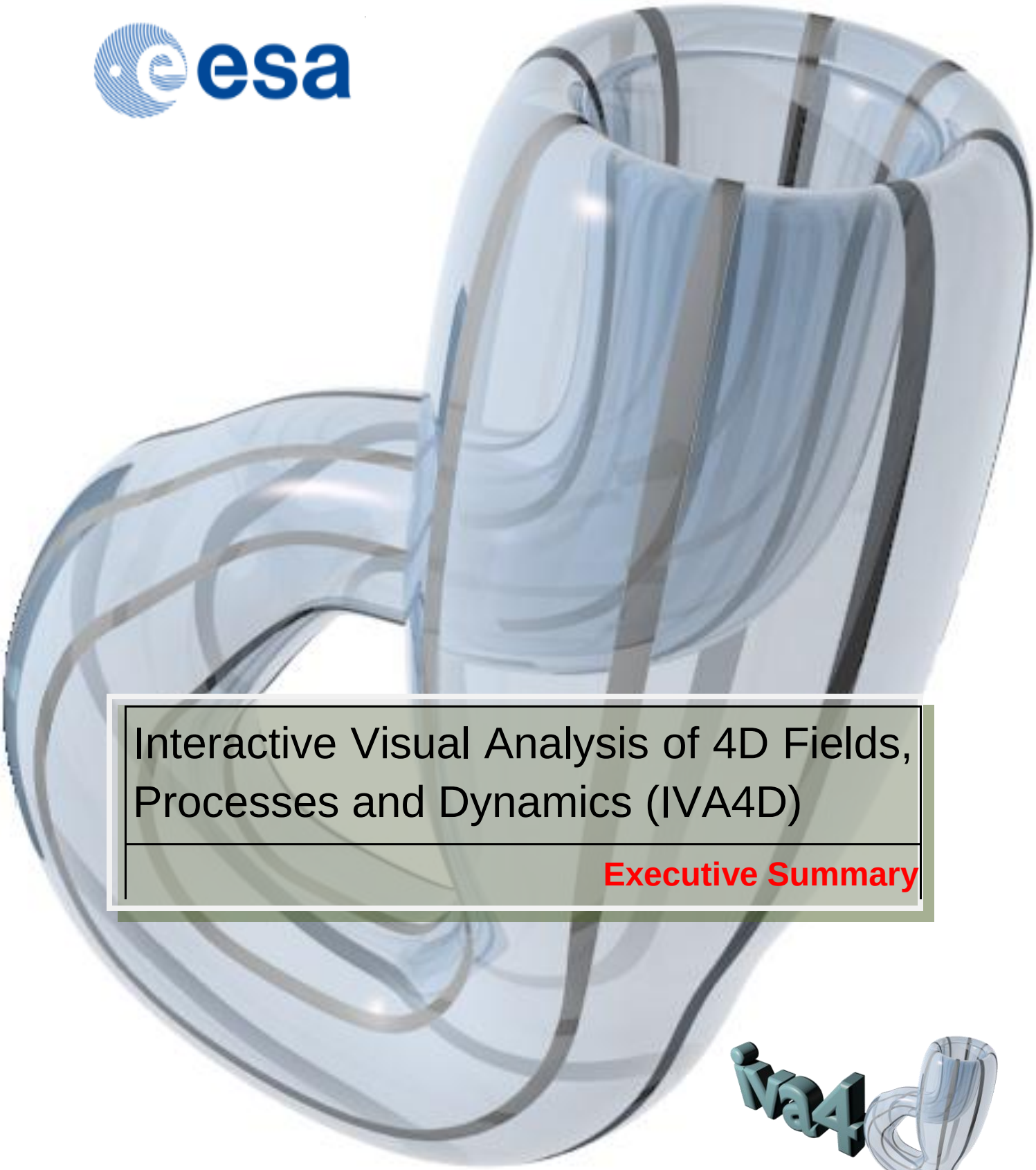


Interactive Visual Analysis of 4D Fields, Processes and Dynamics (IVA4D)

Executive Summary

magellium

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Science & Technology Facilities Council

EXELIS
Visual Information Solutions

The “Klein bottle” on the cover, is a 3D representation of a 2D non-orientable surface without any boundary.
If you follow any path on the surface, continuously, you will never reach a boundary (surface with no limit) and your orientation will change when you come back where you started (reversing of the normal to the surface). For comparison, a sphere is an orientable surface with no boundary.

It illustrates that a smart representation of a concept (data) helps to understand the features of a complex concept (data).

References

Document reference:	ALL_IVA4D_069_ESY
Authors:	D. Petit Magellium A. Doeuvre Magellium J. Blower NCEO D. Clifford NCEO D. Burridge Exelis VIS C. Cook Exelis VIS D. Bailey Exelis VIS B. Matthews STFC S. Nagella STFC
Issue / revision:	0.9
Issue date:	24/12/2012
Addressee:	Pascal Lecomte ESA
ITT reference:	AO/1-6740/11/F/MOS
Contract No:	4000104423/11/F/MOS

History

Issue / rev	Date	Observations/Comments
0.9	24/12/2012	Draft for review

TABLE OF CONTENT

1 - Introduction

1.1 - Purpose of this document

This document is the “Executive Summary” of IVA4D project. It is part of the deliverable DRD14. The purpose of this document to provide highlights of IVA4D Study including aims, overall approach, findings, achievements and conclusions. It includes all the relevant information of the following documents:

- DRD1 Survey analysis of existing tools;
- DRD2 User/Stakeholder Requirements Document for Scientific Data Visualisation System(s);
- DRD3 Gap Assessment based on the analysis of DRD1 and DRD2
- DRD4 User/Stakeholder Requirements & Concepts presentation;
- DRD5 Service Layer Definition Document for Scientific Data Visualisation System(s);
- DRD6 Infrastructure Layer Definition Document for Scientific Data Visualisation System(s);
- DRD7 Preliminary Definition (Service and Infrastructure Layer) presentation;
- DRD8 Traceability Matrix to USRD for Service Layer Definition;
- DRD9 Traceability Matrix to USRD for Infrastructure Layer Definition;
- DRD10 Final (System) Definition (Core Functions, Applications, Methods & Scenarii) Document;
- DRD11 Comparison of Final System Definition with Existing systems;
- DRD12 Final (System) Definition Review Presentation
- DRD13 Minutes of Meetings

Please, refer to these documents for more details.

1.2 - Acronyms and abbreviations

Acronym, abbreviation	Description
CCI	Climate Change Initiative
CEMS	Climate and Environmental Monitoring from Space
CMS	Content Management System
ECV	Essential Climate Variable

EO	Earth Observation
GIS	Geographic Information System
GSP	General Studies Programme
IVA4D	Interactive Visual Analysis of 4D Fields, Processes and Dynamics
NEX	NASA Earth Exchange

List of acronyms and abbreviations

1.3 - Applicable documents

ID.	Ref.	Description
AD01	Contract No. 4000104423/11/F/MOS	Contract with Magellium Ltd Interactive Visual Analysis of 4-dimensional Fields, Processes & Dynamics
AD02	GSP-SOW-10-187	Interactive Visual Analysis of 4-dimensional Fields, Processes & Dynamics: ESA Statement of Work
AD03	MAG-11-PTF-043-v1.0	Interactive Visual Analysis of 4D Fields (IVA4D): Technical Proposal
AD04	MAG-11-PTF-043-v1.0	Interactive Visual Analysis of 4D Fields (IVA4D): Financial, Management and Administrative Proposal

List of applicable documents

1.4 - Reference documents

ID.	Ref.	Description
RD01	Not Applicable	

List of reference documents

2 - Introduction

2.1 - GSP context

IVA4D is a fifteen month project (October 2011 – December 2012) executed in the context of ESA's General Studies Programme (GSP, www.esa.int/gsp). The main role of this programme is to carry out preparatory analysis and act as a "think tank", laying the groundwork for the Agency's future activities.
The objectives of the general studies programme are to:

- Contribute to the formulation of the overall ESA strategy
- Study feasibility for selection of new mission concepts
- Prepare/demonstrate the case for approval and funding of new optional projects/programmes
- Support the evolution of ESA by analysing and testing new working methodologies

A diversity of topics is investigated via GSP undertakings, running across the entire spectrum of the Agency's activities. In average, each study lasts one to two years, sufficient time for in-depth exploration of each subject. The assessment studies undertaken by the GSP provide ESA and its member states with the necessary information on which to base their decisions about the implementation of new programmes and the future direction of space activities.

2.2 - Technical context

Data visualisation is the key to both the understanding of complex science issues and to their communication. Convincing non-expert communities (the public, government, and industry) is increasingly the role of scientists, as research moves into an era where its outputs are of major relevance to our economies, in some cases determining future global policy on the environment, energy, the use of satellite resources, and many others.
The present context is rather fragmented, with a concerted approach to data visualisation slow to emerge. Although there is considerable knowledge and skill related to scientific data visualisation, this tends to be concentrated in single research establishments, and even worse, small teams or individuals in these establishments. Results are often presented effectively, with a sound mastery of the visualisation technology. However, the context lacks coordination, and lacks standard methods of representing 3D datasets for communications purposes. The result is that entities external to this confined scientific community get a disjointed, sometimes contradictory message.
Any new architecture must allow information to be presented in the same way across a domain / community, and must support effective and recognised ways of presenting this information to

non-expert, decision making bodies within, or external to, the community. This is particularly important when the domain in question has a major influence on the way society behaves: energy, remote sensing, climate, meteorology, are only a few of a much longer list.
It is therefore the case that we must try to satisfy at least two aims simultaneously: visualising the datasets of our own domains / communities in a standardised and shareable manner, and explaining the significance of these datasets to the people in society with a decisional role. It is also critical that the visualisation systems of the future make it much easier for different kinds of data from different sources to be combined more easily in order to support research and decision-making. Fulfilling these aims will serve the needs of the community and society as a whole.

2.3 - Aim of the project

More and more often currently and in the future, decision-making is and will be based on the analysis of large amounts of scientific data, which can only be properly evaluated in the form of a multivariate 4-D system. Essential Climate Variables (ECV) in the context of Climate Change research and science are a good example of this new imperative but many other science, engineering and economic disciplines are already addressing this type of issue.

The IVA4D project focuses on

- **Interactive** (Make it easier for different kinds of data from different sources to be combined, explored and shared)
- **Visual** (Visualise multidimensional datasets in a consistent and coherent way)
- **Analysis** (Represent data in a way that helps to understand and to take decisions)
- **of 4D Fields, Processes and Dynamics** (multidimensional, complex, large, multivariate data)

The purposes of IVA4D project itself are:

- to **collect requirements**, best practices and bad practices in various scientific domains
- in order to **define a proper generic system** that can be declined later in specific domains
- to **demonstrate the feasibility and the benefits** of the challenging features of the system

2.4 - Objectives of the Study

This study will principally address three objectives:

1. A strategic one, which will consider the case for such a tool across a number of science domains, and determine the nature of the message that must be conveyed, through the tool, to appropriate decision makers and committees.

2. An architectural one, in which the specifications of such a visualisation tool will be derived, taking into account that it must include a reference architecture, valid and deployable across a number of science domains.
3. A technical one, which will look at the technical feasibility of some of the proposed solutions, through state-of-the-art investigations and practical feasibility studies, which will look at the viability of vertical concepts for specific domains.

The schematic below illustrates the objectives of the system that is the subject of the present study, showing also the likely stakeholders and decision making authorities which might benefit from its use:

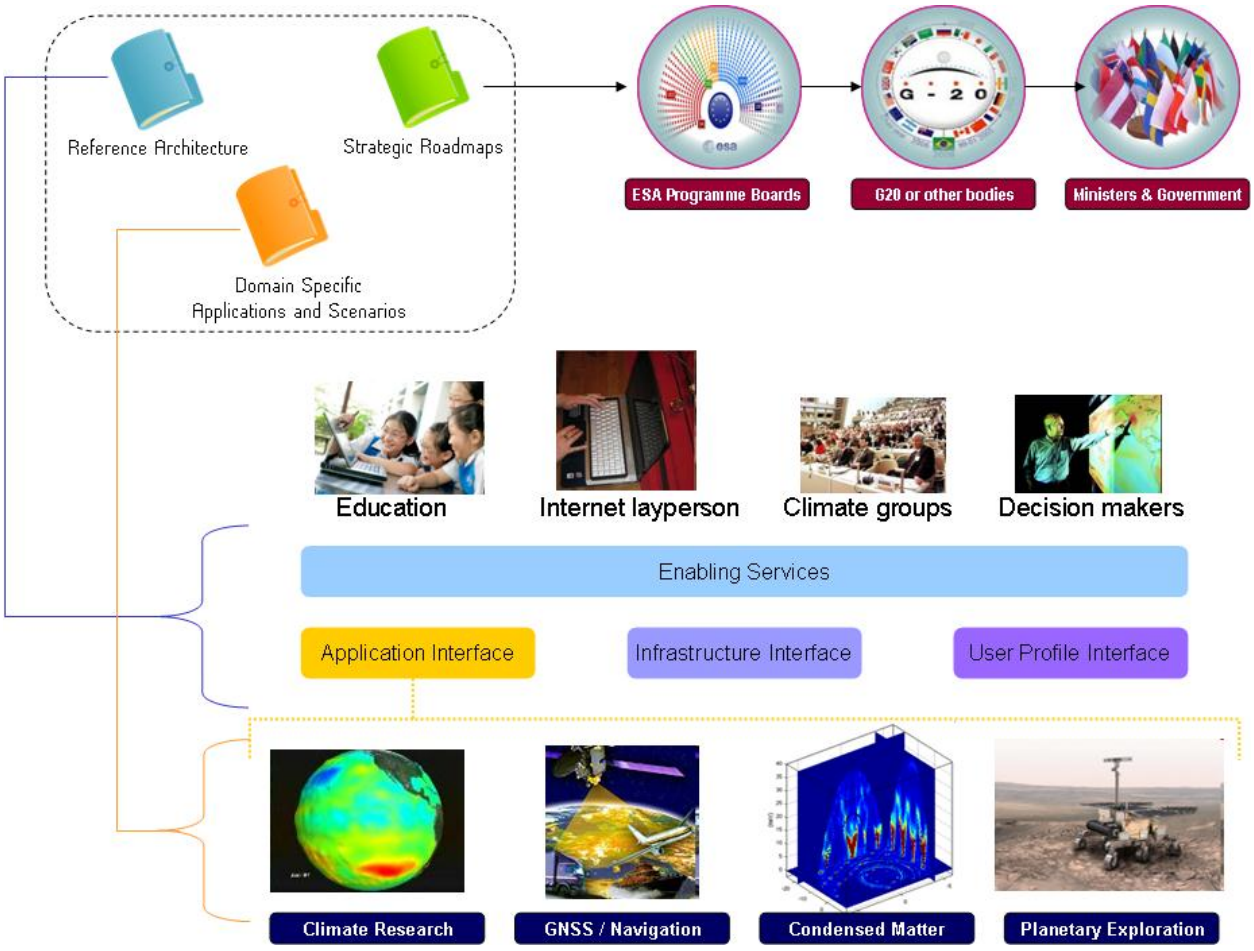


Figure 1. Objectives of the system

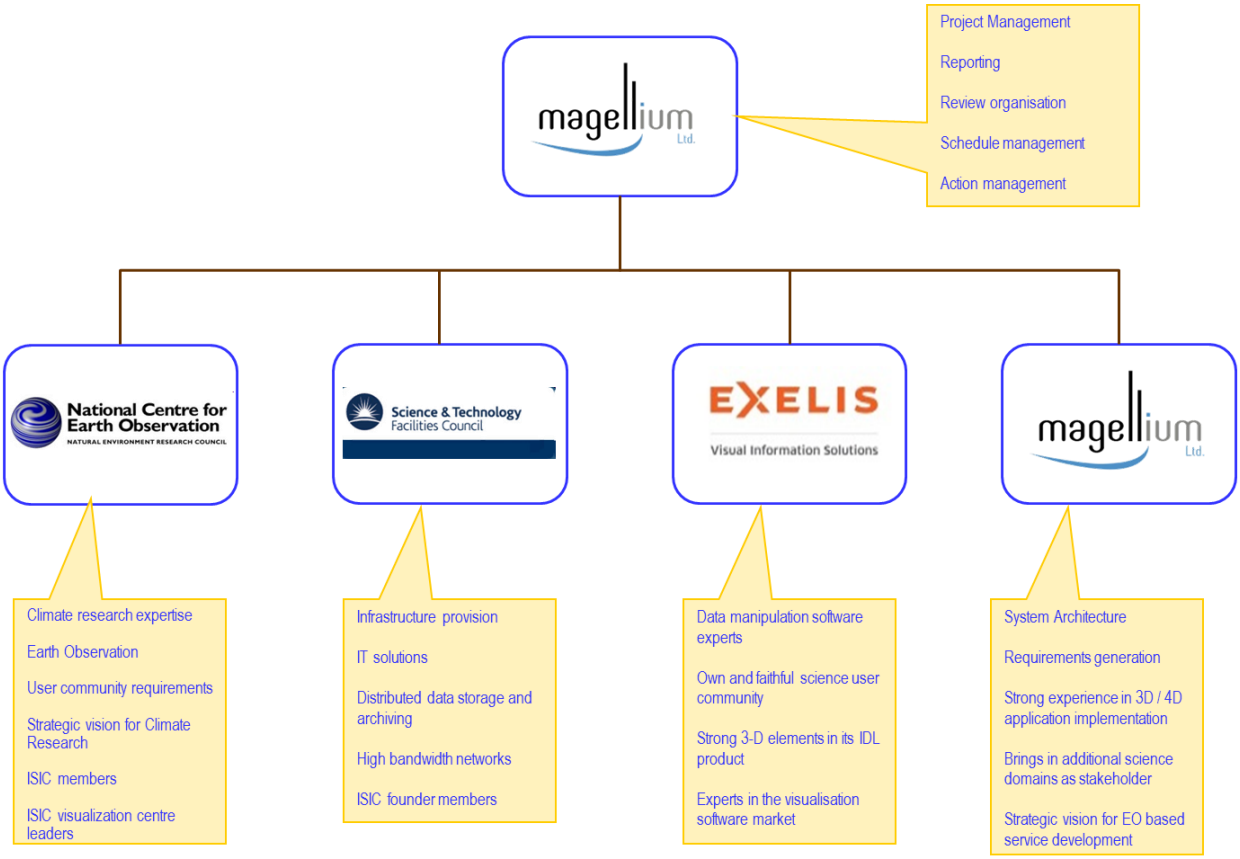
- The Reference Architecture will go towards implementing a domain-independent architecture for data access for visualisation purposes. In the terminology of the study, this is an element of the System Definition Document.
- The Domain Specific Applications and Scenarios will serve to customise the system to particular selected domains, should an implementation be decided on after this study.

- Strategic Roadmaps will be generated (as part of the Strategic Needs Assessment Report deliverable) explaining how the system can be adapted to the different domains through appropriate customisation of the applications and scenarios.
- Several major ESA programmes are shown (at the lower end of the schematic) contributing configurations to the Application Interface, so as to create their domain-specific instances.
- At the top right, a number of European-wide decision makers are shown, which range from national agencies to government departments and pan-European (or worldwide) policy making bodies, e.g. G20.
- Accessing the system via the enabling services are the users of the system: from right to left, climate / science experts, climate groups such as UNFCCC / Kyoto, lay users on the internet with an interest in the domain's subject matter, and education.

3 - Methodology

3.1 - Organisation and team

The organisation is described in the diagram below.



3.2 - Modularity, genericity and specificity

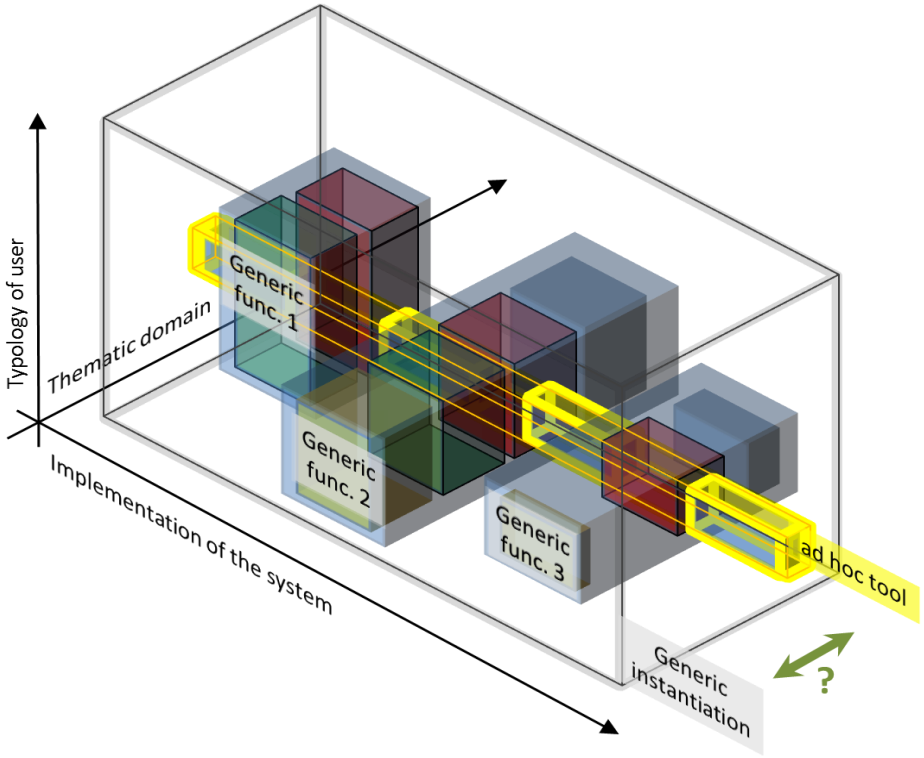
The system will require - as any system - a modular approach. But the eternal question is how to find the best compromise between genericity and specificity. The project is trying to address different domains (Climate change initiative, Earth Observation, Robotics...) and different types of users. They may share common needs, and probably some common tools. The project will define the common needs (or not) and the gap between their needs and the available tools. It could lead to the definition of some generic functions, in order to reuse them in each deployment of the system when these functions are useful.

However, we have to consider the main differences between an instantiation of the system, based on generic functions and an *ad hoc* tool (these aspects are summarized in the figure below).
The system instantiation is composed of generic function, it means that :

- It has more functionalities than required (illustrated by different coloured blocks in the figure);
- It has a higher complexity than necessary, with unused parameters, more interfaces;
- Setup time and computation time might be longer than necessary;
- Cost of development are reduced by the use of existing generic function;
- The reuse of existing function (components) give more insurance on the quality of the results;
- The initial developments are very costly.

The *ad hoc* tool is specifically developed, it means that

- It covers only the requirements;
- Cost of development might be high;
- The needs of validation are high.



Genericity versus specificity. A system built from generic functions offers more functionalities than required whereas an ad hoc tool offers only what is necessary.

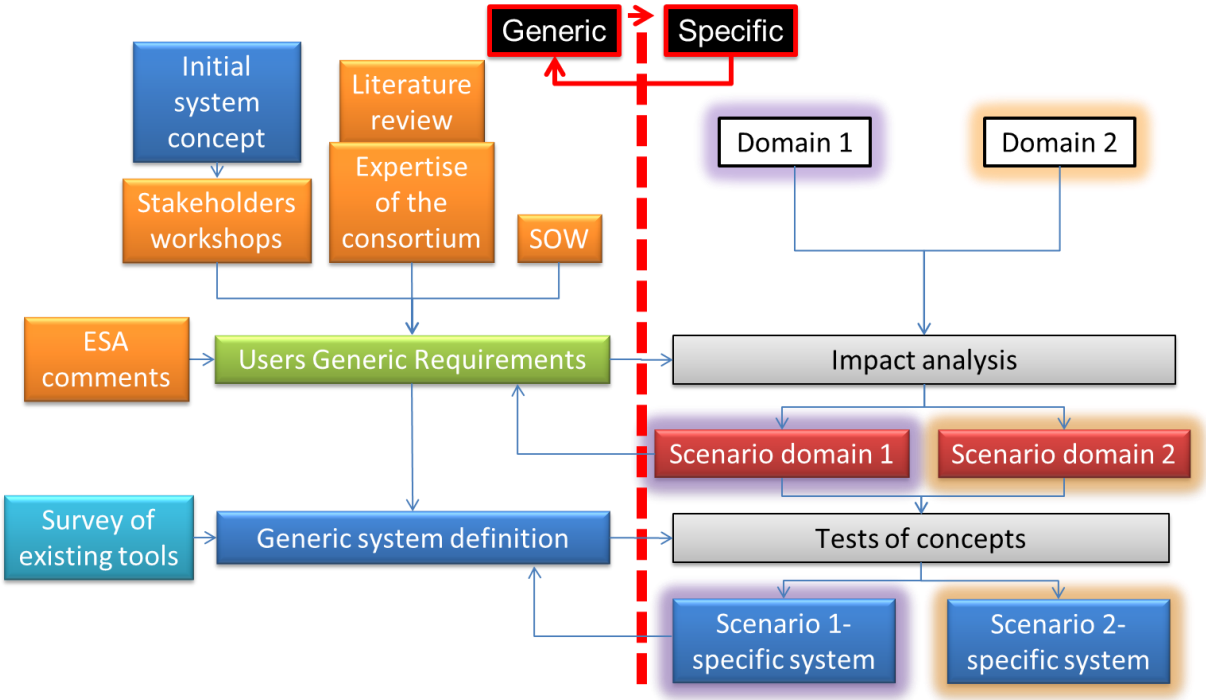
It does not make sense to think about developing a reference architecture for all domains and types of users. But even in a specific domain and for a specific type of user, it is not obvious that a generic architecture would be a solution. As concluded in (RD01) “D1.1 Strategic Needs Analysis Report”, It is more reasonable to think about defining common concepts and recommendations. But one future implementation of a such system could be as generic as possible for the target domain and the target type of user.

We are looking for a cross-domain approach and a generic system definition but not a generic solution.

3.3 - Methodology of definition

The diagram below illustrates the methodology:

- An initial system concept of IVA4D has been presented during workshops as a support for discussion
- The requirements have been collected, coming from
  - the workshops;
  - literature review;
  - expertise of the consortium;
  - Statement of Work.
- They have been improved thanks to:
  - their analysis through scenarios in various domains
  - ESA comments
- The generic requirements and the survey of existing tools lead to a generic system definition
- The generic system has been evaluated through scenarios and demonstrations.



The scenarios are detailed in [RD03]. Each **scenario is too much specific to a domain** or an application into that domain **to be used** as an entry point **for any definition**. Therefore, scenarios have been used as test cases and illustrations. They have been **used**:

- **to test the generic requirements**
  - The requirements are applied and assessed in scenarios
- **to test the generic definition of the infrastructure and the services**
  - The infrastructure and the services are conceptually implemented in scenarios
- **to test the generic definition of the system**.
  - Scenarios are used to give a proof of feasibility and to illustrate the benefits of some system concepts

3.4 - Challenges

During the project we identified a list of challenges for an IVA4D system:

- Ch1: Data representation, summarization
  - In which space do we find the real information, what is the appropriate projection for multidimensional data?
  - How to summarize data but avoid loss of important information (interesting anomalies)?
- Ch2: Dealing with uncertainty
  - What is the proper representation of uncertainty in a domain, specially for a non-scientist or someone who doesn't know the domain?
- Ch3: Correlation and dependence analysis
  - How to analyse the correlation between different kind of data, considering that they may not be correctly co-registered?
- Ch4: Collaborative working
  - How to share results and the path that lead to these results?
- Ch5: Interoperability
  - How combine different data, language and protocol?
- Ch6: User appropriation
  - How to find the proper tools to solve my problem?
- Ch7: Repeatability
  - How to trace efficiently and allow to reproduce a work?
- Ch8: Remotely-held data
  - How to solve the performance and security issues with remote data?
- Ch9: Large data visualisation
  - How to perform a visual analysis of large data?
  - How to visualize large data with a limited memory or/and bandwidth?
- Ch10: New modality of interaction and visualisation
  - What are the benefits of videowall, 3D screen, head tracking...?

Even if we are able to define an IVA4D system we need to prove that a solution exists or at least its feasibility and its benefits. This proof of concept must be focused on the challenges. The demonstrations need to show that the system definition is suitable to solve these challenges.

4 - Summary of the system definition

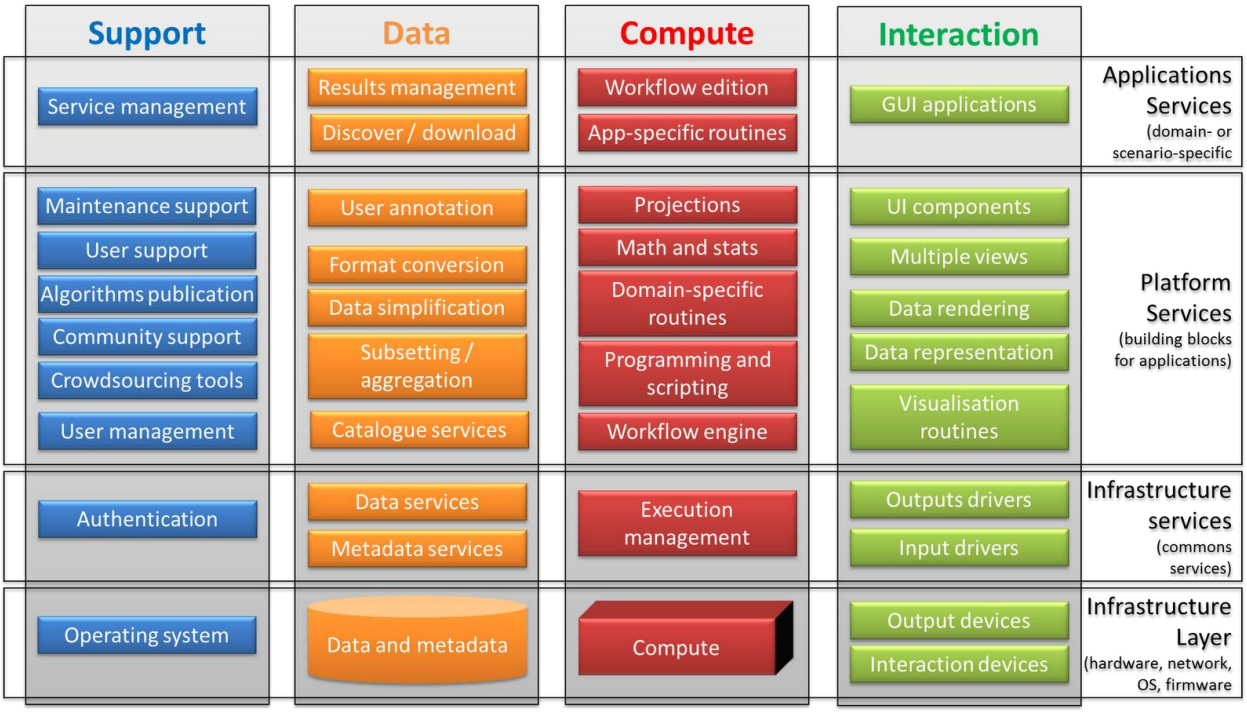
4.1 - Main features

The essential features of IVA4D System:

- IVA4D defines an infrastructure-based solution. This provides:
  - Capacity for handling large datasets
  - A platform for sharing data, results, algorithms and hardware
  - Automation of routine tasks
  - Support for users in various ways
- The infrastructure can be specialized for different domains
- Different modes of interaction and different display techniques are considered
  - From web-based interaction to specialist technology (e.g. head tracking)
  - From mobile devices to desktops to videowalls
- The system may be distributed:
  - Remote data, remote displays

4.2 - Core functions

The Core functions of a generic IVA4D system have been identified in the diagram below.



Overview of generic system architecture, showing the main system components and services. In many cases there will be several services within each category/box.

Note that these diagrams are purely logical in nature. they do not attempt to specify *how* the services are implemented (they could be Web Services, programming libraries or another type of module), or *where* they should be implemented (everything could be installed on the same computer, or the system could be distributed). Such concerns of implementation and deployment will be particular to a specific System instance, and will depend strongly on the particular user requirements within a domain, together with practical concerns.

The system is divided into four layers:

- The **infrastructure layer** comprises the hardware, including data storage, computing facilities, networking hardware and devices that are used for user interaction (keyboard, mouse, head tracking, display facilities).
- The **infrastructure services layer** provides a level of abstraction above the hardware, providing a set of low-level services that are common to all (or most) scenarios.
- The **platform services layer** provides a set of reusable building-blocks or modules, which can be composed to develop particular applications. Services in this layer may range from the highly generic (e.g. mathematical and statistical functions) to services that are designed for particular scenarios (e.g. feature tracking).
- The **application services layer** consists of particular, domain-specific applications that address particular user needs through a specific user interface.

The structure mirrors the definitions used in cloud computing: the upper three layers can be thought of as representing the three cloud services levels: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).

The services within the four architectural layers are further divided among four, broadly-defined vertical “functional categories”.

- **Support** services provide support in using the system and its services.
- **Data** services focus on managing, extracting, transforming and viewing data (and associated metadata). Little or no “processing” of the data is performed.
- **Compute** services focus on composing, managing and executing data processing tasks.
- **Interaction** services focus on capturing user interaction and displaying results.

These categories are illustrative only; particular services may straddle more than one category (for example, some computation is required for data format conversion). In particular, visualization services can require tight coupling between computation and interaction; they are therefore not easy to place neatly into one category or the other.

It is helpful to consider that the above structure is directly analogous to the architecture of a typical computer. A computer has data storage, processing and interaction capabilities, and its software is divided between hardware control and device drivers (analogous to the above "infrastructure services"), reusable libraries (analogous to "platform services") and particular applications.

See [RD01] for a full description of the system.

4.3 - Use cases categories

We started with three reasons for visualization

- **"I know the answer and I want to explain it to you"**
- **"I have a question and I want to find the answer"**
- **"I don't know the question I need to ask"**

These are Bergeron's (1993) three uses: presentation, analytical and discovery

We also considered a fourth reason:

- "**I want to collaborate with others to understand my data"**

We will look at how IVA4D helps with these cases through our scenarios, examining in each case:

- What are the current practices?
- What could be possible with an IVA4D implementation?

5 - Presentation use cases

5.1 - Description

This use case is: **"I know the answer and I want to explain it to you"**

The purpose is presenting scientific results to policymakers, the public or other scientists. The general problem is to produce an honest summary of highly nuanced data to support a conclusion. The interactivity is usually limited (or absent).

The typical problems faced are:

- What is the best visualization technique to use?
- What will make the audience trust my results?
- How can others reproduce my results and verify them?

The relevant scenarios are (see [RD02] for a full description of the scenarios):

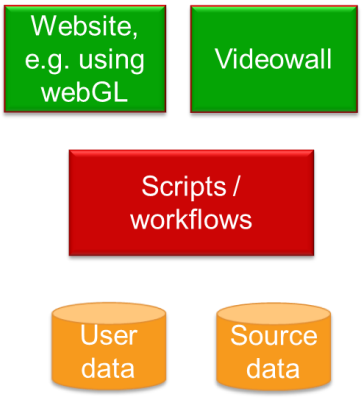
- C1 "Collaboration through sharing of visualizations and commentary"
- G1 "Sharing geographical information with a variety of audiences"

These are not a complete set of problems of course.

5.2 - Presenting using IVA4D

What could do a user to present data with IVA4D? This section briefly illustrates the main components of an IVA4D system (in the red, green and yellow boxes) that are "active" in the use cases in question. These are simplifications of course, and in fact other services and components will play important roles.

- Source data are stored and curated in IVA4D system
  - Or another system accessible to it, e.g. a data centre
- Scientist searches on IVA4D to find published examples of visualizations similar to the one she wishes to produce
  - Perhaps consults experts and peers via the user support system
- He writes (or adapts) a script or workflow to generate the required visualization from the source data



    - High-level functions are available to make this easier
- The system automatically records all steps for traceability (source data and script/workflow)
- He publishes the visualization as a web page with a permanent URL. This can be printed out as a report.
- He links any relevant citations in the literature (and informal online discussions) to the visualization

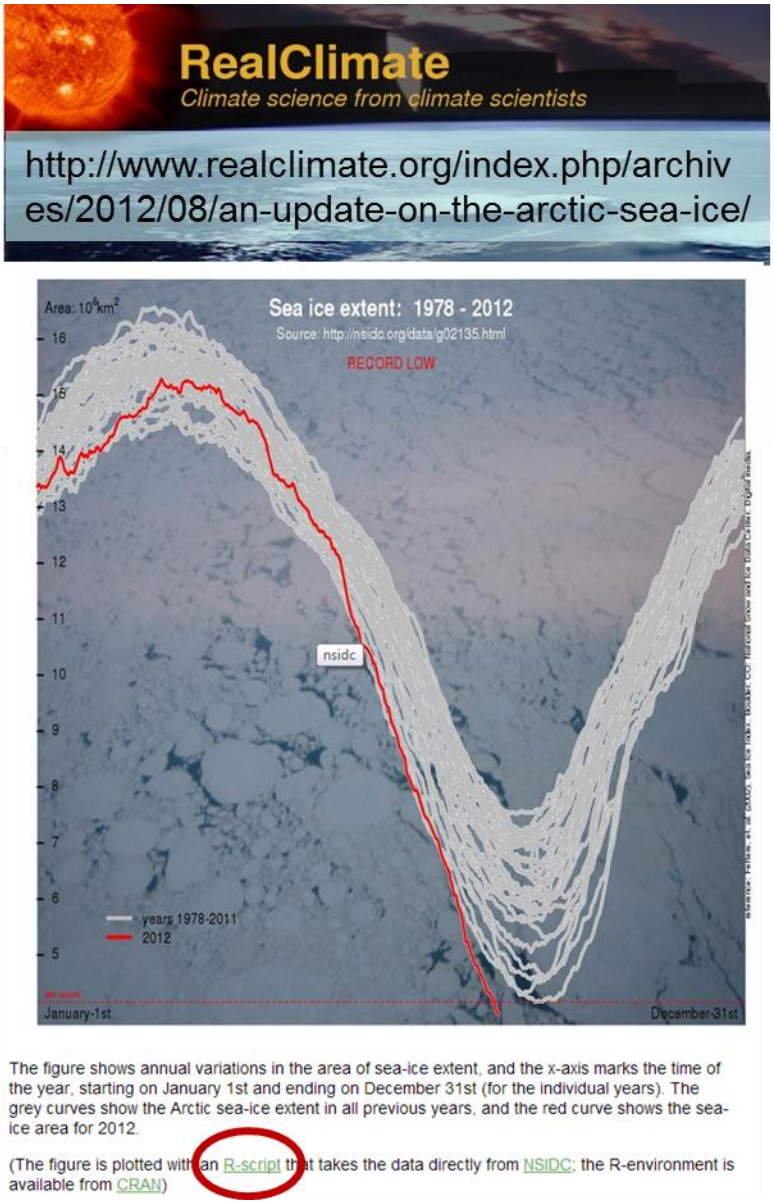
There is an important distinction here between the “source data” (i.e. the shared data pre-installed on the system) and the “user data”, which is the data created by users (visualizations, annotations, derived datasets and other metadata).
Different levels of presentation are possible:

- Static visualizations and reports
- Web pages and “linked data”
- Interactive visualizations based on mostly pre-rendered data but with some limited opportunity to customize and interact (pan, zoom, simple actions). Avoids dead-end images. “Full” interactivity may not usually be desirable if it complicates the message that is being communicated, or requires more complex infrastructure (e.g. large number of client-server interactions).

Demo notes: We can possibly use WebGL and Virtual GL for Communication of research/findings. We have identified that WebGL has valuable potential as a 3D/4D data delivery mechanism for HTML5 enabled browsers. In this demo we show a web-based WebGL scene generated by Google as an example of the potential for this technology.

5.3 - Progress beyond state of the art

- Typical current practices:
    - Generate visualization in common format and insert into document (DOC, PPT)
    - Provide citations to the literature and (if you’re lucky) source data
    - Blogs and websites can provide more functionality (right) but information becomes scattered
- IVA4D advantages:
    - Visualizations are not “dead-ends” but can be adapted through modification of scripts/workflows or use WebGL for interactivity
    - Scientist is able to build upon accumulated expert knowledge
    - Trust is engendered through openness and traceability from data to result
    - Encompasses informal discussions (Scenario C1) and formal citations



Screenshot is from RealClimate and shows the linking of a visualization with the R script that was used to generate it (it works!) This is not common practice and there is no consistent way to do this (so, for example, it's not possible to search RealClimate for all the visualizations that have associated scripts). Similarly, the RealClimate blog supports only simple keywords for searching, and a richer and more precise vocabulary and metadata structure would allow data, visualizations, citations and scripts to be linked in a machine-readable way.

6 - Analytical use cases

6.1 - Description

This use case is: **"I have a question and I want to find the answer"**
In this use case, the processing algorithms have already been devised

The typical problems faced are:

- How can I process large amounts of data?
- How can I handle errors and uncertainties?
- How can others reproduce my results and verify them?
- Are my results directly comparable with those of other groups?
- How can I ensure consistency of analysis and presentation?

The relevant scenarios are (see [RD02] for a full description of the scenarios):

- C2: "Validation of MyOcean global ocean reanalyses"
- M1: "Computer aided detection for PET-CR imaging"
- E1: "Search of Genghis Khan's Tomb"

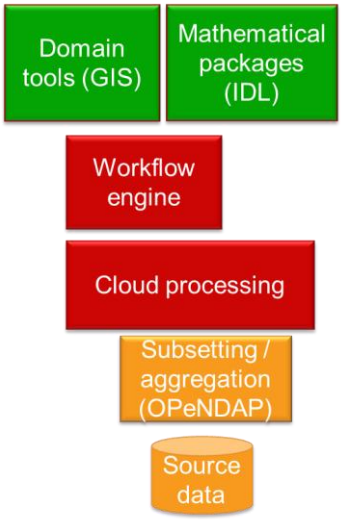
The key here is that using IVA4D analysis is not disconnected from visualization. Although standalone analysis is outside the scope of this visualization project, there are many kinds of analysis that are closely linked to the visualization process – e.g. in the MyOcean case the analysis results in standard visualisations that need to be comparable between different scientific groups. This requires following standard procedure and keeping a record of the process that has been executed. A workflow environment is a suitable tool for this kind of task.
(Simpler analytics – particularly those that can be performed quickly and interactively – are also relevant, but we consider these mainly under data exploration, because the user often does not have a fixed question in mind and there is no predetermined procedure to follow.)
Crowdsourcing is an interesting area in which the visualization almost *is* the analysis because it exploits the brain's capacity for visual processing in order to find a result that a computer could not easily find.

6.2 - Analysis using IVA4D

IVA4D provides a range of library routines at various levels to speed up generation of analytical processes
Two approaches/solutions are possible:

- Solution 1: analysis can be automated

    - Processing algorithms are implemented as workflows that link together the platform services:
        - Declarative description of what needs to be done, not how to do it
    - Executed using a workflow engine that automatically records inputs, intermediate results and outputs
        - Engine handles parallelisation (if possible) and restarts of failed jobs
        - Automatically records provenance information
    - Workflows can be shared among groups to ensure consistency
- Solution 2: analysis requires human intervention
    - Crowdsourcing: participants are trained and then allowed to analyse pieces of the data each
    - Redundancy provides quality control



The following demos have been used to prove the concept and/or illustrate the benefits:

- Workflow
- Service deployment
- Crowd sourcing

Demo notes:
During the “Processing Service Deployment Demo”, we showed how a processing chain developed in IDL can then be readily deployed to a server for accessibility as a web service. The aim here is to show how the gap can be bridged between scientific innovation and research to operationally accessible server-side data processing.

6.3 - Progress beyond state of the art

- Typical current practices:
    - Many workflow tools exist (e.g. Taverna, see below)
    - Workflow-sharing solutions exist for particular communities (e.g. MyExperiment)
    - Cloud-based data processing systems are emerging (e.g. CEMS, JASMIN)
    - Crowdsourcing frameworks (e.g. Zooniverse) also exist
    - But systems not widely used in some domains (e.g. EO/Climate)
- IVA4D advantages:
    - Combines appropriate tooling with a data/compute platform
    - Combines data processing with visualization capability.
    - Automated capture promotes reproducibility

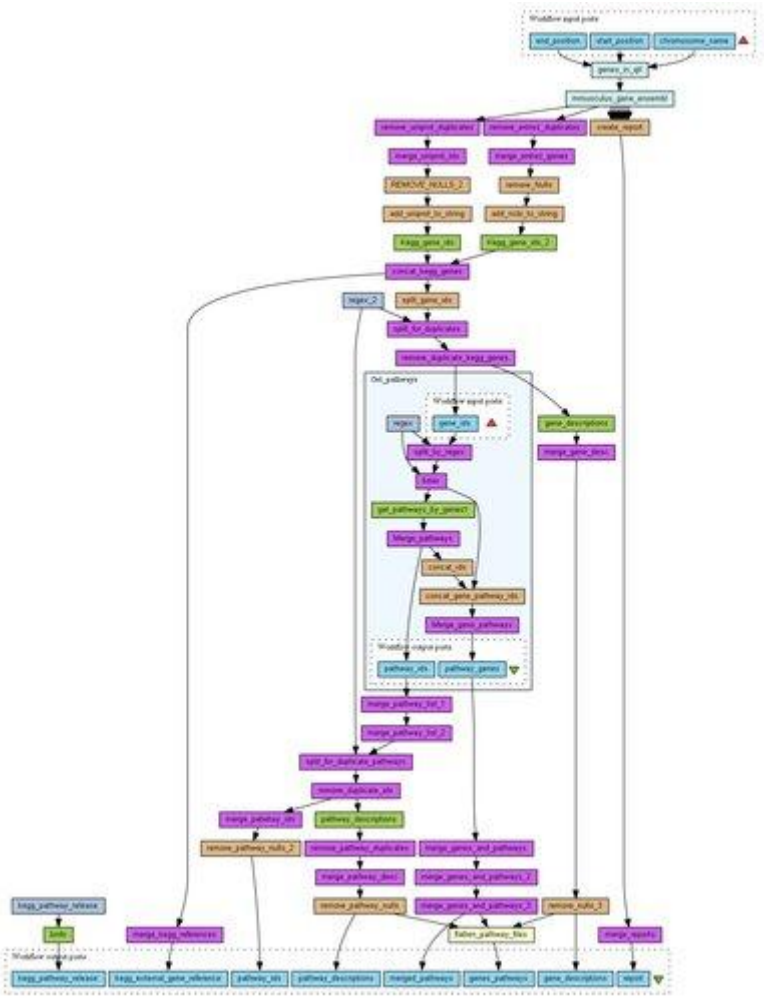


Image shows a bioinformatics workflow taken from MyExperiment, which is executed using the Taverna workflow engine. Many other workflow engines (Kepler, Triana, INGRID) are available. In early versions of MyExperiment, it was only possible to share the workflow definitions. Later versions added the useful capability to share “bundles”, which contain the workflow definition, plus other information such as input and output data.

Note: Taverna is a desktop-based workflow engine. In IVA4D we consider that the workflow engine could be installed on a shared processing resource, close to the data.

The NETMAR project is one of the few projects to exploit workflow technology in environmental data processing. Geospatial data processing functions are wrapped as OGC-compliant Web Processing Services and coordinated using Taverna. It is too early to draw conclusions about how popular the technique is with users.

In IVA4D we consider that processing can be achieved by workflows (data-driven, declarative programs) or scripts (command-driven, imperative programs). Both should be shareable and reusable. A shared execution environment helps reusability greatly, as it mitigates the problem that different users may have different environments (different libraries installed, different versions of programming languages etc).

(The description of the workflow is as follows: “This workflow searches for genes which reside in a QTL (Quantitative Trait Loci) region in the mouse, Mus musculus. The workflow requires an input of: a chromosome name or number; a QTL start base pair position; QTL end base pair position. Data is then extracted from BioMart to annotate each of the genes found in this region. The Entrez and UniProt identifiers are then sent to KEGG to obtain KEGG gene identifiers. The KEGG gene identifiers are then used to searcg for pathways in the KEGG pathway database.” http://www.myexperiment.org/workflows/16.html)

7 - Exploration use cases

7.1 - Description

This use case is: **"I don't know the question I need to ask"**
The user is faced with big data, potentially in an unfamiliar data format. The most appropriate analysis and visualizations may not yet be known.

The typical problems faced are:

- How can I explore large amounts of data?
- What is an appropriate way of summarising this data? What are the key features?

The relevant scenarios are (see [RD02] for a full description of the scenarios):

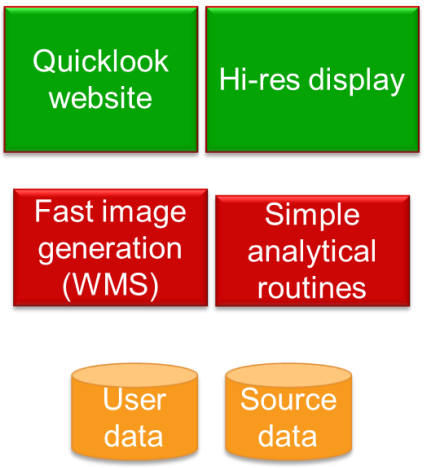
- P1 - Multi Variable Data Exploration of CERN CMS data in Particle Physics
- R1 - Collaborative work for data exploration in realistic 3D environment (Mars Rover)

Emphasis is more on speed, interactivity and intuitive user interfaces, rather than processing power. GIOVANNI is a very widely-used example of a system (for NASA Level 3 data) that permits subsetting, visualization and simple analysis whilst removing from the user the need to understand file formats (or file locations), all accessible through the Web.
Godiva2 enables a different kind of exploration on large gridded datasets. No analysis functions are present, but visualizations are more interactive, enabling the user to create maps, timeseries, vertical sections and other types of interactive plot without following a series of menus (which is the GIOVANNI approach). Note that Godiva2 is not a single installation, but open-source software (http://ncwms.sf.net) that is installed at multiple locations around the world for the exploration of large gridded datasets, held by data centres research institutes or private industry. It is now built into the popular THREDDS Data Server, which is used by met-ocean data providers worldwide (e.g. the MyOcean project).
Although both GIOVANNI and Godiva2 can explore extremely large datasets (both are applied to multi-terabyte archives and could scale much larger than this), the user is given access to the entire archive, without the need for the user to pre-select the data they are interested in. Relevant subsets of data are extracted, processed and visualized on demand – this contrasts with other visualization tools, which do not scale up to handling large datasets and require pre-processing and reduction of large datasets to maintain interactivity; this forces the user to make decisions in advance that may restrict the data exploration.

7.2 - Exploration using IVA4D

What could do a user to present data with IVA4D?

- Rapid exploration of large datasets implies fast data access (potentially to multiple datasets), rough-and-ready visualizations, and a degree of insulation from the technical details of the data and visualization process.
- Quicklooks with library code for simple analytical functions
- Searchable archives of other users' visualizations
- Gallery / "small multiples" demonstration shows the use of both small and large screens, exploration of big data and parameter space, treatment of uncertainties etc.



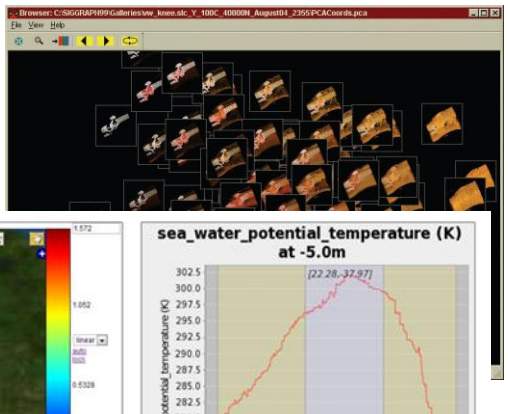
We are showing a video of a gallery application for the ISIC Video Wall. The aim is to show how this medium can be used for collaborative analysis when the scientists are co-located around the wall. It's worth noting that the data being presented is "live" (not Powerpoint or video presentation) and therefore is suitable for interactive discussion and exploration. From a technology point-of-view we are showing use of a cluster of machines for coordinated visualisation of data that could potentially not be shown on a single system.

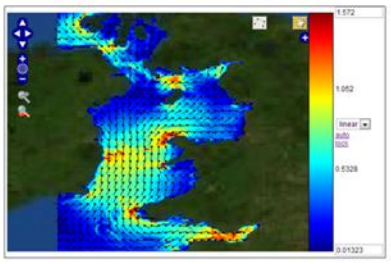
The following demos have been used to prove the concept and/or illustrate the benefits:

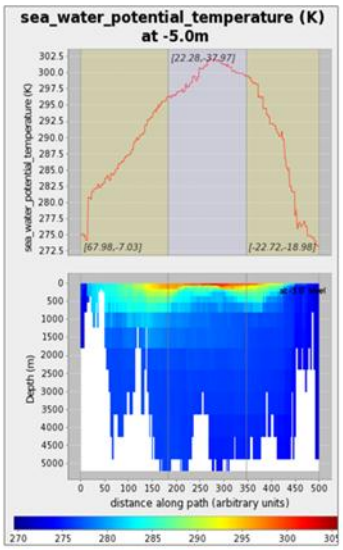
- Ad-hoc Clustered Gallery
- Gallery transparency/threshold
- Mars Rover

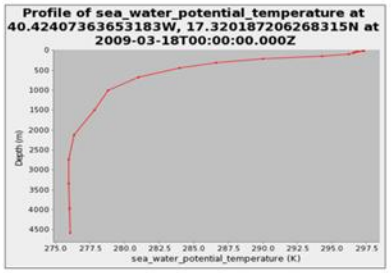
7.3 - Progress beyond state of the art

- Typical current practices:
  - Much time is spent understanding new data formats rather than exploring the data itself
  - Although an experienced user can identify the variables and find the event, it is difficult to look at all the variables in "one go" and identify the best correlated variables.
  - Most of the required techniques already









exist, but not all in one tool or environment
    - Only existing visualization tools are trusted (e.g. ROOT) – difficult to get new tools used

- IVA4D advantages:
    - Quicklook system provides fast route to visualization
    - A scientist can interactively filter the data in the visualisation process by removing the less interesting data.
    - Many options for visualization can be displayed at once thanks to processing resources
    - New visualization tools are integrated into one environment, or an existing environment

8 - Collaboration use cases

8.1 - Description

This use case is: **"I want to collaborate with others to understand my data"**
Potential collaborators may be in the same room, on the other side of the world, or may be viewing the data/visualizations at a different time. They are links to all of the previous types of visualization, depending on the stage at which a user wishes to collaborate.

The typical problems faced are:

- How can I share my results in a way that different people will understand?
- How can collaborators interact with a visualization?
- Ensuring only those with the relevant permissions can access data

The relevant scenarios are (see [RD02] for a full description of the scenarios):

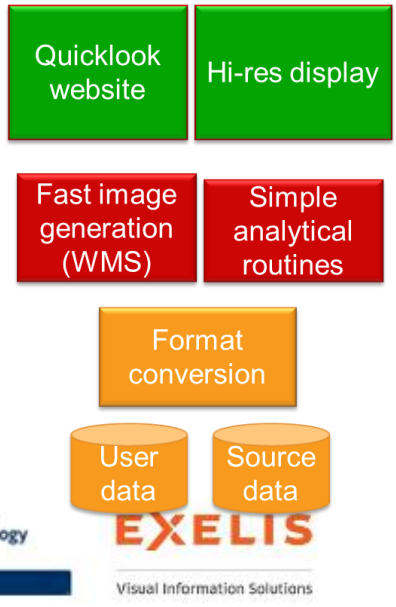
- C1 - Collaboration through traceable sharing of visualizations and commentary (collaborators are elsewhere)
- C2 - Validation of MyOcean global ocean reanalyses (collaborators are in the room)

Collaboration is an important aspect of this project and – because it is a young research area - a strong opportunity to develop scientific practices beyond the current state of the art. Collaboration techniques can share many features in common with exploration techniques, requiring low (apparent) technical complexity, fast routes to visualization, interactivity and low barriers to entry.

8.2 - Collaboration using IVA4D

What could do a user to present data with IVA4D?

- Interactive collaboration requires common environments and data formats.
  - Tools that create visualizations but retain the link to the data behind them
  - Privacy settings also become important when sharing work prior to publication.



Two possible approaches/solutions

- Solution 1: contemporaneous collaboration
  - Use large display surface to bring together multiple datasets
  - Or examine different aspects of same dataset

    - Gallery approach
- Solution 2: remote collaboration
    - Collaborative website/blog used to record visualizations and annotations
    - Entries can be discovered and commented upon, enabling conversations to develop
    - Appropriate permissions (users/groups) must be set

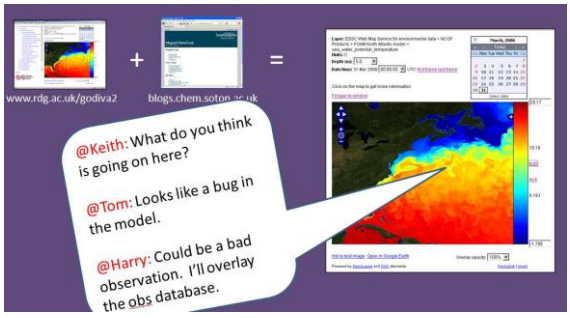
We showed the same application running on a small cluster (4 laptops). This is to show that the principles used for the videowall demo are also applicable in an ordinary network environment and shows the flexibility of the concept tool for different display mediums. It should be noted that the goal here is not the application; rather that we can take an application developed in this way and readily deploy it to different hardware configurations.

The following demos have been used to prove the concept and/or illustrate the benefits:

- Videowall Gallery (co-located)
- Blog My Data (remote / temporally separate)

8.3 - Progress beyond state of the art

- Typical current practices:
    - Visualizations tend to be shared (either in person or remotely) using static images such as PowerPoint files
    - Code to produce visualizations is rarely shared, or written in a way that makes it easily re-used by others
- IVA4D advantages:
    - Consistent visualizations (e.g. domain, colour scale, time period) can be produced across collaborators
    - Visualizations are linked to the source data –not ‘dead-ends’
    - Comments can be captured and searched
    - A mechanism for sharing code



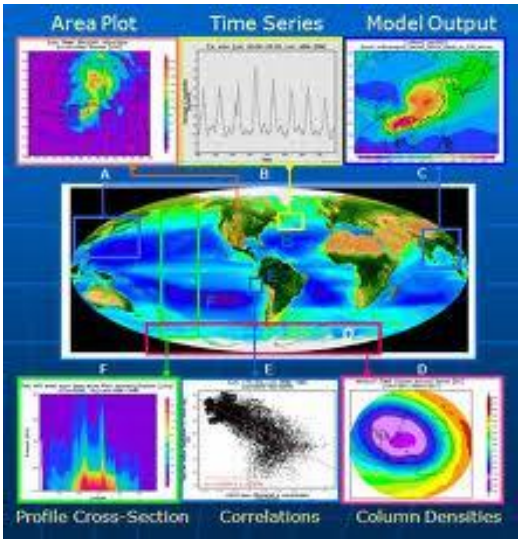
9 - Comparison with existing systems

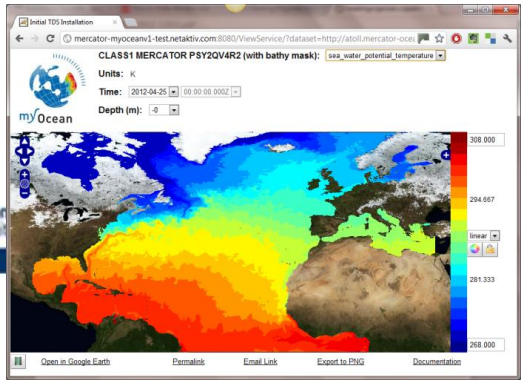
9.1 - General observations

- Common approach is to develop a set of domain-specific software libraries
  - E.g. Virtual Physiological Human, Iris (Met Office) – see “Analysis of existing tools” deliverable
- Data infrastructures often do not include specific visualization components
  - E.g. Data Grids, generic Clouds, most data centres
- Methods for results-sharing and collaboration are emerging, but often domain-specific and not strongly linked to source data or scientific process
  - E.g. generic blogs, MyExperiment (bioinformatics), FigShare, LabTrove (chemistry)
- Very few scientific communities receive formal visualization training, and support mechanisms are generally lacking

9.2 - Specific to Earth observation / Climate science

- **NASA GIOVANNI** (right) provides quicklook visualization and simple analysis through a web interface
  - But little interactivity and limited to NASA Level 3 data
- **CEMS** provides a cloud processing system, user management and support
  - But no specific visualization tools
- **MyOcean** (GMES Marine Core Service, right) provides Discover-View-Download functionality with built-in Quicklook service
  - Uses Godiva2 online visualization system and Web Map Services
  - Provides interface to GIS systems
  - But no data processing facilities
- **GEOSS** (Global Earth Observation System of Systems) provides Discover-View-Download for very large numbers of environmental datasets, with a Geoportal
  - Visualization must largely be done outside the system





- **Google Earth ecosystem** (including desktop client, Enterprise data servers, Engine for EO data processing) provides scalability and ease of use for processing and simple GIS-style visualization
    - Data must be converted to specific formats for visualization
    - Analytical functions very limited
    - Visualizations usually pre-generated and cannot be edited
    - "Gallery" functions very hard to implement
- **NASA Earth Exchange** (NEX) provides means for sharing resources within projects and externally (very new)
    - Unidata RAMADDA provides another kind of shared content repository, with support for ontologies and Linked Data
    - Neither has focus on visualization of resources
    - In future RAMADDA may include quicklook functionality for user datasets
- **Greenland** (52North) is a sophisticated web client for exploring geographic data (raster and vector), supporting uncertainty exploration (cf. UncertWeb)
    - Simple processing only

9.3 - IVA4D-CCI

An "IVA4D-CCI" system could/should:

- Consider spectrum of users:
    - Developers of ECVs (scientists), Users of ECVs (scientists, public, decision-makers…)
- Store data in a consistent format (CF-NetCDF) and model uncertainties accurately (e.g. UncertWeb, NetCDF-U)
- Provide Web Map Service interfaces (cf. MyOcean) to the data to provide:
    - Fast route to interactive visualization (including uncertainty visualization, cf. GeoViQua project)
    - compatibility with multiple clients (Geographic Information Systems, Videowalls, Web-GIS and quicklook systems)
- Support analysis in various forms:
    - In-place (e.g. CEMS)
    - Quick analysis in web browser (e.g. GIOVANNI)
    - In user's system (Matlab, IDL, Python) via OPeNDAP interface (e.g. Earth System Grid)
- Support collaboration and exchange of user data
    - Also provides feedback to ECV and EO data producers
    - E.g. BlogMyData, CHARMe
- Support data exploration through gallery techniques (exploring options, data intercomparison) on large screens (e.g. videowall)
- Support for crowdsourcing tools could be useful for certain types of problem
    - E.g. Cloud detection in EO data

10 - Conclusions

In summary:

- An infrastructure-based solution allows the consolidation and preservation of information, reduces duplication of effort, provides the capacity to enable “big data” discoveries and increases the quality and trustworthiness of science.
- Range of visualization methods supported, from simple quicklooks to precise visual analysis to high-performance graphics
- Such a system would actually become more useful with time, as the knowledge that is shared becomes part of the value of the system.
    - “A likely development path within a community is:
    - from offline visualization,
    - to interactive visualization,
    - to information-assisted visualization,
    - to knowledge-assisted visualization.” (Chen et al 2009)
- The elements of such a system largely exist (frequently as open-source software solutions) but the value and the innovation lies in bringing them together as a common platform

Demos showed how the same ‘back-end’ can deliver different-looking services to different users –flexible, but one infrastructure/investment.
Where would the confidence/uncertainty demo sit? It’s relevant to all of them.