Product Lifecycle Management for scientific data: I2M
an industrial approach to data access, preservation and exploitation

IIM 2012
24-26 October 2012
Speaker Raffaele Guarino - Capgemini

German Aerospace Centre (DLR), Oberpfaffenhofen, Germany
• Rationale of this presentation
• The Earth Science data challenge
• The proposed PDM/I2M solution
• Advantages of the tool
• Next steps
Here we will show a tool (I2M) aimed at support data exploitation.

This tool which is still in development (re)-uses concepts and paradigms matured in the frame of the industrial federated manufacturing processes.

In particular it is using the following concepts:

a) Product (de)composition list and production process (PDM/PLM)

b) Distributed production (federation and layering of the composition-processes) including concepts concerned with logistic data flow chains.

c) Exploitation and implementation of upper layer of usage (e.g. information mining, ontologies, trees of workflows, etc.).

The paper and this introduction are directed to sketch the proposal, to show the status and to trigger comments and observation for improvements.
Science, like other business and even everyday life, is becoming more and more data intensive. In some sciences the data-management challenge already exceeds the computer power challenge in its needed resources.

Moreover science is the motivator for many changes, new paradigms and new technologies. We identify six technology areas that are fundamental to supporting the data management requirements for scientific applications:

- Workflow, data flow, data transformation
- Metadata, data description, logical organization
- Efficient access and queries, data integration
- Distributed data management, data movement, networks
- Storage and caching
- Data analysis, visualization, and integrated environments
The PDM/I2M objective is the rapid identification, access and exploitation of the available scientific data as well as of all companion information like: supporting tools, methods of use, documentation, technology, product/data generation/support, preservation methods and needs.

Targets of proposed services are:
1) Scientific data users (e.g. Principal Investigators, Communities of Interest) who are concerned with access and use of scientific data series independently of their actual location, with all relevant information and relationships with other elements.
2) Data owners gathering scientific data, responsible for their processing (at least the basic processing), concerned with their availability for exploitation, for their mid/long term maintenance and usability.
3) Value Adders, who have the interest to collect more and wider in order to improve their own business and think for new possibilities.

We aim at contribute to the development of the common framework for a Collaborative Data Infrastructure as indicated by the DCCU document of EC, October 2010.
Product Lifecycle Management (PLM) and Product Data Management (PDM) paradigms have the purpose to support the lifecycle of any man-made product and support manufacturing/production of products in order to be on the market (to the user) according to the demand and of the industrial capabilities. Close to this there are other concepts like: time-to-market, just-in-time, holistic production/manufacturing, integrated logistic support (ILS) and many others.

Generally speaking a simple PLM/PDM is a tool aimed at support the assembly of an artifact composed by elements; in the simplest case it is a pair of a list of elements and of a method (or prescription, or process) to compose the artifact plus descriptive documents.

In the simplest formulation a PLM/PDM contains and manages:

1. Data of elements needed to build/make a product
2. Documents concerned elements and processes
3. Processes to build or make

Naturally the process can be see at multiple scale of trees.
Our approach and reuse of PLM/PDM has been based on the following main drivers:

- **a)** Scientific products (data files) are the result of a data processing applied to a set of data (primary, secondary, ancillary);
- **b)** Data and algorithms are subject to changes and evolutions of various kinds and have a different lifecycle;
- **c)** Documents and descriptions of data and processing (data manipulation) are needed in order to allow future evolutions while maintaining the history, context, and provenance;
- **d)** Scientific product can be stacked as results of beneath productions.

Furthermore, we have considered:

- **a.** User Centric building a **services model** rather than a “product”. Vision of this service is to support **users** like a commodity and becoming an important asset class.
- **b.** The service aimed at supporting exploitation and production, while collecting relationships among elements and available manipulations.
- **c.** Also, Product centric and its relationships with other “**products**”.
PDM/I2M is not a new scientific data archive or repository.

What PDM/I2M is:

1. A database of already existing metadata and information, maximizing the reuse of what it is available somehow.
2. The database contains links to elements (e.g. data, processors, algorithms) and relationships among elements;
3. The database contains workflows (prescription, processes descriptors) aimed at execute different activities (product generation, collection, data mining).
4. Conceived as supporting tool for data managers and data users with limited training needs.
5. Conceived to work in a federated environment and utilisation of the latest technologies: data/processing virtualisation, cloud computing.
6. Conceived to be settled as a service oriented environment (ITI services).
Position of I2M in the research and exploitation loop

Research process

- Question and Hypothesis definition
- Information and Resources Study
- Experiment & Results
- Data collection & analysis
- Data interpretation
- Resulting publications

PDM/I2M Services

- Info resources and relationships (bricks)
- Data & Processing virtualization
- New resources and relationships (bricks)

Production process

- Data & Info capture
- Data processing & reprocessing
- Data curation
General data schema

Mission Documentation

Requirements
- Mission Requirements
- Required instruments/sensors
  - Processing methods
  - Qualification models
- Scope & Performance
  - Information relevant orbits
  - Platform position
  - Altitude
  - etc..

Architecture

Reports
- Events
- Studies
- Communities
- Model Review
- Algorithm Changes
- Cal/Val Changes
- Publications
Entity-Relationship Diagram (a detail)
Information to Mining (I2M) prototype is aimed at providing the following services:

1) Management of the product list, of its elements and of relevant relationship. Conceptually the product is anything that can be produced starting from “lower” layer of elements. This include information mining using collections of inputs, information mining algorithms and capable to “create” a products;

2) Management of multiple trees of stacked products;

3) Management of variants (same basic product with some changes)

4) Search&retrieve of products, elements and metadata upon federated archives;

5) (re) Generation of scientific data products through their product generation list. This includes tools supporting the workflow management (creation, modification or testing);

6) Support to virtualization (CPU, OS, Algos, Data) and to the ontology’s engines

7) As future vision the service will help users to search or identify products not available (search for missing).
Prototype Architecture

Technology: Blade, CloudComputing IaaS, Languages :Java, PHP, FLEX and XML Semantic: RDF and OWL

I2M DB works on metadata and pointers in order to create trees of products including all relevant informations and prescriptions creating scientific products.
<table>
<thead>
<tr>
<th>Instrument Category: Medium-resolution spectroradiometer</th>
<th>Products</th>
<th>Data Products</th>
<th>Data Products</th>
<th>Products &amp; Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDCTS, GEOSTATIONARY ENVIRONMENTAL MONITORING SPECTROMETER</td>
<td>Data</td>
<td>Level 0</td>
<td>Orbit parameters</td>
<td>Product Details</td>
</tr>
<tr>
<td>CDCTS, GEOSTATIONARY ENVIRONMENTAL MONITORING SPECTROMETER</td>
<td>Software</td>
<td>Level 1a</td>
<td>Orbit state vectors</td>
<td>Product Details</td>
</tr>
<tr>
<td>CDCTS, GEOSTATIONARY ENVIRONMENTAL MONITORING SPECTROMETER</td>
<td>Documentation</td>
<td>Level 1b</td>
<td>Orbit state vectors</td>
<td>Product Details</td>
</tr>
<tr>
<td>CDCTS, GEOSTATIONARY ENVIRONMENTAL MONITORING SPECTROMETER</td>
<td></td>
<td>Level 2</td>
<td>Orbit state vectors</td>
<td>Product Details</td>
</tr>
<tr>
<td>CDCTS, GEOSTATIONARY ENVIRONMENTAL MONITORING SPECTROMETER</td>
<td></td>
<td>Level 3</td>
<td>Orbit state vectors</td>
<td>Product Details</td>
</tr>
<tr>
<td>CDCTS, GEOSTATIONARY ENVIRONMENTAL MONITORING SPECTROMETER</td>
<td></td>
<td>Auxiliary</td>
<td>Orbit state vectors</td>
<td>Product Details</td>
</tr>
<tr>
<td>CDCTS, GEOSTATIONARY ENVIRONMENTAL MONITORING SPECTROMETER</td>
<td></td>
<td>Ancillary</td>
<td>Orbit state vectors</td>
<td>Product Details</td>
</tr>
<tr>
<td>CDCTS, GEOSTATIONARY ENVIRONMENTAL MONITORING SPECTROMETER</td>
<td></td>
<td>Calib/Val</td>
<td>Orbit state vectors</td>
<td>Product Details</td>
</tr>
<tr>
<td>CDCTS, GEOSTATIONARY ENVIRONMENTAL MONITORING SPECTROMETER</td>
<td></td>
<td>Instru</td>
<td>Orbit state vectors</td>
<td>Product Details</td>
</tr>
<tr>
<td>CDCTS, GEOSTATIONARY ENVIRONMENTAL MONITORING SPECTROMETER</td>
<td></td>
<td>Browse</td>
<td>Orbit state vectors</td>
<td>Product Details</td>
</tr>
</tbody>
</table>

**Data**
- Level 0
- Level 1a
- Level 1b
- Level 2
- Level 3
- Auxiliary
- Ancillary
- Calib/Val
- Instru
- Browse

**Documents**
- PO-TN-MEL-GS-002 MERIS Level 1 Detailed Processing Model - Parameters Data List Issue 8, Revision 0 - 10/09/2011
- ACRI ST - ESA

**Software**
- BEAM

**Documents & Algorithms**
- MERIS-FAQ MERIS Frequently Asked Questions Issue 1, Revision 1 - 07/02/2012
- ESA - VicicTerra

**Documents & Algorithms**
- ESA - SERCO - ELSAG DAMEMAT

**Documents & Algorithms**
- ESA - SERCO - YEGA

**Documents & Algorithms**
- MERIS Spectral Characterisation Issue 1, Revision 3 - December 2005
- ESA - ACRI
Summary of our proposed approach is:

a) Implementation of a PLM/PDM like conceptual paradigm to support output of scientific products starting from their basic components and variants;

b) Providing support to “information mining” methods and algorithms just making available the “framework” for new exploitations.

c) Implementation of relational database(s) aimed at contain “relational” information of scientific products and of their in components or elements.

d) Fill-in trough automatic and semi-automatic capture of relationships and elements (e.g. DOORS, RTC, annotations, etc.)

e) Fully based on federated relationship rather then on a centralised core;

f) Service model rather than product model.
Advantages

The most important advantages can be summarised as:

a) I2M supports generation of outcomes from identified inputs both as from a single scientific product or collections of products;

b) Workflow concepts overcome issues concerned with product generation. Workflow includes identified and available algorithms or user provided ones. Workflows can be standard (provided by the data owner or producer) or user-defined.

c) Working in a federated environment, and in future using forefront technologies, it is possible to reduce at minimum data moving across archives.

d) Increasing and growing the availability of more and relationship among elements and data supports a gradual reduction of knowledge gap, will improve the exploitation of other elements in the network and will trigger an overall reliability among elements and components.

e) It will pave the way for not-ITI skilled users.
We are implementing the prototype’s software (virtualisation, workflow engine) and completing description of dependencies and some relationships.

We have decided to use the Earth Observation context (GENESI-DEC) and we are asking ESA/EOP to use data for testing with first focus on radiometers, we will proceed on with other instruments/sensors/missions just after completing the current prototype (expected by February the end).

The next step will be the extension to other scientific domains (Atmosphere, Oceanography) and the implementation of the service model (ITI infrastructure, service and service desk prototype).
We thank ESA EOP-G Department for the support and availability to provide us comments and suggestions. We thanks colleagues who are working with us in implementing this tool. Our contacts are:

**Barbara Polsinelli and Simona Manailescu** (Capgemini) who are working on metadata and relationships.

**Claudio Di Giulio** (Capgemini) who is working on databases models and federated structures relationships.

**Raffaele Guarino** (Capgemini) who is working on requirements and relationships models (complex network relationships).

**Christian Ceprani** and **Ugo Di Giammatteo** (ACS) who are working on prototype’s software, virtualisation and workflow engines.

Barbara.polsinelli@capgemini.com, simona.manailescu@capgemini.com, claudio.digiulio@capgemini.com, raffaele.guarino@capgemini.com, udig@acs.sys.it, christian.caprani@acs.sys.it
L. Barabasi, CCNR / USA, complex network example Human Disease