Functionality	EAST	XML
Data formats	X (binary)	X (ASCII)
Software infrastructure		X
Persistence back-end layer	X (binary files)	X (native XML, SGBD and XML,…)
Metadata		X
Communication protocol and exchange techniques	X (also embed EAST descriptors into XML documents)	X (SOAP, XML dialects…)
Packaging techniques		X (data package or also encoding binary data in XML documents: brute force, base-64, huffman encoding, …)
Presentation logic		X (XSLT, XFORMS, SVG,…)
Querying		X (XML Query)
Customisation, plug and play mechanisms		X
Web services		X

6 Case studies

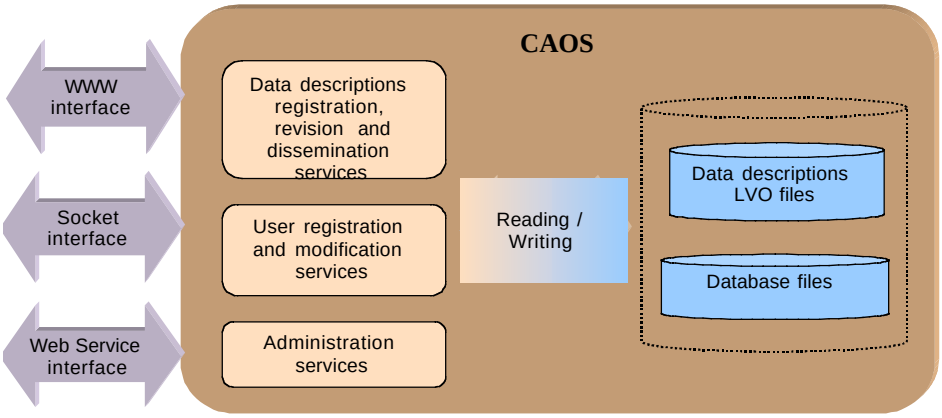
Two case studies are given in the following to illustrate the use of XML/EAST in operational architectures.

6.1 CCSDS Control Authority Office System

6.1.1 Objectives

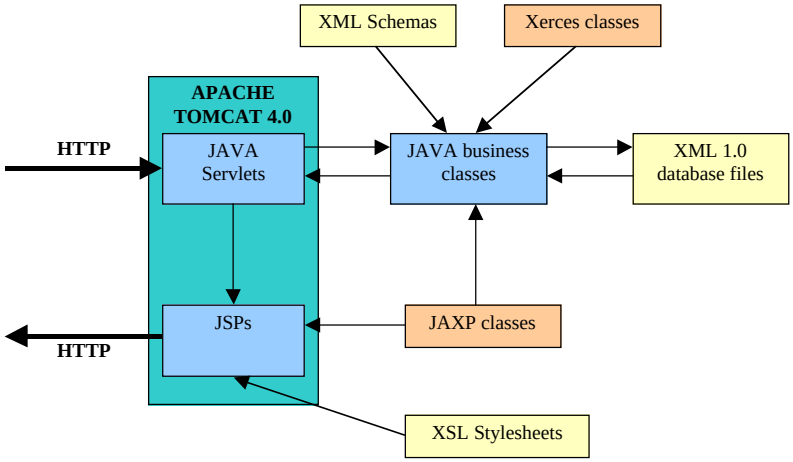
The Control Authority Office system is a World Wide Web based system which provides support for one or more Consultative Committee for Space Data Systems (CCSDS) Member Agency Control Authority Offices (MACAO) in the areas of data description registration, dissemination, revision and annual reporting.
It stands as a generic solution usable by any agencies following the CCSDS standards needing :

- A physical safeguarding of the data descriptions ;
- An advanced and powerful access to these descriptions ;
- Long-life preservation of these data descriptions.



CAOS functional architecture

The goal of the project is to renovate the existing system by the introduction of new technologies (e.g. Java, XML and XML derivatives).



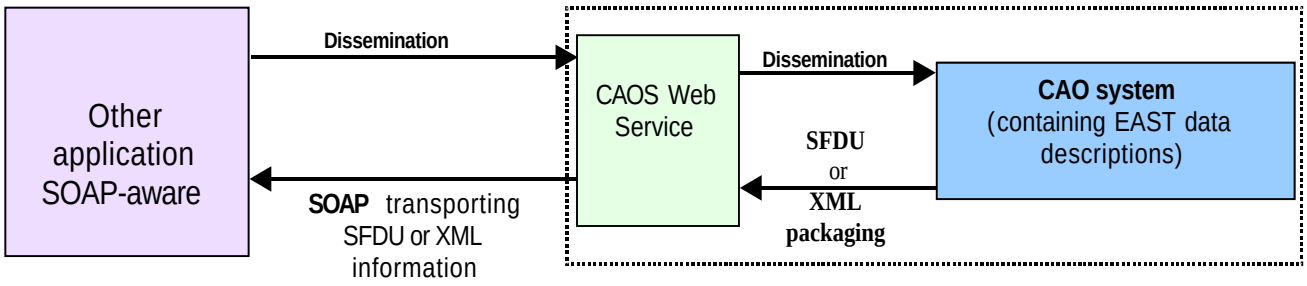
CAOS technical architecture

6.1.2 Use of EAST and XML

The following table summarises the fields of application of EAST and XML for the CAO system:

Functionality	EAST/CCSDS standards	XML
Data descriptions formats	X (or other CCSDS compliant formats)	
Software infrastructure (management database, validation)		X
Presentation layer		X (XSLT, Java)
Web service interface : communication protocol and exchange techniques	X (also SFDU packaging techniques)	X (XML packaging and SOAP)
Queries		X (XML parsing, and XML query)
Administration		X

The data descriptions are delivered **over the SOAP protocol** using **XML** for encapsulating the data descriptions information. SOAP is defined as a lightweight protocol for passing structured and typed data between two peers using XML. It can be used on top of any protocol (e.g. HTTP, SMTP,…) that supports the transmission of XML data from a sender to a receiver. SOAP is then naturally suited to transport data descriptions in their XML packaging forms.



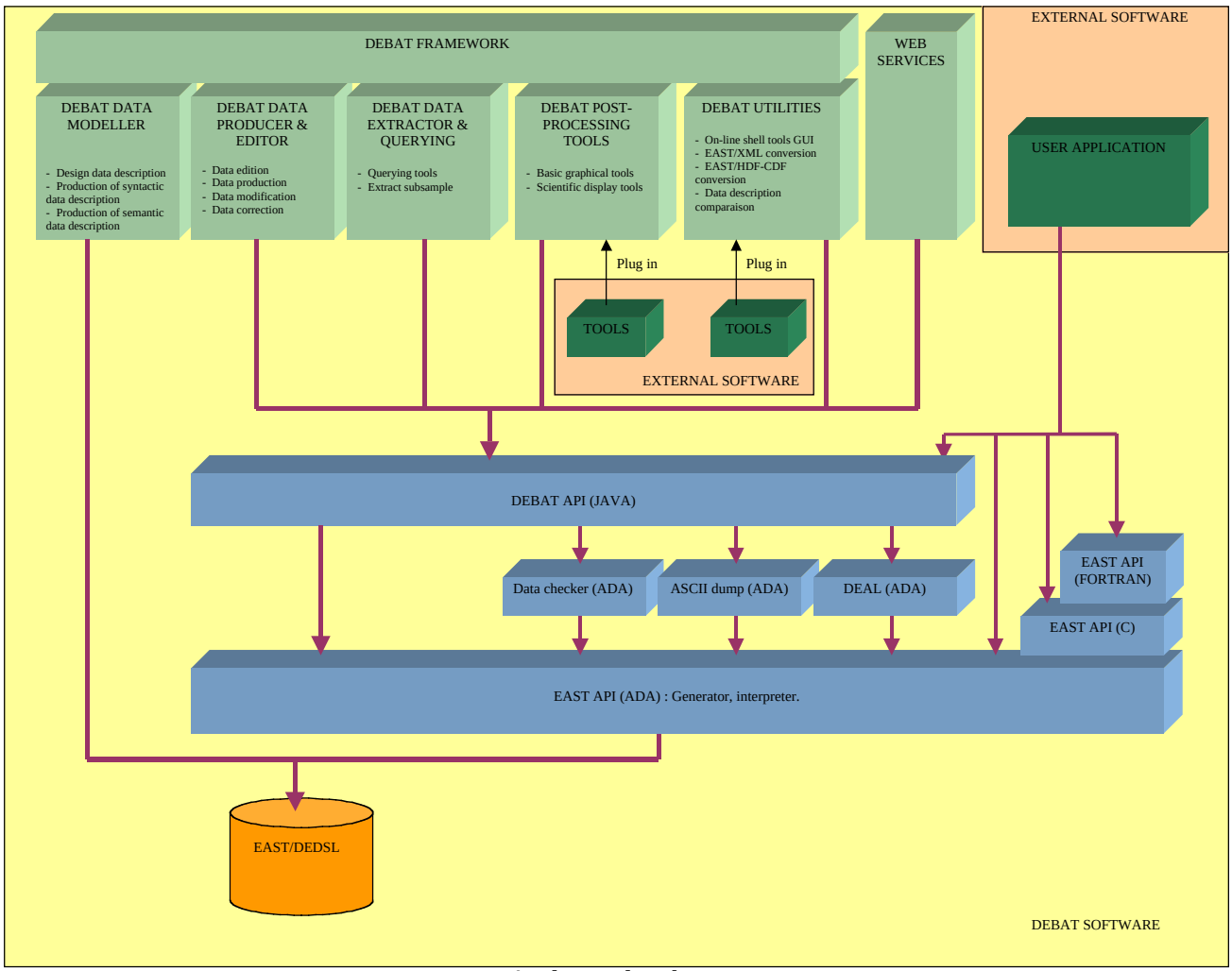
SOAP transport and XML packaging

6.2 DEBAT – Development of EAST Based Access Tools

6.2.1 Objectives

The main objective of "Development of EAST Based Access Tools" is to **build a set of enhanced tools** built upon EAST technologies to provide engineers, scientists and end-users with **powerful tools covering the entire data life cycle** (modelling/definition, production/generation, editing, accessing, checking, processing, extracting, querying, post-processing,…).
Another objective is to **promote and diffuse** DEBAT tools family and concepts to the potential user community (existing fields of application or new ones) by making them aware of the availability and advantages of these enhanced tools (DEBAT workshop, web site, demonstration kit,…).
The underlying philosophy is to build on the knowledge and capitalised experience gained during previous developments (CNES, CS SI and ESA) being carried out for several years, to take advantage of the lessons learned from current/past EAST user projects (limitations, expectations), and to take into account new requirements (coming from the analysis of a range of selected projects/missions) and forthcoming requirements.
There are two major identified axes of work:

- **Improvements (or new developments) to existing tools *within current fields of application***: this covers the extensions and enhancements of EAST language and tools for well-known domains where the degree of confidence is high due to past experience. The final results of these developments are expected to be high quality software: reliable, well engineered, user friendly, properly documented and flexible.
- **TM/TC fields of application**: this covers the application of EAST technologies to TM/TC processing which is a new and critical domain that requires in depth analysis. The main objective for this new field of application is to provide the "**proof of concept**", i.e. to analyse (and potentially demonstrate) the feasibility and advantages gained from applying EAST in the TM/TC arena.



DEBAT layered architecture

6.2.2 Use of EAST and XML

Functionality	EAST	XML
Data descriptions formats	X	
EAST enhanced tools	X	
Software infrastructure		X
Plug and play, customisation		X
Presentation layer		X (XSLT, XFORMS)
Communication protocol and exchange techniques		X (XML packaging and SOAP)
Queries		X (XML parsing, and XML query)
EAST/XML bridge	X	X
Post-processing	X (EAST tools & utilities)	X
Data Web services		X

7 References

[1] The Data Description Language EAST specification CCSDS 644.0-B-2/ISO 15889:2000
[2] Reference Model for an Open Archival Information System CCSDS 650.0-R-2, ISO/DIS 14721-2
[3] XML Extensible Markup Language http://www.w3.org/TR/2000/REC-xml-20001006