FROM LONG-TERM PRESERVATION TO LONG-TERM VALORIZATION OF ARCHIVES. A SEMIOTIC VIEW

Daniel Galarreta

CNES/CST Centre Spatial de Toulouse

ABSTRACT

The issue of long-term preservation of archives is usually considered with respect to fixed designated communities that are identified groups of potential consumers who should be able to understand a particular set of information.

Even if this notion allows that such a community could be composed of several communities, the profile of such these communities should be outline at first. The evolutions of one of these communities because of the evolution of scientific or technical knowledge or because of the advent of new means for operating available archives are not always predictable. And this is complicated all the more so as that evolution can go along with loss of knowledge and know-how within the same community. This is a well-known issue.

Another situation is the case where new communities join later the former ones, but with objectives regarding the archives which are quite different from those of the designated community at the root of the archives.

As far as the initial designated community is concerned, the value that this community gives to the archives is likely to evolve from a high level in the first years of their existence toward a gradual decrease the following years.

It is also likely that a clear evolution of this designated community or the arrival of new communities will raise the question of the valorization of these archives over a long time. That is to say: upon which conditions, can the archives produced at a given time and in a given context, offer later a significant value in new contexts? What are the steps to be taken with respect to these archives in order to give them a chance to still represent a value for new consumers later? These measures could be either technical or organizational.

In this communication we propose to explore this issue by examining several notions that participate to the definition and the management of archives from the point of view of meaning. More precisely we want to switch from the information-processing paradigm to a meaning processing paradigm. This opposition was proposed by Ilkka Tuomi who has been a principal scientist at the Nokia Research Center. He observed that whereas during the last decades, research and development of information technologies has been guided by the information-processing paradigm, in the new meaning-processing paradigm, information and communication technologies are understood as technical systems that are integrated with and embedded in social, cultural and cognitive processes.

We will build our analysis upon a semiotic methodology; semiotics is particularly suited to our goal since it is the study of meaning-making.

Using this approach we will see for instance, that another conception of data is desirable. This conception differs from the view that considers data as the building brick of information and of knowledge. This hierarchy which is represented by the acronym DIKW (Data, Information, Knowledge, Wisdom) was popularized by authors such as Russell Ackoff, theorized by others such as Naim Zins, but also criticized by a few ones such as Rafael Capurro. The core of this criticism is that this notion rests upon a positivist epistemology: the observable things exist in a world ready-made. They are given, i.e. literally data.

An alternative view of digital archives that is based upon a semiotic point of view is possible. As an illustration we will present the outline of digital audiovisual archives such as the archives developed in the Audiovisual Research Archives programs (ARA). ARA is a French R&D program of the FMSH (Institute of Human and Social Sciences) conducted by Peter Stockinger. Although it does not addresses a space community, this project shares with space archives a few features: scientific purposes, multimodal records (images, sounds, texts), metadata management, distributions issues.

As a conclusion we will advocate the evolution of the conception of long-term preservation toward a conception of a long-term valorization of archives congruent with an economy of meaning which is today emerging in our societies.

APPLICATION OF OAIS TO THE PRESERVATION OF LINKED DATA

David Giaretta [1] , **Carlo Meghini** [2] , **Anna Fensel** [3]

(1) Giaretta Associates , (2) CNR Consiglio Nazionale delle Ricerche , (3) STI International Consulting and Research GMBH

ABSTRACT

There is a demand that society reap the benefits of the investment made in creating the growing deluge of data with which we are faced. Data becomes more valuable and exploitable the more it is combined. Linked Data is one of the growing and most flexible ways of doing this. Yet this poses a problem. The very scale and diversity of the data is compounded by the scale, flexibility and diversity of the links, and there is a need to preserve some, if not all, of the data and the links since combining recent data with older data is a vital part of the overall process, not least for longitudinal studies.

This paper summarises the state of the art of the understanding of the fundamental techniques of digital preservation using the concepts introduced by OAIS (ISO 14721), and then systematically discusses these techniques in the context of Linked Data current practices. It derives from the work performed in the PRELIDA (www.prelida.eu) project.

The question addressed is whether these techniques are applicable to Linked Data, whether current Linked Data practices involve new techniques which might be more broadly applicable and finally whether there are improvements to current linked data practices.

The conclusions are that current Linked Data preservation practices can be significantly improved by systematic application of the general digital preservation techniques. Linked Data presents a number of specific challenges in terms of distribution and changeability; while these are not qualitatively unique challenges, nevertheless they are potentially quantitatively unique i.e. Linked Data may involve much more widely distribution and components which are less controlled in terms of persistence and variability.

The state of the art in digital preservation can provide a number of tools and services which could be applied to Linked Data but there is a need for solutions to the challenges of distribution, which is likely to be more widely beneficial, and specific tools and services which fit within the context of the general techniques.

This abstract approach is complemented by PRELIDA D4.3 Roadmap, which takes are more detailed look at the specific use cases.

EXPERIENCE IN PRESERVATION OF LEGACY DATA IN ASTRONOMY

pierre le sidaner , **Jean Guibert**

obspm/Observatoire de Paris

ABSTRACT

The data centre in charge of data distribution and data preservation presented here is VO Paris Data Centre (VOPDC) at Paris Observatory, France. VOPDC is a joint venture by scientists and engineers from all scientific departments of the Observatory with skills spanning the large scientific spectrum of Astronomy, Astrophysics, Planetary Science and Atomic & Molecular physics. Twelve years ago, VOPDC started to use and develop the standards of the International Virtual Observatory Alliance (IVOA) for data publishing. As in all data centres, the question of data preservation arose very early. Technical issues of storage, replication, and preservation have been solved using storage virtualization for tapes. File systems with snapshot and replication facilities at the level of disk blocks were the solution to synchronize billions of files. The ingestion procedure for new data has been enhanced. However the question of context data for legacy archives is still not completely solved. I will present feedback of the work done about digitized images of the sky for legacy surveys, in particular the context information we collect, as well as the organization of Storage Information Packages (SIP) from an Open Archive Information System (OAIS) point of view. I will also discuss the evolution of IVOA data models to describe provenance and context information.

VITO EO-DATA ARCHIVE EVOLUTIONS AND INTEGRATION IN PADUA

Martine Paepen , **Erwin Goor**

VITO

ABSTRACT

In May 2013 the **PROBA-V** satellite was launched to extend the time series of 15 years SPOT-VEGETATION data, a mission that ended in May 2014. PROBA-V is a micro-satellite aiming at monitoring the Earth's vegetation on a daily basis with a spatial resolution of 1 km and 1/3 km. From spring 2015 products at 100m resolution are available as well with a global coverage every five days. VITO developed the PROBA-V user segment and is responsible for the image processing, archiving and distribution of all products. Next to the **VEGETATION operational activities**, VITO hosts several other processing facilities, which e.g. offer hyperspectral images from the airborne APEX instrument or bio-geophysical parameters in the frame of Copernicus GIO Global Land Services.

New approaches, technologies and further investments were/are needed to cope with the heterogeneous nature and the huge increase of available EO-data for the near future including the SENTINEL missions which will be used to provide value-added products. In that context the **VITO PADUA (Product Archiving, Distribution and User-oriented Access)** programme was launched to provide one comprehensive platform for data distribution, data viewing/analysis and on-demand processing services. PADUA integrates different existing components such as the Product Distribution Facility (PDF) and builds further on the ESA/GSTP ESE project.

The **ESE project** designs and demonstrates an **infrastructure to ease the exploitation of massive amounts of EO-data** for both the remote sensing **expert and the non-expert user**. The project aims to provide powerful Web-based tools integrated in a single end-to-end solution for EO data access, visualisation and analysis of EO time series (e.g. vegetation parameters like NDVI, LAI, FAPAR), on-demand data processing and e-collaboration. The on-demand processing capabilities are provided on two levels:

- **Software level**: by providing interactive Web-based processing applications, ready to be used by any non-expert user by only providing some input parameters.
- **Platform level**: by providing a platform for EO-specialists which allows them to design EO-applications as workflows, involving processing services and data from VITO, powerful Open Source EO Processing libraries from third-parties, as well as algorithms and data from the user themselves.

The first gateway to the EO data at VITO is the Product Distribution Facility (PDF) which is integrated in PADUA. Data can be locally available on disk or can be retrieved from the **Long Term Archive (LTA)**. The VITO LTA is a **multi-mission archiving facility** which acts as a long term storage for both data preservation and data retrieval if the data is no longer available on the short or medium term storage within the PDF. Additionally the PDF and the LTA can provide access to data from older, inactive missions.

The LTA is a flexible and independent facility that has been developed according to the OAIS (Open Archival Information System) principles and covers all main archiving functionality i.e. ingestion, storage, retrieval and management of heterogeneous data and metadata. The two-layer service oriented architecture makes the application independent from the selected database and storage facilities. To face the challenges of the exponential growth of EO data volumes to be archived and to be fast retrievable, we recently evolved from the custom made software with the EMC Networker plug-in for tape storage towards a system that utilizes **SIMPANA software from Commvault** integrated in a Tiered Storage Architecture using also high end and low end disk storage. In this way we eliminate the different plug-ins for disk and tape storage and disconnect the storage management functionality from the LTA software, reduce the maintenance and support efforts/costs and increase the generic reporting and monitoring integrated in the VITO data centre. Together with the implementation of the SIMPANA storage management software we have recently **migrated all the archived data to new LTO6 tapes.**

NASA'S EARTH SCIENCE DATA STEWARDSHIP ACTIVITIES

Hampapuram Ramapriyan [1] , **Dawn Lowe** [2] , **Kevin Murphy** [3]

(1) Science Systems and Applications, Inc. & NASA Goddard Space Flight Center , (2) NASA Goddard Space Flight Center , (3) NASA

ABSTRACT

NASA has been collecting Earth observation data for over 40 years using instruments on board satellites, aircraft and ground-based systems. With the inception of the Earth Observing System (EOS) Program in 1990, NASA established the Earth Science Data and Information System (ESDIS) Project and initiated development of the Earth Observing System Data and Information System (EOSDIS). A set of Distributed Active Archive Centers (DAACs) was established, distributed across the United States. Today, EOSDIS, consisting of 12 DAACs and 8 Science Investigator-led Processing Systems, processes data from the EOS and SNPP missions, as well as archives and distributes data from most of NASA's Earth science missions. The data held by EOSDIS are available to all users consistent with NASA's free and open data policy, which has been in effect since 1990. The EOSDIS archives consist of raw instrument data counts (level 0 data), as well as higher level standard products (e.g., geophysical parameters, products mapped to standard spatio-temporal grids, results of Earth system models using multi-instrument observations, and long time series of Earth System Data Records resulting from multiple satellite observations of a given type of phenomenon) . The EOSDIS has been distributing data to a broad, diverse and global user community since 1994. During 2014, the distribution from EOSDIS exceeded one billion files.

The data stewardship responsibilities include ensuring that the data and information content are reliable, of high quality, easily accessible, and usable for as long as they are considered to be of value. To meet these responsibilities, the ESDIS Project is involved with the data producing missions or projects throughout their lifecycles. Interface agreements are developed with varying degrees of formality (appropriate to the missions or projects). The products are produced by science teams with peer-reviewed algorithms, quality assessed and documented with appropriate caveats about usage. The capabilities in EOSDIS ensure easy accessibility to data and associated services for discovery, visualization, and downloading for purposes of science and applications. Near real-time access is available to some of the data. NASA ensures that the active archive capabilities are available as long as there is interest in using the data. This is well beyond the life time of missions. In 2011, NASA developed a document titled Earth Science Data Preservation Content Specifications, which is being used to ensure that the data and all the associated metadata and documentation from missions are collected and archived in preparation for permanent preservation, so that future generations will still be able to understand and uses the data products from today's missions. NASA is taking active roles in the Data Stewardship Committee of the US Federation of Earth Science Information Partners and in the Data Stewardship Interest Group of the Committee on Earth Observing Satellites (CEOS) Working Group on Information Systems and Services (WGISS) to exchange ideas and help evolve best practices in stewardship.

ENVIRONMENTAL DATA AND SERVICES FROM NOAA'S NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION

Nancy Ritchey

NOAA/NESDIS/NCEI

ABSTRACT

In January 2015 NOAA’s National Data Centers (i.e. National Climatic Data Center, National Geophysical Data Center, National Oceanographic Data Center and National Coastal Development Data Center) merged into NOAA’s National Centers for Environmental Information (NCEI). NCEI is the most comprehensive, accessible, and trusted source of environmental data and information and it now supports data from the depths of the ocean to the surface of the sun and from million-year-old sediment records to near real-time satellite observations.

This merger consolidated twenty geographic locations supporting weather (atmospheric and solar), climate, oceanographic and geophysical data disciplines, customer service and data preservation processes, and created the nation’s leading authority for environmental information. The purpose of the merger was to achieve efficiencies in IT, administrative and management infrastructure; provide consistent data management capability for all of NOAA; and enhance interdisciplinary teamwork across multiple science disciplines. This presentation will describe the new organization and progress on merging data stewardship and preservation services

THE INTERNATIONAL PLANETARY DATA ALLIANCE (IPDA) : OVERVIEW OF THE ACTIVITIES.

Gopala Krishna [1] , Daniel Crichton [2] , Yukio Yamamoto [3]

(1) ISRO Indian Space Research Organisation , (2) JPL Jet Propulsion Laboratory , (3) JAXA Japan Aerospace Exploration Agency

ABSTRACT

The IPDA's main emphasis is to ease discovery, access and use of planetary data by world-wide scientists regardless of which agency is collecting and distributing the data. Ensuring proper capture, accessibility and availability of the data is the task of the individual space agencies. The IPDA is focusing on developing an international standard which allows the following capabilities: query, access and usage of data across international planetary data archive sys-tems. While, trends in other areas of space science are concentrating on the sharing of science data from diverse standards and collection methods, the IPDA shall concentrate on promoting standards which drive common methods for collecting and describing planetary science data across the international commu-nity. Such an approach will better support the long term goal of easing data sharing across system and agency boundaries. An initial starting point for deve-loping such a standard will be internationalization of NASA's Planetary Data System standards. The IPDA has grown significantly since its first meetings back in November 2006. The steering committee is composed today of 28 members from 24 countries or international organisations. A technical expert group has been created in 2013, now composed of 20 members from participating countries, in charge of technical, redactionnal and compatibility issues. Also, many projects are still open, including the creation of the Memorendom of Understanding (MOU) template for international missions, the investigation of IVOA/IPDA (International Virtual Observatory Alliance-IVOA) interaction, PDS4 implementation project, the development of international registries to enable registration and search of data, tools and services, Chandrayaan-1 interoperability project with PDAP and many others. Lets indicate that since 2006, a project duration is on average 2-3 years and that 8 standards specifications documents have been pub-lished by the IPDA. IPDA is also maiking effort to-ward outreach activities, trying to be present or represented at all important national and international levels and meetings like Cospar, AGU, EPSC, EGU etc... with on an average 2-3 meeting per year, plus the IPDA annual meeting. Also, the web page www.planetarydata.org contain mainy tools for planetologists (services) and new tools can be submitted freely. The evolving standards from IPDA (ex: PDS4) are being implemented for the up-coming planetary mission archives by the respective agencies.

THE CHALLENGE OF KNOWLEDGE PRESERVATION: THE CASE OF THE ATV CONTROL CENTRE.

Mike Steinkopf [1] , **Emiliano Micaloni** [1] , **Jean-Michel Bois** [1] , **Eric Conquet** [1] , **Roberta Mugellesi-Dow** [2]

(1) ESA , (2) ESA/ESOC

ABSTRACT

In the last decades the problem of preserving knowledge at the end of a project represents an issue on which organizations put more and more attention.

Initially the topic of knowledge preservation has been totally put on the back of the project manager in the framework of the overall project closure activities without specific support dedicated by the organisation at the scope.

As a result the outcome of knowledge preservation was often negatively affected by different factors such as: the release of project human resources leaving the project with their baggage of implicit knowledge, the lack of general knowledge management culture, scarce attention on the topic of preserving knowledge since the early phases of the project. This all leads to results in the field of knowledge preservation not in line with expectations.

The above is also representing a major challenge for the European Space Agency (ESA) at the end of the Automated Transfer Vehicle (ATV) Programme. It was not foreseen to have any immediate follow-on mission, which would have guaranteed the preservation and straightforward re-use of the agency know-how built up over twenty years of preparation and execution of 5 ATV missions. ESA Management therefore requested the ESA Operations Team in the ATV Control Centre to put in place a strategy to preserve the ATV operations expertise also with the objective to develop a methodology to become potentially an ESA corporate reference tool for other projects.

In this framework, among the different existing systems in use in the agency, it was decided to invest in the further development of an existing, embryonic ESA Knowledge Preservation system, for which ATV represented the pilot project and first real use case. Main emphasis was put on a user-friendly, dynamic and interactive knowledge management system in order to not only preserve the project knowledge but also to make the knowledge database attractive enough to be used for future applications in similar projects. A knowledge data base and knowledge transfer is only efficient if it can be used by other projects.

This paper explains the specific knowledge preservation needs of the ATV projects, how this could be achieved through the existing ESA Knowledge Management approach and how the ESA Knowledge Management has been developed to fit the needs of the ATV case. This is based on a combination of a powerful search engine and ATV documentation, photos, lessons learned, list of ATV experts coupled with ATV (WIKI) pedia articles.

INTEGRATING DIGITAL PRESERVATION INTO EXPERIMENTAL WORKFLOWS FOR SPACE SCIENCE

Simon Waddington [1] , **Emma Tonkin** [1] , **Charaka Palansuriya** [2] , **Christian Muller** [3] , **Praveen Pandey** [3]

(1) King's College London , (2) The University of Edinburgh , (3) B.USOC

ABSTRACT

PERICLES (Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics) is a four year EU FP7 integrated project on digital preservation that began in February 2013. PERICLES aims to ensure the ongoing accessibility of digital material in a constantly changing environment. The project focuses on two case studies, one from space science and the other from digital media and art. Through communities of practice, the project is also exploring how these specific studies can be generalised across a wider set of domains. The space science case study is based on the SOLAR experiment, which generates raw observations of the spectrum of the sun, using instruments based on the International Space Station (ISS). This raw data are analysed and calibrated by a team of scientists using a complex set of scripts. Calibration is an iterative process running over a number of years. Increased understanding of the behaviour of the instruments and phenomena which may bias the observations (e.g. arrival of vehicles at the ISS) mean that previous calibrations need to be recomputed. The software platforms on which the data gathering and processing are performed are themselves subject to change. Results obtained from SOLAR are also compared to observations made using different instruments, which are generally based on different technologies. The long duration of the SOLAR experiment, the need to continuously review and reprocess previous results, and the ongoing evolution of software platforms collectively pose a critical challenge for traditional end-of-life preservation approaches. In this respect, the boundary between active life and end-of-life is not clearly defined. We instead apply a continuum approach where preservation tasks such as appraisal and policy management are much more integrated into active life. The paper will consider two specific issues. [bullet] Appraisal of experiments, where newer calibrations are compared with older results, and some results are marked as less accurate or reliable. [bullet] Technology change such as updates to software platforms. For example, a new version of a maths library may increase the accuracy of a data analysis operation. It may consequentially be desirable to refactor older calibration results to take advantage of this.

FENGYUN-3 SERIES METEOROLOGICAL SATELLITE DATA ARCHIVING AND SERVICE SYSTEM

Zhe Xu [1] **, Jianmei Qian** [2] **, Di Xian** [2] **, Yonggang Qi** [2]

(1) CMA/NSMC National Satellite Meteorological Center , (2) CMA/NMSC National Meteorological Satellite Center

ABSTRACT

The Fengyun-3 series is the second generation of Chinese sun-synchronous meteorological satellites. They are designed to enhance China's three dimensional atmospheric sounding capability and global data acquisition capability, in an effort to collect more cloud and surface characteristics data. Fengyun-3C, the first operational satellite of FY-3 series, was launched on Sep 23, 2013. Together with the 2 former experimental FY-3 series satellites which are still in orbit, National Satellite Meteorological Center (NSMC) of CMA operates 3 sun-synchronous FY-3 meteorological satellites.

The daily archive volume of NSMC has increased from 280GB to 4000GB due to the three FY-3 series satellites' global observation data generation. The Fengyun-3 satellite data has following features: Huge amount of daily data volume and archived data volume, high frequency of the historical data download, and the archive data users distribute around the world. In order to meet the needs, NSMC has implemented new generation satellite data archive and service system for Fengyun-3 series, focusing on high frequency data archiving and management for high capacity.

This paper introduces the characteristics and the new technologies applied in Fengyun-3 Series Meteorological Satellite Data Archiving and Service System. The system integrates several high-performance servers, high availability disk array, large-scale automated tape library as well as operation system, database, storage software and application software, all of which constitutes archive and service application clusters. The application of configured dynamic load container technology and multi-server processing load balancing and parallel computing technology makes the cluster a highly scalable, available, efficient operational system to realize all levels of satellite data archiving, customization, retrieval, dynamic information release, operation monitoring, management and other functions. Internet-based full dataset sharing, distributed file system, spatiotemporal integration of database storage, WebGIS global satellite image distribution, visual processing and display technology, multi-source data fusion technology are adopted in the system to improve data sharing and service.

Completion of the system meets the needs of operational meteorological satellite, achieving daily 4 TB data storing and daily 1 TB data download service. The system running in NSMC/CMA is the first one to provide internet-based full archive dataset (online, nearline and offline) sharing service in the area of Chinese civil remote sensing satellite data services. National Satellite Meteorological Center has archived 92 kinds of data and products for 23 Chinese and foreign satellites, the total storage is up to 6820 TB, and various satellite data sharing service are available for 32009 registered users worldwide.

POLAR THEMATIC EXPLOITATION PLATFORM

David Arthurs [1] , **Andrew Fleming** [2] , **Ola Grabak** [3]

(1) Polar View Earth Observation Limited , (2) +++OTHER+++ , (3) ESA/ESRIN

ABSTRACT

The European Space Agency (ESA) and other satellite operators have been at the forefront of collecting, analysing, processing and disseminating new data and information from EO satellites for over four decades. Throughout this time, the volume of data and range of applications has increased dramatically, challenging us to find new methods to fully utilise this capacity.

ESA is therefore establishing a series of Thematic Exploitation Platforms (TEP) providing the necessary advanced ICT platform to allow new ways of working. These projects will address present challenges and opportunities in scientific data exploitation by collocating data, processing capabilities and ICT infrastructure, providing a complete cloud based work-environment for users performing scientific exploitation of EO data. One of the TEPs will focus on the challenges and requirements of the polar user community.

The TEP concept aims to provide a working environment where users can bring their algorithms, applications and development activities to the data, avoiding the need to download and store large volumes of data. This represents a new way of working with EO data to encourage and allow easier processing, sharing and wider exploitation of EO data.

The Polar TEP (P-TEP) is being created by a team lead by Polar View Earth Observation Limited. It will deliver a rich set of polar themed EO and complimentary datasets, plus functionality to easily establish interfaces to new local or distributed data resources. P-TEP will also include appropriate toolboxes and processing capabilities, allowing deployment of user defined workflows. Use of these resources will be accessed through a web portal which will also allow users to setup their own development and processing environments.

An initial pilot project will demonstrate the potential of P-TEP to investigate current and future iceberg risk in Baffin Bay. A diverse set of data, processors and models will be deployed and integrated to allow users to investigate linkages between iceberg populations, observed and modelled changes in ice sheet movement and calving rates, ocean circulation and iceberg trajectories. The integration of these components and toolsets will allow P-TEP users to ask questions about changing iceberg populations, e.g., as input to climate change studies in the region, or inform infrastructure and ship routing decisions.

BIG DATA CUBES AT YOUR FINGERTIPS

peter baumann

jacobs university

ABSTRACT

Big Data in the Earth sciences, the Tera- to Exabyte archives, mostly are made up from coverage data, according to ISO and OGC defined as the digital representation of some space-time varying phenomenon. Common examples include 1-D sensor timeseries, 2-D remote sensing imagery, 3D x/y/t image timeseries and x/y/z geology data, and 4-D x/y/z/t atmosphere and ocean data. Analytics on such data requires on-demand processing of sometimes significant complexity, such as getting the Fourier transform of satellite images. As network bandwidth limits prohibit transfer of such Big Data it is indispensable to devise protocols allowing clients to task flexible and fast processing on the server.

The transatlantic EarthServer initiative, running from 2011 through 2014, has united 11 partners to establish Big Earth Data Analytics. A key ingredient has been flexibility for users to ask whatever they want, not impeded and complicated by system internals. The EarthServer answer to this is to use high-level, standards-based query languages which unify data and metadata search in a simple, yet powerful way.

A second key ingredient is scalability. Without any doubt, scalability ultimately can only be achieved through parallelization. In the past, parallelizing code has been done at compile time and usually with manual intervention. The EarthServer approach is to perform a semantic-based dynamic distribution of queries fragments based on networks optimization and further criteria.

The EarthServer platform is comprised by rasdaman, the pioneer and leading Array DBMS built for any-size multi-dimensional raster data being extended with support for irregular grids and general meshes; in-situ retrieval (evaluation of database queries on existing archive structures, avoiding data import and, hence, duplication); the aforementioned distributed query processing. Additionally, Web clients for multi-dimensional data visualization are being established. Client/server interfaces are strictly based on OGC and W3C standards, in particular the Web Coverage Processing Service (WCPS) which defines a high-level coverage query language.

Reviewers have attested EarthServer that "With no doubt the project has been shaping the Big Earth Data landscape through the standardization activities within OGC, ISO and beyond".

We present the project approach, its outcomes and impact on standardization and Big Data technology, and vistas for the future.

REGARDS : THE NEW CNES GENERIC SYSTEM TO ACCESS AND ARCHIVE SPACE DATA

Aurélie Bellucci , **Dominique Heulet** , **Jean-Christophe Malapert**

CNES Centre National d'Etudes Spatiales

ABSTRACT

REGARDS (« REnewal of Generic tools to Access and aRchive Space Data) is a new framework for long term preservation and diffusion of space data from CNES heritage or scientific laboratories. CNES will develop REGARDS relying on two existing products: SIPAD-NG and SiTools2. SIPAD-NG is a generic system providing three of the main OAIS functions: “data ingest”, “data management” and “data access”. It rests on CNES infrastructures to answer the needs of operational data centers. SiTools2 is an open source generic system providing a self-manageable data access layer to be deployed on already existing databases in scientific laboratories. REGARDS will replace both tools while fulfilling all their current functions but avoiding duplicate developments. The future system will be able to cope with the huge data volumes expected from space missions in the 2020 and beyond. With its modular design and its open source license, REGARDS will answer present and future needs of CNES operational data centers as well as scientific laboratories. With proper plugins, REGARDS will be compliant with main interoperability protocols in astronomy and Earth observation. Due to its architecture, REGARDS will be an extensible framework able to follow the evolutions of interoperability standards. Development will begin at the end of the year.

VIRTUAL OBSERVATORY TOOLS AND AMATEUR RADIO OBSERVATIONS SUPPORTING SCIENTIFIC ANALYSIS OF JUPITER RADIO EMISSIONS

Baptiste Cecconi [1] , **Sébastien L G Hess** [2] , **Pierre Le Sidaner** [1] , **Renaud Savalle** [1] , **Stéphane Erard** [1] , **Andrée Coffre** [3] , **Emmanuel Thétas** [3] , **Nicolas André** [4] , **Vincent Génot** [5] , **Jim Thieman** [6] , **Dave Typinski** [6] , **Jim Sky** [6] , **Chuck Higgins** [6]

(1) Observatoire de Paris , (2) ONERA, Toulouse , (3) Station de Radioastronomie de Nançay , (4) IRAP, CNRS-Université Paul Sabatier , (5) IRAP, CNRS-Université Paul Sabatier , (6) RadioJOVE

ABSTRACT

In the frame of the preparation of the NASA/JUNO and ESA/JUICE (Jupiter Icy Moon Explorer) missions, and the development of a planetary sciences virtual observatory (VO), we are proposing a new set of tools directed to data providers as well as users, in order to ease data sharing and discovery. We will focus on ground based planetary radio observations (thus mainly Jupiter radio emissions), trying for instance to enhance the temporal coverage of jovian decametric emission. The data service we will be using is EPN-TAP, a planetary science data access protocol developed by Europlanet-VESPA (Virtual European Solar and Planetary Access). This protocol is derived from IVOA (International Virtual Observatory Alliance) standards. The Jupiter Routine Observations from the Nancay Decameter Array are already shared on the planetary science VO using this protocol. Amateur radio data from the RadioJOVE project is also available. We will first introduce the VO tools and concepts of interest for the planetary radioastronomy community. We will then present the various data formats now used for such data services, as well as their associated metadata. We will finally show various prototypical tools that make use of this shared datasets. A preliminary study based on January-February 2014 data will also be presented.

LONG-TERM PRESERVATION OF DATA ANALYSIS SOFTWARE WITH OPERATING-SYSTEM-LEVEL VIRTUALIZATION

Helge Eichhorn , Thomas Trinkel , Reiner Anderl

Technische Universität Darmstadt

ABSTRACT

Many of today's computer-aided engineering tasks, such as data analysis, are computationally expensive, highly domain- or problem-specific, and of high complexity. Due to this fact commercial off-the-shelf software solutions do not satisfy the requirements of many organizations and custom software tools are developed. For the sake of knowledge management and traceability it should be possible to reproduce and modify such analyses with minimal effort in future engineering processes. It is not sufficient though to archive source code together with input data since every program generally relies on an extensive dependency tree. This includes shared libraries, runtime environments or compilers and even operating system (OS) features. A simple yet inefficient approach is the archiving of the complete operating system together with all required software components as a virtual machine (VM) image. An emerging trend in the cloud computing industry is the move towards OS-level virtualization technologies, so-called containers. Within such containers applications and their dependencies are packaged. The resulting files require significantly less storage space and computational overhead and can be run on any machine with the container runtime installed. While these tools are intended for software deployment expanding the scope to archiving purposes appears reasonable. This publication compares the available OS-level virtualization technologies and evaluates their potential for utilization within archiving processes. A use case based on a real world example is developed and implemented with the most promising container technologies. Finally performance overhead and storage requirements are compared to VM-based solutions.

DATA MINING AND KNOWLEDGE DISCOVERY FOR TERRASAR-X PAYLOAD GROUND SEGMENT

Daniela Espinoza Molina [1] , **Vlad Manilici** [2] , **Shiyong Cui** [1] , **Christoph Reck** [2] , **Mathias Hofmann** [2] , **Octavian Dumitru** [1] , **Gottfried Schwarz** [1] , **Henry Rotzoll** [2] , **Mihai Datcu** [1]

(1) DLR-IMF Deutsches Zentrum für Luft- und Raumfahrt/ Institut für Methodik der Fernerkundung , (2) DLR-DFD Deutsches Zentrum für Luft- und Raumfahrt/ Deutsches Fernerkundungsdatenzentrum

ABSTRACT

Earth Observation (EO) imaging satellites continuously acquire huge volumes of high resolution scenes increasing the size of image archives and the variety and complexity of EO image content. Thus, new methodologies and tools that allow the end-user to access a large image repository, to find and retrieve dynamically a collection of desired images, and to extract and to infer knowledge about the patterns hidden in the image archives are required. In this context, this paper presents Earth Observation Image Librarian (EOLib), which is a modular system offering data mining and knowledge discovery functionalities for the TerraSAR-X Payload Ground Segment and serving to setup the next-generation of Image Information Mining (IIM) systems. It implements novel techniques for image content exploration and exploitation. The main goal of EOLib is to create a communication channel between Payload Ground Segments and the end-user who receives the image content enriched with annotations and metadata as well as coded in an understandable format associated with semantic categories that is ready for immediate exploitation. EOLib is composed of several components to offer functionalities such as ingestion and feature extraction of TerraSAR-X images, metadata extraction, semantic definition of the image content based on machine learning and data mining methods, advanced queries of the image archive utilizing content, metadata and semantic categories, and 3D visualization of the huge and complex image archives. EOLib will be interfaced to and operated in the DLR Multi-Mission Payload Ground Segment of the DLR Remote Sensing Data Center premises at Oberpfaffenhofen, representing at the same time a general new concept for the operations of Ground Segment infrastructures.

ENSURING DATA IS PLANNED FROM THE START TO BE RE-USABLE AND PRESERVATION READY

David Giaretta [1] , **Helen Glaves** [2] , **Daniele Boucon** [3] , **John Garrett** [3] , **Robert Downs** [4] , **Mike Martin** [5]

(1) Giaretta Associates , (2) British Geological Survey , (3) CNES Centre National d'Etudes Spatiales , (4) CIESIN, Columbia University , (5) NASA

ABSTRACT

A tsunami of scientific and other data is being created and some of it will be worth preserving for future re-use. The fundamental standards are in place (e.g. OAIS, ISO 14721 and ISO 16363) and the fundamental techniques for digital preservation are well understood, at least in principle. Some of these techniques are being adopted for the preservation of scientific data and related research information in digital form. However, in many cases, the “metadata” which needs to accompany data is often not captured. Unfortunately, where metadata is incomplete or absent, the data are not as useful as they could/should be.

In an effort to enable the re-use of data that have been developed with public funds, funding agencies are increasingly requiring that the projects they fund submit Data Management Plans as a condition of the grants. However, these plans often tend to be relatively rudimentary documents, which do not necessarily evolve during the course of the project, and adherence to such plans is seldom monitored.

This paper reports on work carried out in the Research Data Alliance Interest Group on Active Data Management Plans and in the Consultative Committee for Space Data Systems (CCSDS) Data Archive Ingest (DAI) working group on a standard for an Information Curation Framework to address these shortcomings. Other concerns, including costing, risk management, policies and workflow, value-adding and service architectures, are addressed at a high-level.

The goals of these groups complement each other to provide a high level definition of the data curation and preservation activities that are needed during each stage of a project for the purpose of data management planning.

The focus is on defining the requirements for metadata content that are necessary to ensure preservation. In particular, these include identifying, for projects generating data, the stages during which specific metadata elements would need to be created, identifying the roles that would be responsible for creating it, and the roles that would receive it.

In addition, the ambition is to specify the kinds of tools that will be required to help data providers efficiently create the required metadata and help the funders monitor and confirm that the metadata is of suitable quality to enable future discovery, exploration, use and, in particular, exploitation, of the data.

ESA ASTRONOMY MULTIMISSION INTERFACE (MMI) AT ESAC: IMPLEMENTATION AND SERVICES

Fabrizio Giordano , Deborah Baines , Ignacio Leon , Bruno Merin , Jesus Salgado

ESA/ESAC

ABSTRACT

ESDC (ESAC Science Data Centre) is on charge of the archiving and distribution of data from ESA astronomical missions, handling with different wavelength ranges, from infrared to high energy and, also, in totally different stages of life, like missions in development, operations or legacy. Historically, individual per project archives have been developed for the different missions: Herschel, Planck, XMM-Newton, Hubble, Exosat, ISO, Gaia etc. Also, a close integration with the science operations centres is also done during the development phase like, e.g. in the Euclid case for the integration of the archive with the operations phase.

However, although this heterogeneous access is needed to fully exploit the access for project experts, a multi-mission and project agnostic interface would be quite useful for the general scientific community and to enable multi-wavelenght science use cases. Following this approach, ESDC has created the first version of the MultiMission Interface (MMI) that enables a fast access to data from different ESA astronomical missions, removing whenever possible the project dependent language and using only scientific terms.

MMI proposes a very endearing interface through which scientists can access to scientific data present into different ESAC Science Archives by simply accessing to a single web page. MMI aims to be a useful starting point for scientific analysis on astronomic objects.

It allows the user to navigate through different maps (HiPS) generated for different missions and/or satellite instruments. Through those maps the user can perform search for multiple targets using their names or coordinates. The results are directly drawn on the full sky as interactive and precise footprints.

Once a search is performed it is possible to navigate through observation data divided by missions, through TAP connections, retrieve the real data directly from each archives and displaying postcards if any. On the full sky, on the top of maps, the user can choose to display targets coming from relevant catalogs.

MMI has been developed using last technologies and integrates third party software like Aladin Lite and Simbad search engine. By now MMI integrates maps, observation and catalogs coming from XMM, HST, Herschel, Plank and ISO missions.

The integration and interaction with a very big amount of metadata, which can increase daily, coming from different missions by minimizing the coupling on the availability of other archives. This is why MMI is essentially a big work of integration and data mining of metadata.

As a next steps for MMI we will focus on the future archives integration, the generation of new accurate and interesting maps and some other features like a online dynamic spectra preview viewer.

NORWEGIAN SATELLITE EARTH OBSERVATION DATABASE FOR MARINE AND POLAR RESEARCH

Øystein Godøy [1] , **Johnny Johannessen** [2]

(1) Norwegian Meteorological Institute , (2) Nansen Environmental and Remote Sensing Center

ABSTRACT

The Norwegian Satellite Earth Observation Database for Marine and Polar Research (NORMAP) is creating a data repository for the Nordic Seas and Arctic Ocean from polar orbiting Earth Observing satellites, serving research and application in marine, polar and climate sciences. The main focus of the NORMAP infrastructure (http://normap.nersc.no/) is to facilitate easy and seamless access to remote sensing products for the scientific community who may not be expert satellite users.

NORMAP is developing a distributed data management system which integrates physically separated data repositories (currently 3 in Norway, one in France). These data repositories are filled with existing and new products generated from satellite remote sensing data. To support the scientists working with remote sensing products, a Python toolbox, NANSAT, has been developed (https://github.com/nansencenter/nansat/wiki).

NORMAP is a metadata powered system relying on standardised metadata. Metadata are used for two purposes. First for data discovery, i.e. the process of finding the relevant products for a scientific purpose. Second for data utilisation, i.e. understanding the data accessed.

Following discussions with user communities, users articulate a clear priority on functionality. Their primary request is to be able to find and access relevant products. In the process of determining whether a dataset is relevant or not for their scientific work, they state that quick visual inspection of the products is important. Traditionally users of remote sensing products have had strong skills in computer science. Generally these users are capable of tweaking the remote sensing raw data and products to the form suitable for their purpose. NORMAP is trying to extend the user community of remote sensing products with new scientific users that are not experts. In order to achieve this the web portal is offering functionality that allows users to transform products prior to downloading and analyses using tools they prefer. This process is currently focused on data reduction (in time and space as well as variable space).

To further exploit the potential of extended user services, NORMAP has started to test integration with the national e-infrastructure in Norway. This is aligned with European Grid Initiative and EUDAT efforts. If successful, this will ensure a truly interdisciplinary laboratory.

Current status and challenges experienced during the development of NORMAP as a distributed data management system are presented and discussed.

PRESERVING THE LONG TAIL IN A BIG DATA WORLD: FRAMEWORKS FOR E-INFRASTRUCTURES IN RESEARCH LIBRARIES

Angelina Kraft , Peter Löwe , Margret Plank , Lambert Heller

German National Library of Science and Technology TIB

ABSTRACT

Interoperable e-infrastructures for scientific data allow new degrees of openness for research and its resources, accompanied by a huge potential for scientists, inventors, industry and citizens for reuse and value adding. Ideally, e-infrastructures allow research data to be stored, found, managed, annotated, cited and curated in a digital platform available 24/7 and to be used by multiple (specialized) disciplines. A major challenge represents the fragmentation of the European and global research data landscape: While appropriate and innovative preservation strategies and systems are in place for the big data domains (e.g. environmental sciences, space, climate) the stewardship for long tail research data remains uncertain. Consequently, universities and research institutions are becoming more interested in collecting and providing access to data produced at their institution that do not fall within the scope of big data or discipline-based repositories. Moreover, researchers themselves start to look for services which facilitate data management processes. Apart from research libraries, data- and computation centers and their respective interactions are also being addressed. This condition brings new opportunities for libraries to provide data services, e.g. by forming co-operations with data centers. To develop sustainable e-infrastructures for the long tail, the right balance between standardization and invention, public and private, freedom and control, performance and cost, global and local is needed. Therefore, the German National Library of Science and Technology (TIB) co-operates with diverse data management initiatives that engage multiple players:

- Data generators that host small or medium sized scientific laboratories.
- Providers of technical know-how and architectures for e-infrastructures.
- Researchers who, being at the centre of the whole scientific enterprise, use data for their discoveries.

Here we present three examples of such co-operations within the academic framework of the TIB:

- The TIB|AV-Portal (www.av.getinfo.de) provides a web-based platform for quality-tested scientific videos from the realms of technology/engineering, architecture, chemistry, information technology, mathematics and physics. Available videos include computer visualizations, simulations, experiments, interviews as well as recordings of lectures and conferences. All video content can be reliably cited and shared via Digital Object Identifiers (DOI). The portal was developed in cooperation between the Competence Centre for Non-Textual Materials at TIB and the Hasso Plattner Institute for Software Systems Engineering. A key feature of the portal is the use of automated video and semantic analyses enabling pinpoint and cross lingual searches within the video content and to cite and publish such videos in a simple manner.
- The RADAR - Research Data Repository - collaboration project (www.radar-projekt.org). As robust, generic end-point data repository RADAR enables clients to preserve research results up to 15 years and assign well-graded access rights, or to publish data with a DOI assignment for an unlimited period of time. Potential clients include libraries, research institutions, publishers and open platforms which desire an adaptable digital infrastructure to archive and publish data according to their institutional requirements and workflows.
- The VIVO initiative uses the open source software VIVO to connect information about research activities across institutions according to linked open data (LOD) standards. In 2013, the Open Science Lab at TIB and other partners established a network of European institutions that expressed an interest in using VIVO. Its goal is to set up a VIVO web service with interfaces for processing and visualizing structured research information in close consultation with a small specialist community.

Connected interfaces to share research information (including metadata, citations and supplementary content) will ensure the interoperability of the systems listed above, adding credit and value to research data, along with new trust in libraries as preservers of knowledge (and the associated data).

EUMETSAT DATA CENTRE ARCHIVE OPERATIONS AND SERVICES

Peter Miu , **Harald Rothfuss**

EUMETSAT

ABSTRACT

This paper summarises the operational activities at EUMESTAT in the context of the Long Term Data Preservation and provides an overview on the offered services adding value to the mission and climate data contained in the EUMETSAT Data Centre.

The EUMETSAT Data Centre has moved from a secure Long Time Archive to a valuable long term data source accessible by the user community through the evolution of its operations and services.

The Data Access services offered by the Data Centre transform the “data stored” into “information”. Information is generated from the data through the transformation of EUMETSAT proprietary formats (referred to as “Native”) to “user friendly” formats. These user friendly formats offer visualisation, geospatial and aggregation capabilities to support data processing through the exploitation of existing meta-data conventions and standards. Examples of these formats are netCDF, HDF and McIDAS where many programming languages and tools are available “for free” from the Internet.

To further add value to the data stored, EUMETSAT as a leading member of the Global Space-based Inter-Calibration System international collaboration effort focuses on the inter-calibration of satellite sensors to improve the accuracy of their measurements. This directly benefits the generation of Climate Data Records (CDRs) which can be used in reanalysis and Climate studies.

To ensure unambiguous identification of the generated CDR’s, the EUMETSAT Data Centre will administer Digital Object Identifiers (DOIs), allowing easy citation of the CDR’s in research literature.

Additional elements to enhance the user friendliness and data access will briefly be addressed in this paper.

THE MTG MDAF PROCESSING FUNCTIONS AND IDPF - THE WAY FROM GROUND STATION TO DATA ARCHIVE

Maren Mohler-Fischer [1] **, Jens Auer** [1] **, Simon Hutton** [2]

(1) CGI Deutschland Ltd. & Co. KG , (2) CGI IT UK Ltd

ABSTRACT

This paper will discuss approaches and architectures for the data lifecycle, considering the Meteosat Third Generation (MTG) ground segment. In particular it considers the ingestion of data via the Mission Data acquisition Facility (MDAF), and the archival, retrieval and reprocessing of L0 by the Image Data Processing Facility (IDPF).

The MDAF is composed of two redundant ground stations which both simultaneously receive the payload downlink from the MTG spacecraft and the WAN connections to the EUMETSAT HQ in Darmstadt and to further MTG facilities including the IDPF. The IDPF formats the data into chunks for near real time processing and archives these as SIPs. The archived data can then be extracted from the archive and reprocessed by the IDPF.

In the MDAF framework CGI develops the MDAF Processing Functions. At first, the frames from the satellites are demultiplexed to the different virtual channels at both ground stations. Afterwards, the frames are sent channel wise from one ground station to the other via a WAN connection to allow the generation of a most complete data stream by consolidating the two data streams from both ground stations. This way frames missing only in one data stream can be replaced by frames from the other data stream. Afterwards, the frames are sent from both ground stations to the EUMETSAT headquarter, where another consolidation process takes place. After decryption the frames are distributed by the Protocol-End-Point (PEP) to the different data destinations as IDPF and MOF.

In the IDPF, the L0 processing divides the received packets into chunk files for data processing, according to flags in the instrument packets or by time. This is performed for diverse virtual channels in near real time. The resulting chunks for each virtual channel are also archived as SIPs. The archived data can be extracted as collections of archived chunks and reprocessed by the L0 processing to regenerate the data for L1 processing.

We will report on the challenges of processing at high data rate, on the complexity of the consolidation algorithm to cover all eventualities and on keeping the timeliness of the data by guaranteeing a fast data processing from the baseband receiver at the ground station to distributing frames to different facilities, to increase the integrity of the archived data. Also we will report on the experience of formatting near real time SIPs out of the resulting consolidated data and reprocessing archived versions of the data.

THE INTEGRATED CLIMATE DATA CENTER AT THE CENTER FOR EARTH SYSTEM RESEARCH AND SUSTAINABILITY (CEN), UNIVERSITY OF HAMBURG

Remon Sadikni , Viktor Gouretski , Annika Jahnke-Bornemann , Stefan Kern , Felix Ament , Detlef Stammer

Universität Hamburg

ABSTRACT

Easy access to easy-to-use and quality controlled climate and Earth observations is an important prerequisite for Earth System research. If these data sets are stored in a self-explanatory, easy-to-use format, their usefulness and scientific value increase. This is the guideline for the Integrated Climate Data Center (ICDC) at the Center for Earth System Research and Sustainability (CEN) at the University of Hamburg, which provides a reliable, quick and easy data access along with expert support for users and data providers.

The ICDC provides several types of world wide accessible in situ and satellite Earth observation data of the atmosphere, ocean, land surface, and cryosphere via the web portal http://icdc.zmaw.de. In recent months, data from socio-economic sciences has been integrated into the ICDC data base to enhance the interdisciplinary collaboration.

On ICDC's web portal, each data set has its own page. It contains the data access points, a short data description, the spatial-temporal coverage and resolution, and information about the data quality. Each web page closes with reference documents, contact information and information about how to cite the data. If necessary, data are converted in an easy to use format – netCDF or ASCII - after consistency and quality checks. These data sets can be accessed through the web page via FTP, HTTP or OPeNDAP. Using the Live Access Server, the user can visualize data as maps, along transects and profiles, zoom into key regions, and create time series. In both fields, visualization and data access, ICDC tries to provide fast response times and a high reliability.

The ICDC team supports users and data providers, e.g., regarding selection of data set and format. ICDC is closely linked to the CERA data base at the German Scientific Computing Center (DKRZ). This ensures long-term data storage and provision of Digital Object Identifiers. Additionally, ICDC contributes to studies related to data product quality assessment and improvement on international level.

The Integrated Climate Data Center offers a number of unique data sets and has received growing attention on national and international level. Since its redesign in 2011, the number of visitors of the web site and of data users is increasing. ICDC's vision is to become one of the major national providers for Earth observation data in cooperation with organizations such as, e.g., the German Meteorological Service (DWD) and Eumetsat, and to enhance its growing international reputation.

GAIA ARCHIVE AT ESAC: HANDLING BIG ASTRONOMICAL CATALOGUES AND EXTENDED SERVICES

Jesus Salgado , **Juan Gonzalez** , **Raul Gutierrez** , **Juan Carlos Segovia**

ESA/ESAC

ABSTRACT

Gaia consortium is in the process of creation of one of the more extended and accurated source catalogues in astronomy, to be released at the end of 2015. That represents a challenge on different areas: validation, big data management, comparison with other catalogues, visualization and the effort to create relevant web services to make accessible this information to the astronomers in an efficient way.

ESAC (European Space Astronomy Center), as main custodian of Gaia data, apart of being the main data center for distribution of Gaia data to the general community, is also in charge of the creation of services to share Gaia data with all the different groups inside the Gaia Consortium what represents of the biggest cooperation projects in astronomy,

We will present the architecture, design, current status and plans of the Gaia Archive at ESAC, that already allows complex queries based on an extended ADQL, combining data from different catalogues of 1 billion sources, the extended TAP interface (that includes user schemas and VOSpace areas), interactive crossmatch capabilities with user catalogues and Gaia data, full catalogue operations like, e.g. full sky histograms generation and other advanced features. As part of the design and implementation process, we have explored and tested different available modules, databases and techniques and, also, we will describe some possible enhancements that could be needed in the future.

Finally, we will present the new paradigm for future astronomical missions that will require the access to computing resources for the community close to the data, what can be considered a requirement for all future big scale astronomical missions.

ENRICHING THE NASA CURIOSITY ROVER DATA ARCHIVE WITH CONTEXT MOSAICS

Tom Stein [1] **, Matt Uyttendaele** [2]

(1) Washington University in St. Louis , (2) Microsoft Research

ABSTRACT

Single frame images are returned from NASA's Mars Science Laboratory Curiosity rover in the course of daily science operations. Some image mosaics (panoramas) from the grayscale Navigation Camera have been produced by the science operations team and released as part of the mission data archive in NASA's Planetary Data System (PDS). However, no mosaics have been archived for the color cameras (Mastcam and MAHLI) on board despite the presence of single frame images acquired expressly as sources for mosaicking by the science team.

The PDS Geosciences Node produces mosaics for the cases in which source images exist but no mosaic has been released in the data archive. These products, called context mosaics, add value to the mission archive by offering a contextual view of the Martian landscape along the rover's traverse. To date, more than 1500 context mosaics have been created and are available within NASA's Analyst's Notebook for Mars Science Laboratory (http://an.rsl.wustl.edu/msl).

Context mosaics are created with Microsoft Research's Image Composite Editor
(http://research.microsoft.com/en-us/um/redmond/projects/ice/) using well-understood techniques from the field of computer vision to solve for camera intrinsic parameters and camera pose based on matching features across the source images.

Navcam context mosaics are created by stitching radiometrically calibrated images and then applying a linear 2% stretch. Mastcam and MAHLI context mosaics use calibrated, linearized products as sources. Projection information for the context mosaics is available in the EXIF data that are part of the embedded JPG file header with each mosaic. Both the archive and context mosaics may show artifacts due to the camera positions on the rover mast, especially at distances relatively close to the camera.

The Analyst's Notebook incorporates an online mosaic viewer that delivers image data on demand. Context mosaics may also be viewed using the Microsoft Research HD View browser plugin that reprojects the mosaic on the fly. Mosaics may also be downloaded for offline use.

TOWARDS AN INTEROPERABLE DATA ARCHIVE

Ahmad Hammad [1] , Tobias Kurze [1] , Peter Krauss [1] , Joerg Meyer [1] , Jan Pothoff [1] , Bjoern Schembera [2] , Jos van Wezel [1]

(1) Karlsruher Institut für Technologie , (2) HLRS

ABSTRACT

Scientific and cultural organisations, international collaborations and projects have a need to preserve and maintain access to large volumes of digital data for several decennia. Existing systems supporting these requirements span from simple databases at libraries to complex multi-tier software environments developed by scientific communities. All communities see an increasing volume of data that must be stored efficiently and economically which today, is usually a combination of a dynamic proportion of storage on magnetic disk and on magnetic tape. The bwDataArchiv project at KIT and HLRS is developing an infrastructure for secure and reliable archival storage that functions as a uniform platform for multiple scientific domains and international projects. Access to the actual storage in the data center is through an abstracted bit preservation layer that offers features selected for long time storage such as special metadata tags, takes into account the higher latencies of tape or cloud storage and can be used for infrastructure as a service (IaaS) offerings. At the same time access to the storage remains backward compatible for existing applications. Several projects serving different communities i.e. HPC users, libraries, archives, using the interface are presented as are the collection of requirements and the architecture of the prototype implementation.

ADDING VALUES TO PETABYTES - IN-STORAGE PROCESSING, CONTINUOUS DIFFUSION AND USAGE-DRIVEN OPTIMISATION FOR THE MWA AND GLEAM ARCHIVE

Chen Wu , Andreas Wicenec , Dave Pallot

The University of Western Australia

ABSTRACT

The Murchison Widefield Array (MWA) data system manages over four Petabytes of data since its operation started in July 2013. On a single day, it serves more than 40 Terabytes of interferometry datasets to low frequency radio astronomers across four continents. One of MWA's science projects - the Galactic and Extragalactic MWA Survey (GLEAM) - has also curated over 500,000 images (30 TB) and 12,000 calibrated measurement sets (135 TB) available to MWA scientists through the IOVA query interface. We have developed both MWA and GLEAM data systems based on the Next-Generation Archive System (NGAS). In this paper, we describe techniques we have developed in the past two years for adding values to large volumes of MWA/GLEAM data within the archival storage. In particular, we will discuss three important value-creation activities and review lessons learnt on the following topics: 1. In-storage processing: Lessons learnt after we have developed an in-archive data processing framework that supports both batch and incremental processing of selected, continuously ingested file sets. 2. Continuous data diffusion: Techniques to manage complex, asynchronous data distribution and to deal with network optimisation for long-haul data transfer 3. User-driven Optimisation: techniques to extract valuable access patterns from real users in order to facilitate optimal file placement and disk cache strategies (sizing and eviction policies) using analytical methods

STANDARDS AND BEST PRACTICES USED BY NOAA'S CLIMATE DATA RECORD (CDR) PROGRAM

Daniel Wunder , W. J. Glance , X. Zhao

NOAA/NESDIS/NCDC

ABSTRACT

NOAA established a satellite Climate Data Record Program (CDRP) at its National Centers for Environmental Information (NCEI, formerly NCDC) to provide a systematic process flow to generate sustained and authoritative climate information from satellite data. The CDRP implements a unique approach in archiving not only the data products themselves, but also the software, ancillary data, and documentation which allow full transparency into how the CDR was created. CDRP guidelines align to production guidelines from Global Climate Observing System (GCOS) and WMO's Sustained and Coordinated Processing of Environmental Satellite Data for Climate Monitoring (SCOPE-CM) activity. Best practices, such as common maturity assessments, guidelines, and standards, are employed to facilitate both the transition of research algorithms to operational software, and the long-term preservation of the data. The sustained production of multi-decadal inter-calibrated satellite CDRs feed directly into climate applications and tools for ease of use across a broad community of users.

BEST PRACTICES FOR PERSISTENT IDENTIFIERS IN EARTH OBSERVATION ARCHIVES

Tyler Christensen [1] **, Mirko Albani** [2] **, Andrew Mitchell** [3] **, Iolanda Maggio** [4] **, Razvan Cosac** [4] **, Katrin Molch** [5]

(1) DLR-DFD Deutsches Zentrum fÃŒr Luft- und Raumfahrt/ Deutsches Fernerkundungsdatenzentrum , (2) ESA/ESRIN , (3) NASA Goddard Space Flight Center , (4) RHEA System, S.A. , (5) DLR-DFD Deutsches Zentrum für Luft- und Raumfahrt/ Deutsches Fernerkundungsdatenzentrum

ABSTRACT

Sharing and citing scientific data sources is becoming the global standard. Persistent identifiers (PIDs) allow data providers to permanently and uniquely identify a resource, track its use by the scientific community, and increase its visibility. To encourage and coordinate the implementation of PIDs in Earth Observation, the CEOS Working Group on Information Systems and Services (WGISS) recently developed a best practices document. The Best Practices include advice on choosing a PID system, assigning the IDs themselves, ensuring permanence, resolving the ID to the data resource, choosing an appropriate granularity, and documenting the PIDs properly. The recommendations are being applied in pilot implementations in DLR, ESA ESRIN, and NASA. Establishing common protocols and practices will help ensure interoperability among the global community of earth observation data providers. This paper will present the best practices, a few use cases, and the results of the pilot studies.

PDS4: THE NEXT GENERATION PLANETARY DATA SYSTEM IN THE BIG DATA ERA

Daniel Crichton [1] , **Steve Hughes** [1] , **Tom Stein** [2] , **Reta Beebe** [3]

(1) JPL Jet Propulsion Laboratory , (2) Washington University in St. Louis , (3) NMSU New Mexico State University

ABSTRACT

PDS4 is a Planetary Data System (PDS)-wide project to modernize from PDS version 3 to version 4 in order to support both the acquisition and distribution of data from a wide variety of providers and users, as an international platform for planetary science archives. The goals are to deliver high quality science products to the PDS, preserve and ensure the stability and integrity of the data in the PDS, and improve user support and usability of the data. The system is a distributed information services architecture that uses online registries to catalog data and web services across the PDS and international archives to track and provide its holdings to the world-wide planetary science community. Based on an information model-driven architecture its operational capabilities have been configured and deployed across its nodes providing access to both PDS3 and PDS4 data. Five software and data standards builds have iteratively increased capability and stability of the system. The International Planetary Data Alliance (IPDA) has endorsed PDS4 and data distribution from the first PDS4 mission, LADEE, is underway. Ten missions across multiple space agencies are now developing archives using the PDS4 standards and progress is being made towards international interoperability using the PDS4 software infrastructure. This paper will expand on PDS4 and its successful development and emergence as the science data archive for a highly diverse science community and how it builds a foundation for enabling management and access to massive planetary science data holdings.

RECENT PROGRESSES OF THE LOTAR INTERNATIONAL PROJECT TO SUPPORT LONG TERM ARCHIVING AND RETRIEVAL OF MODEL BASED DESIGN INFORMATION FOR THE AEROSPACE AND DEFENCE INDUSTRIES.

Jean-Yves Delaunay

Airbus Operations

ABSTRACT

Model Based Design becomes increasingly the general practice of all new aerospace and defence systems. This presentation will sum up the activities of the LOTAR international project to develop standards for long term archiving and retrieval of aerospace digital information, such as CAD, CAE or PDM information.

The LOTAR international project is finalizing the standards for long term preservation of CAD 3D Product and Manufacturing Information (PMI) representation information, allowing supporting retrieval for reuse and support in operation; pilots are also ongoing for LT archiving of CAD assembly with PMI graphic presentations.

Significant progress has been done with recommendations for long term archiving of 3D visualization information. A short status of progress of LOTAR standards for long term preservation of advances manufacturing will be also done, covering 3D composite design and additive manufacturing.

The presentation will then provide an overview of the LOTAR standards for long term archiving of product management information, with a focus on long term preservation of “as designed” configured product structure information.

It will also provide a summary of the activities of a new WG for LT archiving of Engineering Analysis & simulation information, focused first on structural analysis and loads information.

This presentation will sum up the 5 years roadmap of development of the LOTAR standards. It will emphasize the need to manage the LOTAR standards as part of the PLM interoperability architecture framework of the Aerospace and Defence industries, with the links to the associated implementer forums. Then, It will highlight the needs of governance and monitoring of interdependencies between the different standardization projects and associated organizations.

TEXT MINING FOR CONTENT ENHANCEMENT & COMPLIANCE

Raphael Hubain , Laurence Maroye

Université Libre de Bruxelles

ABSTRACT

In a context where information systems are increasingly interconnected, interoperability is a crucial challenge. Standards, the common basis for systems development, have a major role to play as an interoperability facilitator. Records management has not been left behind and our presentation focuses on this particular aspect: both ISO 15489 and MoReq2010 — two references in the field — recommend the use of classification plans, also called filing plans. Classifying a record gives it a default disposal schedule and helps to manage its retention rules. In so doing, the filing plan is a vital component and its quality could imply strategic legal, political or administrative impacts. However, developing and managing a classification plan are time-consuming tasks which require a large amount of resources. Since these tasks have no directly perceptible benefits, organisations tend to allocate them little funding. Automation could be a way to reduce the financing costs of filing plan management. But the point is that building and using a classification scheme requires a detailed human analysis; so making it automated is undoubtedly a very complex task. However, on the one hand, the natural language processing research area has made great leap forwards towards information extraction from textual data; and the information needed to develop classification plans can be found, most of the time, within textual documents. On the other hand, automated classification systems based on textual features extraction and machine learning algorithms are beginning to achieve outstanding results. Gartner predicts that those technologies will reach their maturity peak within two years. The presentation aims at presenting, (a) a state of the art in the records and document management requirements suggested by regulatory bodies, and (b) a literature review on how text mining and automated classification methods could make this compliance more financially accessible for small and medium-sized enterprises. In addition, a case study (c) will be carried out in order to assess the effectiveness of such solutions.

THE PDS4 INFORMATION MODEL DESIGN PRINCIPLES - HOW WELL DID THEY WORK?

John S Hughes [1] , Daniel Crichton [1] , Richard Simpson [2] , Mitchell Gordon [3] , Ronald Joyner [1] , Anne Raugh [4] , Edward Guinness [5] , Michael Martin [6]

(1) JPL Jet Propulsion Laboratory , (2) Stanford University , (3) SETI Institute , (4) University of Maryland , (5) Washington University , (6) Contractor

ABSTRACT

The Planetary Data System (PDS) has released Version 1.4 of the PDS4 Information Model, the primary component of the PDS4 Information Architecture. The information model is now stable and has been used by two active missions and several missions in various phases of development. The information model drives the PDS4 Information System using a multi-level governance structure that provides for common, discipline, and mission level management of the system's information standards. Information model design principles adopted include the following. The information model is defined using a formal language. It remains independent of implementation technologies. It defines a few fundamental data structures that do not evolve over time. Within data products, those data structures should be organized so the physical structure parallels the logical structure of the contents. It is extensible, allowing more complex structures of related fundamental structures. Archive data formats are independent of contemporary data provider and data consumer formats. It incorporates a standard data dictionary reference model. This paper will describe how well the design principles worked, problems encountered during development and their solutions, and how well the results meet the requirements of the multi-discipline planetary science community.

Government sponsorship acknowledged

EVOLVING METADATA IN NASA EARTH SCIENCE DATA SYSTEMS WITH THE COMMON METADATA REPOSITORY (CMR) AND UNIFIED METADATA MODEL (UMM)

Andrew Mitchell [1] **, Daniel Pilone** [2]

(1) NASA Goddard Space Flight Center , (2) Raytheon/Element 84

ABSTRACT

NASA's Earth Observing System (EOS) is a coordinated series of satellites for long term global observations. NASA's Earth Observing System Data and Information System (EOSDIS) is a multi-petabyte-scale archive of environmental data that supports global climate change research by providing end-to-end services from EOS instrument data collection to science data processing to full access to EOS and other earth science data. On a daily basis, the EOSDIS ingests, processes, archives and distributes over 3 terabytes of data from NASA's Earth Science missions representing over 3500 data products ranging from various types of science disciplines. EOSDIS is currently comprised of 12 discipline specific data centers that are collocated with centers of science discipline expertise. EOSDIS has continually evolved to improve the discoverability, accessibility, and usability of high-impact NASA data spanning the multi-petabyte-scale archive of Earth science data products.

Metadata is used in all aspects of NASA's Earth Science data lifecycle from the initial measurement gathering to the accessing of data products by end-user. Missions use metadata in their science data products when describing information such as the instrument/sensor, operational plan, and geographic region. Acting as the curator of the data products, data centers employ metadata for preservation, access and manipulation of data.

NASA has recently begun the development of a single, shared, scalable Earth Science Metadata repository, the Common Metadata Repository (CMR), as a high-performance, high-quality metadata engine for next-generation Earth Observing System Data and Information System (EOSDIS). The CMR's need for dealing with multiple metadata formats and underlying data models led to the development of the Unified Metadata Model (UMM), a common data model across metadata held in the CMR. The current version of the UMM identifies Science Collections, Science Granules, Science Services and Meta-Metadata concepts. To foster interoperability with other agencies and international partners, NASA is in the process of applying uniform international standards to Earth science metadata represented in the UMM using ISO 19115 "Geographic Information – Metadata format.

A common metadata standard across NASA's Earth Science data systems promotes interoperability, enhances data utilization and removes levels of uncertainty found in data products. With this approach, the Earth Scientist is provided with a consistent data representation as they interact with a variety of datasets that utilize multiple metadata formats. This presentation discusses the lessons learned and successes from NASA metadata catalogs, which directly influenced the design decisions taken the CMR and the UMM.

FROM DISCOVERY TO DOWNLOAD - THE EOWEB® GEOPORTAL (EGP)

Henry Rotzoll , Daniele Dietrich , Klaus Dengler , Bernhard Buckl , Stephan Kiemle , Torsten Heinen

DLR-DFD Deutsches Zentrum für Luft- und Raumfahrt/ Deutsches Fernerkundungsdatenzentrum

ABSTRACT

Interactive, responsive, user friendly and appealing state-of-the-art web portals help users to search, view and retrieve geospatial data, thus promoting the use of Earth Observation (EO) data in a wide-range application community. DLR's German Remote Sensing Data Center (DFD) has a long tradition of providing web-based access to EO data, starting nearly 15 years ago with the Earth Observation on the Web (EOWEB®) portal, followed by the EOWEB®-Next Generation (NG) user interface. In 2015, the existing EOWEB®-NG portal will be replaced with a completely new developed user interface that utilizes up-to-date technologies in order to provide better user experiences.

The EOWEB® GeoPortal (EGP) is the new DLR multi-mission web portal for interactive access to long-term archived earth observation data. EGP implements two innovative approaches. First, it seamlessly integrates the classic catalog-and-order functions of the current EOWEB®-NG portal with new browse-and-download features. Thus, the whole process from "Discovery" to "Download" is combined in one single tool. Secondly, it uses common interoperability standards of the Open Geospatial Consortium (OGC): the Catalogue Service Web (CSW) for data discovery, the Web Feature Service (WFS) and Web Map Service (WMS) for data viewing, as well as the Web Coverage Service (WCS) for data retrieval in its backend interface to underlying services. These standards enable EGP to easily communicate with the OGC interfaces of the DLR-DFD-Geoservice, as well as external OGC-compliant services. This allows the end user to individually visualize EO data products with additional geodata from any source on the same web platform. In addition, the EOWEB Heterogeneous Mission Accessibility (HMA) services are supported via the HMA CSW and HMA Ordering standards. HMA standards have been developed by the European Space Agency (ESA) in cooperation with other Earth Observation data providers and the OGC.

EGP is designed to offer an interactive single-page user interface on the web, similar to a desktop application. To reach this goal, the front end is built with the Java-to-JavaScript Framework Google Web Toolkit (GWT), and SmartGWT is used as a widget library. On the server side, EGP benefits from Spring as the primary Java framework for dependency injection, security and design pattern implementation. Expandability and customizability was emphasized early in the design phase of EGP. The core functionality of EGP is separated into detached modules, which are further subdivided into three different tiers (client, domain model and server). New functionality can therefore be easily added without changing the existing code base. Furthermore, EGP is customizable for the use of exclusive satellite missions or project portals.

This presentation will give an overview of the architecture, the integration of standards and the usability of the EOWEB® GeoPortal (EGP).

EXPECTED ADDED VALUE THROUGH PUBLISHING EO METADATA AS LINKED OPEN DATA - EXAMPLE EUMETSAT METADATA

Uwe Voges

con terra GmbH

ABSTRACT

EO data is mostly already described by metadata based on EO/Geo specific metadata schemas like ISO19115(-2) for EO collection metadata or the Earth Observation Profile of Observations & Measurements (O&M) for EO product metadata. This metadata is often made accessible for search and discovery by services implementing EO/Geo specific interfaces e.g. OGC Catalogue Services and profiles thereof.

But there are different problems with all those specific metadata models and discovery mechanism especially when outside the EO/Geo community it is tried to use it to find and get access to EO data.

Linked Open Data (LOD) is about linking arbitrary things (resources) described by the Resource Description Framework RDF. Statements about the resources are described by Subject, Predicate and Object Triples. Any kind of these objects or concepts is identified by Unique Resource Identifiers (URIŽs) ideally by HTTP URLs allowing someone to directly look up those objects. The most important thing is to include links to other URIs from different sources, allowing the discovery of more things. For a common understanding of the semantics, e.g. of the predicates, those must be based on common controlled vocabularies or ontologies (defined in RDFS / OWL). Important examples here are the Dublin Core, DCAT, DCAT-AP or GeoSPARQL (an extension to SPARQL providing a vocabulary for representing geo-data in RDF).

EUMETSAT is already providing metadata for their weather- and climate-related satellite data by ISO and HMA based metadata standards and catalogue services. This can be explored by using the EUMETSAT ProductNavigator and the EOPortal. The HMA metadata standards and catalogue services are not only of advantage for internal information structures and processes but also for the integration into external EO-/GEO-infrastructures as Copernicus/GMES, GEOSS, CWIC and INSPIRE.

Additional added value is expected when the visibility and usability of the EUMETSAT data outside the EO-/GEO-Communities would be significantly improved. A very good opportunity to do this would be the provision of metadata (describing and pointing to EUMETSAT EO collections and products) as Linked Open Data (LOD) based on the common controlled vocabularies and ontologies listed above. This would enable exposing the metadata in public accessible RDF-based (Open Data Portals, through SPARQL endpoints, linking with third-party features (e.g. natural disasters, traffic, population statistics) or other data (e.g., DBpedia, GeoNames), reasoning, getting easier access to the data (e.g. download or access to OGC WMSs services).

The paper will provide an example showing added value with LOD.

EUDAT B2FIND A PAN-EUROPEAN AND CROSS-DISCIPLINE METADATA PORTAL

Heinrich Widmann

DKRZ Deutsches Klimarechenzentrum

ABSTRACT

In recent years, significant investments have been made to create a pan-European e-infra-structure supporting multiple and diverse research communities. This led to the establishment of the community-driven European Data Infrastructure (EUDAT) project that implements services to tackle the specific challenges of international and interdisciplinary research data management.

The EUDAT metadata service B2FIND plays a central role in this context as a repository and a search portal for the diverse metadata collected from heterogeneous sources. For this we built up a comprehensive joint metadata catalogue and an open data portal and offer support for new communities interested in publishing their data within EUDAT.

The implemented metadata ingestion workflow consists in three steps. First the metadata records - provided either by various research communities or via other EUDAT services - are harvested. Afterwards the raw metadata records are converted and mapped to unified key-value dictionaries. The semantic mapping of the non-uniform, community specific metadata to homogenous structured datasets is hereby the most subtle and challenging task. Finally the mapped records are uploaded as datasets to the catalogue.

The homogenisation of the community specific data models and vocabularies enables not only the unique presentation of these datasets as tables of field-value pairs but also the faceted, spatial and temporal search in the B2FIND metadata portal. Furthermore the service provides transparent access to the scientific data objects through the given references in the metadata.

We present here the functionality and the features of the B2FIND service and give an outlook of further developments.

EUMETSAT ROAD TOWARDS TRUE INTEROPERABILITY: LESSONS LEARNED FROM THE WMO INFORMATION SYSTEM INTEGRATION

Michael Schick , **Guillaume Aubert**

EUMETSAT

ABSTRACT

The WMO Information System (WIS) is an infrastructure based and developed on international cooperation since early 2000 and declared operational early 2014. The WIS is a data access and retrieval infrastructure allowing to discover and retrieve on demand or via subscription datasets from multiple meteorological centres. The WIS is offering a set of product catalogue web portals which builds on ISO metadata standards, interoperable protocols and practices for exchanging those information. EUMETSAT has been an active WIS member and would like to share multiple lessons in particular with metadata standards and management that have been drawn from its integration within the WIS to achieve interoperability. Here are some of points that will be discussed within this talk: Interoperability is not just about providing technical interoperable services and information, managing current product metadata standards is complex, interoperable services should be made for the benefit of end-users, linking metadata and services for accessing the products.

LESSONS LEARNED FROM IMPLEMENTING A RESEARCH DATA MANAGEMENT SYSTEM FOR AN INTERDISCIPLINARY, LONG-TERM RESEARCH PROJECT

Constanze Curdt [1] , **Georg Bareth** [1] , **Dirk Hoffmeister** [1] , **Ulrich Lang** [2]

(1) Universität Köln , (2) Regional Computing Centre (RRZK), Universität Köln

ABSTRACT

The importance of research data management (RDM) has increased in many fields (e.g. earth sciences) in recent years, since new technologies and methods have facilitated the rapid creation of digital information and (research) data. Besides a proper storage and backup, also sharing and re-use of research data are getting more important. Especially, research that is conducted in the context of collaborative, interdisciplinary research projects is dependent on secure data exchange between all involved scientists. Consequently, the establishment of user-friendly RDM systems is essential, which support needs of the scientist (e.g. data collection, storage, sharing, data documentation). Metadata have an important role within RDM systems. They enable the accurate description of all involved research data and consequently, ensure their searchability and re-use. Available metadata schemas and standards should be considered.

In this contribution, we will present lessons learned from designing and implementing an RDM system for the DFG-funded (2007-2018) Collaborative Research Centre / Transregio 32 ‘Patterns in Soil-Vegetation-Atmosphere Systems: Monitoring, Modelling, and Data Assimilation’ (CRC/TR32). The CRC/TR32 is an interdisciplinary, long-term research project between several research groups of the German Universities of Cologne, Bonn, and Aachen, as well as the Research Centre Jülich. The main research aim of the scientists is to yield improved numerical Soil-Vegetation-Atmosphere models to predict CO2-, water- and energy transfer by calculating the patterns at various spatial and temporal scales in the study area of the river Rur catchment. To achieve this research goal, several scientists from various fields are involved (e.g. soil and plant sciences, hydrology, geography, geophysics, meteorology, remote sensing, and mathematics). They create various, heterogeneous data in different spatial and temporal scales. These result from several field measurement campaigns, meteorological monitoring, laboratory studies, and modelling approaches. Additionally, all scientists create a lot of publications, conference contributions or PhD reports, which also have to be handled.

The self-designed CRC/TR32 project database (TR32DB, www.tr32db.de) has been online since early 2008. Since then the TR32DB has been continuously developed further. The main goal of the TR32DB is the accurate collection, storage, and backup of all created project data with accurate metadata. This should support the data sharing between all involved project participants during the project duration, as well as re-use of all created data beyond the project funding in 2018. Besides the framework conditions of the CRC/TR32 participants, also requirements of the DFG had to be considered (e.g. cooperation with a computing centre or library, Good Scientific Practice etc.). Consequently, a sustainable system was designed. The system is set-up in cooperation with the Regional Computing Centre (RRZK) of the University of Cologne, were it is also hosted. The TR32DB is organised in a three-tier architecture applying available hardware and software components of the RRZK. All project data are stored in a file-based data storage, operated by the Andrew File System. Corresponding metadata and administrative data of the TR32DB are managed in a MySQL database. A webinterface including web mapping components is provided for user-access. The centrepiece of the TR32DB is the self-designed and implemented multi-level metadata schema. This schema supports the accurate, interoperable description of all considered data types in the TR32DB (e.g. geodata, data, publications, reports, pictures). Besides Dublin Core as a basis schema, also other metadata standards and schemes are involved (e.g. ISO 19115, INSPIRE, DataCite), as well as CRC/TR32 specific properties (e.g. keywords, themes). All properties and sub-properties are described in detail in the ‘TR32DB Metadata Schema for the Description of Research Data in the TR32DB’.

Overall, the TR32DB supports common features of RDM systems. These include data storage, backup, exchange, publishing, search (e.g. combined search, map search), and provision of DOIs (Digital Object Identifier). The self-designed metadata schema enables the accurate description of all TR32DB data. A userfriendly wizard enables the input and modification of the metadata via the web-interface. Finally, the establishment of the TR32DB in the hardware environment of the RRZK ensures the long-term accessibility, availability, and consequently the re-usability of all project data beyond the CRC/TR32 funding.

KEYWORDS

Research data management, metadata, metadata schema, case study, data repository, interdisciplinary project

CERTIFICATION OF DIGITAL ARCHIVES - STATE OF THE ART

John Garrett [1] , David Giaretta [2] , Robert Downs [3] , Steve Hughes [4] , Mark Conrad [5] , Bruce Ambacher [6] , Simon Lambert [7]

(1) Garrett Software , (2) Giaretta Associates , (3) CIESIN Columbia University , (4) NASA , (5) NARA , (6) NARA retired & University of MD retired , (7) STFC UK Science and Technology Facilities Council

ABSTRACT

A great volume and variety of scientific data is being produced. Funders, both public and private, are keen to ensure that the funds they put into this is not wasted and one important way is to ensure that the data is re-used, immediately and into the future. Firstly this requires that adequate “metadata” is created and accompanies the data. Secondly it requires that the data and metadata are preserved.

Other efforts, e.g. RDA and CCSDS, are trying to put in place standards and tools which will guide and help the production of the “metadata” up to and during the production of data. Then the information must be handed over to a repository, but significant questions must be answered. Which repository should be used? Is the repository doing digital preservation well enough? Can it be trusted?

The Reference Model for an Open Archival Information System (OAIS) standard is almost universally recognized and since it was originally published in – 2002 - is a veritable grey beard of the digital archiving community. Most digital archives claim to be compliant and in fact most are making capable and good intentioned efforts to do so. However, managers and funders of the archives may want those claims verified by outside observers.

There are a number of review processes in this area. However there is only one which allows funders to have the confidence that a repository has been evaluated using the same well established procedures and degree of cross-checking that we all rely on in much of the rest of our lives, from the safety of our food to the security of our personal data held by governments.

Recently two additional standards were published. ISO 16363:2012 provides a set a specific metrics to verify adherence to OAIS intents. ISO 16919:2014 specifies how audits should be conducted and the competences audit teams need to allow them to make consistent and reliable judgements about the capabilities of repositories. This ensures that these audits are carried out using ISO’s long standing and internationally recognized auditing process.

This paper will describe the efforts that are underway to develop nationally recognized digital archiving auditing bodies and the individual auditors those bodies would require. Initial contacts have been made with a number of ISO National Bodies in order to work with them to accredit the auditors and the companies performing the audits. In the meantime, several of the authors of the above standards, through Primary Trustworthy Digital Repository Authorisation Body (www.iso16363.org) are conducting high-level classes aimed at professionals in current auditing organizations and existing archives that are interested in performing audits or undergoing an audit to obtain certification.

DEVELOPMENT, PRODUCTION, PRESERVATION, AND DISSEMINATION OF NOAA CLIMATE DATA RECORDS (CDRS)

Walter Glance [1] , Xuepeng Zhao [1] , Edward Kearns [1] , Daniel Wunder [2]

(1) NOAA/NESDIS/NCDC , (2) +++OTHER+++

ABSTRACT

Climate Data Records (CDRs) reveal Earth’s short and longer-term environmental changes and variations, allowing scientists and decision makers across society to better understand the climate system; assess the state of the climate on regional, national, and global scales; project future climate states; and inform economic decisions impacted by future weather and climate. NOAA’s CDR Program (CDRP) at the National Centers for Environmental Information (NCEI) is leading the end-to-end management for NOAA’s generation of operational climate data records of the atmosphere, oceans, and land. The CDRP mission objective is to develop and implement a robust, transparent, sustainable, and scientifically defensible approach for developing, producing, preserving, and provisioning CDRs generated from NOAA (and NOAA partner) operational satellite observations and in-situ measurements that span decades. In this paper, we will share our experience and lessons learned from developing and performing data stewardship for NOAA’s operational CDRs over the past six years. The experiences and lessons learned are gleaned from CDR development, sustained production, preservation, curation, and distribution as well as our most recent challenges and current approach to adding value for users. Note: The newly formed NCEI is a merger of NOAA's National Oceanographic (NODC), National Climatic (NCDC), and National Geophysical (NGDC) Data Centers. https://www.facebook.com/NOAANCEIclimate

LONG-TERM DATA PRESERVATION IN HIGH ENERGY PHYSICS: STATUS, LESSONS LEARNED AND 2020 OUTLOOK

Jamie Shiers

CERN

ABSTRACT

The world's major High Energy Physics (HEP) laboratories, institutes and experiments are collaborating together on addressing the problems of long-term data preservation for future re-use. This presentation will summarise the results of the first DPHEP Collaboration Workshop - to be held at CERN on June 8th and 9th 2015 - the first DPHEP workshop since a core set of institutes have signed a Collaboration Agreement. The draft agenda for the workshop can be found at https://indico.cern.ch/event/377026/other-view?view=standard. CERN's LHC experiments have recently adopted data preservation and access policies that include making meaningful subsets of their data available after an embargo period. They have also adopted a common set of Use Cases driving the need for preservation. A simple cost model has been elaborated, discussed and approved with the funding agencies and now forms part of the medium and long-term plan for these experiments. Some open issues still exist, not least of which concerns the effort and resources required for open access to significant and growing volumes of data. There are also significant data sets from other experiments at HEP laboratories around the world: these data are often unique in one or more key scientific aspect. The strategies adopted by the various experiments and laboratories vary depending on local and external factors (e.g. some formerly key laboratories are moving out of the HEP field, raising important questions concerning long-term funding. These aspects and others - particularly those that focus on collaboration both within HEP and with other projects and disciplines - will be summarised, including what technologies and experience we can bring to the table.

THE MISSION TO PRESERVE LONG-TERM

ROBERTA SVANETTI [1] **, Luciano Ammenti** [2] **, Luca Dominici** [3] **,**

(1) Dedagroup ICT Network , (2) Biblioteca Apostolica Vaticana , (3) DEDAGROUP Spa

ABSTRACT

The target of the Vatican Library in the last 500 years has been to preserve for future generations the collections of manuscripts held. With the project keeps long-term digital started in 2012 we laid the groundwork to allow the use of the manuscripts of the Vatican Library in digital format thus ensuring both the conservation of the collections that their disclosure.

The recurring theme in recent times is the management and long-term storage of big data generated by numerous sources of information or the need to analyze and store data from analog sources from which it is crucial to extract logical sequences comparative.

The traditional methods offer rigid procedures governed by protocols that have not been upgraded with the same evolution suffered by the possibility of processing data.There is no standard format for storing images and there is no standard format for storing text and to ensure the enjoyment even after time spans that last more than 50 years.

Try to think for a moment to have the option of not having a priori choose which documents to preserve and which to discard, not to worry about their size and to have clear how to ensure future access to those documents. Try to imagine this "big sea" of documents can generate synapse junction thematic or logic that increase day by day and that populate the search indices allowing the user an innovative result in the expression of deductive logic of research.The research not only to better designed themes expressed in the user-defined parameters, but the data suggested by the system result of other related issues and projections relating to research arising from the same memories of previous consultation processes.

Think finally the ability to use a standard format for storing images with which to keep all the processed data and all documentation digitized also those derived from an analog process, tie any information to an alphanumeric key inserted in the PREMIS it possible to rework the search before processing the data, and also to have an instrument of "PATTERN SEARCH RICOGNICTION" capable of analyzing all the images and select them according to a chosen theme.