An overview on the reuse of data in the science domain

Community of Practice on Reuse of Science Data

A report written by PERICLES Project partners in collaboration with participants of the Community of Practice (CoP) related to the Reuse of Science Data.

Motivation behind the Community of Practice - an Introduction

The “Re-Use of Science Data” Community of Practice (CoP) was inspired by the Brussels workshop 'Promoting the reuse of science data' which took place on 29 October 2015. We refer to the published deliverable D9.2.2 “Updated Version of Dissemination Plan”[1] for the notion we accepted here for defining a CoP. In contrast to a CoI (Community of Interest, which in fact comprise the whole community interested in the topic of reuse of science data) or to a SC (Stakeholder community who are actively engaged with granting reuse or actually reusing scientific data), the CoP is a small group composed of selected experts with the objective to discuss a current fundamental question or set of issues. The purpose is the exchange of ideas, research and perspectives amongst experts working on a topic closely related to the reuse of data, to find inspiration for one’s own work and help “thinking out of the box”.

The PERICLES project instigated the CoP as part of their engagement with experts outside the project. The CoP was led by the use case partner B.USOC who combined two key aspects that distinguished them from the second use case partner TATE in that the latter has a different understanding of “re-use” when thinking of art objects than the science world has when “re-using” research data. And while TATE’s core remit is preserving their collections for exhibition and safeguarding their value by protecting them from damage, B.USOC as an operations centre does not currently employ preservation techniques as they are not in its ESA mandate. Preservation and access for reuse is limited to specific sets of data and currently this limitation and practice is highly disputed and widely discussed between the science teams and ESA. Reuse of science data has also a fundamental difference with restitution or adaptation of art data, it produces new intellectual property as new products are generated from the reused data. this extends the intellectual property rights of the original data providers and owners by adding the IPR’s of the data re-interpreters. In art, reusing an object by putting it in a different art context as for example the Marcel Duchamp fountain creates an intellectual property on its own and the “Armitage Shanks” bathroom supplies company never contested the Marcel Duchamp claim. The preservation and new installation of the Marcel Duchamp fountain cannot be counted as work of art in its own. The same reasoning should apply to digital objects associated with art.

PERICLES has researched ways to meet the challenge of impact of change in digital environments and its potential impact on reuse. This would strongly support the remit of preservation as it is being understood in the preservation community, in which assets are deposited in the custody of an archive to support long retention. However, the PERICLES approach is much broader, applicable to any type of digital ecosystem whether a preservation management system is in place or not. Thus the discussion on reuse of science data is not primarily driven as a preservation issue. The concerns addressed in these discussions were fed back into the project to inform the research and get a clearer understanding of the challenges of complex digital objects provided by and of concern to the science community.

[1] http://pericles-project.eu/uploads/files/PERICLES_WP9_D922_Updated_Dissemination_Plan_V01.pdf, p.6

The CoP members

In recruiting members for the CoP, we not only invited participants at the above mentioned workshop on 29 October, but also used the USOC network (User Support and operation Centre, put in place by ESA for the ISS support), the IPCC solar forcing community and the BigSkyEarth (BSE) COST action (TD1503), which joins big data users in astronomy and earth sciences. We spread the word, and presented the PERICLES project at the Lyon workshop of BSE on October 19 and 20 2015 to interest experts to join the CoP. Though the audience expressed a keen interest after the presentation, it needs to be said that joining a CoP is not really common practice in the science community. As many people are very busy, finding a date when more than a small group could regularly meet proved difficult.
Eventually, the CoP brought together science data experts from EURIX, BIOTESC, DANS, CERN, IPCC and STFC with PERICLES partners B.USOC, King's College London and University of Liverpool.

B.USOC (Brussels, BE)
Christian Muller (Belgian User Support and Operations Centre, Knowledge Manager)

By its specific nature as an operations centre dealing with space payloads on the ISS, B.USOC covers almost all aspects of science in space from physical experiments, life science to the more familiar space and earth observations. Experiments in the pressurised modules have a lot in common with laboratory experiments as the conditions are controlled and if they can be proven to be independent from the external environment, they can be reproduced. Their main specificity is microgravity which can only be achieved in orbit over long periods of time. However, the costs involved make the re-enactment of flown experiments is prohibitive and thus lead to a need for comprehensively archiving the data even if the main data lie in the processed samples returned to earth. Ancillary data are usually transmitted to the operation centres as well as digital science data. In the case of earth and science observations, each science data element is unique as in nature. There are no elements which are not subjected to change. This is especially true for earth observations, where since the beginning of the space age human activity has been rapidly modifying atmosphere, land and ocean. The preservation of time series becomes essential to find new phenomena which nobody envisaged at the outset by analysis of periods and trends. The use case which served through the project was the SOLAR set of instruments, an optical package flown on the ISS between February 2008 and February 2017 in the ISS and providing a monitoring of the total and spectral energy output of the sun as received at the earth's orbit. Some components of SOLAR are part of a series beginning its flights in 1976 so that together with comparable NASA instruments, the solar output is now monitored for around forty years.

King's College London (London, UK)
Simon Waddington (King's College London, Research Fellow working on PERICLES)

The Centre for e-Research (CeRch), which is leading PERICLES at King's College London, is a research centre situated within the Department of Digital Humanities, an academic department that is an international leader in the application of technology in the arts, humanities and social sciences. CeRch brings to the Department its own overlapping and complementary focus on topics from information science and e-Research more broadly, addressing methodological crossovers between humanities, sciences, information science and computer science.

University of Liverpool (Liverpool, UK)
Fabio Corubolo (University of Liverpool, Research Associate working on PERICLES)

The University of Liverpool holds particular specialisation in the development and integration of data management (high performance computing), digital library, and persistent archives technologies. The faculty represented in PERICLES founded the UK National Text Mining Centre (NaCTeM), and as a result of this initiative demonstrated the integration of client and server workflows within data management systems to manage transformations needed to display, search, and manipulate collections of records held in the integrated Rule Oriented Data System (iRODS).

FORGET-IT (FP7 project)
Walter Allasia (Chief Research Officer at EURIX Group, Turin, Italy)

ForgetIT is a EU-funded project (2013-2016), which brought together an interdisciplinary team of experts in preservation, information management, information extraction, multimedia analysis, personal information management, storage computing, and cloud computing, as well as in cognitive psychology, law, and economics, who developed three new concepts to ease the adoption of preservation in the personal and organisational context: “Managed Forgetting”, “Synergetic Preservation” and “Contextualised Remembering”. The implemented Framework embodies a first step towards a promising alternative to the prevailing “keep it all” approach in our digital society.

BIOTESC (Luzern, CH)
Fabian Ille (BIOTESC, Senior Research Associate, life science data in space)

BIOTESC is the service branch of the Center of Competence in Aerospace Biomedical Science and Technology. BIOTESC is the Facility Responsible Center for KUBIK, a transportable incubator, and the Facility Support Center for BIOLAB, a facility for biological experiments in the Columbus module. The centre is located at the University of Lucerne.

DANS (The Hague, NL)
Ingrid Dillo (Data Archiving and Networked Services, Deputy Director)

DANS promotes sustained access to digital research data. For this, DANS encourages scientific researchers to archive and reuse data in a sustained form, for instance via the online archiving system EASY and DataverseNL. With NARCIS, DANS also provides access to thousands of scientific datasets, publications and other research information in the Netherlands. The institute furthermore provides training and consultancy and carries out research on sustained access to digital information. Driven by data, DANS ensures the further improvement of access to digital research data with its services and participation in (inter)national projects and networks. DANS is an institute of KNAW and NWO.

CERN (Geneva, CH)
Jamie Shiers (CERN, Data Preservation Project Manager)

CERN, the European Organisation for Nuclear Research, is one of the world's largest and most respected centres for scientific research. Founded in 1954, CERN is situated just outside Geneva,

extending into France and now has 22 member states. The organisation has been leading some key work in probing the fundamental structure of the universe and uses the world's largest and most complex scientific instruments to study the basic constituents of matter – the fundamental particles. The particles are made to collide together at close to the speed of light using purpose-built accelerators. The Large Hadron Collider (LHC) is by far the largest (27 km) and most powerful particle accelerator built to date. CERN has also been conducting some exceptional work in developing data mining and data preservation.

IPCC (Geneva, CH)
Benjamin Laken (Research Software Engineer in the Information Services Division of UCL, London)

The Intergovernmental Panel on Climate Change (IPCC) is the leading international body for the assessment of climate change. It was established by the United Nations Environment Programme (UNEP) and the World Meteorological Organization (WMO) in 1988 to provide the world with a clear scientific view on the current state of knowledge in climate change and its potential environmental and socio-economic impacts. In the same year, the UN General Assembly endorsed the action by WMO and UNEP in jointly establishing the IPCC.
Benjamin Laken joined the CoP whilst working at University of Oslo in the role of Researcher in climate data. He changed affiliation during the duration of the CoP but remained an IPCC climate expert.

STFC (Oxford, UK)
Esther Conway (Senior Earth Observation Data Scientist at the Centre for Environmental Data Analysis (CEDA) based within the RAL SPACE department at STFC) and Catherine Jones (Software Engineering Group Leader at STFC)

The Science & Technology Facilities Council (STFC) is one of Europe's largest multidisciplinary research organisations supporting scientists and engineers world-wide. STFC support an academic community of around 1,700 in particle physics, nuclear physics, and astronomy including space science, who work at more than 50 universities and research institutes in the UK, Europe, Japan and the United States, including a rolling cohort of more than 900 PhD students.
All represent different aspects of data preservation and reuse. There was a special focus on climate data and reports, as they present an encyclopaedic character and are based on long term observations beginning before the digital age. The climate data born in the space age grow each year in complexity as exemplified by the PERICLES use case SOLAR. Climate data from several sources must thus be merged in coherent series resolving the horizontal (missing data) and vertical gaps (two data sets for the same period have a systematic difference). The climate problem requires thus constantly the reuse of data acquired in various times and locations. Life sciences require also data preservation, certainly in space as space experiments are not likely to be repeated and also in the case of clinical data, old data might contain information about pathogens which were unsuspected at the time of acquisition.

The discussions

Four online meetings were organised on January 18, March 1, April 27 and June 21 of 2016. Each time a loose agenda was proposed including presentations by participants. The PERICLES team

avoided to be directive and led the participants define their priorities leading sometime to repetitions from one session to the other.
Three themes were recurring: the comparison between an already existing Long Term Data Preservation (LTDP) approach and the one proposed in PERICLES, this theme includes a recurrent discussion on OAIS compliance and other models and their usefulness for reuse of science data. The other two themes were appraisal with an emphasis on the question of selecting the data to be preserved, the second was the role of data policy in supporting or constraining the reuse of data. The very important point of linking data acquisition and the operational chain with data preservation and reuse was unfortunately never directly fully addressed. This is due that all our experts with the exception of Fabian Ille work with the downstream data flux and that we could not convince operational personnel to participate in our Community of Practice.
Three large formal presentations were discussed and compared with PERICLES research: the CEDA (Centre for Environmental Data Archiving, UK) approach, the DANS (Data Archiving and Networking Services, The Netherlands), and the ForgetIT FP-7 project. The meetings centred around ideas and statements made in the respective presentation by one member.

Long-term preservation

This topic was introduced through a presentation by Esther Conway of the CEDA archive of STFC. CEDA is especially interesting as it merges several data sources including ESA, EUMETSAT and the relevant data archives of the German Max-Planck institution. CEDA operates in both a computing and an archiving environment and satisfies the needs of the present British scientists. It manages a large number of formats, around 140. However, the data remains in the same form as provided by the original sources and as in the ESA databases, the amount of metadata is insufficient to fully provide replays of the acquisition. Most of the data is unfortunately high level, due to the funding situation. Low level data is not necessarily archived. Despite its limits CEDA presents the advantage of providing an operational service. Maintenance and distribution of this data set ensures by itself a form of preservation.
DANS also puts considerable effort into digital preservation and was presented by Ingrid Dillo. As DANS is not specifically tuned to geophysical data, it is less confronted with “organically” growing data sets. As in CEDA the main advantages are existence and data distribution. The presentation delivered by Ingrid Dillo can be found here.
The question of long term stability of the CEDA and DANS archives was not considered, space agencies have rules which limit the time duration of their archives and we assume that on the contrary, memory institutions would keep data for longer periods. Our panel could have included institutions which find their origin in binding international treaties as the World Data Centres set by the International Geophysical Year (1957-1958) and the Antarctic treaties. Due to their long duration and continuation of original procedures (printed records), they are now threatened by obsolescence and would benefit very much from at least the user oriented approach of CEDA or the even more dynamical approach proposed by PERICLES. This could help turn them from data graveyards to actual virtual geophysical observatories. This discussion could lead to a project in itself designing a roadmap for the evolution of data memory institutions.
When presenting the model-driven approach proposed by PERICLES, the discussion commonly leads to comparison the OAIS (Open Archival Information System) to the continuum viewpoint represented in PERICLES by the LRM (Linked Resource Model). In the case of the Science Data CoP, this was less apparent and was only briefly dealt with at the occasion of the CEDA presentation. In fact, most practitioners do not frequently refer to these models despite the fact that OAIS was introduced as part of the space agencies’ agreement managed by the CCSDS

(Consultative Committee for Space Data System). Since 1982, more the 800 space missions have flown using the CCSDS standards but most users downstream in the space segment use other accepted standards such as ISO norms or the E.U. INSPIRE directive. This question of models would have probably been more important if the CCSDS could have been implicated on the CoP. This would have been impractical as CCSDS regroups 11 member agencies and 30 observer agencies, such an extension beyond the limits of the E.U. should again be a project in itself.

Appraisal

With regard to appraisal, data selection and quality control, among our participants, only the CERN policy mandates the preservation de facto of all raw data for which it provides state-of-the-art services. In addition, CERN provides services for the preservation of the associated documentation, software and the environment in which it runs. As in the B.USOC space case, the derived data are preserved under the control of the scientists. Since missions are very expensive to run, it is very important for B.USOC and CERN to preserve not just the data, but the whole processing chain including i.e. all the documents, telemetry, parameters etc. which can allow the future validation and correct interpretation of data.
CERN hosts this scientific archive while space agencies put it more downstream in specialised data centres. The other CoP participants were very curious to see how the PERICLES proposition of preserving all data and documents could work, and they saw it as particularly valuable to other domains like biological data. However, they were also realistic about the constraints that such an approach might have and the big amount of work and resources required to create a functional prototype and a viable product.
The PERICLES research focused on technical appraisal as the first candidate for developing an automated support for this type of appraisal. The discussions on appraisal led to Walter Allasia presenting the work done in the ForgetIT project (04/2014 – 3/2016).
The ForgetIT approach gives a fresh view of appraisal as this subject is its main objective. The FP7 project is based on the organisation of human memory and opened a new possibility for PERICLES application to the space case: the inclusion of data appraisal in the data acquisition process. Artificial intelligence is important for the development of space systems in which the data downlink is limited and for which control by a remote operation centre would be ineffective. The ideal solution would be to have a team of humans on location which is unrealistic even at some locations on the earth (volcanic craters, remote Antarctic location …). Introducing ForgetIT elements in the PERICLES appraisal tools would extend the PERICLES process to the entire data chain and not only to the a posteriori archiving and preservation of the data.
The memory institutions (DANS and CEDA) do not have as much interest in appraisal as the data has already been screened by the data originators.
The ForgetIT approach applied to planetary exploration led to a common PERICLES-ForgetIT oral presentation at the EANA (European Astrobiology Network Association) 2016 meeting which was devoted to space technology. More information on the ForgetIT Information Model can be found in this presentation by Walter Allasia.

Data policies

Data policies were discussed at each meeting as they impact a lot the mechanisms of data distribution and preservation. Two aspects exist for space data: first, the embargoes which limit data use to the original investigator team during a fixed period or which prevent data release

before validation and quality control; second: general archiving policy which limits the time the elements of the data chain have to keep the data they collect. In the case of ESA HRE (Human spaceflight and Robotic Exploration) which is the agency contracting and supervising B.USOC, it is limited to 10 years. The participants always agreed that these restrictions would limit both distribution and preservation capabilities, but did not make a recommendation for alternative policies. In particular, the time limitation implies that if the PERICLES objective of following semantic change are to be implemented, they would require a data management plan performing the necessary data manipulation and archiving before the end of the retention period.
Specific policy aspects were not discussed, in particular, the anonymisation of medical or personal data. Commercial aspects were also not considered as well as problems related to government secrets, for example, the geographical coordinates of a scene from a dual use earth observation satellite might reveal an ongoing military operation. All these restrictions should be dealt with in the case of long term preservation of these data.

Conclusions and recommendations

The Reuse of Science Data Community of Practice aimed to listen and record the opinions of practitioners in various fields. It was not easy to arrange due to the lack of freedom of personnel to take time out of their schedule. This is evident for operators whose activities and schedules are bound to their contracts. It was also the case for scientists who must balance their time between teaching assignments and research contracts. The biggest contributions were made by the two representatives of memory institutions and one data researcher.
PERICLES was well accepted in the group finally assembled as giving solutions to problems which they already had encountered. The different tools which were demonstrated in the final event of December 2016 could not be shown to the group in operation. Thus the generally positive opinion on PERICLES and its tools could not be substantiated by running a benchmark. Overall, the CoP evaluated the project positively, but does not make recommendations for the future except the ones which are already in most similar studies: standardisation, use of open source for sustainability, and avoidance of hardware bound solutions…
As the CoP did not influence the overall direction of the project, it might make sense in future similar projects to place similar CoPs at the end of the project as intensive workshops of several days taking place when most of the tools and deliverables are available
This CoP recommended proposing a new project in order to define a roadmap for data preservation based on a survey of users and practitioners both inside E.U. and on a world scale with an inventory of all living archives in earth and space science. This study would enable to draft a set of general recommendation which would have influences on standards.

This project has received funding from the European Union's Seventh Framework Programme
for research, technological development and demonstration under grant agreement no 601138.