Scientific Data Preservation 2014

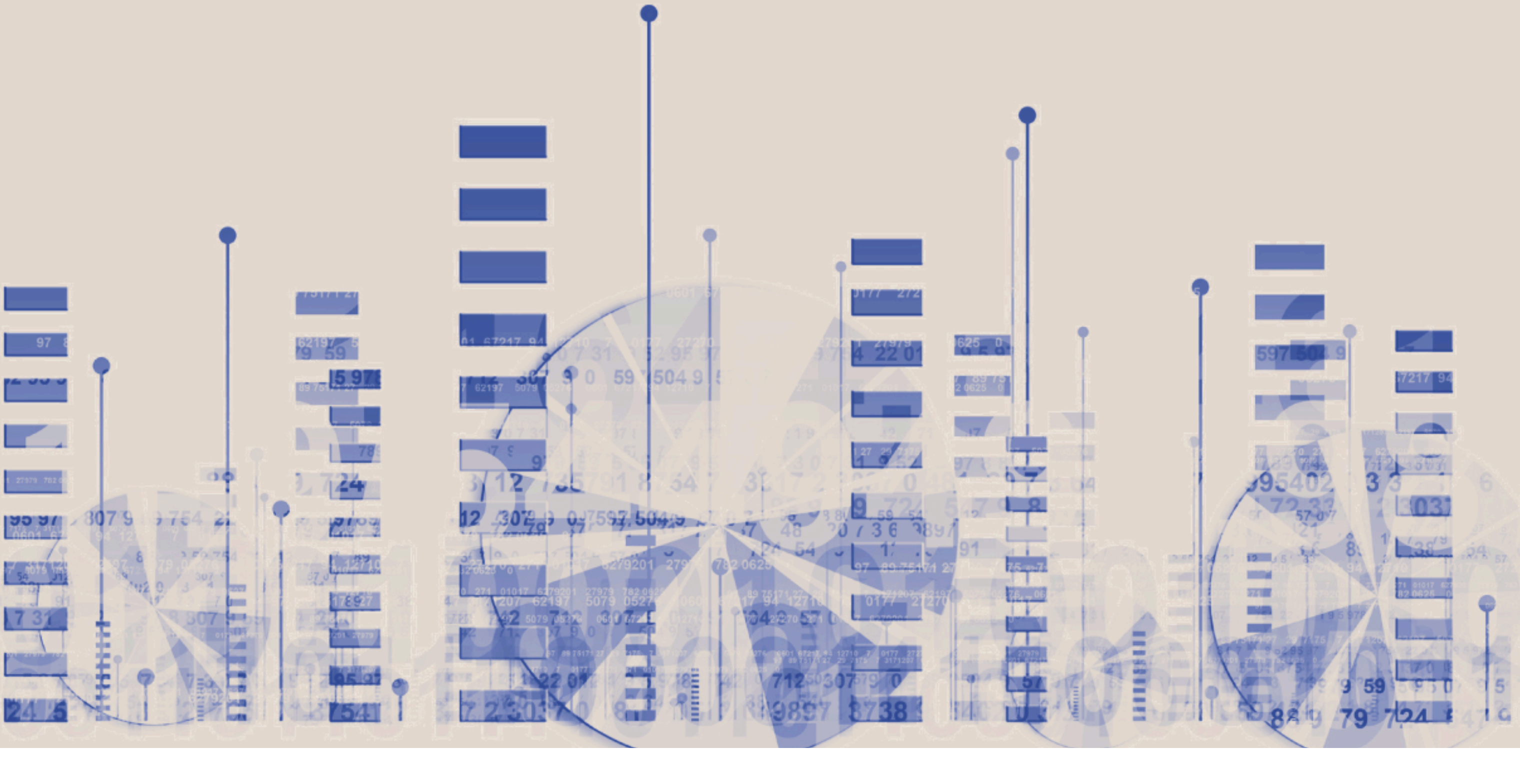
MASTODONS DATA

SECURITY SCIENTIFIC

PREDON BIG DATA WORKFLOWS DATA MINING

ARCHIVE PRESERVATION PERSISTENCY SCIENTIFIC

PERSISTENCY SCIENTIFIC DATA MINING BIG DATA



Summary

Introduction

Scientific data collected with modern sensors or dedicated detectors exceed very often the perimeter of the initial scientific design. These data is obtained more and more frequently with large material and human efforts. A large class of scientific experiments are in fact unique because of their large scale, with very small chances to be repeated and to superseded by new experiments in the same domain: for instance high energy physics and astrophysics experiments involve multi-annual developments and a simple duplication of efforts in order to reproduce old data is simply not affordable. Other scientific experiments are in fact unique by nature: earth science, medical sciences etc. since the collected data is "time-stamped" and thereby non-reproducible by new experiments or observations. In addition, scientific data collection increased dramatically in the recent years, participating to the so-called "data deluge" and inviting for common reflection in the context of "big data" investigations.

The new knowledge obtained using these data should be preserved long term such that the access and the re-use are made possible and lead to an enhancement of the initial investment. Data observatories, based on open access policies and coupled with multi-disciplinary techniques for indexing and mining may lead to truly new paradigms in science. It is therefore of outmost importance to pursue a coherent and vigorous approach to preserve the scientific data at long term. The preservation remains nevertheless a challenge due to the complexity of the data structure, the fragility of the custom-made software environments as well as the lack of rigorous approaches in workflows and algorithms.

To address this challenge, the PREDON project has been initiated in France in 2012 within the MASTODONS program - a Big Data scientific challenge, initiated and supported by the Interdisciplinary Mission of the National Centre for Scientific Research (CNRS). PREDON is a study group[1] formed by researchers from different disciplines and institutes. Several meetings and workshops lead to a rich exchange in ideas, paradigms and methods.

The present document includes contributions form the participants to the PREDON Study Group, as well as invited papers, related to the scientific case, methodology and technology. This document should be read as a "facts finding" resource pointing to a concrete and significant scientific interest for long term research data preservation, as well as to cutting edge methods and technologies to achieve this goal. A sustained, coherent and long term action in the area of scientific data preservation would be highly beneficial.

[1] https://martwiki.in2p3.fr/PREDON

Chapter 1: Scientific Case

Data Preservation in High Energy Physics

Cristinel Diaconu and Sabine Kraml

Abstract: The quest for matter intimate structure have required increasingly powerful experimental devices, stimulated by the experimental discoveries and technological advances. In most cases the next generation collider operates at a higher energy frontier or intensity than the previous one. With the increasing costs and complexity of the experimental installation, the produced data became unique and non-reproducible. In turn, the re-use of old data may lead to original results when new paradigms and hypothesis can be cross-checked against the previous experimental conditions. The data preservation in high energy physics appears to be a complicated though necessary task and an international effort is being developed since a several years.

Size and circumstances

At the end of the first decade of the 21st century, the focus in high energy physics (HEP) research is firmly on the Large Collider (LHC) at CERN, which operates mainly as a proton-proton (pp) collider, and currently at a centre–of–mass energy of up to 14 TeV. At the same time, a generation of other high-energy physics (HEP) experiments are concluding their data taking and winding up their physics programmes. These include H1 and ZEUS experiments at the world's only electron-proton (ep) collider HERA (data taking ended July 2007), BaBar at the PEP-II e^+e^- collider at SLAC (ended April 2008) and the Tevatron experiments DØ and CDF (ended September 2011). The Belle experiment also recently concluded data taking at the KEK e^+e^- collider, where upgrades are now on going until 2014.

These experiments and their host laboratories have supported the installation of an International Study Group on Data Preservation in High Energy Physics (DPHEP, http://dphep.org) . The situation has been summarised in the recent report [1] of the DPHEP Study Group. One of the main recommendations is to proceed to a complete preservation of the data, software and metadata, and to install an international organisation in charge with Data Preservation in HEP. Following this recommendation, the European Organisation for Nuclear Research CERN has appointed a Project Manager and now elaborates the collaboration agreements to be signed by the major HEP centres and funding agencies.

The size of each of the data sets of the DPHEP founding experiments vary from 1 to 10 Pb, with LHC data expected to reach several hundreds of Petabytes. Figure 1 displays only a few examples of High Energy Physics experiments taking data in the past few decades. It is clear that, similarly to other scientific fields and with the digital data overall, the increase in data sets has literally exploded in the last few years.

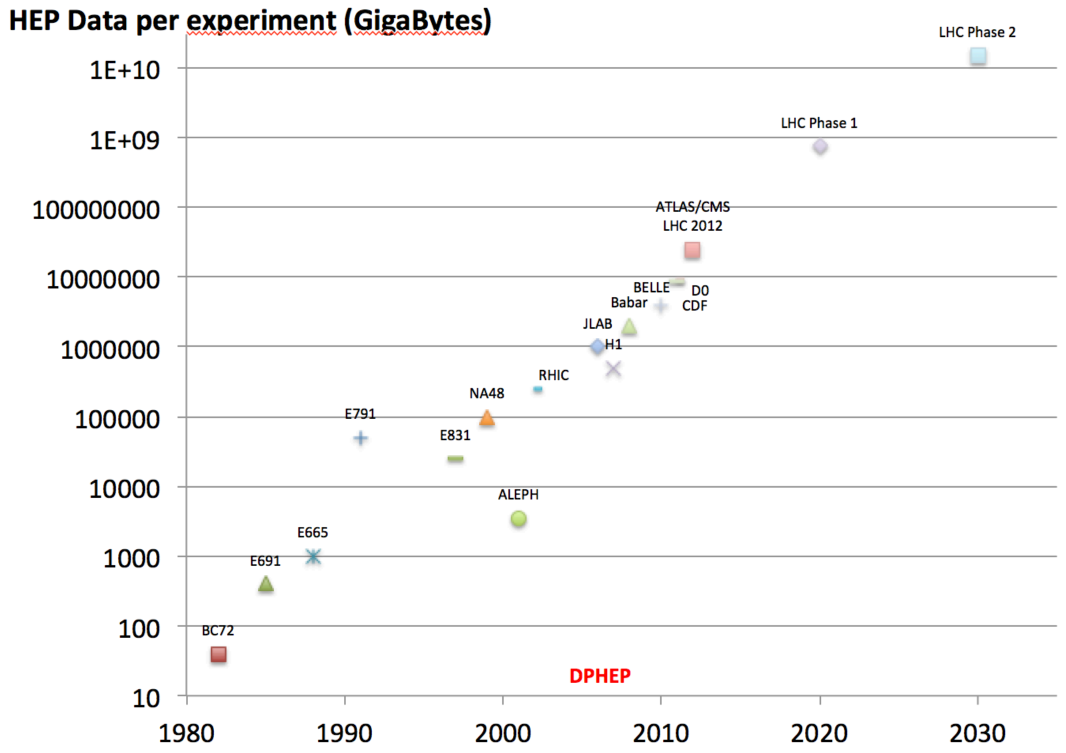


Figure 1: Data collected by a sample of HEP experiments in the last decades.

The high-energy physics (HEP) data is structured on several complexity levels (from “raw” to “ntuples”) where superior data sets are smaller and are obtained via large campaign of processing which may last up to a few months. The preservation should include not only the data itself, but also the associated software, which may include systems of several lines of code custom made and specific to the collaborations that have built and ran the respective detectors. In addition, external dependencies and a massive, complex and rather unstructured amount of meta-information is essential to the understanding of the ”collision” data.

The experimental data from these experiments still has much to tell us from the on going analyses that remain to be completed, but it may also contain things we do not yet know about. The scientific value of long-term analysis was examined in a recent survey by the PARSE-Insight project (PARSE-Insight FP7 Project: http://www.parse-insight.eu), where around 70% of over a thousand HEP physicists regarded data preservation as very important or even crucial, as shown in figure 2. Moreover, the data from in particular the HERA and Tevatron experiments are unique in terms of the initial state particles and are unlikely to be superseded anytime soon.

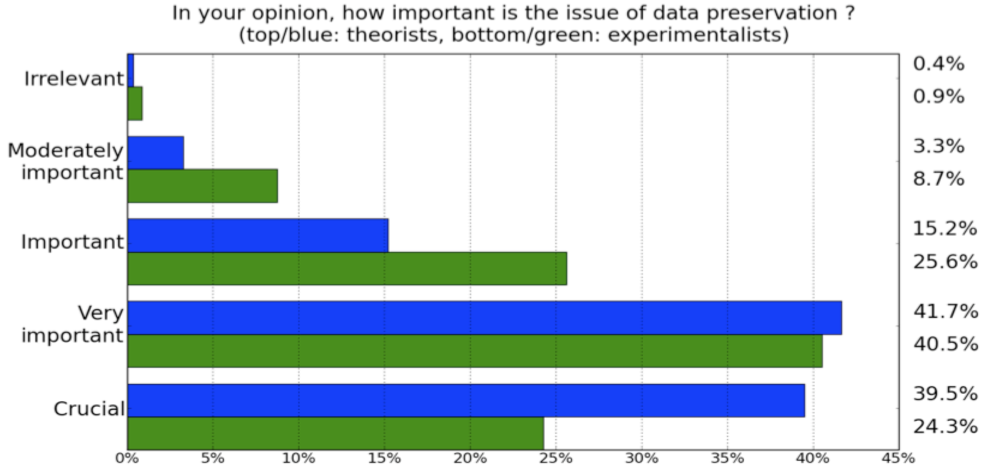


Figure 2: One of the results of the PARSE-Insight survey of particle physicists on the subject of data preservation. The opinions of theorists and experimentalists are displayed separately.

It would therefore be prudent for such experiments to envisage some form of conservation of their respective data sets. However, HEP had until recently little or no tradition or clear current model of long-term preservation of data.

Costs, benefits and technical solutions

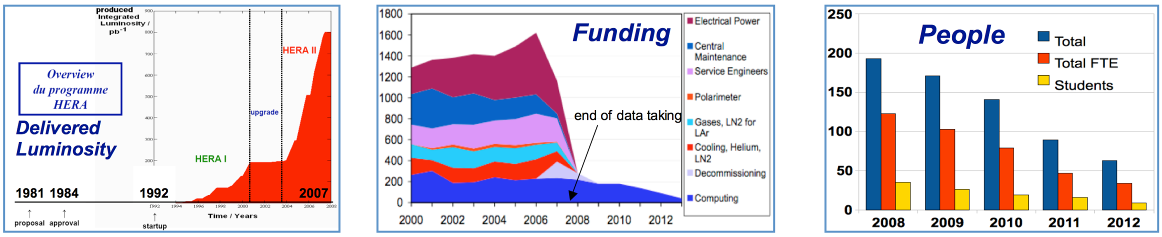


Figure 3: Illustrative luminosity profile (left), funding (centre) and person-power (right) resources available to a high-energy physics experiment.

The preservation of and supported long term access to the data is generally not part of the planning, software design or budget of a HEP experiment. This results in a lack of available key resources just as they are needed, as illustrated in figure 3. Accelerators typically deliver the best data towards the end of data taking, which can be seen in the left figure example for the HERA accelerator. However, as the centre and right figures show, this contrasts with the reduction in overall funding and available person-power. Any attempts to allocate already limited resources to data preservation at this point have most often proven to be unsuccessful.

For the few known preserved HEP data examples, in general the exercise has not been a planned initiative by the collaboration but a push by knowledgeable people, usually at a later date. The distribution of the data complicates the task, with potential headaches arising from ageing hardware where the data themselves are stored, as well as from unmaintained and out-dated software, which tends to be under the control of the (defunct) experiments rather than the associated HEP computing centres. Indeed past attempts of data preservation by the LEP experiments, of SLD data at SLAC and of JADE data from the PETRA collider at DESY have had mixed results, where technical and practical difficulties have not always been insurmountable. It appears therefore as mandatory that a consolidated data management plan including long term data preservation be presented in the very initial stages of an experiment proposal and adopted (for the running experiments) as soon as possible and well before the end of the data taking.

Technologies and organisation

Due to a solid and pioneering practice with large data sets, the HEP data centres have the necessary technology and skills to preserve large amounts of data (up to few Pb and beyond). This has been indeed proven over several decades in several laboratories. With the advent of the grid computing techniques, the management of large data sets became even more common. However, it is by now full recognized that data preservation in HEP have several components, in increasing order of complexity:

- *Bits preservation:* the reliable conservation of digital files need a rigorous organisation but it is otherwise manageable in the computing centres. It is nevertheless understood that a moderate amount of development is needed to increase the reliability and the cost-effectiveness of the digital preservation for large data sets.
- *Documentation:* a rigorous documentation policy has not always been adopted and large efforts to recover essential documents (thesis, technical notes, drawings etc.) had to be pursued at the end of some experiments. While the scientific papers are well preserved, more information around the publications, including high level data sets used for the final stages may be useful to maintain the long term scientific ability of the preserved data sets. New services and user functionalities are provided by INSPIRE (http://www.projecthepinspire.net/).
- *Software preservation*: HEP experiments rely on large software systems (few millions of lines of code and distributed systems over as much as several thousand of cores), used to reconstruct, calibrate and reduce the "raw" data towards final scientific results. The preservation of these systems implies that the functionality is not broken by technological steps (changes in core technology, migration to new operating systems or middleware, outdating of external libraries etc.). The problem is quite complicated and has lead to innovative solution, combining the so called "freezing" approach, based on the conservation of a running environment using virtual machines, to the "full migration" approach, where the framework is prepared for a continuous migration with prompt correction of issues "as-you-go", minimizing therefore the risk of a major glitch. The software preservation is strongly linked to automatic validation systems, for which an innovative, multi-experiment solution has been proposed [3] and is illustrated in figure 4.
- *Community knowledge*: The lively scientific exchanges of several-hundred scientific communities familiar with a given HEP data sets lead to a community wide

knowledge. This information is sometimes not fully captured in the standard documentation, but can emerge from the electronic communications (for example hypernews, emails etc.).

- *Organisation and long term supervision:* The technical systems installed to preserve the computing systems cannot be fully effective in absence of the necessary scientific feedback of the experts that participated to the real data taking. In addition, the supervision of various systems (like the validation illustrated in figure 4) require human and expert action. It is therefore mandatory that large collaboration of have a specific organisation for the long term period, adapted to a less intensive common scientific life but sufficiently structured in order to cope with all scientific and technological issues that may arise around the preserved data.

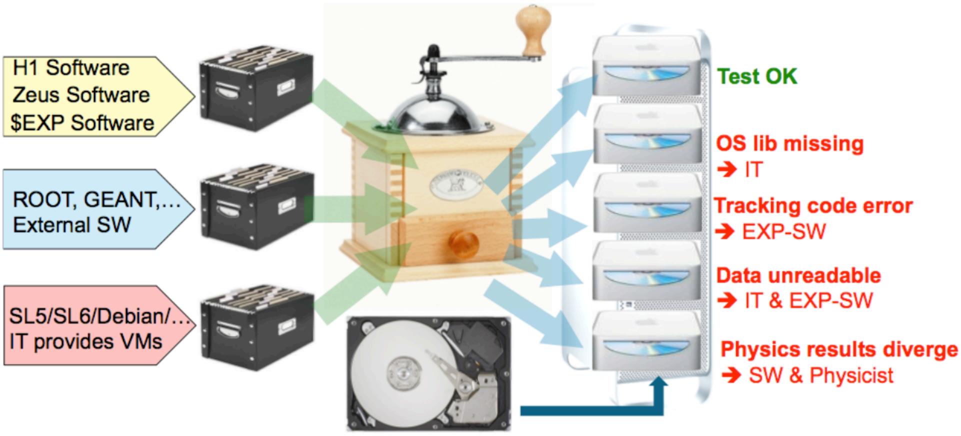


Figure 4: Scheme of the software validation system studied at DESY for HERA experiments (from [3]).

The preservation of a complex computing system as the ones used in HEP is therefore technologically challenging and includes many different aspects of the scientific collaborative work at an unprecedented level of complexity.

Data preservation and the open access

The preservation of HEP data may well be done using the appropriate infrastructure and tools which are necessary to give a broad , open access to data. Common standards and tools for reliable data preservation frameworks may well be coupled with a global analysis and interpretation tools. The data preservation and the open access are therefore intimately related.

In a recent document [2] a set of recommendations for the presentation of LHC results on searches for new physics, which are aimed at providing a more efficient flow of scientific information and at facilitating the interpretation of the results in wide classes of models. The target of these recommendations are physicists (experimentalists and theorists alike) both within and outside the LHC experiments, interested in the best exploitation of the BSM search analyses. The tools needed to provide, archive and interpret extended experimental information will require dedicated efforts by both the experimental and the theory communities. Concretely, the actions to be taken in the near future are:

- Develop a public analysis database that can collect all relevant information, like cuts, object definitions, efficiencies, etc (including well-encapsulated functions), necessary to reproduce or use the results of the LHC analyses.
- Validate and maintain a public fast detector simulation, reproducing the basic response of the LHC detectors. This is one of the key pre-requisits to make the (re-) interpretation of LHC results possible, and thus allow a wide range of BSM theories to be tested.
- Develop the means to publish the likelihood functions of LHC analyses**,** both as mathematical descriptions and in a digital form (e.g. as RooStats objects), in which experimental data and parameters are clearly distinguished. Here, key issues are e.g. the treatment of tails in distributions, and to reliably define the ranges of validity, when publishing only the *final* likelihood of an analysis. Alternatively, one could publish the complete data model.
- Develop and maintain a coherent analysis framework that collects all LHC results as they are made public and allows for testing of a large variety of models and including results from a large spectrum of research areas. Future versions of this platform are expected to include a user-friendly fast detector simulation module.
- The open access to preserved data is obviously a great opportunity for education and outreach, since the sound advances in HEP can be exposed to a large audience in an attractive way and using effective educational methods.

Outlook

The activity of the DPHEP study group over the last four years has lead to an overall awareness of the data preservation issue in HEP, but also made evident to all its members and for the community at large that there is a need for more action to be taken, in particular:

- **Coordination**: There is a clear need, expressed since the very beginning for international coordination. In fact, all local efforts profit from an inter-laboratory dialog, from exchange in information at all levels: technological, organisational, sociological and financial.
- **Standards**: There is a strong need for more standard approaches, for instance in what concerns data formats, simulation, massive calculation and analysis techniques. An increased standardisation will increase the overall efficiency of HEP computing systems and it will also be beneficial in securing long-term data preservation.
- **Technology:** The usage of some of the cutting edge paradigms like virtualisation methods and cloud computing have been probed systematically in the context of data preservation projects. These new techniques seem to fit well within the context of large scale and long-term data preservation and access.
- **Experiments**: The main issues revealed by the DPHEP study group are easily extendable to other experiments. Conversely, the recent experience shows that new aspects revealed by different computing philosophies in general do improve the

overall coherence and completeness of the data preservation models. Therefore the expansion of the DPHEP organisation to include more experiments is one of the goals of the next period.

- **Cooperation:** High-energy physics has been at the frontier of data analysis techniques and has initiated many new IT paradigms (web, farms, grid). In the context of an explosion of scientific data and of the recent or imminent funding initiatives that stimulate concepts as "big data", the large HEP laboratories will need to collaborate and propose common projects with units from other fields. Cooperation in data management: access, mining, analysis and preservation; appears to be unavoidable and will also dramatically change the management of HEP data in the future.

The new results from LHC and the decisions to be taken in the next few years concerning LHC upgrades and other future projects will have a significant impact on the HEP landscape. The initial efforts of the DPHEP study group will hopefully be beneficial for improving the new or upgraded computing environments as well as the overall organisation of HEP collaborations, such that data preservation becomes one of the necessary specifications for the next generation of experiments.

References

[1] "Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics", DPHEP Study Group, (text overlap) http://arxiv.org/abs/1205.4667
[2] F. Boudjema et all. ``On the presentation of the LHC Higgs Results,'' http://arxiv.org/abs/arXiv:1307.5865 .
[3] D. South and D. Ozerov, A Validation Framework for the Long Term Preservation of High Energy Physics Data, http://arxiv.org/abs/arXiv:1310.7814 .

Contact

Cristinel Diaconu, Centre de Physique des Particules de Marseille, CNRS/IN2P3 et Aix-Marseille Université; diaconu@cppm.in2p3.fr
Sabine Kraml, Laboratoire de Physique Subatomique et Cosmologie de Grenoble, CNRS/IN2P3 et Université Paul Sabatier; sabine.kraml@lpsc.in2p3.fr

Virtual Observatory in Astrophysics

Christian Surace

Abstract: In astrophysics, data preservation is really important, mainly because objects are far away and that an observational project (satellite, telescopes) is very expensive, time consuming and very difficult to redo when it is over. Nevertheless any projects have been undertaken and more and more data are available to the scientific community. The International Virtual Observatory Alliance (IVOA), formed in 2002, gather efforts on data standardisation and dissemination in Astrophysics. Endorsed by the International Astronomical Union (IAU), the Virtual Observatory consists on describing all validated astrophysical data, the format of them, the way to disseminate these data, protocols and software. This new way of data analysis is rather in advance in the interoperability of tools. In a context of cooperation between countries, the Virtual Observatory shows a good example of the new era of astrophysical analysis in massive and distributed data exchange.

Introduction

Astrophysics was one of the first science gathering information to create catalogs and atlases. From the first steps of drawings and measuring, astrophysics brought to light the existence of new far-away objects and pushed back the frontier of knowledge. Every piece of information was written down and classified. From the firsts catalogs to the deeps surveys that are undertaken today, the spirit of data dissemination is still present and is the crucial corner stone of scientific collaboration and advances.

Virtual Observatory for data preservation

Nowadays the astrophysical surveys are covering the entire sky, leading to a large amount of data. For example the Very Large Telescope brings every year 20 Terabytes of data, LSST : 3 billions pixels every 17 seconds, Pan-Starrs will produce Several Terabytes of data per night. By 2020, in one year the astrophysical instruments will deliver to the community more that 1000 Peta bytes of data. Due to this still growing amount of data, it is very difficult to perform deep study of individual objects. Statistical approaches, group selection and analysis became the necessary steps for data analysis. However some questions are still being debated on the nature of data to be preserved. Is it better to preserve final products, or raw data with all the infrastructure and pipelines needed to create final data? The question of repeatability, reproducibility and reliability of results are the keys to the data preservation for the future.

In astrophysics, data set are distinguished by origins: extracted from Space Satellites (see fig 1), ground based telescopes or provided using simulators and simulation programs (see fig 2). Then, data can be « images » from the sky taken with specific filters (ima = *f*(x,y)). Data can be « spectra » the light of which has been dispersed by a specific prism or grism (spec=*f*(*wavelength*,y)). Data can be « Time series » of an object, measurements of the flux

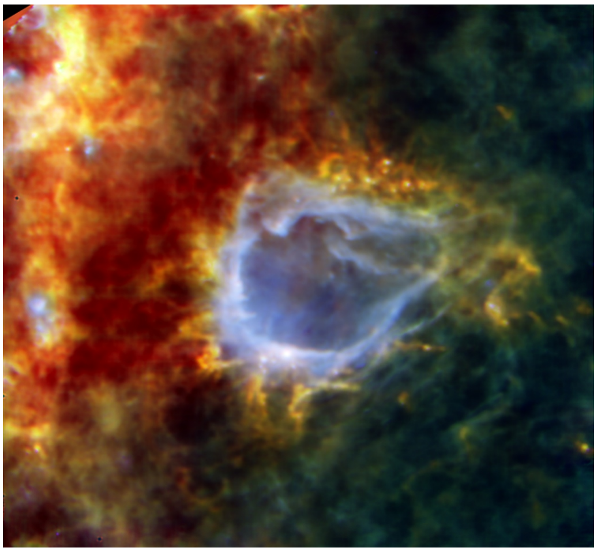


Figure 1. Herschel combined image (false colours) of a star forming region

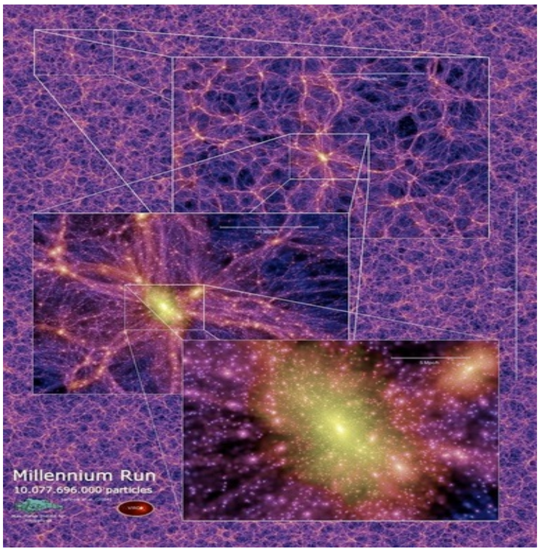


Figure 2. view from the simulation millennium run from Volker Soringel et al., Nature, 2005

emitted at periodic time during a long observation run (TimeSerie=*f*(time, flux)). Data can be cubes observed with specific instruments (Fabry Perot, IFUs) (cube=f(x,y,wavelength)). Data can be also simulated data built up from cosmological, galactic, stellar simulation programs that include quite a lot of information on particles. And finally data can be tabular, results of astrophysical analysis to complete the observations with added value like for example redshift, nature, velocity fields, velocity of the object . All these kinds of data are of interest for the astrophysical community.

Surveys are the main providers of astrophysical data. They offer the ability to gather many sources with the same instrument and the same environment. They are mainly conducted by the international agencies like the European Space Agency (ESA), European Southern Observatory (ESO), National Aeronautics and Space Administration (NASA) among others. They cover the overall range of the wavelength spectrum covering the range of radio waves, infra red, Optical, Ultraviolet, X-rays. Some surveys are considered as references for the overall astrophysical community. Here are some examples : « IRAS » is the first Infrared all sky survey giving the first image of the universe in several bands of the infrared range. « 2dF » is the first optical spectroscopic survey of the local Universe drawing for the first time the position of the galaxies of the local Universe. « CoRoT » an « Kepler » are providing information, images, time series of the exo-planets (planets orbiting around stars of the Galaxy). Apart from these reference data, more and more data are now available throughout the Virtual Observatory, numerical simulations : GalMer, 3D Data : IFU, FabryPerot, Radio Cubes, transient events : Bursts, SuperNovae, Exoplanet data : CoRoT, Kepler.

As the Universe is caring on its own evolution, and because the observations can be redone easily, astronomers have already carried out a first approach to homogenise the data back in 1986, with the « FITS » format. Most of the observational data are serialised in FITS. FITS format includes data and metadata describing the data in a same file. « GADGET » format is dedicated to simulation files. VO Format is the next generation for the description of data. In addition to the data description and storage. The preservation of accessibility of data is also very important with the Webservices available in the VO Portals and VO tools. But moreover, there should be preservation of knowledge, astrophysical pipelines and patterns as initiated by WF4ever.

VO technical implementation (models and protocols)

Since 2002, the virtual Observatory started the work to disseminate validated data all over the world. The goal is to make the data available, searchable, downloadable by any member of the scientific community and in open access. The focus was firstly put on definitions and technical overall infrastructure. Based on this infrastructure, and on the development of specific tools, data are now accessible and usable. The Virtual Observatory is focused on the scientific outputs from the usage of the Virtual Observatory. It is really a new age with new discovery technics for astronomy.

In order to exchange data between the members of the community, and to make it widely available, standards have been defined, from the definition of products and access protocols to the discovery process of the data. All exchanges and formats are based on the Extensible Markup Language (XML) format (see XML on w3). One of the first activities of the IVOA has been to define the data models to describe most, if not all, data that can be provided by the astronomical community. The definition of metadata is also essential to describe resources available in different sites, a dictionary to define the data and standards to define VO registries.

Moreover, further standards used for storage and data processing are being defined. It covers virtual storage addressing, single sign on, semantics and web service definition. But at the beginning, basic standards have been adopted as data models such as « VOTable » : a table exchange format, « Space Time Coordinates (STC) » that defines the coordinates of an event or an object, « ObsDMCore » : Core components for the Observation Data Model, « Astronomical Dataset Characterisation » that defines the overall characterisation of an observation, « Spectral lines » data model : defines environment and nature of a spectral line, « Spectrum » data model : defines an observational spectrum, « VOEVENT » : describes a tansient event, « Theory » : describes and access any numerical simulation.

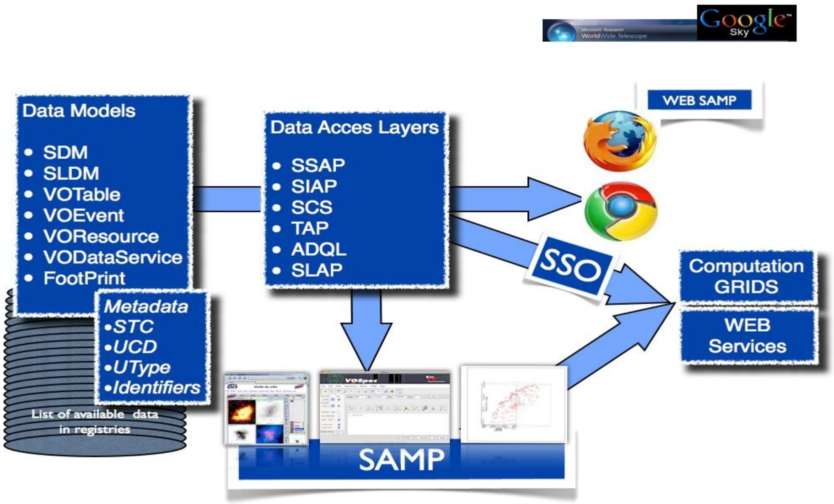


Figure 3. Example of implementation of VO standards in a scientific use case.

Several years have been needed to settle down these standards and some new standards are still being defined in order to cover all new data. Defining the data with models and standard is not sufficient as accessing the data is also one of the key of the scientific usage of data. Access protocols have then been defined. Such basic protocols have been first agreed : « SIA » : Simple Image Access : to access images and part of images depending on coordinates and observational band, « ConeSearch » : Position related search : to access any sources depending on its coordinates and « VOQL : VO Query Languages » : to setup queries with defined parameters linked to astronomical searches. More complicated protocols have been used afterwards: « SLAP » : Simple Line Access : to access spectral lines depending on atomic data and environment, « SSA » : Single Spectrum Access : to access spectrum type depending on coordinates and spectral range, « TAP » : Table and catalog Access : to access data with selection possible on any criteria whatever the way of storing data.

Using the VO

To get best advantages of the data disposal, tools have been developed. More are dedicated to the discoveries like portals and queries. Several portals exist to access data of the virtual observatory , like « Datascope », « CDS Portal ». Database Discovery Tool are also developed to explore Catalogs like « VOCAT » which is a catalog data interface tool to transform astronomical data in databases or « SAADA ». Several tools offer plotting and analysis functionalities like « VOPlot », designed for Large data sets, or « TOPCAT » for table/VOTable manipulation or « VOSTAT », tool for statistical analysis. « Aladin » is one of the tools for Image and Catalogue displaying, « VisIVO » as well offers a visualisation Interface. Data Mining tools already offer data mining studies and analysis : « MIRAGE », Bell Labs Mirage offers multidimensional visualisation of data, « VOSTAT » is a VOIndia tool for statistical analysis. « VOSpec » and « SPECVIew » from the STSCI are dealing with spectral data. Such tools are really powerfull as they hide the complexity of the VO infrastructure throughout easy interfaces and extend the VO capabilities with dedicated functionalities. These tools provide basic access to VO formatted data, FITS and ASCII data. All these tools offer great functionalities to analyse any kind of data but when combined they form a really powerful software suite. This combination is possible using SAMP (Simple Application Messaging). SAMP is a messaging protocol for interoperability to enable individual tools to work together. It is based on XML-RPC. Messages are standardised using keywords defined as standard for exchanges using ”mtypes” (message types) and ”params” (parameters) . As an extend of applications messaging protocol, a new WEBSAMP has been defined to connect web applications. The global usage is defined in figure 3. Fluxes are described with blue arrows while standards are shown in boxes. The goal is to use these standards and protocols in a transparent way to provide scientific outputs.

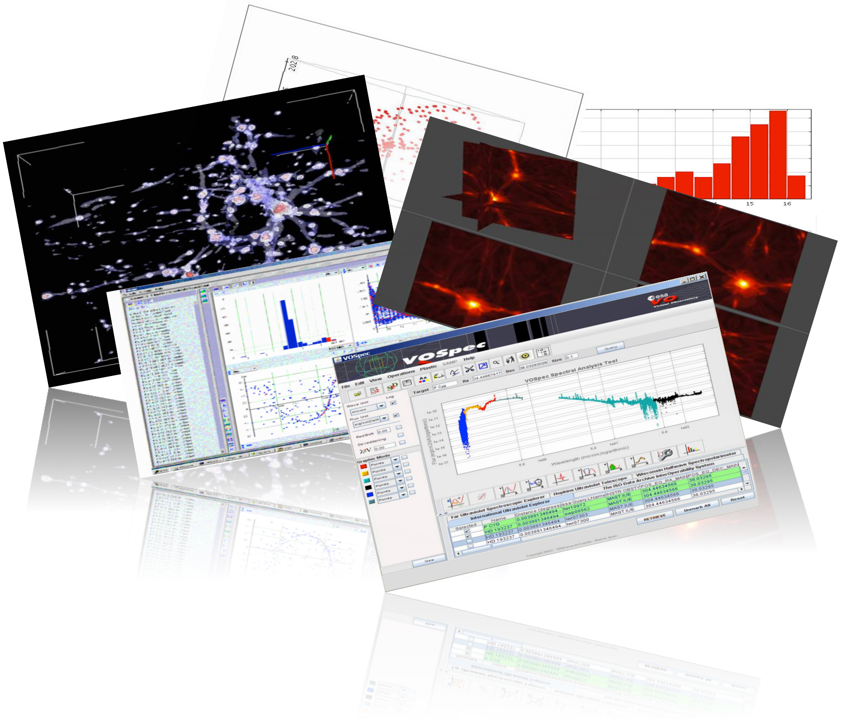


Figure 4. Screen shots of VO-compliant softwares able to retrieve, visualise and analyse VO data (List available at http://ivoa.net/astronomers/applications.html .

Trying to use the infrastructure, one can define its own use case. As seen in figure 2, the astrophysicist may, for example, try to compute the statistical properties of two populations of galaxies derived from different observational fields or using two different wavelength range. Starting from extracted images from the Virtual Observatory (using Data Models and SIA), sources catalog are created (using Web Services SExtractor), then VO tables are created, to be cross identified with other catalogs (using TAP, ConeSearch) and data fusion is operated (using tools TOPCAT, ALADIN, Web Services). Then a comparative study can be performed and statistical results can be derived.

The Virtual Observatory environment becomes easier and easier to use with more and more data available. After focusing on technical parts, this is a new era where the scientific goals drive technical developments. Of course there is still some things to improve (check for data quality, the curation of data, new standard to finalise). Of course there is still some things to do (include ALL data, define even easier-to-use portals). Of course there is new area to work on (Data Mining, ObsTAP, cloud computing). But Science can be done more easily using the quick access to data, in an homogenised way

References

IVOA : International Virtual Observatory Alliance : http://www.ivoa.net
Aladin:http://aladin.u-strasbg.fr/; VOPLOT:http://vo.iucaa.ernet.in/voi/voplot.htm
SAMP : Taylor et al. 2010 , http://www.ivoa.net/Documents/latest/SAMP.html
WF4ever : http://www.wf4ever-project.org/

Contact:

Christian Surace, Laboratoire d'Astrophysique de Marseille, christian.surace@lam.fr

Crystallography Open Databases and Preservation: a World-Wide Initiative

Daniel Chateigner

Abstract: In 2003, an international team of crystallographers proposed the Crystallography Open Database (COD), a fully-free collection of crystal structure data, in the aim of ensuring their preservation. With nearly 250000 entries, this database represents a large open set of data for crystallographers, academics and industrials, located at five different places world-wide, and included in Thomson-Reuters' ISI. As a large step towards data preservation, raw data can now be uploaded along with "digested" structure files, and COD can be questioned by most of the crystallography-linked industrial software. The COD initiative work deserves several other open developments.

Crystallography and Data Preservation

Crystallographic data acquisition relies on 0D-, 1D- and 2D-detector patterns, in scattering, diffraction, tomography ... experiments using x-rays, neutrons or electrons scattering. It has increased in volume in the past decades as never before. With the advent of new high-resolution detectors, large datasets are acquired within shorter and shorter times, reaching less than a millisecond per pattern with high brilliance ray sources. Large facilities like synchrotron, neutron and X-ray Free Electron Laser centres are daily producing data for thousands of users, each generating Gb to Tb data volumes. At laboratory scales, newer diffractometer generations using image acquisitions also generate non negligible data volumes requiring specific backups. The costs (sometimes very large) associated either to large facilities or to laboratory instruments by themselves impose data preservation. Large facilities are financially matters of, often, collaborative actions, like European or United States tools, and data produced by such institutions must be maintained. But also individual laboratory tools represent non negligible financial masses at a global scale (the cost for one diffractometer ranges from 100k€ to 1M€, a price which reaches several M€ for an electron microscope), in view of the number of equipped laboratories (if we imagine 50 laboratories equipped with several diffractometers and microscopes ... only in France).

A specificity of crystallographic data is then its geographic dissemination over the world. Any single scientific University, academic Centre or Institution possesses at least several instruments, if not several tens, usually relying to different laboratories. If large facilities can usually afford for large backup systems (data are one of their "products"), individual laboratories sometimes face backup problems on a long-term basis, particularly in front of new data acquisition experiments and data maintenance. Furthermore, but this can be true for other fields of science, scientific progresses, new developments in analysis tools, approaches and methodologies, also find interests in crystallographic data preservation. Newer concepts bring new analysis ways with new treatment capabilities allowing to provide more information and/or accuracy from older data. In such cases an incomparable value-addition comes from newer analysis of old data, at negligible cost. Data preservation becomes a "must-do".

More recent concerns for crystallographers, dating from early 2010, are frauds and plagiarisms. Several tens of scientific papers went through retraction procedures initiated by the publishers, because of proved frauds. Modified or purely invented data or results were detected after irreproducibility of the results by separated teams or clean examination of the scientific procedures. Such characteristically non scientific behaviour could have been stopped at an early step if original data deposition had been required with paper submissions, allowing (forcing) serious peer-reviewing. If not detected under peer-review, such an unwholesome behaviour could have easily been detected a posteriori using automated analyses of repository data.

Data Preservation became a major concern of the International Union of Crystallography (IUCr, www.iucr.org) for the recent past years, with the creation of a Diffraction Data Deposition Working Group (DDD WG) focused on diffraction images, though with other relevancies than solely images. Concerning long-term storage of diffraction images, this group concluded (*www.codata.org/exec/ga2012/iucrRep2012.pdf*):

i) there is not yet sufficient coherence of experimental metadata standards or national policy to rely on instrumental facilities to act as permanent archives;

ii) there is not sufficient funding for existing crystallographic database organisations (which maintain curated archives of processed experimental data and derived structural data sets) to act as centralised stores of raw data, although they could effectively act as centralised metadata catalogues;

iii) few institutional data repositories yet have the expertise or resources to store the large quantities of data involved with the appropriate level of discoverability and linking to derived publications.

Unfortunately, scientific literature via periodicals and books cannot maintain a sufficient level of scientific data preservation on a long term. Publishers are subjected to strong financial fluctuations and can decide to stop the edition of whole bunches of too-low profitability materials, which can contain invaluable scientific data. This is also true for open literature as far as this latter is kept under publishers' authority and maintenance. In particular, newly and small publishing houses that numerously pop up in the recent years are irradiating with very large panels of open titles and scopes, with no warranty of data survival after titles cancellation. Supplementary materials are more and more developed as a substantial material under article submissions, and could be thought helping data preservation. However they suffer the same uncertainties as the articles to which they are belonging and as such cannot be considered more stable over time.

Teaching is also an important aspect linked to data preservation. Many institutions cannot afford for renewal of scientific databases, materials, literature ... neither every year nor even every several years. Well preserved data allow at negligible costs to accommodate for this unfortunate financial lack and work on real case studies for a better student formation.

Finally, data preservation has to manage with older data supports, which can become unreadable with time. Old magnetic tapes, DAT bytes and other supports are no longer in use, and newer storage systems will reach obsolescence inevitably. We all suffered once this difficult situation of non-readable old data, that data preservation would ideally avoid using periodic reading tests and backup upgrading.

Crystallography Open Database as a Model

COD [1, 2, www.crystallography.net] choose from the beginning a fully open, collaborative way of working. With 14 advisory board members from 10 different countries, this project is definitely international and internationally recognized. At the present time, around 250000 structure files are made available for search and download (the whole database can be downloaded !) using various standard communication protocols (Figure 1). From 2012, the site allows all registered users to deposit published, pre-published and personal communications structure data, enabling COD extension by many users simultaneously.

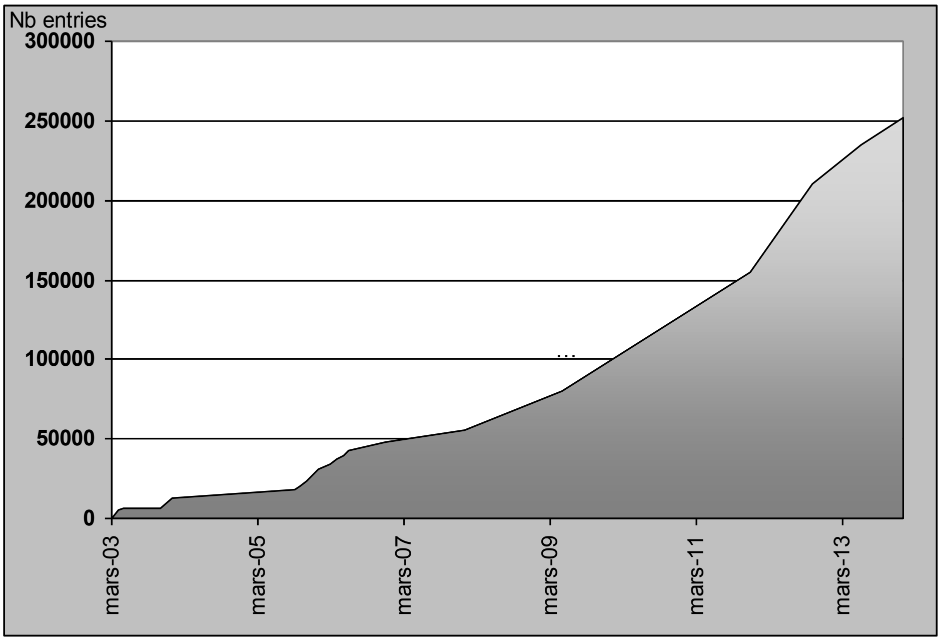


Figure 1: Number of structure items archived into COD since launching in 2003

The data in COD are stored in the Crystallographic Information File/Framework (CIF) format, created and developed by the IUCr in 1990, today a broad system of exchange protocols based on data dictionaries and relational rules expressible in different machine-readable manifestations, including, but not restricted to, CIF and XML. CIF is now a world-wide established standard to archive and distribute crystallographic information, used in related software and often cited as a model example of integrating data and textual information for data-centric scientific communication. Importantly, CIF is an evolving language, users being able to create their own dictionaries suited to their fields, and relying on a core dictionary for already defined concepts. Accompanied by the *checkCIF* utility, this framework was recognised by the Award for Publishing Innovation of the Association of Learned and Professional Society Publishers in 2006. The Jury was "*impressed with the way in which CIF and checkCIF are easily accessible and have served to make critical crystallographic data more consistently reliable and accessible at all stages of the information chain, from authors, reviewers and editors through to readers and researchers. In doing so, the system takes away the donkeywork from ensuring that the results of scientific research are trustworthy without detracting from the value of* *human judgement in the research and publication process*". This, one year or so before the advent of Internet HTLM !

Originally, new data entries were collected manually, by the advisory board and international volunteer scientists. Now mainly operated by our Lithuanian representative team in Vilnius, COD uploads are more and more automated using harvesting procedures from scientific supplementary materials. Some publishers, and among them IUCr, agree on such practices for the best scientific knowledge and its sustainability.

Data preservation is ensured in COD via mirroring. Four mirrors are actually settled, in Lithuania (Vilnius: http://cod.ibt.lt/), France (Caen: http://cod.ensicaen.fr/), Spain (Granada: http://qiserver.ugr.es/cod/) and USA (Portland/Oregon: http://nanocrystallography.org/), and one registered domain, www.crystallography.net. Additional regular backups are made on DVD-ROM. Mirroring is an efficient way to keep data on long time scales, independently of national, regional or local politics, institutions closing or reorganisation, scientists move or change of activity. In this respect one big data centre is considered less sustainable over time than an international network of mirrors. Also, unlike closed databases for which data preservation depends solely on the owner of the database, open databases can be backed up flexibly, balancing backup costs against the value of data for the stakeholders.

The COD data items will be indefinitely maintained as available over designated URIs. Thus, an URI containing a COD number in a form http://www.crystallography.net/<COD-number>.cif (e.g. http://www.crystallography.net/1000000.cif), is permanently mapped to the corresponding CIF, no matter what file layout or internal representation the COD is using. So far we have maintained the described URIs since 2003, and researchers can rely on the web services provided by the COD server, and on the possibility to obtain local copies or restore previous data in a standard way if needed. Further developments are envisioned towards clustering of the COD mirrors, including incorporation and/or linking of other open databases for larger data sharing and inter-operability.

COD also receives much attention from industrials in the crystallography field (mainly diffractometers and software companies), but also from Thomson Reuters. The formers found in COD an invaluable way of getting free, ready-to-use and high quality scientific data. They incorporate COD subversions in their own software and for their client purposes. The latter incorporated a new member to the Web of Knowledge family of databases: the Data Citation Index (DCI) in which COD took not less than the fifth place in 2013.

More than COD

Several other open databases exist in the field of crystallography, actually curating, delivering and archiving independently structural data, more or less not redundantly. Among the prominent ones, we find the American Mineralogist Crystal Structure Database (http://rruff.geo.arizona.edu/AMS/amcsd.php), the Protein Data Bank (http://www.wwpdb.org/), the Bilbao Crystallographic Server (http://www.cryst.ehu.es/), the International Zeolite Association Database (http://www.iza-structure.org/databases/), the Raman Spectra of Minerals (http://minerals.gps.caltech.edu/files/raman/), ..., a full list being at http://nanocrystallography.net/.

The AMCSD is fully incorporated in COD from the beginning, while the protein-target of the PDB makes it not redundant with COD. The Bilbao server is oriented towards special

structures like aperiodics, incommensurates, modulated ... which are not still incorporated in COD. The IZA database is periodically harvested for new zeolite structures which have been approved by the zeolite structure commission.
COD also deserved inspiration for other Open Database developments (Figure 2). The Predicted COD (http://www.crystallography.net/pcod/ and http://sdpd.univ-lemans.fr/cod/pcod/), a resource containing inorganic compounds (silicates, phosphates, sulfates of Al, Ti, V, Ga, Nb, Zr, zeolites, fluorides, etc) predicted using various software, is the largest structure data set with over 1 million entries. The Theoretical COD (http://www.crystallography.net/tcod/), is a collection of theoretically refined of calculated from first-principle calculations or optimisations. Both PCOD and TCOD are not based on experimentally measured data that would necessitate preservation. However, they require large calculation times and as such can be considered as experimental-like value-added, and benefit from data storage of the results.

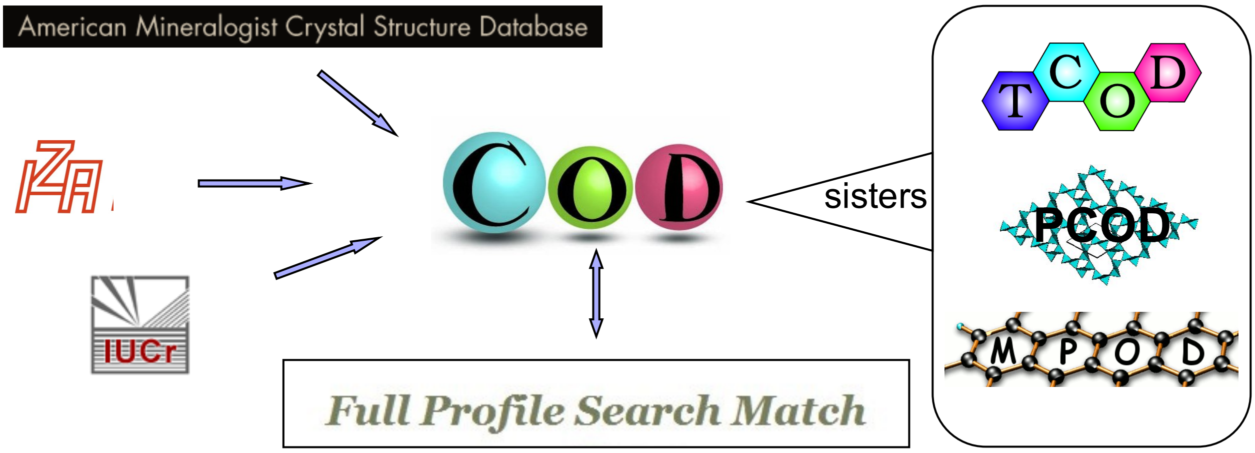

American Mineralogist Crystal Structure Database
IZA
IUCr
COD
sisters
TCOD
PCOD
MPOD
Full Profile Search Match

Figure 2: Actual Open Databases landscape directly surrounding COD

Materials exhibit specific properties which are expressed as tensors and depend on structures. The Material Properties Open Database (http://www.materialproperties.org/), linked to COD entries [3], offers a place to get property tensors of various kinds, and will be soon mirrored on a Mexican site to develop tensor surfaces representations and an automated search of new properties data.

Finally, a recent effort to exploit open data has been launched. The Full-Profile Search-Match tool (http://cod.iutcaen.unicaen.fr/ and http://nanoair.ing.unitn.it:8080/sfpm) uses COD to identify and quantify phases from powder diffraction patterns, freely accessible to everybody. Such an application really opens a new delocalised mode for treating data. Associated to more developed numeric preservation it could allow real breakthrough in data analysis, Combined Analyses of multiple datasets (eventually measured by different techniques and other peoples), automated cross-checking of results, including easy statistical distribution of results. This would also allow to concentrate human and financial efforts in a more efficient way (experimental efforts are best used where recognised instrumentalists are, analysis efforts with analysis experts' hands), enhancing collaborative actions.

As a conclusion, one can see that crystallographers are building progressively a complex network of tools, backups, digested and operational data, with clearly in mind Scientific Data Preservation. Languages, syntaxes, formats and software were developed for now more than 23 years, in the view of establishing interactive architectures in the future. As far as crystallographic data are of concern, proper preservation appears more ensured using geographic dissemination modes to warranty stable backups, not depending on local issues. In January 2014 the International Year of Crystallography begins as mandated by UNESCO (http://www.iycr2014.org/). In August 2014 the 23rd World Congress of the International Union of Crystallography will take place, with major meetings of the CIF and data preservation commissions.

References

[1] S. Grazulis, D. Chateigner, R.T. Downs, A.F.T. Yokochi, M. Quiros, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, A. Le Bail: Crystallography Open Database - an open-access collection of crystal structures: *Journal of Applied Crystallography* **42(4)**, 2009, 726-729

[2] Saulius Grazulis, Adriana Daskevic, Andrius Merkys, Daniel Chateigner, Luca Lutterotti, Miguel Quiros, Nadezhda R. Serebryanaya, Peter Moeck, Robert T. Downs, Armel Le Bail. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research* **40** Database Issue, 2012, D420-D427

[3] G. Pepponi, S. Grazulis, D. Chateigner: MPOD: a Material Property Open Database linked to structural information. *Nuclear Instruments and Methods in Physics Research B* **284**, 2012, 10-14

Contact:

Daniel Chateigner (for the COD Advisory Board), Institut Universitaire de Technologie (IUT-Caen), Université de Caen Basse-Normandie (UCBN) and Laboratoire de CRIstallographie et Sciences des MATériaux (CRISMAT) – Ecole Nationale Supérieure d'Ingénieurs de CAEN (ENSICAEN); daniel.chateigner@ensicaen.fr

Satellite Data Management and Preservation

Therese Libourel , Anne Laurent, Yuan Lin

Abstract: We describe here how satellite earth data are managed and preserved. They represent a huge volume of data, collected with specific sensors. The satellite missions require national and/or international cooperation and are more and more pooled so as to ease the access to resources and to reduce the costs. Maintaining, treating and preserving these satellite earth data is of prime importance as they are essential for dealing with many current environmental challenges (climate change, littoral, etc.). In order to be able to treat and reuse them, many issues must be tackled regarding both technical and semantic problems. In particular, we show how important metadata are.

Introduction

Earth observation is essential for environmental issues: erosion of coastline, natural hazards, evolution of biodiversity,...

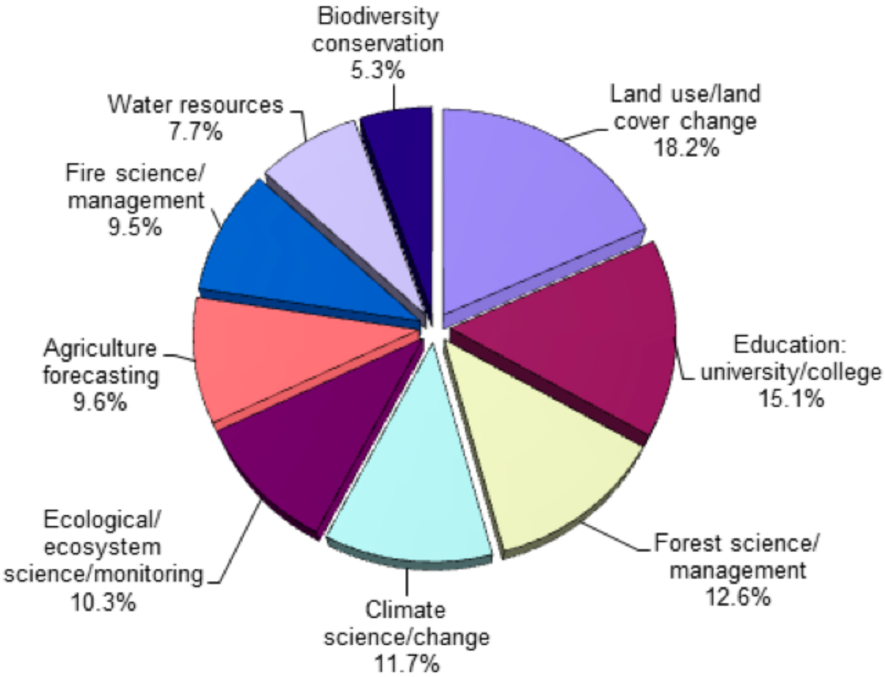


Figure 1: Top ten primary data uses (source: http://landsat.usgs.gov/Landsat_Project_Statistics.php)

Satellite observations come as an essential complementarity to *in situ* observations at various scales. The first missions date back to the end of the 20th Century. Since then, several satellites have been launched and their number keeps increasing, thus leading to a huge (and rapidly increasing) volume of satellite data being available. The most famous missions include Landsat, SPOT, Pleiades,... The number of images distributed is indeed increasing very fast: 14 million of Landsat scenes in 2013, 6.811.918 SPOT4 images acquired between 1998 and 2013.

Observational data are produced by using different sensors, which can generally be divided into two main categories: *in-situ* sensors and those which are carried by satellites (*passive optical sensors* and *active radar sensors*).

As explained in [1], roughly speaking, *“the detail discernible in an image is dependent on the spatial resolution of the sensor and refers to the size of the smallest possible feature that can be detected. Spatial resolution of passive sensors (we will look at the special case of active microwave sensors later) depends primarily on their Instantaneous Field of View (IFOV). The IFOV is the angular cone of visibility of the sensor and determines the area on the Earth's surface which is "seen" from a given altitude at one particular moment in time. This area on the ground is called the resolution cell and determines a sensor's maximum spatial resolution”.*

Spatial resolution has evolved over time from low resolution images (e.g., 300m for MERIS, 80m for the first Landsat images) to medium resolution (e.g., 20m for SPOT-1 images) and to high resolution images (e.g., 1.5m for SPOT-6 images, 0.5m for Pleiades images).

Several initiatives are currently undertaken in order to pool resources and services. As an example, the Landsat project (http://landsat.usgs.gov) is a joint initiative between the US Geological Survey (USGS) and NASA. It is one of the world's longest continuously acquired collection of space-based moderate-resolution land remote sensing data representing four decades of imagery.

The SEAS project (www.seasnet.org) is a technology platform network for earth satellite observation data reception and exploitation. SEASnet is implemented in european and french universities (Guyane, La Réunion, Canaries, Nouvelle-Calédonie, Polynésie Française) and aims at participating at the management of the environment and the sustainable development in tropical areas.

The GEOSUD project, funded by the French ANR National Agency for Research, stands for GEOinformation for SUstainable Development. It aims at building a National satellite data infrastructure for environmental and territorial research and its application to management and public policies. The project includes many actors. The CINES (Centre Informatique National de l’Enseignement Supérieur) contributes for data preservation, while researchers provide their expertise on scientific data workflows. A center for high performance computation is involved (HPC@LR) in order to provide supercomputing resources that are necessary to deal with such voluminous and complex data. This project is used below to describe how preservation is complex.

Satellite Missions: Specificities and Production Workflow

Satellite image producers organise the acquisition and production of such data as described by Fig. 2.

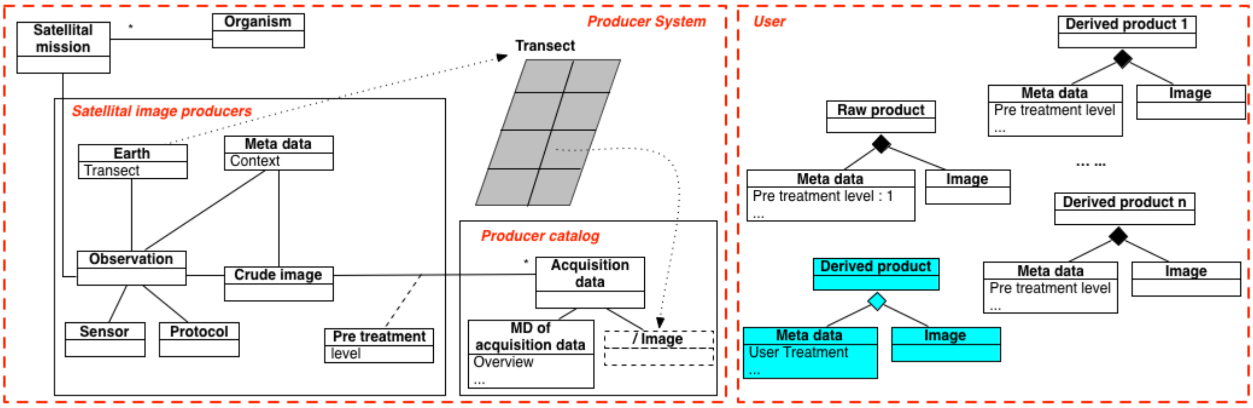


Fig. 2 - Acquisition and Production of Satellite Images (UML Formalism)

One mission performs observations on earth transects with specific sensors and protocols. Every observation results in a batch of raw data coupled with contextual metadata (description of observation parameters: viewing angle, spectral band, timestamps, etc.). Starting from this batch, the producer performs a set of remedial treatments. These treatments are undertaken at different levels. For instance, SPOT missions perform the following treatments at levels 1A (radiometric correction), 1B (radiometric and geometric correction), 2A (radiometric and geometric corrections regarding map projection standards). The producer provides the users with the so-called *acquisition data* within a catalog associated to this producer. Every item in the catalog corresponds to a virtual image from the transect together with a set of metadata associated to it.

End users choose an item from the catalog and a level of pretreatment. From raw data, metadata and pretreatments having been chosen, the producer system generates a product. This product is denoted by *raw product* if the level of pretreatment is basic, and by *derived product* otherwise. The product always contains an image together with contextual metadata. It should be noted that the user can also generate derived products from treatments he can define by himself.

In this process, metadata are crucial. First, to be relevant, treatments require the user to know them. Second, metadata are essential for indexing and reusing data. However, metadata also raise problems as they are not yet well standardized. Moreover, producers change metadata descriptions from one mission to another.

In the GEOSUD project, the targeted infrastructure for spatial data management is service-oriented. Every retrieving service and data access service relies on standardized metadata and data.

Satellite Data Preservation Workflow

The first idea in the GEOSUD project is to preserve in a short-term vision the raw products in a repository and the metadata of the raw product will be completed and standardized for feeding a catalog which users can select relevant products from. The second idea is to preserve in a long-term vision both the raw and derived products by relying on the services provided by the CINES. The processes that implement these two ideas are modeled as shown in Fig. 3.

Workflows refer to a sequence of treatments applied to some data. Treatments can be chained if the result of the previous treatment is consistent with the next treatment. In the workflow for data preservation in GEOSUD, the raw product feeds a complex service for metadata generation (including sub-steps like: extraction, completion, standardisation, etc.) which separates the metadata for the retrieving service from the metadata that are inherent to the image. The first ones are provided for the catalog while the second ones feed an image repository.

High performance computing is used to apply time consuming treatments that transform raw products to derived products. The HPC service is complex as it includes the generation of the metadata describing the applied treatment and the derived product which contains the processed image together with the associated metadata.

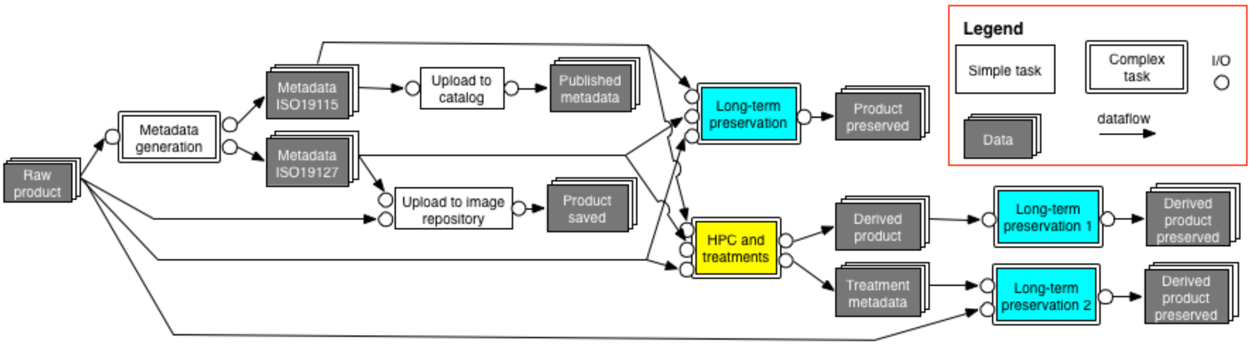


Fig. 3 - The GEOSUD Workflow for Short- and Long-Term Preservation

The long-term preservation services are also complex depending on the targeted product, should it be raw or derived product. Regarding the raw products, the preservation concerns all the metadata and the raw product. It should be noted that two current standards have been used for the metadata, i.e. ISO19115 and ISO19127 which may evolve, thus requiring the need for a maintenance of the various versions.

Regarding the derived products, two possibilities are provided for preservation. The first one is similar to the raw product long-preservation. The second one uses metadata treatments for rebuilding the derived product from the raw product. It should be noted that both raw and derived products combine images and metadata. As images are voluminous, it would be interesting to preserve the image from the raw product independently from the various metadata associated to it. The intelligent management of metadata will allow to manage any need without replicating images, which is a main challenge.

References

[1] Spatial Resolution, Pixel Size, and Scale. Natural Resources Canada. http://www.nrcan.gc.ca/earth-sciences/geomatics/satellite-imagery-air-photos/satellite-imagery-products/educational-resources/9407

[2] Yuan Lin, Christelle Pierkot, Isabelle Mougenot, Jean-Christophe Desconnets, Thérèse Libourel: A Framework to Assist Environmental Information Processing. ICEIS 2010: 76-89

[3] James B. Campbell and Randolph H. Wynne. *Introduction to Remote Sensing*, Fifth Edition. The Guilford Press. 2011

Contact

Therese Libourel*,** therese.libourel@univ-montp2.fr
Anne Laurent* laurent@lirmm.fr
Yuan Lin** Yuan.Lin@lirmm.fr
* LIRMM, University Montpellier 2, CNRS
** EspaceDev - IRD, University Montpellier 2, UAG, UR.

Seismic Data Preservation

Marc Schaming

Abstract : Seismic methods are used to investigate the subsurface of the Earth to image the sedimentary layers and the tectonic structures, for hydrocarbon exploration, near-surface applications, or crustal studies. Since the 1st half of the 20th century, data are acquired and stored on paper, film, tapes or disks. The preservation of these unique data is of outmost importance, and has to deal with favorable and unfavorable aspects. Some recent European projects demonstrated that it is possible to preserve and re-use the seismic data, but that this is to be done at national or European level.

Introduction

Seismic data are used to image the sedimentary layers and the tectonic structures of the Earth, for hydrocarbon exploration, near-surface applications (engineering and environmental surveys), or crustal studies (figure 1). Exploration and production companies as well as academia use these methods on land and on sea since the 1st half of the 20th century.

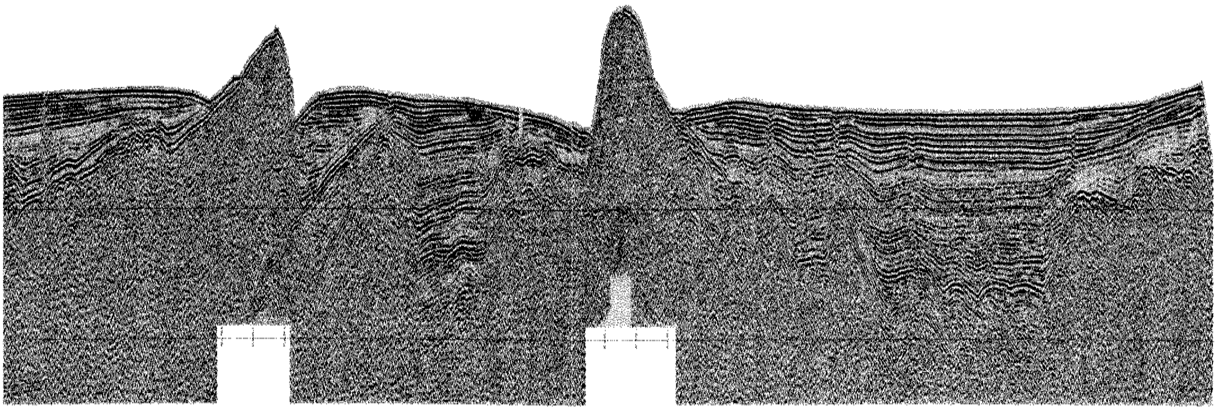


Figure 1: Typical seismic section (vertical scale: two-way travel time 0-4.5s, horizontal scale: distance) showing a rift with opposite tilted blocks.

Data were first on paper or film (figure 2), than digitally on magnetic supports, and represent many valuable datasets. Preservation of these patrimonial data is of highest importance.

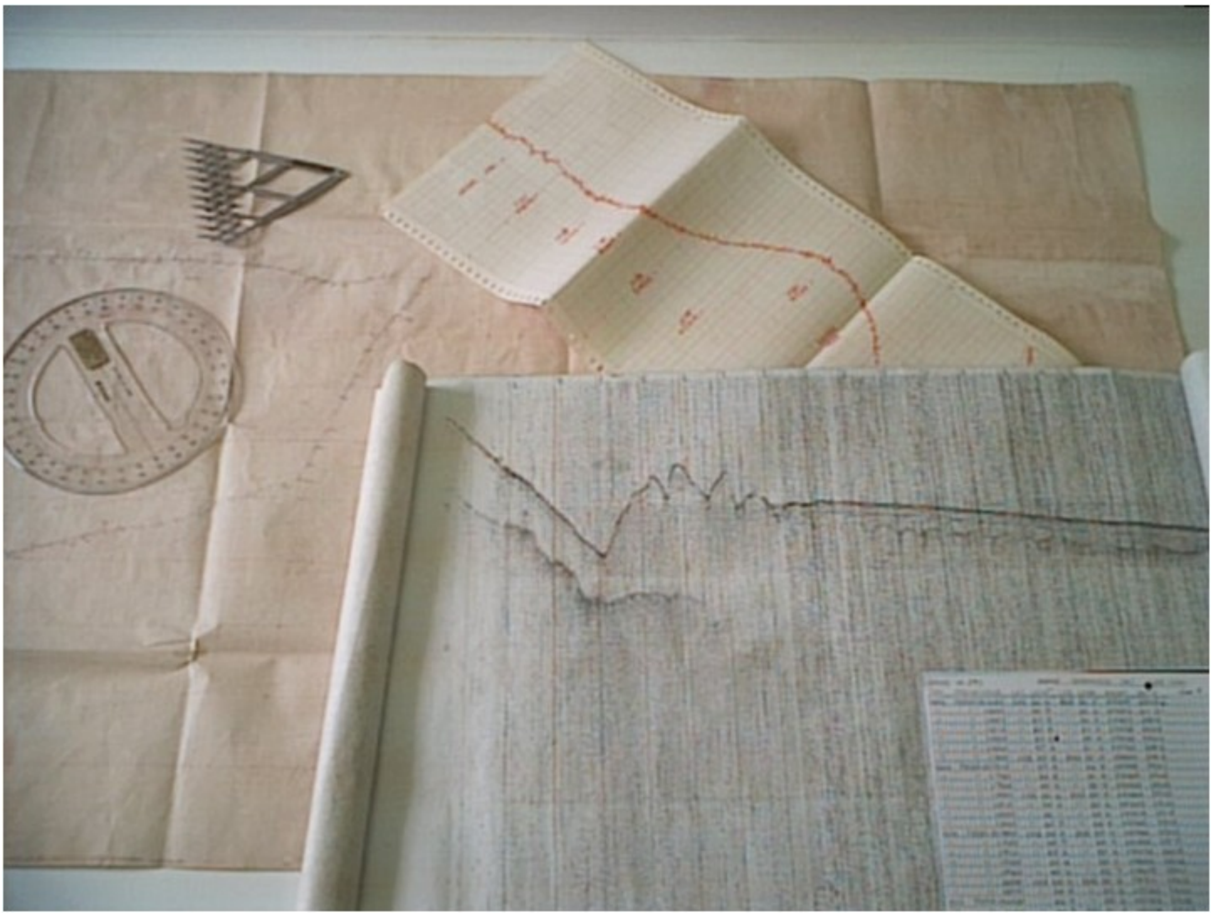


Figure 2: Old original documents: seismic section on paper, navigation chart, logbook, etc.

Why preservation of seismic data is essential

« Geophysical data is preserved *forever* ». That's what wrote Savage [1], chair of the SEG (Society of Exploration Geophysicists) Archiving and Storage Standards Subcommittee in 1994, and he compared the seismic data hoards to « family jewels ». Several reasons explain this:

- Acquisition or resurvey costs are high, because of duration of surveys, necessary personnel, immobilization of hardware, platforms, etc. Typical costs range from ~$3,000/km in onshore 2D, $20,000/km2 in onshore 3D, and marine seismic surveys can cost upward $200,000 per day.

- resurveying may be infeasible due to cultural build-up, or political changes in countries. Moreover, it is sometimes interesting to compare several brands of surveys that were acquired along time for 4D studies.

- Older data are reused with or without reprocessing using new algorithms (like PSDM, pre stack depth migration), for newer geophysical/geological studies. Recent examples are the use of legacy data to support national claims for ZEE extensions (UNCLOS, United Nations Convention on the Law of the Sea) where data offshore Mozambique, Kenya, Seychelles, Madagascar, Bay of Biscay, were of first value ; contributions to academic research like for

ANR TOPOAFRICA ; acquisition of oil industry to prepare new bids in the Mozambique Channel, etc. Data from the french ECORS program (Etude Continentale et Océanique par Réflexion et réfraction Sismique, 1984-1990) are regularly requested by academic researchers as well as by the industry.

Obstacles and advantages

Preservation of seismic data have to deal with favorable aspects (+), but also have to cover unfavorable ones (-).

- **Permanent increase of data volume and archive size** (-)

Along time there is a permanent increase of data volume during acquisition and processing phases.

Year	#Streamer	#Traces/streamer	Recording length (s)	Sample rate (ms)	Samples/Shot
~1980	1	24	6	4	36,000
~1990	1	96	12	4	288,000
~2000	1	360	15	2	2,700,700
~2013 (industry)	16	1024	12	2	98,304,000

Table 1: Typical volumes of datasets at acquisition

To give some typical volumes : between 1981 and 1993, french academia acquired about 50,000 km of seismic data representing 250 days of cruise, 7000+ 9-track tapes, but only 1Gb of data; Ecors data (1984-1990) represents about 4,000 9-track tapes.

- **Variety of media/devices and formats** (-)

An important problem is related to the variety of media/devices. Before the digital area things were quite easy, records were on paper and/or films, sometimes on microfilms. These media have a long expected lifetime, and have the advantage of being human readable. Since the digital revolution E&P industry tried to used media of best capacity and transfer speed ; from 7-tracks or 9-tracks magnetic tapes to IBM 3490/3590 to 4mm-DAT or 8mm-Exabytes to DLT to LTO to ... Many data are also written temporarily on hard-disks (mainly during processing & interpretation steps).

Tape formats are generally record-oriented, and depending the recording format shots are in individual files or not, channels are multiplexed or not, description and informative records are added or not, and all these records have different block sizes. A simple copy to disk is not correct, therefore some specific formats (e.g. TIF, Tape Interchange Format, or RODE, Record Oriented Data Encapsulation [2]) were defined.

Older data are often poorly stored, in boxes in a cupboard or on an office shelf, and of limited access. Regularly, some collections are thrown. Things are not better with digital data if no conservative measures are taken: older tapes are no more readable or only partly and with difficulties, or also thrown during office moves, companies merging, etc.

- **Some standardization of formats and media/devices** (+)

Exploration geophysicists defined quite rapidly technical standards for exchanging data. A major actor is the SEG, Society of Exploration Geophysics. First standards were published in 1967 for digital tape formats, and SEG is in a permanent process of updating or adding formats. These give a good frame to data exchange, even each company may tune them to their usage.

Also, some devices were used by the E&P industry and became de-facto standard devices (7-tracks or 9-tracks magnetic tapes, IBM 3490/3590 cartridges, 4mm-DAT or 8mm-Exabyte, DLT, and LTO) and had/have therefore a quite long product life and support.

- **Patrimonial and market value** (+)

Seismic datasets have a patrimonial value as explained above, because reacquisition is expensive and sometimes impossible. The Seiscan/Seiscanex [3] European projects (rescue the early paper seismic reflection profiles using long large format scanning with archive to a CD-ROM database of image files and minimal metadata) concluded that the 11,000 A0-images scanned (1,400,000 line kilometers) would cost over 30 million euros to re-survey at current rates.

They may have also a commercial value, when useful for E&P industry. Academia datasets were often acquired for fundamental research in places with no industrial interest; but E&P industry explores now newer regions (e.g. in deeper water, or closer to oceanic domains, or simply at places that were politically closed for exploration) and are really interested in accessing the datasets to assess their new projects.

Data access

A way to convince of importance of seismic data preservation is the show that they are of interest. Some recent projects at the European scale dealt with this topic. A prerequisite is to describe the datasets with metadata, than give easy but controlled access to them.

Metadata

Metadata describes the datasets by answering basic questions such as what, where, when, who, how, where to find data, etc. Seiscan/Seiscanex projects provided minimal metadata, but Geo-Seas used more complete ISO compliant metadata to describe the datasets. For seismic datasets, additional records had to be added: an O&M record (Observations and Measurements) with informations for data visualization, aggregation of segments of seismic lines and navigation, and a SensorML record that holds domain-specific parameters (figure 3).

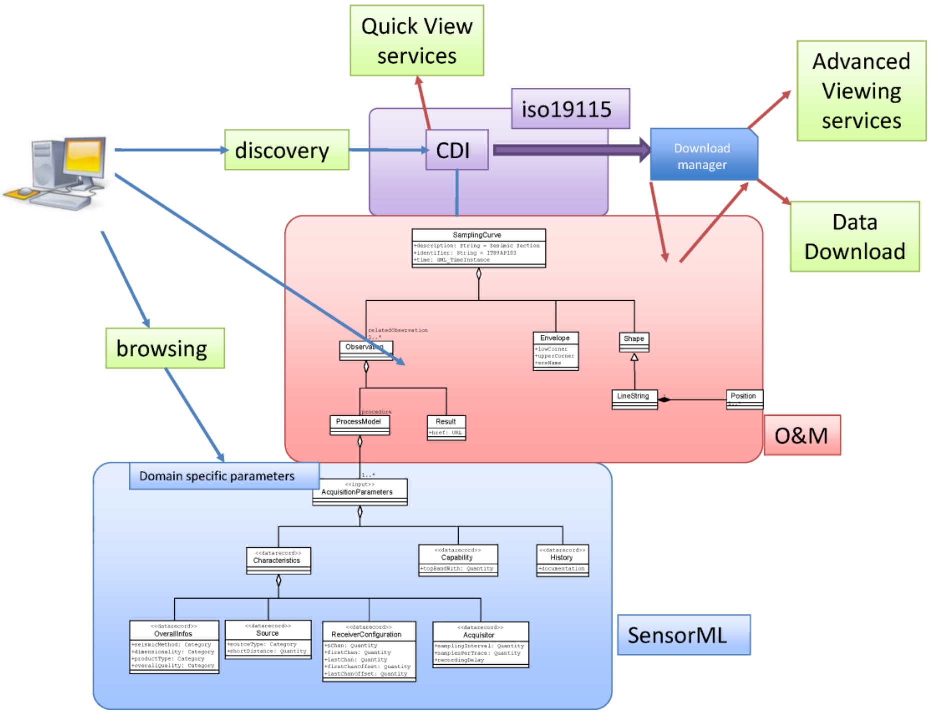


Figure 3: Geo-Seas metadata schema

Data accessibility

Data valorization can be improved through accessibility. Therefore publishing metadata is very important, as well as giving a quick-view of the data. This is done in the Geo-Seas [4] portal: metadata allow users to browse through datasets and select/query some of them. It is also possible to retrieve a thumbnail of the seismic data. After that, only registered users that accepted the data licenses can go further and have either a high-resolution view of the seismic data (figure 4), or retrieve them.

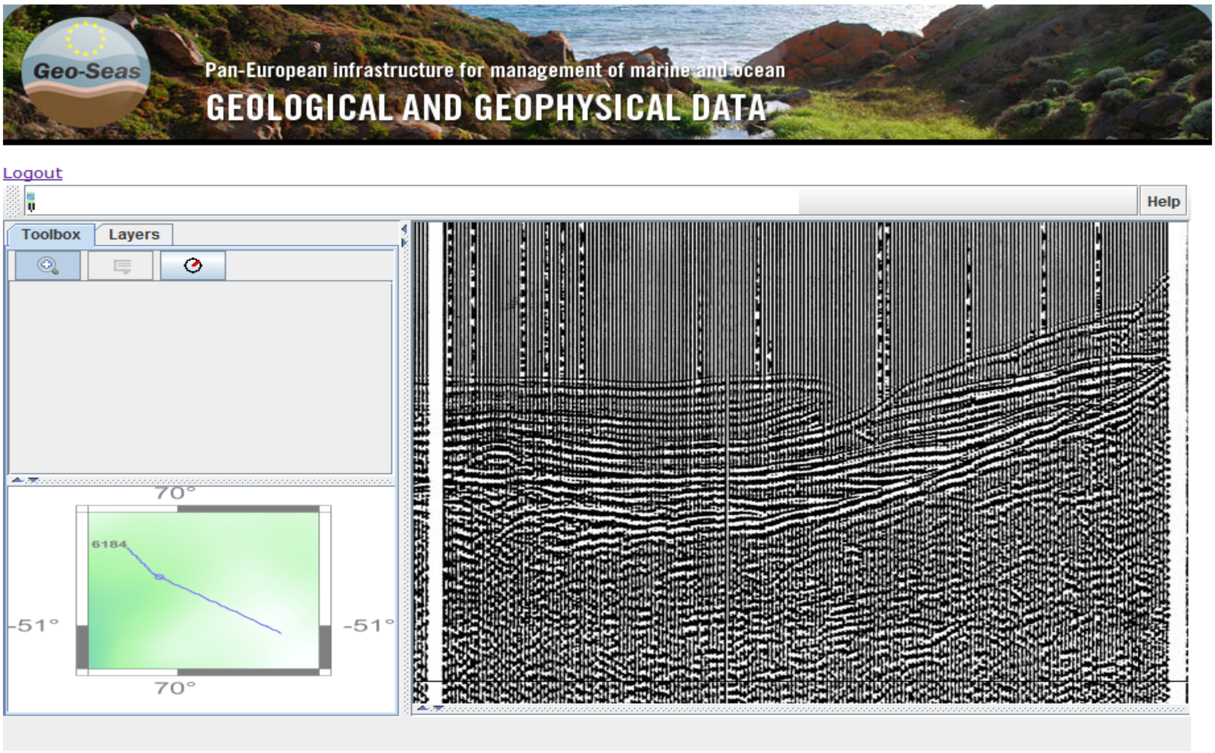

Geo-Seas
Pan-European infrastructure for management of marine and ocean
GEOLOGICAL AND GEOPHYSICAL DATA
Logout
Help
Toolbox
Layers
70°
6184
-51°
-51°
70°

Figure 4: Seismic Image from High Resolution Seismic Viewing Service

Conclusion

Preservation of seismic data is essential, but usually not considered by scientists, because it takes resources to document metadata, to read and copy tapes, to convert formats, etc. These tasks should be addressed at national and/or European level. Some European projects (Seiscan/Seiscanex, Geo-Seas) demonstrated that it is possible and useful. Repositories at national level should pursue this task with geophysical skills.

References

[1] Savage, P, 1994 – Recommended practices for storage and archiving of exploration data. The Leading Edge, 102-104.

[2] Booth, Algan, Duke, Guyton, Norris, Stainsby, Theriot, Wildgoose and Wilhelmsen, 1996 - SEG Rode Format Record Oriented Data Encapsulation Geophysics, 61, no. 05, 1545-1558.

[3] Seiscanex – Developing a European Facility to re-use seismic data: http://cats.u-strasbg.fr/seiscanex.html

[4] Geo-Seas – Pan-European infrastructure for management of marine and ocean geological and geophysical data http://www.geoseas.eu

Contact:

Marc SCHAMING, Institut de Physique du Globe (CNRS/UNISTRA), Strasbourg ; Marc.Schaming@unistra.fr

Chapter 2: Methodologies

Workflows and scientific big data preservation

Salima Benbernou and Mustapha Lebbah

Abstract : The scientific data landscape is expanding rapidly in both scale and diversity. Consequently, to handle the scalability of data generation, it is needed a scalable processing methods for managing and analysing the data. The workflow are widely recognised as a useful paradigm to describe, manage, and share complex scientific analyses, simulations and experiments. However, the long term preservation of scientific workflow and the methods used for executing it faces challenges due to the vulnerability and volatility of data and services required for its execution. Changes can be made in the workflow environment because Web services may evolve over the time. Consequently, it will alter the original workflow and hinder the reusability of output from workflow execution. In this chapter we will give an overview of scientific workflow and present some challenges and analysis methods for long term data preservation.

Data preservation: representation

Today, computation has become a very important aspect of science alongside theory and experiment using big data (scalable). Hence, scalable computational tools are needed in applications that involve complex tasks for scientific data representation, analysis and visualization. A typical scenario is a repetitive process of moving data to a supercomputer for simulation, launching the computations and managing the representation of data and storage the output results that are generally beyond the competencies of many scientists.

Scientific workflow systems aim at automating this process in a way to make it easier for scientists to focus on their research and not on computation management i.e. methods for extracting data, visualizing data, predicting data, validating data, reproducing complex tasks, reusing data results etc. Therefore, the workflow is becoming a powerful paradigm for scientists to manage big scientific data [1]. A scientific workflow describes a scientific procedure requiring a series of step process to coordinate multiple tasks. Each task represents the execution of a computational process, such as running a program, querying a database, submitting a job to a compute, invoking a service over the Web to use a remote resource. An example of scientific workflow is depicted in Figure 1.
Scientific workflows help in designing, managing, monitoring, and executing in-silico experiments. Moreover, workflow orchestration refers to the activity of defining the sequence of tasks needed to manage a business or computational science or engineering process.

A workflow that utilizes Web services (WSs) as implementations of tasks is usually called service composition [2]. Web services are the most prominent implementation of the Service-Oriented Architecture. The Web Service technology is an approach to provide and request services in distributed environments independent of programming languages, platforms, and operating systems. It is applied in a very wide range of applications where integration of heterogeneous systems is a must. Scientific workflow systems have become a necessary tool for many applications, enabling the composition and execution of complex tasks as web services for analysis on distributed resources.

Scientific workflow and preservation: challenges

Reusability/reproducibility

Many scientists are using workflows to systematically design and run computational experiments and simulations. Once the workflow is executed, the scientists would like to reuse the dataset generated as a result to be reused by other scientists as input to their experiments [3]. In fact, it may be possible to re-execute workflows many years later and obtain the same results.

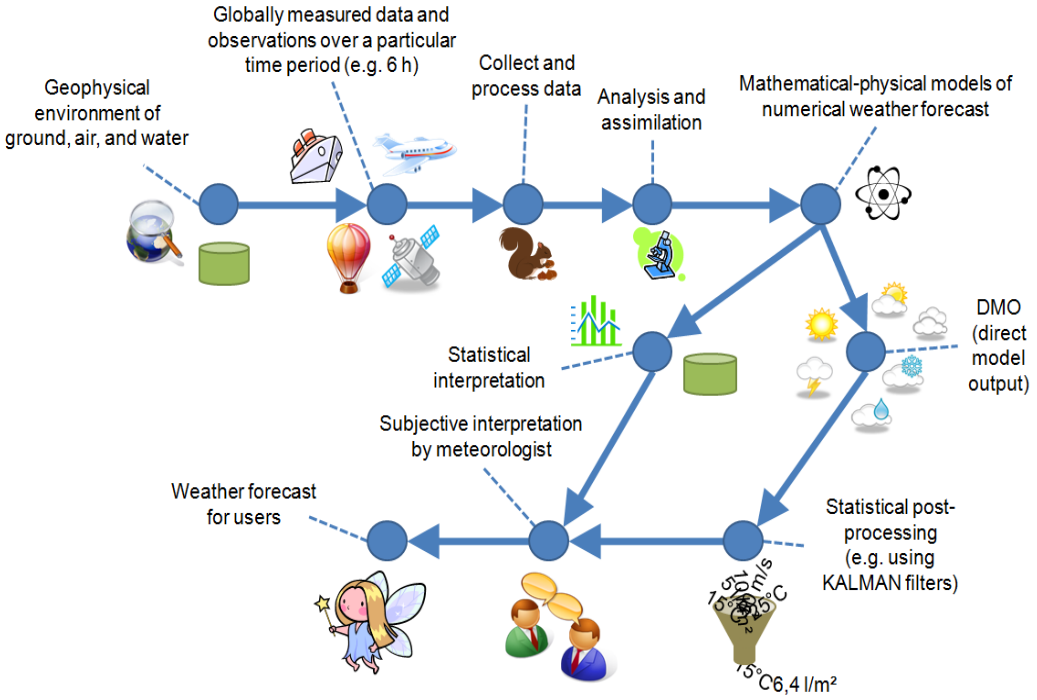

Globally measured data and observations over a particular time period (e.g. 6 h)
Collect and process data
Analysis and assimilation
Mathematical-physical models of numerical weather forecast
Geophysical environment of ground, air, and water
DMO (direct model output)
Statistical interpretation
Subjective interpretation by meteorologist
Weather forecast for users
Statistical post-processing (e.g. using KALMAN filters)
10 m/s
5°C
15°C
15°C6,4 l/m²

Figure 1: An example of scientific workflow (borrowed from [2]).

In doing that, the scientists need to curate such data sets by specifying metadata information that describe it. Hence, the workflow needs to address the evolving requirements and application because both service specification and implementation will evolve over the time [4]. Therefore, the workflow should support new capabilities in future during the preservation.

Partial reusability/fragment reusability

Not only the datasets obtained as output from running workflows can be reused by scientists in future, but also the part of the scientific workflow called « fragment » can be re-executed. It is not always feasible and not needed to execute a workflow in its original environment. Only parts of it are useful for new applications. In doing, the original workflow can be split/fragmented in many fragments, where some of them can be available for scientists for reusability [5].

Provenance quality

For long term preservation, it is necessary to ensure: the integrity of the workflow referring to the condition of being completed and unaltered workflow, and the authenticity of the workflow. Such relevant quality of information will be studied through investigating the workflow provenance tackling the space evolution.

Data preservation: Analysis

Workflow systems are increasingly used to define various scientific experiments. The number of new or reused workflows and the volume has increased significantly.

Workflow reuse can be seen in the following ways:
-Personal reuse: Building large workflows can be a long time process and use more complex functions. Keeping track and path of the relationships between workflow parts become a challenge, so versioning support is required for personal reuse.
-Reuse by collaborators: Researchers are often a member of community research group or collaborative project, inside of which they exchange knowledge.
-Reuse by third party: The research group is distributed across the world, and people get insight and input from experiments done by colleagues they never met. Indeed, scientists have a lot of work already modeled as workflows.

A large part of these workflows could be derived from existing workflow. Thus, if we could compare/analyse existing workflows, we would be able to structure the experience knowledge. The availability of these processing chains or workflows creates new opportunities for the total or partial exploration and visualization. The ability to group similar workflows together has many important applications. Clustering can be used to automatically to partition and organize workflows. To better preserve the chain of treatment, it is necessary to organize them into homogeneous groups. The most obvious solution is to associate each workflow a keywords.

Therefore, the use of any search engine will provide a long list where users must examine the results sequentially to identify those that are relevant. Clustering the "workflows" in homogeneous clusters, allow users to have more comprehensive results and will quickly identify the information. Clustering and partitioning techniques are widely used in many different fields. These areas include, but are not limited to: document retrieval, image segmentation, graph mining and data mining. Clustering has also been applied in the context of business workflows to derive workflow specifications from sequences of execution log entries. The analysing workflows problem, however, remains largely unexplored.

In summary, three key elements are needed to analyse workflows:
- A model to represent workflow elements: A workflow can be represented as a graph structure where each node is associated to the information (input, output data). Hence the workflow is considered as complex and mixed data.
- A similarity measure: according to the representation of the workflows as a graph and multidimensional data or mixed data, a specific distance can be used or redefined especially these workflow.
- An analysing algorithm: The algorithm selection is important. The challenge in the context of workflow preservation is to adapt some existing algorithms.

References :

[1] Carole A. Goble, and David De Roure The impact of workflow tools on data-centric research.. The Fourth Paradigm,Microsoft Research, (2009)

[2] Mirko Sontag, Dimka Karastoyanova: Model-as-you-go: An Approach for an Advanced Infrastructure for Scientific Workflows. J. Grid Comput. 11(3): 553-583 (2013).

[3]David De Roure, Khalid Belhajjame, Paolo Missier, José Manuel Gómez-Pérez, Raúl Palma, José Enrique Ruiz, Kristina Hettne, Marco Roos, Graham Klyne, Carole Goble (2011). Towards the Preservation of Scientific Workflows. in Proc 8th International Conference on Preservation of Digital Objects (iPRES 2011)

[4] Vasilios Andrikopoulos, Salima Benbernou, Michael P. Papazoglou: On the Evolution of Services. IEEE Trans. Software Eng. 38(3): 609-628 (2012)

[5]Mehdi Bentounsi, Salima Benbernou, Cheikh S. Deme, Mikhail J. Atallah: Anonyfrag: anonymization-based approach for privacy-preserving BPaaS. Cloud-I 2012: 9

[6] Emanuele Santos, Lauro Lins, James P. Ahrens, Juliana Freire, and Claudio T. Silva. 2008. A First Study on Clustering Collections of Workflow Graphs. In Provenance and Annotation of Data and Processes, Juliana Freire, David Koop, and Luc Moreau (Eds.). Lecture Notes In Computer Science, Vol. 5272. Springer-Verlag, Berlin, Heidelberg 160-173.

[7] V Silva, F Chirigati, K Maia, E Ogasawara, D Oliveira, V Braganholo, L Murta, M Mattoso. Similarity-based Workflow Clustering. Journal of Computational Interdisciplinary Science (2011) 2(1): 23-35

Contact:

Salima Benbernou, Laboratoire d'Informatique Paris Descartes LIPADE, Université Paris 5;
salima.benbernou@parisdescartes.fr
Mustapha Lebbah , Laboratoire d'Informatique Paris Nord, Université Paris 13;
mustapha.lebbah@univ-paris13.fr

Long Term Archiving and CCSDS standards

Danièle Boucon

Abstract: This article[2] presents some conceptual and implementation CCSDS –Consultative Committee for Space Data Systems- standards for long term archiving. It focuses on the most recent one, the Producer Archive Interface Specification (PAIS) standard. This standard, currently available as a draft on the CCSDS web site, will be published by the beginning of 2014. It will enable the Producer to share with the Archive a sufficiently precise and unambiguous formal definition of the Digital Objects to be produced and transferred, by means of a model. It will also enable a precise definition of the packaging of these objects in the form of Submission Information Packages (SIPs), including the order in which they should be transferred.

Context for space scientific data

For 40 years, in CNES (Centre National d'Etudes Spatiales, French Space Agency), a large number of space missions have been producing a huge amount of data (hundreds of Tb). These data constitute a valuable heritage that must be preserved because many of them are unique - related to an event that will never happen again or for a very long time (e.g. Halley comet period is 76 years!). These data could be integrated in long cycles of observations, including cycles for climate change observation and may be mandatory to prepare future missions (e.g. GAIA benefits from HIPPARCOS experience). With the arrival of new missions, this amount of data will further increase in volume and complexity.

In the space sector, archiving can be set up at different levels depending on the organizational structure implemented, such as within a mission control system or with a multi-mission Archive of scientific data such as the NSSDC (National Space Science Data Center), the PDS (Planetary Data System) or the CDPP (Plasma Physics Data Center). The context is complex, most often involving international cooperation with an ever increasing diversity of the Producers. Even if the lifespan of a space project is between ten and twenty years, data have to be preserved over an unlimited period. Furthermore, the data Producers are located all over the world.

Overview on standards

In this context, the CNES, with other organizations (NASA, ESA, BnF, …), actively participates in the CCSDS - Consultative Committee for Space Data Systems. The CCSDS has produced major standards such as (all are available on the CCSDS website at www.ccsds.org):

- the OAIS - Open Archival Information System- (http://public.ccsds.org/publications/archive/650x0m2.pdf)
- the Audit and Certification of Trustworthy Digital repositories (http://public.ccsds.org/publications/archive/652x0m1.pdf)

[2] Invited contribution.

- the PAIMAS -Producer Archive Interface Methodology Abstract Standard- (http://public.ccsds.org/publications/archive/651x0m1.pdf)
- the PAIS – Producer Archive Interface Specification (publication planned for March 2014, before ask daniele.boucon@cnes.fr)
- the XFDU -XML Formatted Data Unit- (http://public.ccsds.org/publications/archive/661x0b1.pdf), and
- the DEDSL -Data Entity Dictionary and Specification Language- (http://public.ccsds.org/publications/archive/647x1b1.pdf),

The Reference Model for an OAIS identifies, defines, and provides structure to the relationships and interactions between an information Producer and an Archive. The PAIMAS, PAIS, and the certification standards are linked to the concepts and functions introduced in the OAIS.

The Audit and Certification of Trustworthy Digital Repositories defines a standard which provides metrics on which to base an audit for assessing the trustworthiness of digital repositories. The scope of application of this document is the entire range of digital repositories.

The PAIMAS is a methodological standard that identifies four phases of a Producer-Archive Project (i.e, the set of activities and the means used by the Producer as well as the Archive to ingest a given set of information into the Archive): Preliminary, Formal Definition, Transfer, and Validation phases. The phases follow one another in a chronological order. The Preliminary Phase includes a preliminary definition of the objects to be archived, a first definition of the SIPs, and finally a draft submission agreement. The Formal Definition Phase includes a complete SIP definition with precise definition of the objects to be delivered, and results in a Submission Agreement. The Transfer Phase performs the actual transfer of the SIPs between the Producer and the Archive. The Validation Phase includes the actual validation of the SIPs by the Archive and any required follow-up action with the Producer. Each phase is itself further broken down to end up with series of actions, some of which can be performed independently of one another: the methodology comprises some thirty action tables taking into account many possible factors in the negotiation. This standard is at the interface between the Producer and the ingest OAIS functional entity.

The PAIS implements part of the PAIMAS. It implements the model of the data to be transferred, SIP specification and creation. The PAIS is described in more detail in the remainder of this article.

The XFDU is a packaging standard for data, metadata, and software, into a single package (e.g., file, document or message) to facilitate information transfer and archiving. It provides a full XML schema.

The DEDSL, Data Entity Dictionary and Specification Language, defines the abstract definition of the semantic information that is required to be conveyed. The DEDSL standard presents the specification in a layered manner (attributes, entities, dictionaries). This is done so that the actual technique used to convey the information is independent of the information content and, therefore, the same abstract standard can be used within different formatting environments. The DEDSL standard also specifies the way to extend the language itself (e.g. how to add attributes and preserve interoperability). This also permits the semantic information to be translated to different representations as may be needed when data are transferred across different domains.

CNES has developed methods, requirements on data and Archive, and tools for data preservation based on the CCSDS standards. These tools include:

BEST framework, available at http://logiciels.cnes.fr/BEST/FR/best.htm,
SITools2, available at http://sitools2.sourceforge.net/, and
the SIPAD-NG a generic Electronic Archiving System for accessing scientific data.

PAIS, the new CCSDS standard for transferring data between a Producer and an Archive

The primary objective of the Producer-Archive Interface Specification (PAIS) standard is to provide concrete XML files supporting the description and the control of transfers from a Producer to an Archive.

A transfer, as seen by the PAIS standard, is the movement of Data Objects from a Producer to an Archive. The Data Objects are not transferred as independent plain items but rather they are grouped and encapsulated in higher level objects known as Submission Information Packages (SIPs) thereby providing better control in term of content types, fixity information, inter-relationships and sequencing as outlined in the following figure 1.

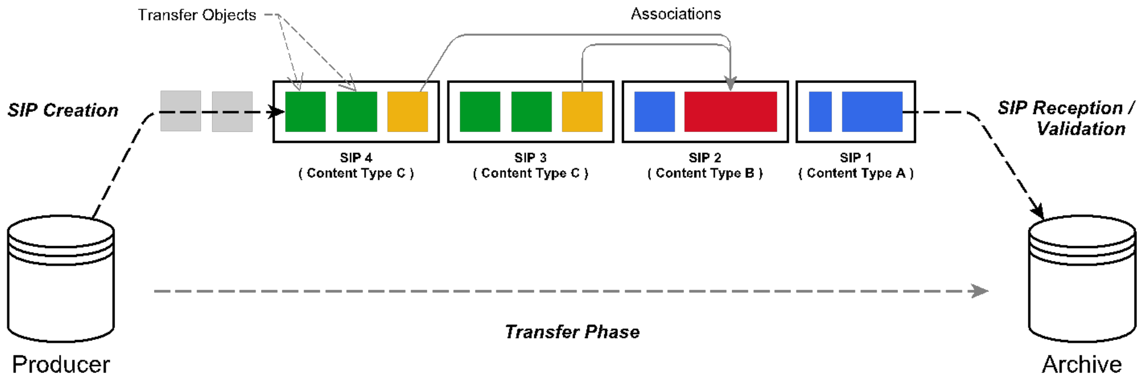


Figure 1: Example of Transfer

The Producer is responsible for the creation of SIPs according to content types agreed with the Archive and for their submission in a sequencing order that may also have been negotiated with the receiving Archive. In the example above, the Producer has generated and submitted four SIPs, one of Content Type A, the second of Content Type B and the remainders of Content Type C. As suggested by their names, the Content Types govern the actual content allowed for a SIP in terms of structure and data format.

According to the PAIS standard the content of the SIPs are decomposed in Transfer Objects (depicted as colored boxes in the figure 1 above) holding one or more trees of Groups (usually denoting folders) organizing the Data Objects (usually a single file or a small set of files) that are the subject of the transfer. A typical example of Transfer Object could be an Earth Observation product composed of various metadata and data files (i.e. the Data

Objects) organized in a tree of folders (i.e. the Groups). The PAIS standard supports the control of these objects through the description of their types, namely the Transfer Object Types, Group Types and Data Object Types.
According to the PAIS, the definition of these Content Types is given by a “SIP Constraints” XML document that can be as short as the following:

```
<sipConstraints xmlns="urn:ccsds:schema:pais:1">
  <producerArchiveProjectID>MyProject</producerArchiveProjectID>
  <sipContentType>
    <sipContentTypeID>Content Type A❶</sipContentTypeID>
    <authorizedDescriptor>
      <descriptorID>Blue Descriptor ID❷</descriptorID>
      <occurrence>❸
        <minOccurrence>2</minOccurrence>
        <maxOccurrence>2</maxOccurrence>
      </occurrence>
    </authorizedDescriptor>
  </sipContentType>
</sipConstraints>
```

This “SIP Constraints” document shall include all the Content Type definitions although only the Content Type A ❶ has been described in the example for simplicity. This Content Type A accepts only one Transfer Object Type identified as “Blue Descriptor ID” ❷. The example also defines that two and only two objects of this type are expected per SIP of this Content Type ❸. The “SIP Constraints” document can also define the sequencing constraints, for example, to force the transfer of SIPs of Content Type B prior to those of Content Type C.

The “Blue Descriptor ID” ❷ refers to a Transfer Object Type that has to be defined in a separate “Transfer Object Type descriptor” XML document as the following one:

```
<transferObjectTypeDescriptor xmlns="urn:ccsds:schema:pais:1">
  <identification>
    <descriptorModelID>CCSD0014</descriptorModelID>
    <descriptorModelVersion>V1.0</descriptorModelVersion>
    <descriptorID>Blue Descriptor ID❶</descriptorID>
  </identification>
  ...
  <groupType>❷
    <groupTypeID>Blue Group</groupTypeID>
    ...
    <dataObjectType>❸
      <dataObjectTypeID>Blue Data Object</dataObjectTypeID>
      ...
    </dataObjectType>
  </groupType>
</transferObjectTypeDescriptor>
```

The descriptor clearly declares the “Blue Descriptor ID” ❶ and the content tree composed of one “Blue Group” Group Type❷ holding one “Blue Data Objet” Data Object Type❸. Some parts of the example have been truncated and replaced by “…” for simplicity. Those parts are dedicated to the control of the occurrences, sizes and associations between the types. Some but not all of those parts are optional.
In addition, the PAIS standard specifies the minimal set of metadata that shall be attached to a SIP for the complete typing of all the objects it contains i.e. the mapping of the objects to

the PAIS descriptor types. The PAIS standard also defines a default SIP format based on the CCSDS XFDU recommended standard. in the XFDU implementation, the SIPs are containers of any type (i.e. usually a ZIP archive or a root folder), that hold the Data Object files organized in an arbitrary number of nested folders. This structured dataset is accompanied by an XFDU Manifest XML document that registers all the Data Objects and, when specialized as defined by the PAIS, univocally identifies their types in the PAIS Producer-Archive Project i.e. the PAIS Data Object Types, Group Types, Transfer Object Types, SIP Content Type, etc.

The list of methods for writing PAIS descriptors is endless and none may fit with all contexts as for many standards. Nevertheless, the following workflow gives an overview of the major steps that are usually addressed during a project definition:

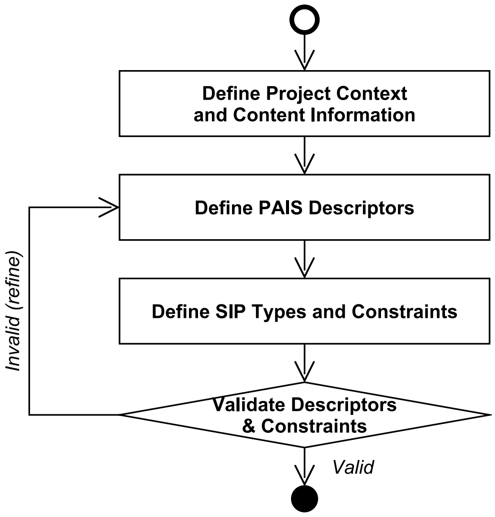

Define Project Context and Content Information
Define PAIS Descriptors
Define SIP Types and Constraints
Validate Descriptors & Constraints
Invalid (refine)
Valid

Figure 2: Typical steps driving a PAIS Producer-Archive Project definition

Finally, a Producer-Archive Project can benefit from the PAIS standard by writing a set of XML documents according to a formal XML language, validate these descriptors against XML Schema documents provided in annex of the standard and develop or reuse tools for building, transferring, receiving and validating SIPs.

PAIS, preservation process and data lifecycle

The previously cited standards are used at different steps of a data preservation process, either on the Producer side or on the Archive side. The data lifecycle is covered by the 3 main phases: preparation for data production, data production (generally beginning with the satellite launch), and data preservation.

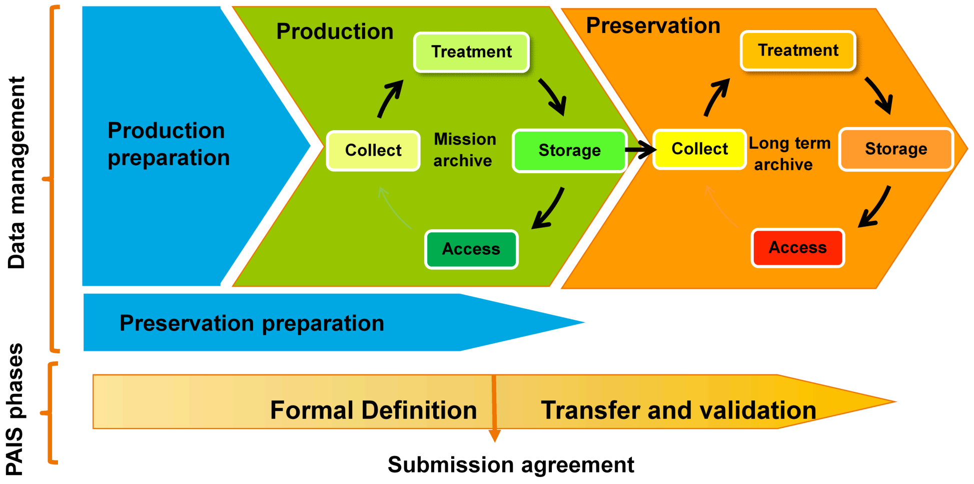


Figure 3: PAIS, preservation process and data lifecycle

Figure 3 is a high level view of the data lifecycle . It is recommended that a data preservation plan be prepared early in the data lifecycle, rather than at the point of withdrawal from active systems. During the production, data are created from a collection of data (for example raw data with orbit data and other parameters), processed, stored in the mission archive, and according to legal constraints, they are published. Once they are stabilized and have been validated, the data items planned to be preserved may be transferred to the long term archive. The treatments in this phase may be conversions to other formats for example.

In this schema, the PAIS Formal Definition phase takes place during the preparation of the preservation (model of the objects to be transferred, SIPs specification), while the Transfer and Validation phases are the first steps of data preservation. The last step is the archive maintenance in order than the data remain usable on the long term, even if the user community or the systems evolve.

All this should be defined in the future CCSDS project on Data Preservation Process. Its purpose is to provide a standard method structured as a complete process to formally define the steps and the associated activities required to preserve digital information objects. The process thus defined along with the activities, is linked with the data lifecycle. This project is planned to begin in January 2014.

Conclusion

The CCSDS standards provide methods, concepts and implementation for long term archiving. Among them, the PAIS provides an implementation to help in the negotiation between the Producer and the Archive, in the automation and management of the transfer and in validation of the Digital Objects by an Archive. It should be published as a CCSDS standard in the beginning of 2014 and should become an ISO standard later in the same year. The use of these standards should provide better quality for archived data, and should reduce the cost of the operation.

References

[1] Reference Model for an Open Archive Information System (OAIS), Recommendation for Space Data Systems Standards, CCSDS 650.0-M-2, Magenta Book. Issue 2, May 2012. [Equivalent to ISO 14721:2012].

[2] Producer Archive Interface Specification, the new CCSDS standard for modeling the data to be transferred to, and validated by, an Archive, PV 2013, Danièle Boucon, http://www.congrexprojects.com/2013-events/pv2013/welcome

Contact :

Danièle Boucon – Centre National d'Etudes Spatiales CNES, Centre spatial de Toulouse; Daniele.Boucon@cnes.fr

Cloud and grid methodologies for data management and preservation

Christophe Cérin, Mustapha Lebbah, Hanane Azzag

Abstract: As data sets are being generated at exponential rate all over the world whatever the disciplinary field (science, engineering, commercial), Big Data has become a Big issue for everybody. While IT organizations are capturing more and more data than ever, they have to rethink about and figure out what to keep for a long time and what to permanently archive. Moreover, the meaning to give to data can be obtained through novel and evolving algorithms, analytic techniques, and innovative and effective use of hardware and software platforms. In this contribution we investigate the coupling between Grid and Cloud architectures as well as the impact of machine learning on programming languages for harnessing the data, discovering hidden patterns, and using newly acquired knowledge that has to be preserved. We do not consider only the archive stage but examine the life cycle of data as a whole; one of the last effort is to decide what we need to preserve.

The landscape

Even if we restrict our concern to the field of scientific data, we first need to consider the 'business process' and to accept that we all share a common interest: first, putting the data close to the computation. Second, mine, analyse... the data. Third, archive what we think to be imporant.

The last 15 years have taught us that, because of the 'business process', data are traveling from infrastructures (clusters) to infrastructures (clusters again) in order to be calibrated, analyzed, visualized... From an architectural point of view we have built Grids and the notion became a success story, we are thinking about the EGEE project for instance (see http://en.wikipedia.org/wiki/European_Grid_Infrastructure).

In this contribution, we analyze the life cycle of data as we understand it nowadays in E-sciences but with the novel architectures and programming styles in mind. We do not specialize our comments on the archive part of the 'business process' but we consider the life cycle as a whole, the preservation of data being one aspect of the problem.

Beyond the architecture, we (the computer scientists) shall also notice that, for a while, we have switched from the design of programs (sorting, searching...) to the design of middleware. We remind here, as quoted by wikipedia, that "Middleware is computer software that provides services to software applications beyond those available from the operating system. It can be described as software glue".

Among the success stories in Grids, we need to specifically mention the Globus toolkit project (http://en.wikipedia.org/wiki/Globus_Toolkit). The Globus Toolkit is an implementation of the following standards (their names give the type of services they offer) that need to be addressed with Grids: Open Grid Services Architecture (OGSA), Open Grid

Services Infrastructure (OGSI), Web Services Resource Framework (WSRF), Job Submission Description Language (JSDL), Distributed Resource Management Application API (DRMAA), Web-Service-Management, Web-Service-BaseNotification, Simple Object Access Protocol (SOAP), Web Services Description Language (WSDL), Grid Security Infrastructure (GSI).

All these services run concurrently but few of them have distributed/parallel implementations. Parallel implementations are still for programs (numerical algorithms for instance). Regarding the data movement and management, the Globus Toolkit implements also the OGF-defined protocols to provide Global Access to Secondary Storage (GASS) and GridFTP (file transfers).

A new context: mixing Grid an Cloud ideas

As noticed previously, several tools and frameworks have been developed to manage and handle the big amount of data for the Grid platforms. However, the use of these tools by the basic scientist and the Grid computing community is not well adopted by 'basic' users because of the complexity of the installation and configuration processes.

In order to process large data-sets, users need to access, process and transfer large data sets stored in distributed repositories. The users get difficulties to manage easily their data. For instance, to move data from its site to the experimental platform (cluster or computational grids), the user must install client software tools and place data by hand, using simple scripts through the command line interface. To accomplish this task, the user must necessarily have a knowledge about data management technologies and transfer protocols such as scp, rsync, FTP, SRM tools, Globus GridFTP, GridTorrent etc.

This is our first observation. Our second observation is that Cloud is more than a buzz word, in particular if we consider the following matter of concern. We assume first that the queries about "HPC clouds" are a little vain because the Cloud is always analyzed from HPC point of views which is a bias, not fair, and we always conclude that Clusters are far superior to Clouds (the sole metric for analyzing the architecture is performance hence the conclusion!). Second, we think that "HPC in the cloud" is the real concern. But, in this case we have to explain the metric used to compare the possible options or, better, to explain the utility that we can give to Clouds.

We would like to emphasize here that the big challenge with Clouds is to put the user in the middle of our concerns and to automate, as much as possible, the tasks of the 'business processes'. To our view, this is the essence of Clouds.

We now discuss our experiences with this vision in mind. In other words, we suggest to give more control to the user, motivated by the fact that he really needs to control his experiment without being disturbed by a system administrator that imposes his vision.

Since basic users lack the fundamental IT and networking knowledge, they spend too much time to download, install, configure and to run the experiment. Hence our arguments:

- To achieve data management on demand, the users need a resilient service that moves data transparently;

- No IT knowledge required, No software download/installation/configuration steps.

With the above requirements in mind, we have implemented a system based on the following technologies:

- Stork (http://stork.cse.buffalo.edu/): Stork is a batch scheduler specialized in data placement and data movement, which is based on the concept and ideal of making data placement a first class entity in a distributed computing environment. Stork understands the semantics and characteristics of data placement tasks and implements techniques specific to queuing, scheduling, and optimization of these type of tasks.
- Bitdew (http://www.bitdew.net/):The BitDew framework is a programmable environment for management and distribution of data for Grid, Desktop Grid and Cloud Systems. BitDew is a subsystem which can be easily integrated into large scale computational systems (XtremWeb, BOINC, Hadoop, Condor, Glite, Unicore etc..). Our approach is to break the "data wall" by providing in single package the key P2P technologies (DHT, BitTorrent) and high level programming interfaces.
- SlapOS (http://www.slapos.org): the SlapOS Cloud presents a configurable environment in terms of the OS and the software stack to manage without the need of virtualization techniques. By the way, it is a bit strange that people coming from HPC, optimization and sober resource management defend at this point the concept of virtual machine. We do prefer to count on the operating system rather building more and more software layers. SlapOS reuses, in part, some concepts of Desktop Grids: optionally, machines at home may host services and data, a master contains a catalog of services and publishes them in a directory on a slave node. The SlapOS vision of a Cloud is a) an ERP (Enterprise Resource Planning) b) a model of deployment c) nodes to host and run services (among them, a compute service if we need it). This is an orthogonal vision of Cloud computing, meaning that we anchor it in the field of Services rather than HPC.

Thus, our data management system is made of the following components (see figure 1) for an architectural overview):

- Stork data scheduler: manage data movement over wide area networks, using intermediate data grid storage systems and different protocols.
- Bitdew: make data accessible and shared from other resources including end-user desktops and servers.
- SlapOS: with only a one-click process, instantiate, configure data managers (Stork+Bitdew) and deploy them over the Internet.

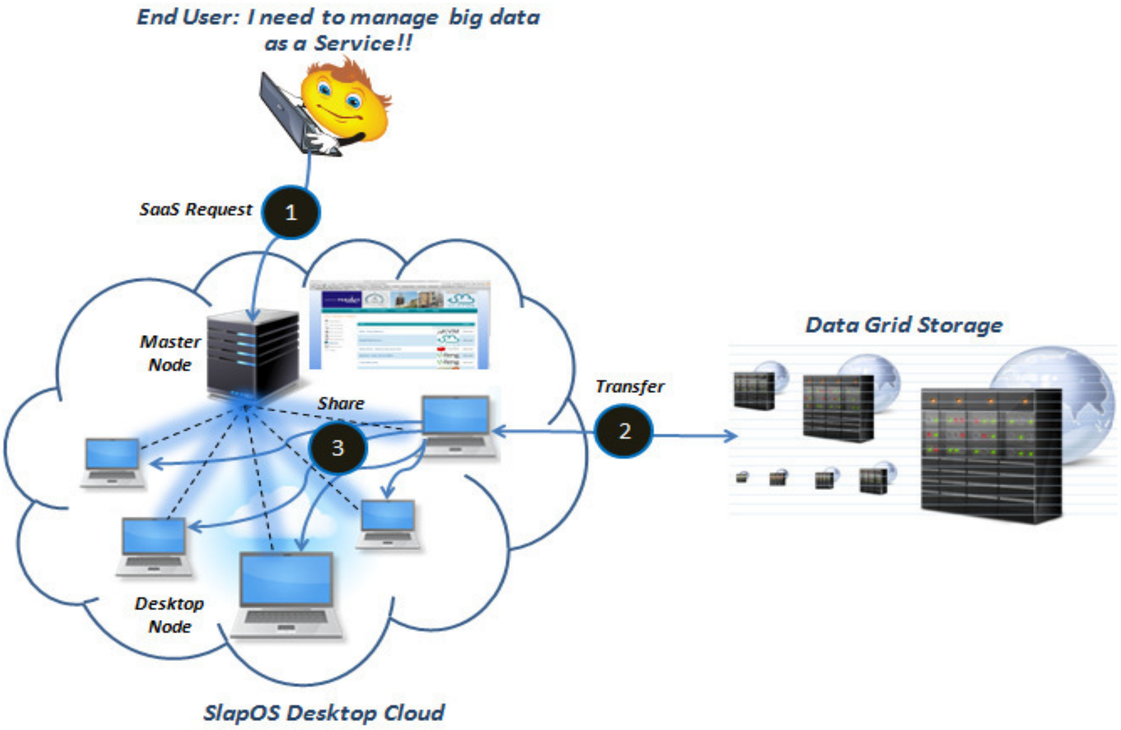


Figure1: Our approach overview: The user utilizes web interface to (a) interact with SlapOS master; (b) deploy data transfer tools (Stork) to move data from remote grid storage to SlapOS (c) Share data inside SlapOS cloud and (d) perform simulations (or a specific processing) on data already published.

Our system allows, for distributed data-intensive applications, to deal with recovering data from outside (remote storage server) and sharing data with a large number of nodes inside Cloud infrastructure with the least effort. Our design and implementation of these two services, make the users to request and install automatically any data movement and sharing tools like Stork and Bitdew without any intervention of a system administrator.

Impact of architectures on data mining and machine learning

Data mining problems have numerous applications and are becoming more challenging as the size of the data increases. Nevertheless, good mining algorithms are still extremely valuable, because we can (and should) rewrite them for making them as parallel algorithms using the MapReduce paradigm for instance.

In situations where the amount of data is prohibitively large, the MapReduce (MR) programming paradigm is used to overcome this problem. Thus, in recent years, an increasing number of programmers have migrated to the MapReduce programming model. The MR programming model was designed to simplify the processing of large files on a parallel system through user-defined Map and Reduce functions. A MR function consists of two phases : a *Map* phase and a *Reduce* phase. During the Map phase, the user-defined Map primitive transforms the input data into (key, value) pairs in parallel. These pairs are stored and then sorted by the system so as to accumulate all values for each key. During the Reduce phase, the user-defined Reduce primitive is invoked on each unique key with a list of all the values for that key; usually, this phase is used to perform aggregations. Finally, the results are output in the form of (key, value) pairs. Each key can be processed in parallel during the Reducephase.

Hadoop (http://www.hadoop.com), an open-source implementation of the MR programming model, has emerged as a popular platform for parallelization. A user can perform parallel computations by submitting MR jobs to Hadoop. While the Hadoop framework is very popular in their particular domains, we believe that it has a set of limitations that make it ill-suited to the implementation of parallel data mining algorithms. Many common data mining algorithms apply a single primitive repeatedly to the same dataset to optimize a parameter. Thus the Map/Reduce primitives need to reload the data, incurring a significant performance penalty.

Existing programming paradigms for dealing large-scale parallelism such as MapReduce and the Message Passing Interface (MPI) have been the choices for implementing these machine learning algorithms. MapReduce is the most popular suited for data already stored on a distributed file system, which offers data replication as well as the ability to execute computations locally on each data node. However, the existing parallel programming paradigms are too low-level and ill-suited for implementing machine learning algorithms.

To address the challenge some authors present a portable infrastructure that has been specifically designed to enable the rapid implementation of parallel machine learning algorithms. Recently, a MapReduce-MPI library was made available by Sandia Lab to ease porting of a large class of serial applications to the High Performance Computing (HPC) architectures dominating large federated resources such as NSF TeraGrid, which is used to create two open-source bioinformatics applications and to explore MapReduce for clustering task.

In our case, we use another emerging open-source implementation named Spark (http://spark-project.org/), which is adapted to machine learning algorithms and supports applications with working sets while providing similar scalability and fault tolerance properties to MapReduce. The main questions are (a) how to minimize the I/O cost, taking into account the already existing data partition (e.g., on disks), and (b) how to minimize the networking cost among processing nodes.

Conclusion

In some way, we believe that we enter to a post-era where cores/CPU/Nodes are unlimited in number as well as storage. We need to pay attention where the data are stored - we hope that big companies will not capture (all) the markets related to data. We also need to pay attention to the user and make sure that he will be able to imagine and deploy experimental scenarios on large scale distributed infrastructures in a simple and natural manner. It is a necessary condition for the adoption of the new paradigms, both architectural and programming paradigms, by large communities of users, in particular those that, one day, decide to preserve one part of data.

References

[1] Walid Saad, Heithem Abbes, Christophe Cérin, Mohamed Jemni: *A Self-Configurable Desktop Grid System On-Demand*. 3PGCIC, Victoria, BC, Canada; 2012: pp 196-203.

[2] Christophe Cérin, Walid Saad, Heithem Abbes and Mohamed Jemni, *Designing and Implementing a Cloud-Hosted SaaS for Data Movement and Sharing with SlapOS*, in submission.

[3] Tugdual Sarazin, Mustapha Lebbah, Hanane Azzag. *SOM Clustering at Scale using Spark-MapReduce*. in submission.

[4] Nhat-Quang Doan, Hanane Azzag, and Mustapha Lebbah. *Growing Self-organizing Trees for Autonomous Hierarchical Clustering,* Neural Networks. Special Issue on Autonomous Learning. Volume 41, May 2013, Pages 85–95. Elsevier.

Contact:

Christophe Cérin, Mustapha Lebbah, Hanane Azzag;
Laboratoire d'Informatique Paris Nord, Université de Paris 13;
christophe.cerin@lipn.univ-paris13.fr
mustapha.lebbah@lipn.univ-paris13.fr
hanane.azzag@lipn.univ-paris13.fr

Scientific Data Preservation, Copyright and *Open Science*

Philippe Mouron

Abstract: The purpose of this paper is to sum up the terms of a discussion about the legal aspects of scientific data preservation. This discussion was presented at the Marseille workshop organized on November 14th. This paper is only a basis for forthcoming works about the main project of preserving scientific data (PREDONx). The paper is focused on intellectual property rights, such as copyright or patent, and their effect on the use of scientific data. *Open Science* appears to be the best way to ensure the preservation, but also the publication, of scientific data.

The use of information technologies has significantly improved the preservation of scientific data. The development and networking of digital storage spaces can ensure the integrity of these data, as well as their access to the researchers interested in the results of the scientific research. However, the will of preserving scientific data on a long term is not so new. The work archiving policy has in fact always existed but only for tangible formats (paper, samples,...). However, the physical size of the archives is not infinitely expandable. This limit could jeopardize their preservation, because it implies the necessity to select data which are going to be preserved. Those which remain are mostly lost. Moreover, the access to these data is limited in regard to the scarcity of copies and the cost of their public release. Modern technologies have therefore been remedied these disadvantages, but also moved the heart of the problem on another ground. Thus, it is less a question of preserving than giving access to data that arises. Conservation is finalized by the purposes of research, involving pooling of works and sharing their results.

However, this point leads to legal considerations. For lawyers, 'Preservation' means 'reservation'. The best guarantee for ensuring the integrity of a resource is based on property. Affecting a property right to tangible and intangible things tends to optimize their conservation and especially their exploitation. These results seem easiest to obtain with a private ownership model. Historically, the enhancement of tangible goods was placed into the fold of this model, for reasons of efficiency. The same reasons have pushed to affect the results of the scientific work of intellectual property rights. That’s why copyright law and patent law are such mobilized in the field of research. This was the subject of a very recent act in France, which purpose is to make easier the use of these property rights, from a purely economic perspective.

However, isn’t there a public ownership of scientific research? We know that these works depend on public funding, even through the status of institutions and researchers. It seems logical that the research results have to belong to the community, which is their main fundraiser. In truth, even if the public authorities may fundamentally participate in the scientific research, this does not mean, *ipso facto*, that they own its results. Of course, the reference to a particularly renowned institution may increase the value of its work, but it will be purely moral in legal terms. The idea is based on another fundamental principle: freedom of research. This freedom, so essential to the scientific field, applies primarily to the natural persons who take part in research. It founds the confrontation of ideas and works between researchers. Such freedom would not be effective without the private ownership of their work. That's why public authorities have only a promoting or incentive role in that field.

The goal of digital preservation of scientific data must therefore be reconciled with intellectual property rights. In the first part, we will examine the kinds of data that shall be concerned. Then, we will see how the intellectual property rights shall be managed in order to facilitate the preservation of scientific data.

Typology of scientific data under intellectual property rights

There are several types of scientific data. Each category is the matter of a specific legal regime, which may have different effects on the use of these data.

The first set is composed of elements that do not fall under any intellectual property right. 'Raw' data are mostly concerned in that first type of data. It can be described as the results of the scientific research regardless of any treatment. In other words, this category consists of objective data resulting from observation of nature and not from a creation of the mind. For example, statistics and mathematical data shall be considered as such. It is the same with geography or astrophysics, because the form and the relief of continents and the layout of the stars are dictated by nature itself. Other elements could be added to theses data as subjects of research, but which are legally considered as discoveries and not creations. Scientific theories, algorithms and other mathematical formulas are included in that field. Programming languages are also classified in this category. All these scientific data are insusceptible to exclusive ownership, for the main reason that has just been set out: they are not creations of the mind. Moreover, the interests of scientific research justifies that they remain completely free. Everyone can access these data and use it for new works. Therefore, the preservation of these data in an open circuit, or open space, is perfectly free, as long as they remain in their original form. They are some kind of common goods, without any ownership. However, theses data are never completely free. Whenever a human treatment can be found in their processing, we come to the second category of data. These include all creations of the mind. In scientific matters, they can be of two kinds.

On one hand, there are intellectual works that are subject to a copyright, on the other hand, patentable inventions. The first are works of mind; they were traditionally classified into the literary and artistic fields. Now, purely technical creations can also be classified in that category, such as software and documents required for their use. Beyond that, all the scientific works are concerned. By 'scientific work' we mean every work including a personal treatment of the first kind of data. So, any paper, article, report, record, thesis, book, graphic, map,... conducting personal choices of a researcher, or expressing his own personality, will be considered as a work of mind. It is the same with databases, whose architecture can also be the result of a creative work. That's why all these works are copyrightable. As such, it is important to understand the purpose of copyright law. The raw data contained in these documents are not the subject of copyright; it is only the formatting of these data, the personal treatment embodied in a document. If we take the example of a doctoral thesis, the theories, statistics and formulas which are employed will remain free, but the text, the plan of the work and all of its arrangement will constitute the work of mind created by the candidate. As we will see in the second part, these elements will be subject to an intellectual property statute, so that their conservation and public display are under the owner’s right.

The same conclusion is implied from the other category of data we have cited, that is to say those that are part of a patentable invention. All technical creations that bring a solution to a technical problem can be the object of a patent if they are capable of industrial or commercial application. It is not unusual that such inventions are developed by academic institutions. It implies again the issue of a long-term preservation policy for data. If the communication of some kind of data related to the invention seems acquired by the effect of the patent, their reuse may be limited under the monopoly granted to the holder. In some cases, the publication may even be blocked if it could threaten the exploitation of the invention.

These considerations require to consider the impact of the intellectual property rights on the objective of scientific data preservation.

Management of intellectual property rights and scientific data preservation

Digital archiving presents, as we have seen, a great interest for research. It would be however meaningless in the absence of access to the data. This imperative must be conciliated with the rights of intellectual property that we have just mentioned.

Ab initio, it is indispensable to collect the authorization of rights holders, but several ways can be used. These can be both physical and legal persons. It will be either from the authors of the considered works, either from institutions or companies for which property rights have been legally or contractually transferred. It is necessary to obtain an authorization in all cases, or the owner’s right will be infringed. For example, in 2012, a researcher from the French National Scientific Research Center has been convicted for counterfeiting, because he uploaded the first draft of a doctoral thesis without the authorization of its author. Despite the purpose of the researcher, who just intended to comment this work in a scientific framework, it wasn’t considered as a fair use.

As such, some Governments or institutions can sometimes encourage flexible models of diffusion. But they cannot legally compel holders of rights, which remain the only ones to accept or refuse such models. Therefore, respecting intellectual property rights requires a case by case approval of the rights holders. Only the raw data should be able to be displayed, but it is difficult to separate them from the research works in which they are included and treated. It would be the same if these data were presented in a database, because its own architecture may be the subject of related right. Most of all, we shall not forget that such authorizations may not be combined with some economic purposes. This is particularly the case in regard to the modern importance of promoting scientific research. By “promoting”, we mean “commercialization”, implying an economic exploitation of scientific results. This objective has specifically been the subject of a legal act in France (law n° 2013-660 of 22 July 2013). It is exclusive of open access to the data that is contained in these scientific works.

However, open access has also been the subject of different Acts, for public data in general and scientific data in particular. For example, the US Government launched the Open Government Initiative in 2009, in order to give more transparency to public affairs. In France, the same frame is applied to public data, with the website *data.gouv.fr*. But it has also been applied, for a long time, to scientific data. We refer to the open model of management of intellectual property rights. The main tool of this model lies in the so-called open licenses, which the most successful are *Creative Commons* licenses. This is all the more important that these models have sometimes been created by researchers. The interest of these licenses is to ensure conservation of the data in an open circuit, allowing access for a wide range of people, including researchers. Moreover, it allows reusing data for new works, which shall be available under the same rules. Different movements, based on these licenses, have been aimed at the creation of open archives for scientific research. Open science is a goal promoted on an international level, with famous achievements. The projects developed under the *Science Commons Foundation* are good examples. This model perfectly fits with the practices of researchers. Sharing works is the best way to ensure their preservation. Copying scientific data is free under theses licenses. It makes easier their dissemination among the scientific communities. As we have seen before, the use of these flexible models is now increasingly recommended by government institutions. In France, Section L 112-1 of the Research Code specifies the objectives of public research, which include “sharing and display of its own results, with giving priority to open access formats”.

To conclude, we understand that the use of this model of management is therefore an interesting perspective for the long term preservation of scientific data, beyond the official exceptions to property rights. However, it is important to conciliate open access with the promoting objective. It seems possible to consider a management of these rights, by applying both copyright, patent and open access tools to the different kind of data. Of course, the interest of this management implies to forget the two kinds of data we’ve presented in the first part. New criterias shall be found, especially for the second range of works. For example, we could consider the nature of data, or the duration of their publication. Some may be freely displayed on open archives, even at the time of their first publication. Others may be the subject of an exclusive exploitation for a while, to be later disseminated in an open access format. Similarly, it is possible to consider “circles” of publication around the same data. Different versions may be established according to their degree of treatment or precision. The more precise or personal it is, the more exclusive it would be. These circles would also include different numbers of associated researchers, according to the usefulness of data. A research can be conducted by a limited number of researchers, in order to keep some kind of secret on it, but the less significant data could be publicly released in open access.

Finally, this management is a gain of freedom for researchers for two reasons. First, they can use their own property rights by different ways and not only the legal one, which doesn’t really fit with the need of research. Then, the use of license opens the access to a wide range of data, at least raw data, but also personal works which would be copyrightable or patentable. Whatever the policy that shall be applied, these tools are essential to ensure the preservation of scientific data with internet technology.

References

[1] “Open science et marchandisation des connaissances” – Cahiers Droits sciences et technologies, n° 3, CNRS éditions, Paris, 2010, 444p.

[2] LAMBERT T., « La valorisation de la recherche publique en sciences humaines et sociales face au droit d’auteur des universitaires », D., 2008, pp. 3021-3027

[3] ROBIN A., “Créations immatérielles et technologies numériques: la recherche en mode open science”, PI, n° 48, juillet 2013, pp. 260-270

Contact:

Philippe Mouron, Aix-Marseille University, Faculté de droit et de science politique d’Aix-Marseille; philippe.mouron@univ-amu.fr

Chapter 3: Technologies

Storage technology for data preservation

Jean-Yves Nief

Abstract: Preservation of scientific data aims at storing data for many years or even decades. This is a challenge as hardware and software technologies are changing at a high rate with respect to the time scale involved in data preservation. Moreover, scientific data can be preserved in a distributed and heterogeneous environment involving several data centers. Storage and data policy virtualizations are strongly needed in such an environment, in order to achieve this endeavor. We will show that iRODS middleware can provide a suitable solution to the data storage and policy virtualization.

Storage challenges for data preservation

Computing centers dedicated to scientific projects usually provides storage services within a distributed environment. The experimental sites where data are produced by detectors, telescopes, microscopes etc…, as well as the collaborators and computing facilities involved in the data processing, can be spread around the world.

Data management in such an environment is a challenge for data centers as they may have to provide storage services within a multidisciplinary context. Each scientific field may have its own set of requirements and needs for data preservation. For instance, biomedical applications will require that medical records are kept anonymous. Documents stored in digital libraries may be subject to copyright rules. Whereas for other domains (such as astrophysics and even high energy physics), scientific data can be available in open access mode after a certain time. In any case, Dublin core metadata or other kind of metadata will be attached to the data in order to keep useful information on preserved data such as data provenance, checksum (for data integrity checks), creation time, ownership, experimental conditions etc…

Also, each data center relies upon its own set of technologies for data storage (file systems, mass storage systems, proprietary or homegrown storage solutions etc…). Storage media can be hard drive disks, SSD, tapes. The stored data may have to be migrated to new storage media several times during their lifetime. Numerous operating systems can be used both on the server and the client sides. Hence, the ecosystem on which data preservation has to be made can be very heterogeneous.

Moreover data can be stored in various formats: it can be flat files, databases, data streams, any kind of standard or homemade file formats. These file formats may evolve (and sometimes disappear) in the future. Hence file format transformation is part of the data preservation process. Storage systems involved must take this need into account: read access should be scaled properly in order to proceed to the reprocessing of the data.

Additionally, in order to insure safety and consistency in time, data should be replicated on several media or storage systems as well as in different data centers.

Towards storage data and data policy virtualization

Based on all the above constraints, it appears that there is a clear need for storage virtualization, in order to provide a unique logical view of the data and of its organization. This logical view should be totally independent on the data location, the kind of storage technology that is used underneath and protocols used to interact with these storage systems. This logical view should also be independent on the users or data preservation applications location, so they can navigate through directories content without having to bother about the files location (in a directory, the files could be located in different data centers or storage systems).

Users have to be organized within a virtual organization where each user has a unique identity. This virtual organization should also include groups and should be able to differentiate users' role depending on their privileges (e.g.: simple user, data curator, administrator etc…). Data accesses rights management is also mandatory.

But storage virtualization is not enough. For client applications relying on data virtualization middleware, there are no safeguards and no guarantee of a strict application of the data preservation policy. There are various pitfalls such as having several data management applications (or several versions of it) coexisting at the same time, each of them having their own set of policies (for data replication, data handling etc…): this can end up with potential inconsistencies in the policies applied on the data. The solution to these problems is to virtualize the data preservation policies: policies are expressed in terms of rules and are being defined on the middleware side, hence on the serve side. The management policies are then centralized and will be applied in a consistent way whatever applications are being used and wherever they are located. For example, let us suppose that one wants to replicate a certain type of data on three sites with one copy on tape; this policy will be expressed on the middleware side, therefore if an application ingest new data through the middleware, it won't be able to choose and override the replication strategy which has been set on the server side. A centralized and virtualized data management policy that nobody can overcome is a key point to the success of a data preservation project.

Middleware solution

A middleware solution for data management and long term preservation has to provide both storage virtualization and policy data management virtualization. Very few tools provide this kind of features or part of them. Among these tools, iRODS answers to all the requirements described above.

iRODS (iRule Oriented Data System) is an open source middleware developed by the DICE team collocated at UC San Diego and University of North Carolina at Chapel Hill, with contributions from external collaborators such as CC-IN2P3. Its scalability and flexibility allows it to be customized to fit a wide variety of use cases and is a good match to be part of a long term data preservation system.

With the help of the storage virtualization, one can move data to new storage devices in a transparent way from the point of view of the end applications. iRODS can be interfaced with a wide variety of storage and information systems (which can be distributed), providing a lot of freedom in the choice of the technologies one might want to use for data preservation projects. Metadata, access rights, auditing are also important features provided by iRODS for data preservation.

Data management policies are expressed in terms of rules on the server side: they are described using a language which allows creating a wide variety of policies. These data management policies can be triggered automatically in the background when for example someone ingests new files: hence, complex data workflow can be created that way.

iRODS is being used by a wide variety of projects. For instance, NASA and the French National Library are using iRODS. CC-IN2P3 is managing 7 Petabytes of data for High Energy Physics, Astrophysics, biology and Arts and Humanities with the help of iRODS.

Data preservation can span over decades. As storage technologies evolve on a much shorter timescale, it is important to provide storage virtualization and a rule oriented system such as iRODS, to get rid of technologies dependence at the higher layers of data management. Obviously, tools like this may also change or disappear in the future. These middlewares are only one layer in the data preservation process. The data managers have to provide an architecture with pluggable interfaces in order to easily switch from one data management tool to a newer one.

References

https://www.irods.org/index.php/Main_Page

Contact:

Jean-Yves Nief, Centre de Calcul de l’IN2P3, Lyon-Villeurbanne; nief@cc.in2p3.fr

Requirements and solutions for archiving scientific data at CINES

Stephane Coutin

Abstract: Historically an high-performance computing datacenter, the "Centre Informatique National de l'Enseignement Supérieur" (CINES) has also a mission of long term digital preservation. By coupling those two areas, it became obvious that CINES had to understand and take into account the requirements of scientific communities regarding their data life cycle, and more specifically their archiving requirements. We will present those requirements and describe the platforms CINES proposes for each service class.

CINES mission on long term preservation

Information in a digital form is now omnipresent in our society, with huge volumes and multiple formats. Despite its complexity and volatility, it's a genuine testimony of activities, an archive, for which preservation is a concern. Thus, the stakes of digital preservation are high, as they reside in the deployment of the means required to guarantee the heritage conservation, from the short term to the long term. The risks associated to digital information have now been identified for quite a while, and can be summarized in four main threats:

- The deterioration and ageing of storage media,
- The disappearance of read hardware or software,
- The impossibility of reading the format of the files containing the data,
- The lost knowledge of the content of digital objects.

Preserving digital data consists of preserving the document (while guaranteeing its integrity and authenticity), while keeping it accessible and understandable. The complexity of such a task is tightly bound to the preservation timescale. Within a few years period, the problem is relatively easy to deal with. Good quality and secure IT storage guarantees against accidental loss of the document. Technologies won't have changed so much that the document will have become irremediably unreadable. Finally, the community of potential users of the document will most likely be scientifically and culturally similar to the one which created the document a couple of years earlier, so the need for an exhaustive information description is not so strong. Within a wider period however, none of this is a foregone conclusion, unless someone has thought of accompanying the document over time, and requirements for comprehensiveness and legibility become mandatory.

For these reasons, the French ministry for Higher Education and Research gave the Centre Informatique National de l'Enseignement Supérieur (CINES) the mandate to implement and experiment a project in long-term preservation of records and data. CINES is a state administration institution based in Montpellier (France) which employs about 50 engineers and which is known worldwide for its HPC (high performance computing) activities. The whole CINES infrastructure and means is made available for all the French researchers, who are split up into scientific domains. The largest communities to use the CINES computing services are the fluid mechanics, chemistry and climatology research communities. As part of this initial mission, CINES hosts advanced computers which include Jade (SGI ICE 8200 EX with 267 TFlops peak, 23 040 cores and 700TB of disks).

The CINES mission for digital preservation eventually resulted in the deployment of one of the very few operational long-term preservation platforms in France for the Higher Education and Research community. In 2006, just two years after the first activities on digital preservation had begun, a first repository, which had been developed internally, was rolled out with the objective to preserve the electronic PhD theses. This infrastructure is called PAC (Plateforme d'Archivage du CINES – the CINES digital preservation system). Since March 2008, the documents have been preserved on PAC-V2, which relies on the Arcsys software published by Infotel as well as on specific additional modules (ingest, data integrity control, statistics tool modules, representation information library…) developed in-house.

At the moment, the preservation team is made of eleven people with different profiles, skills and experiences: I/T manager, archivist, file formats experts, I/T developers, system administrators, XML specialist, hardware and OS specialists, service support and monitoring specialists. The scope of data to be preserved is pretty wide, as it covers the digital heritage of the whole Higher Education and Research community. This includes educational data (courses, digitized books, theses, etc.) as well as research data (papers, simulation or computational results, etc.), or even administrative data from universities (personal/students records, financial records, etc.). Currently, PAC stores about 30 TB of data in the production environment:

- Digital PhD theses;
- Scientific papers uploaded in the open repository HAL (Hyper Article on Line) managed by CCSD;
- Digitized publications as part of the Humanities and Social Sciences program « Persée »;
- SLDR Multimedia collection (sound files of ethnographic recordings in various languages) as a pilot project of the Humanity and Social Sciences program « TGE-Adonis »;
- Digitized collection of the history of law of CUJAS university library;
- Digitized collection of books about the History of Medicine (BIU Santé - Inter-university library of healthcare);
- Digitized works in medicine, biology, geology and physics, chemistry (BUPMC - University Library "Pierre and Marie Curie");
- Digitized collection of books of the Sainte Geneviève library ;
- Library of photos of the French School of Asian Studies (EFEO).

CINES has other preservation projects: data produced by INSERM (National Institute of Health and Medical Research) as part of medical research, administrative records extracted from CNRS applications, as well as a couple of projects as part of the Humanity and Social Sciences program « TGE-Adonis » such as archeological data or language research data.

Survey on scientific data and archiving requirements

There is a strong link between the computing power of an institution and the amount of data it consumes and produces. By offering to the French scientific community, a power of 267 Tflop , CINES must manage a huge amount of scientific data. Recognizing the importance of this data, and beyond simple storage, CINES remains committed to consolidate its expertise on issues inherent in the life cycle of this data. Thus it proposes to the scientific community tools for their valorisation and preservation. This expertise is based on the historical management of scientific data related to its supercomputing mission and its long term preservation mission.

To collect additional information about requirements, we launched in 2011 a survey with 150 French laboratories using our supercomputer for their data. All these elements allow us to offer an overview of scientific data in France and Europe.

Whether at the level of consumption, production or operation, the life cycle of scientific data involves scientific libraries, software applications or sometimes "house" which often determines format. A majority of projects using supercomputing have output binary format data . ASCII and text files are used in a third of projects to complement data . The data in HDF5 and NETCDF are also present. Other formats are rarely used as FITS, Grib, CGNS.

It is not surprising to see that a common problem is the standardization of data. A majority of the projects need to share their data, but initially in a limited circle of known collaborators. Willingness to share with the entire scientific community is rare or it could be done in a second time, eg after publication. Data are not exploitable from one software to another, it is necessary to convert them systematically using pivot formats, specified and sufficiently generic to be understandable and interoperable within a same community while answering to constraints of a problem often linked to a discipline.

Part of the survey allowed us to draw up a panorama of the most popular formats.

HDF5, NetCDF are open formats, royalty-free and very general. They are self-describing to the extent that data and metadata are contained in the file itself. They are designed to hold and manipulate matrices as a mesh.

FITS is an open format, royalty-free adapted to scientific images, it provides advanced description of the image using metadata contained in its header in ASCII format. Each data block can then be described by a couple attribute / value. A number of attributes is available in FITS format, apart from this, the user has the option to define their own.

Scientific data are usually, with a certain complexity, phenomena very accurate. Any description has its limits and if it has not been maturely reflected, it may appear a risk of loss of knowledge in case of departure of a person for example. It is very important to make a description of at least two levels to mitigate this risk.

Syntactic description allows knowing the organization of data in the file. (e.g. primitive types, size, position in a table etc.). This information is usually part of the header files and primarily for computer systems which exploit them.

A semantic description will provide information on the correspondence between the data and the meaning attributed to it. Such value corresponds to a temperature, a pressure etc. This metadata can also be directly contained in the data file or described in an external file.

What should be done to ensure data could be used by a third party and in a few years?

As far as we are not aiming to deliver long term preservation, we don't need to impose a standard format or a laborious process data description to a lab that does not have the means to implement them. We would like to make them aware of the problematic and lead producers to a good risk management and to consequences of the loss of operability of their data.

We would start by asking a number of questions about the life cycle of the data:

- For what purpose has been produced this data? (Sharing with the wider community, or only for a specific job in the laboratory?)
- Who are the recipients of the data and have they a knowledge base and tools sufficient to use this data? (e.g. members of the laboratory or all of the scientists working on this thematic will they be able to understand what binary file?)
- How long are this data relevant? Would the preservation cost be less than the cost of producing them again?

The main goal of getting those answers is simple: to describe the data so that they are usable by all people to whom they are intended. To assess the importance of this, imagine the consequences if there were no reference language i.e. English to describe scientific articles!

To achieve this, actions at several levels are set up:

- Organizational: find interlocutors, articulate exchanges depending on the qualifications.
- Computing: implement infrastructure software, hardware and protocols to process metadata itself.
- Archivistics: find the standards and exchange formats relevant in the domain, to allow the information to emerge unscathed from the ravages of time.
- Methodological: It is important to identify the recipients of the data (target community) and measure their ability to understand this data (knowledge base). Then it is necessary to establish a set of information representation which will constitute a semantic link between the data and the community.

Regarding the formats, we can define together if it is relevant to engage, as an example, a home format migration without description to a standard format and if it is not sufficient to associate a set of metadata understood by recipients. Note here that our survey reveals that 40% of the binary files do not pose a priori migrating problems to a standard format like HDF5 or NetCDF. Obviously, a binary format with a good description of its contents would remain easily readable by a third party and perennial in time. Tool BEST is a solution proposed by CNES to describe binary files whether at the syntactic level with the EAST language ,or semantic level with internal NASA standard (DEDSL) which now enjoyed an international reputation.

Most laboratories do not have standard for metadata sets. They describe data mainly using references to text files, notes, publications, theses, web pages, source code, simulation parameters, or even just using a mnemonic naming system files. The communities of researchers are very different, laboratories are highly specialized in their domain of activity and the description of the data is not necessarily a priority. However, data is a representation of a basic reality, it is not self descriptive and does not has necessarily an obvious sense. The purpose of metadata is to add a descriptive level relevant enough to allow its exploitation and its sharing in the best conditions. Figure following diagram summarizes these aspects.

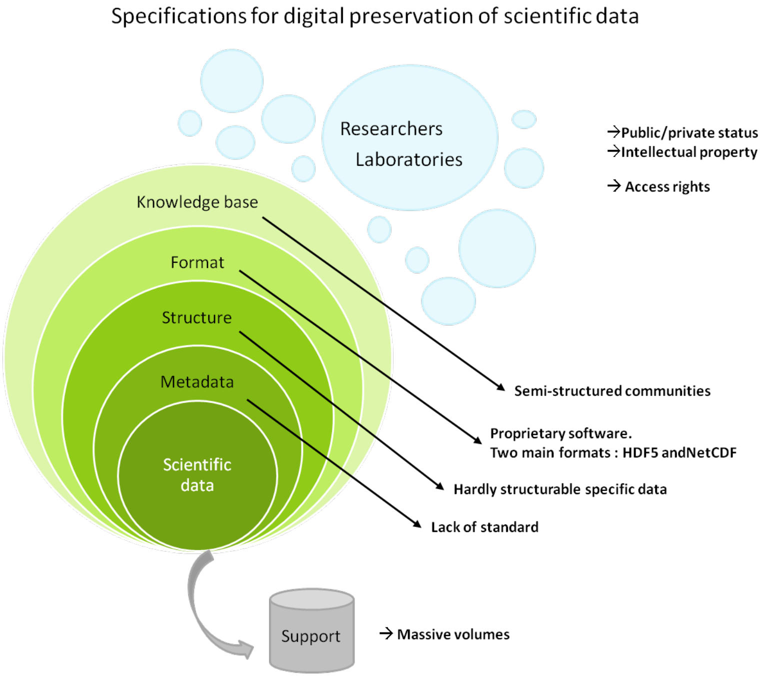


Figure 1: Illustration of the specifications for digital preservation of scientific data.

Solutions proposed by CINES

In order to propose a mutualised solution, we have defined, based on scientific communities requirements for data preservation, three main service classes, and for each of them we propose a solution, as illustrated in figure 2.

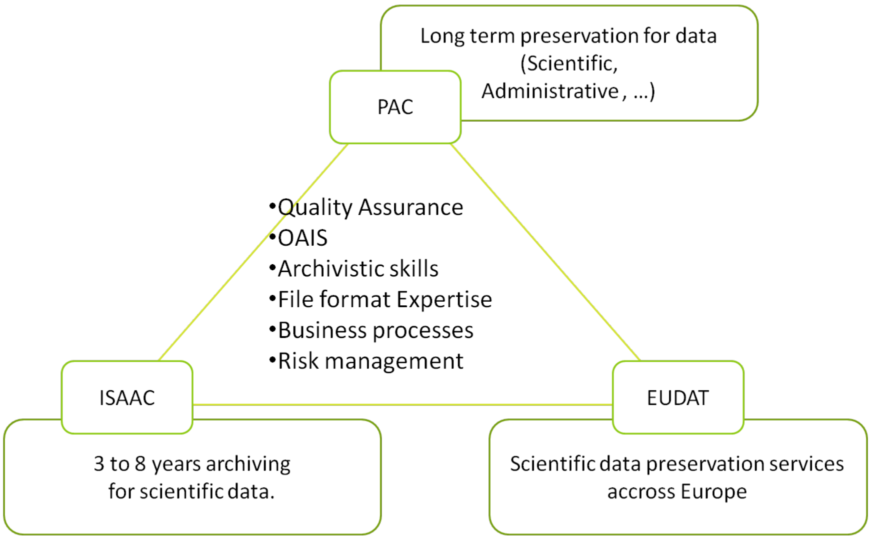

Long term preservation for data (Scientific, Administrative , …)
PAC
•Quality Assurance
•OAIS
•Archivistic skills
•File format Expertise
•Business processes
•Risk management
ISAAC
3 to 8 years archiving for scientific data.
EUDAT
Scientific data preservation services accross Europe

Figure 2: Solutions for scientific data preservation at CINES

We already presented PAC, and this platform remains the solution to fulfil long term data preservation. CINES has developed a strategy and will continue its effort to certify PAC and the related processes against the most advanced standards for digital preservation. CINES hold s the following certification:

- Data Seal of approval (DSA). CINES is a member of DSA board.
- ISO 16363 (ongoing as certification is not officially available)
- Risk management compliant with DRAMBORA methodology

CINES is also involved in the EUDAT project. This project, funded by European Commission as part of the FP7 program, aims to implement a distributed infrastructure for sustainability data to meet the requirements expressed by communities of researchers. CINES is one of the 15 European data centers implementing the Common Data Services. Currently, the main service requested by communities and implemented is a bit stream data replication service in one or more of the datacenters (B2SAFE). It assigns a unique persistent Id assigned to each replica and will perform the necessary checks to guarantee that each of the replicas is equivalent to original data object. This service is operational at CINES. EUDAT is about to deliver other services, for example B2FIND, a service providing a cross community search based on a common set of metadata.

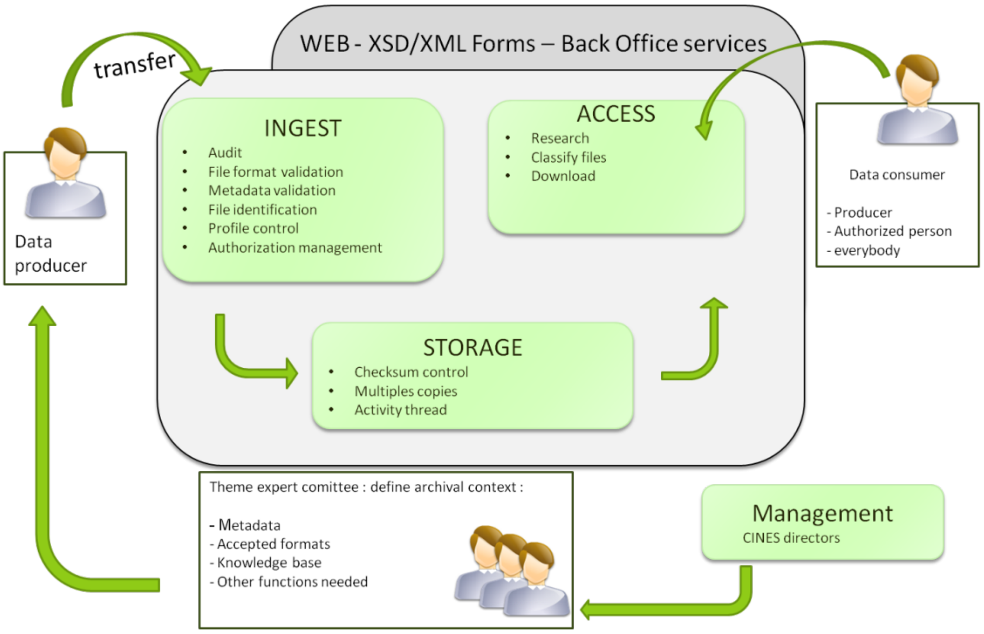


Figure 3: ISAAC system workflow.

The results of the survey we described earlier show us that we need to propose an archiving platform going further than the simple bit stream preservation, but offering more flexibility than PAC as the objective is 3 to 8 years data preservation. CINES launched a project to develop ISAAC (workshop illustrated in figure 3), which will deliver a service class compliant with DSA (Data Seal of Approval). This flexibility is defined upfront and would allow, depending on the specific requirements, to reduce the constraints about metadata or to accept a format without validation as we assume we will be able to read it at the end of the preservation. After the agreed duration, data will be either given back to the producer or moved to the long term preservation system (thus potentially requiring additional metadata or format validation).

One of the main challenges of the ISAAC project is in its organizational and administrative aspects. Indeed, any data produced must be linked to a structured and recognized context, which can ensure its suitability and its integrity. So, in the same spirit as "the thematic committees for scientific supercomputing" at CINES, the ISAAC project proposes the creation of "Thematic Committees of Archiving."Each CTA (Thematic Committees of Archiving) is composed of a chairman, a representative of the archiving platform, and one or more experts in the scientific field.

The key roles of CTA are:

- The study and the choice of file formats accepted into the archive system
- the study and the choice of metadata used to describe data
- The study and the acceptance of project archive

References

This document reproduces large extracts of:

- CINES internet site at www.cines.fr

- Quality and accreditation in a French digital repository -Lorène BECHARD, Marion MASSOL (CINES)

Contact :
Stéphane Coutin, Centre Informatique National de l'Enseignement Supérieur, Montpellier ;
coutin@cines.fr

Virtual environments for data preservation

Volker Beckmann

Abstract: Data preservation in a wider sense includes also the ability to analyse data of past experiments. Because operation systems, such as Linux and Windows, are evolving rapidly, software packages can be outdated and not usable anymore already a few years after they have been written. Creating an image of the operation system is a way to be able to launch the analysis software on a computing infrastructure independent on the local operation system used. At the same time, virtualization also allows to launch the same software in collaborations across several institutes with very different computing infrastructure. At the François Arago Centre of the APC in Paris we provide user support for virtualization and computing environment access to the scientific community.

Why go virtual?

The ability to use scientific data on the long term and to be able to extract scientific results years after an experiment has been finished, relies not only on the accessibility of the data itself. An important aspect is going to be whether it will be possible to apply analysis software that had been written in order to process the data in the first place. While the software package itself might be well documented and developed and, as far as possible, free of bugs, it is always a challenge to install such software on a current day operations system.

An example for this was the recent effort of the European Space Agency (ESA) to make the data of the EXOSAT X-ray satellite available together with the analysis software. Although the data had been preserved in standard format (FITs) used by the astrophysical community, and the software package was available and also well documented, the operation system to run this software on had long ceased to exist.

A large amount of work had to be invested in order to re-organise the source code in a way that it could be re-compiled on a modern day operation system.
Adapting computing code to a new operation system can be work intensive or even impossible, depending on the resources and also on the capabilities of the integrating team. In some experiments, old computers with ancient operation systems are used in order to maintain the ability to analyse the data. But obviously, this also poses only a temporary solution, as one day the hardware will die and replacement will be hard to get.

A way to preserve the ability to use analysis software packages of past experiments is to virtualize the processing. This means, to create a snapshot image that includes not only the analysis software itself, but also the operation system. Then, this whole package can be instantiated on a virtual machine. Figure 1 illustrates this step from a direct installation on a physical machine towards installation using virtual machines. This can be done locally on a personal computer private cluster or in a private cloud, at a larger computing center such as the CC-IN2P3 in Lyon, or on a scientific cloud or even on a commercial cloud environment.

Providing a customized “image” of the operating system together with the analysis software also has advantages in the early phases of an experiment. For example, during the software development phase, this facilitates the coordination of a project between several (international) partners. A team at one partner institute can provide to the consortium of an experiment the software together with the operations system as one package, to be installed elsewhere and being independent of the computational infrastructure available at the site of the partner institutes and ready to run immediately.

Thus, the true advantage of virtualization is the portability. This can be portability to future computing infrastructure, or to contemporaneous computing systems that run on a different operation system than the one used for the development of the software.
This portability of the analysis software poses an important aspect in the context of preserving the information contained in scientific data of past experiments.

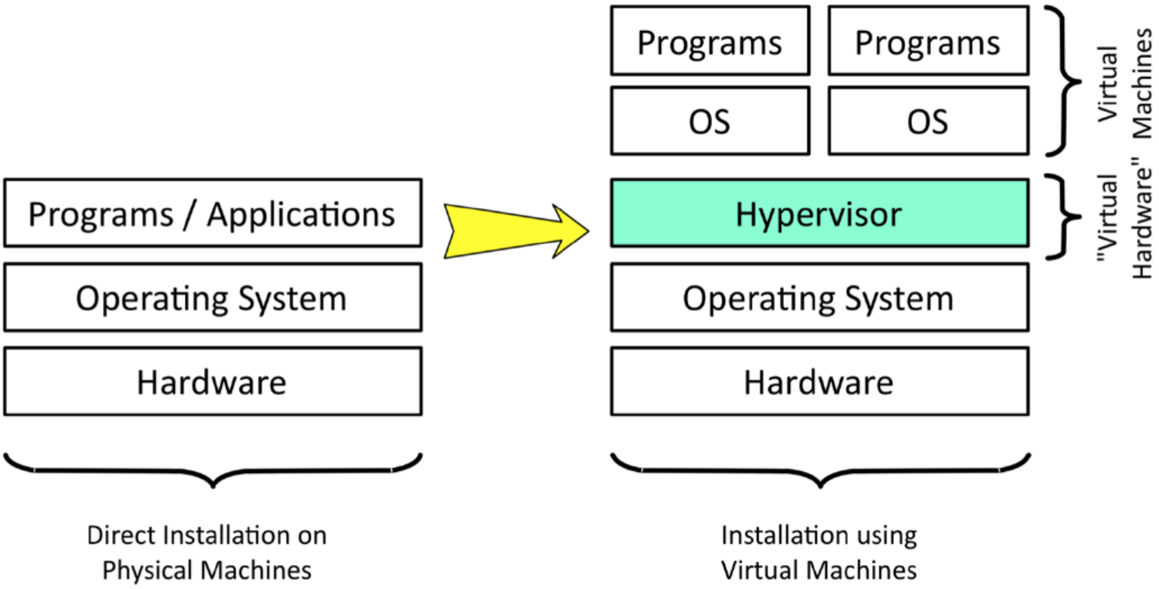


Figure 2: Virtualization of physical resources are managed by hypervisor. The virtual hardware layer encapsulates ressources to create individual entities, i.e. the virtual machines. On each virtual machines, a different OS can be running. Graphic : Charles Loomis (LAL).

The François Arago Centre: building expertise to support the community

Starting a virtual machine or even a cluster of virtual machines on a scientific, private, or commercial cloud environment is a relatively easy task, especially because MarketPlaces of disk images are developed within Cloud solutions. The main difficulty users find is to create the customized disk image of the operating system and including all the packages necessary for running the software. Once this image is created, it is used in the cloud environment as a virtual machine, enabling to run the analysis software the same way as on a single computer with the same operating system. The same way also a cluster can be virtualized. In this context a cloud environment provides the advantage of being flexible in assigning resources.

The main concern with the usage of virtual environments is the potentially reduced performance. In order to evaluate this aspect, we performed a series of tests at the François Arago Centre (FACe)[1]. This centre at the APC in Paris is part of the answer to the challenges described above. The FACe provides moderate computing power through a cluster of 600 nodes and storage facilities (at the time being 100 TByte) to projects which have a specific need in data centre activities. The FACe also provides meeting rooms and video conference facilities, and is embedded in a scientific environment through its connections to the APC, the Université Paris Diderot and it is part of the Space Campus, bringing together research, development, students, and industrial partners. The services provided by the FACe can be as different as the projects hosted. The FACe should not be understood as a classical mission data centre, serving all the ground segment needs of e.g. one satellite mission, but rather as a multi-mission toolbox, out of which each project takes the necessary resources. In many cases the FACe is a interface between the research group and the large computing facilities, like the CC-IN2P3 or the GRID.

Concerning the impact of virtualization on the computing performance, several bench mark tests have been performed at the FACe, comparing standard tests on a classical cluster with the performance on cloud environments (Cavet et al. 2012). For this purpose we used the StratusLab[3] scientific cloud. The results can be summarized as follows:

- Cloud and cluster both approach memory band-width saturation in a similar fashion
- Cloud environments under-perform for processes with large inter-node message transfer
- Cloud environments perform similar for CPU- and memory-bound processes

On the often stated concern that there is a large overhead in converting to cloud environments (e.g. Berriman et al. 2013), it was found that the most difficult part for the users is the creation of the disk image of the virtual machine, i.e. of the operation system. In practice, we found that working with the colleagues interested in virtualizing their processing, the support was not more labor intensive and time demanding than training new users on a cluster.
In addition, cloud systems can provide pre-fabricated disk images using some standard operation system set-ups, which then can be pulled off the shelf by the user.

Finally, security can be an issue in cloud systems in which the user does not have control over where the data are stored. In a commercial cloud, if data are really sensible it is necessary to encrypt or anonymize data to prevent problems. The use of a private cloud environment solves the network problem (restricted access for people of a consortium and restricted exchange with the outside world).

For an internal attack, i.e. from the cloud provider (giving information and data to e.g. the government or to a private company) the problem is based on trust. To avoid this problem, academic cloud system should be further advanced to reach sufficient resources so that the scientific community does not have to rely on commercial cloud system.

The new service of the FACe to virtualize the processing environment for software packages of projects has up to now been used by three space missions. One is the LISA-Pathfinder (LISA-PF) mission. LISA-PF is a technology demonstrator mission by ESA in collaboration with

[3] http://stratuslab.eu

NASA, in order to test technologies needed in the large eLISA project. This satellite will be launched in 2015 and will be placed at the Lagrange point 2 (L2) of the Sun-Earth system. The satellite is basically a work bench in space, in which two free-falling masses are kept within the satellite which is navigating around these test masses.

The community behind this mission is fairly small, but distributed over a number of countries and institutes. Therefore, for testing purposes providing virtual environments that can be easily installed at the partner institute where essential, because little manpower was available for the IT support at some places.

Based on the positive feedback on the virtualization for LISA-Pathfinder software, the same approach was now adopted also for the software development and simulation for ESA's large mission to study gravitational waves.

Finally, the virtualization has successfully been tested for software used in the preparation of ESA's Euclid mission (to be launched in 2020), by running large-scale simulations on a virtual cluster on the StratusLab cloud.

The future of data preservation and virtualisation

For obvious reasons, disk images that enable to use a virtual machine or cluster can only be created as long as the operating system on which the software runs is locally available. It is therefore necessary to prepare these images as long as the experiment or the space mission is still active and the data processing is well supported and understood. Advisable would be a data base of both, standard and specific disk images for virtual machines to be used by the experiments. Such an archive would indeed need little technical support and maintenance, but will be invaluable in the future. One essential requirement for such an archive is indeed that it has to be maintained over many years, and that it should be openly accessible to everyone wanting to re-analyse data from past experiments.

The François Arago Centre would be a logical place to install such an archive, with its multi-mission and multi-experiment expertise combined with the know-how of how to prepare virtual machines and how to train scientists on their usage.

References:

[1] François Arago Centre (FACe) web site: http://www.apc.univ-paris7.fr/FACe

[2] Stratus Lab Project: http://stratuslab.eu ; Cavet et al. 2012, « Utilisation du Cloud StratusLab : tests de performance des clusters virtuels », *Journées scientifiques mésocentres et France Grilles 2012*, Paris, http://hal.archives-ouvertes.fr/hal-00766067

[3] Berriman et al. 2013, « The application of cloud computing to scientific workflows: a study of cost and performance », Phil. Trans. R. Soc. A 2013 371

Contact:

Volker Beckmann, François Arago Centre, Laboratoire Astroparticules et Cosmologie, Université Paris VII; Beckmann@apc.univ-paris7.fr