Guide lines for observational big data preservation and reuse from the PERICLES project.

Christian Muller christian.muller@busoc.be
Belgian Royal Institute for Space Aeronomy
B.USOC
Avenue Circulaire, 3, B-1180 Brussels,
Belgium

Abstract

Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics (PERICLES) – was an Integrated Project funded by the European Union under its Seventh Framework Programme (ICT Call 9) and addressed work programme objective ICT-2011.4.3 Digital Preservation. It ran from 2013 to April 2017. B.USOC was in charge of managing a Community of Practice group regrouping several big data users and providers in the science domain including members of the BSE COST action. B.USOC performed also the detailed analysis of the SOLAR package of instruments which was active on the International Space Station from 2008 to 2018. As B.USOC was Facility Responsible Centre for this package, all aspects of the data flow were accessible to the PERICLES project. Also, due to the B.USOC relation to the Belgian Royal Institute for Space Aeronomy, precursor instruments flying on the Space Shuttle since 1983 were also available for study building thus a 35 years series. During this period, the paradigm of the experiment changed leading to semantic change.
The PERICLES recommendations will be presented as well as the actual data preservation procedures taken by the space agencies now that the space segment is deactivated.

Keywords: space data, preservation, data reuse

1. Introduction

PERICLES (Promoting and Enhancing the Reuse of Information throughout the Content Lifecycle exploiting Evolving Semantics) aimed at preserving by design large and complex data sets. PERICLES was coordinated by King’s College London, UK and its partners are University of Borås (Sweden), CERTH-ITI (Greece), DotSoft (Greece), GeorgAugust-Universität Göttingen (Germany), University of Liverpool (UK), Space Application Services (Belgium), XEROX France and University of Edinburgh (UK). Two additional partners provide the two case studies: Tate Gallery (UK) brings the digital art and media case study and B.USOC (Belgian Users Support and Operations Centre) brings the space science case study.
PERICLES addresses the life-cycle of large and complex data sets to cater for the evolution of context of data sets and user communities, including groups unanticipated when the data was created. Semantics of data sets are thus also expected to evolve, and the project includes elements

which could address the reuse of data sets at periods where the data providers and even their institutions are not available any more.

2. Choice of the space case: solar spectral irradiance

B.USOC supports experiments on the International Space Station and is the curator of the collected data and operation history for ten years. The B.USOC operation team includes B.USOC civil service personnel and Space Applications Services personnel. As a first test of the concept, B.USOC has chosen to analyse the SOLAR payload operating since 2008 on the ESA COLUMBUS module of the ISS. Observation data are prime candidates for long term data preservation as variabilities of the solar spectral irradiance have an influence on earth climate. The paradigm of these observations has already changed a lot in the last fifty years from a time where scientists were aiming at determining with high accuracy the “solar constant” which was the total solar energy per surface unit received at the top of the earth’s atmosphere to the present situation where the same quantity is known as the total solar irradiance and has been shown by thirty years of space observations to vary of about one tenth of a per cent in synchronism with the solar cycle. Right now, larger variations have been detected at UV wavelengths but their effects on climate and atmospheric chemistry are still a matter of scientific discussion.

By creating semantic links between various data bases, the PERICLES process can be applied to already linked data bases as the current set of earth observation data managed by ESA in ESRIN to optimise their future use as an element of the observational data base of future earth’s system models. PERICLES also presents a fundamental reflexion on the reuse and long-term preservation of data which corresponds to the needs of climate research. These arguments on long term data use apply also to the space science data bases hosted by ESA in ESAC.

The SOLAR payload is built from three complementary space science instruments that measure the solar spectral irradiance with an unprecedented accuracy across almost the whole spectrum: 17-3000 nm. This range carries 99% of the Sun's energy emission. Apart from the contributions to solar and stellar physics, knowledge of the solar energy flux (and its variations) entering the Earth's atmosphere is of great importance for atmospheric modelling, atmospheric chemistry and climatology. The three instruments are: SOLSPEC (Solar Spectra Irradiance Measurements, developed by CNRS,France and IASB/BIRA, Belgium) (Thuillier et al, 2010), SOL-ACES (Auto-Calibrating Extreme Ultraviolet and Ultraviolet Spectrophotometers, developed by the Fraunhofer Institute, Germany) (Schmidtke et al, 2006), SOVIM (Solar Variable and Irradiance Monitor, jointly developed by the Observatory of DAVOS, Switzerland and the Royal Meteorological Institute, Belgium). (Mekaoui et al, 2010) The three original PI’s agreed before flight to a synergistic treatment of the data [4].

SOLAR has in fact a much longer history than its flight on COLUMBUS. The precise measurement of the solar irradiance as input to the earth system began one hundred years ago when this parameter was known as the “solar constant”, space borne instruments in the last thirty years have shown variations of the total solar irradiance while spectral irradiance especially in the UV and have confirmed sporadic early balloon and rocket observations. The SOLAR instruments SOLSPEC and SOVIM were first designed for the SPACELAB 1 payload which flew on the US space shuttle in 1983, the decision to fly and the first design studies dating from 1975. After SPACELAB-1, ESA transferred the SPACELAB equipment to NASA and NASA flew these payloads several times to cover the solar cycle until the last COLUMBIA mission in 2003. Ideally, at least this set of missions should be regrouped with the SOLAR ISS data set to build a coherent series.

This task has been performed up to now by scientists from reviews of the published results in peer reviewed journals solving two different types of discrepancies: horizontal gaps as no data are

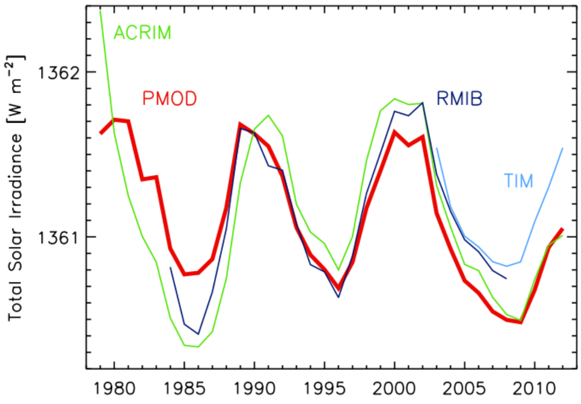

Figure 1: Annual average composites of measured Total Solar Irradiance: The Active Cavity Radiometer Irradiance Monitor (ACRIM) (Willson and Mordvinov, 2003), the Physikalisch-Meteorologisches Observatorium Davos (PMOD), (Frohlich, 2006) and the Royal Meteorological Institute of Belgium (RMIB) (Dewitte et al., 2004).These composites are standardized to the annual average (2003–2012) Total Solar Irradiance Monitor (TIM) (Kopp and Lean, 2011) measurements that are also shown. The RMIB and PMOD series correspond to SOLAR. (Composite from the IPCC report 5, 2013).

available in time and vertical gaps which correspond to disagreements between the series obtained by different instruments at the same time. These comparison and verification processes take place in formal bodies as COSPAR (ICSU committee on space research), the IPCC (Intergovernmental Panel on Climate Change) (Fig. 1) and other bodies essentially led by U.S. agencies or the European framework programme. None of these use a standard data reuse procedure as the one studied by the PERICLES programme.

3. Role of B.USOC in the data flow from the space segment to the science teams.

The ISS European payload operations are conducted through the distributed USOC network. As Facility Responsible Center, B.USOC monitors the data flow in order to assess the quality of the chain and to detect any malfunction that could impact the final data.
Beside sending nominal commands to the instrument in flight, B.USOC executes remedial actions to prevent failure in conjunction with the other elements of the command chain. Instrument monitoring is performed using a series of control screens which inform the operators in real time, this function requires physical presence on console 24 hours a day when the instrument is powered up. Such a requirement is also related to the specifics of manned flight where the safety of human crews requires the control of all on going processes including those for which astronaut intervention is not nominally involved.

The complex data path between the instrument and the scientist is illustrated on figure 2, this scheme involves both ESA and NASA and thus data comes to B.USOC from several sources. B.USOC uses its own software (YAMCS, Sela et al, 2012) to parse the data of a specific instrument and send it to its operators screens and to the scientists.

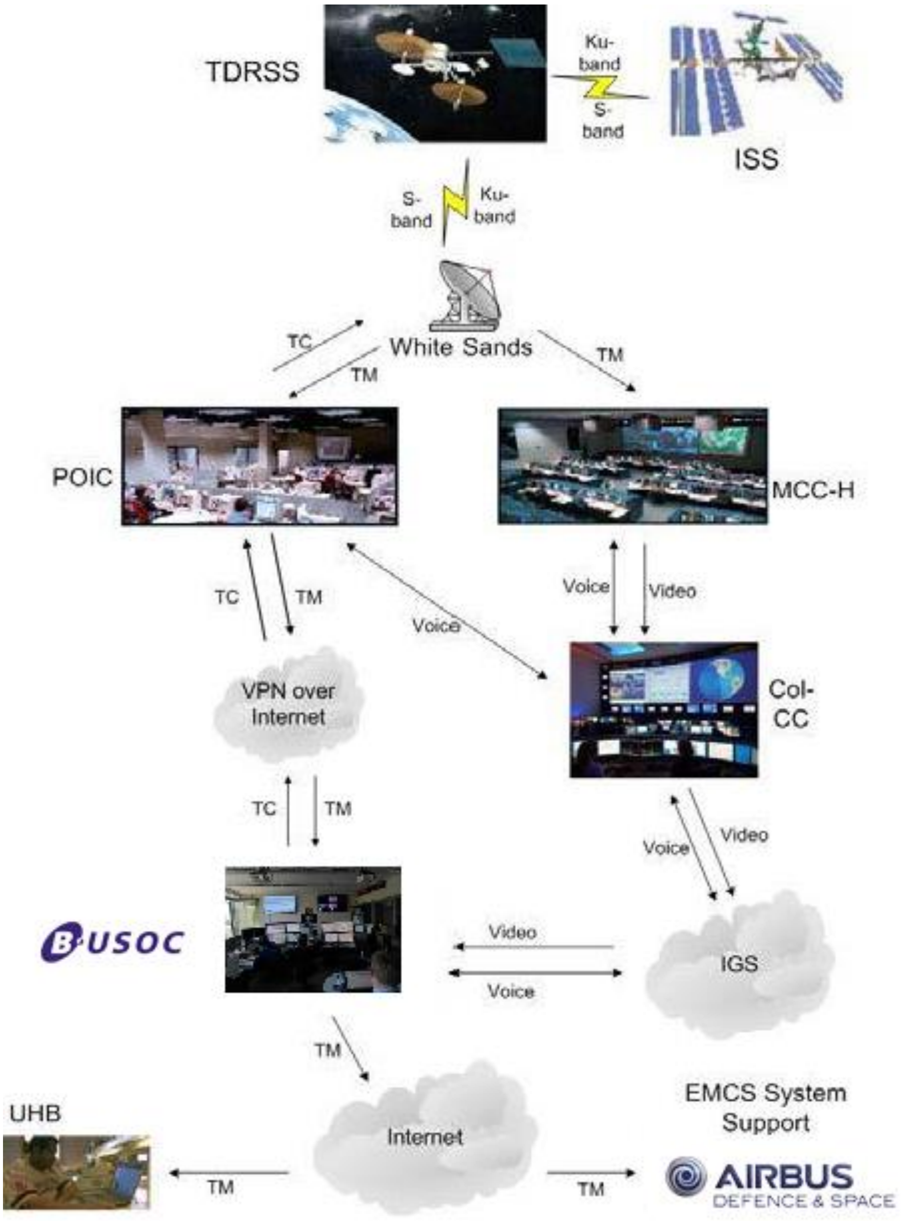

Figure 2: Data flow from the ISS to the UHB (User Home base) of the scientist, the scientist receives only the data he requested for to derive the final science product. All the elements generate data which could be used to replay the mission and generate a new final data flux.

The PERICLES tools of data management introduce a way to ease this process by automatizing it. YAMCS performs already a lot of automatization tasks. Ideally for a SOLAR scientist, the data flow should correspond to the figure 3 scheme. It would be much more complex if all user communities were considered. The user communities have different perceptions of the data collection, these communities are currently: the engineer’s user group, operations user group, sciences users and agencies as data owners. PERICLES has also to address a fifth community: the scientists of the future which is a very prospective task as those may not be consulted.
More than 50 elements of data have been identified in order to describe the complete operation of the instrument from the pre-flight design documents and test data to the recovered flight hardware.

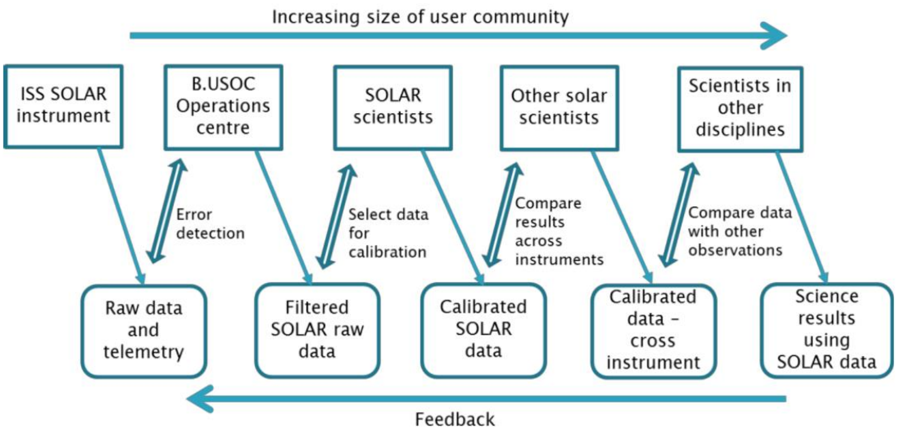
Increasing size of user community
ISS SOLAR instrument
B.USOC Operations centre
SOLAR scientists
Other solar scientists
Scientists in other disciplines
Error detection
Select data for calibration
Compare results across instruments
Compare data with other observations
Raw data and telemetry
Filtered SOLAR raw data
Calibrated SOLAR data
Calibrated data – cross instrument
Science results using SOLAR data
Feedback

Figure 3: Data flow from a space science user point of view, the role of data generated outside the SOLAR instrument during operations is transparent. (Waddington et al, 2016)

The SOLAR data can be divided in several categories: Engineering documents, Operation documents prepared in advance of the operations, Living operation documents generated by B.USOC, Living operation documents generated by the agencies managing the ISS, Telemetry data originating from the payload and ISS, Science data and finally: Data generated by B.USOC during operations. These elements are described and analysed in the PERICLES data survey and ontologies document (PERICLES, 2015).

The different categories have been analysed from the point of view of semantic change using the following grid treating them as a digital eco-system. Each of them impacts the reuse of the data.

Table 1. Types of change occurring on digital ecosystems and their impact.

Type of change	Description	Impact
Knowledge and terminology	Changes in semantics that originate from a designated user community.	Different user communities using the same underlying datasets with different understanding and goals.
Technology	This includes hardware availability, software obsolescence, and changes in formats, protocols and interfaces.	Requires replacement of hardware and software components, transcoding of files, redesign of interfaces etc.
Policy	Changes in permissions, legal requirements, quality assurance and strategy.	This can impact how and where digital objects are stored, quality processes they are subjected to, retention periods etc.

Organisation	Change to the organisation due for example to political, financial or strategic reasons. Often organisational changes can be manifested as policy changes.	This can result in different priorities for retaining or maintaining the reusability of digital objects.
Practice	This change originates from new or changed habits of the designated user community (not necessarily related to knowledge and terminology changes). It is an indicator that user requirements may change.	This can result in changes to the form in which digital objects are retained, reflecting the changing ways in which they are to be reused.
Requirements	This can include business requirements, functional requirements that a system should fulfil, quality of service and user requirements.	This again reflects the way that digital objects are reused and hence how they should be stored and maintained.
Dependency	Either characteristic attributes of a dependency are changed (e.g. quicker, faster, more flexible, cheaper) or the dependency itself changes.	Evolution in dependencies can reflect different views on the types of change that are being considered.

4. Role of B.USOC in the data use and preservation.

The B.USOC ESA mandate is to distribute to the scientists all the data they require to produce scientific results, B.USOC is required to keep all obtained data including ancillary data used only for the operators consoles for ten years, the ESA (Human and Space Operations) data policy at the beginning of the project did not allow B.USOC and its contactors or technical partners to use this data for developing products which had not been requested by the scientists. However, space data use was several times authorized by ESA for technological research on specified time slots for knowledge management projects well distinct from the initial science objectives of the project. An example was for the CUBIST business intelligence project of malfunction analysis and predictions during operations (Klaï et al, 2012). Moreover, this research has to be conducted within the ESA secure servers inside B.USOC. Such a permission was granted to PERICLES for very limited data slices and allowed to develop a prototype of anomaly detector, however the full test of the different tools developed by the PERICLES partners at the final PERICLES benchmarks was made using publicly accessible EUMETSAT data. Now the ESA HSO data policy has been updated and would allow USOC's and their partners to develop products under ESA approval and supervision.

5 COST Action TD1403 role in the PERICLES project.

The B.USOC PERICLES work packages included the organisation of a science data community of practice study while the digital art case was managed by the Tate Gallery. This community was started at the BigSkyEarth Lyon conference and continued for two years increasing in participation at each of its meetings. The 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 PERICLES 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 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.

The limited survey made by the CoP showed however that the communities described in this text were only adequately represented in CERN archiving where data reuse for research beyond the initial objectives of the experiments is expected after the main results, sometime based on less than twenty events in several millions have been published.

6 PERICLES final architecture and set of tools.

The PERICLES project built a complex architecture to preserve and extract data and processes from the entire data of a scientific experiment.

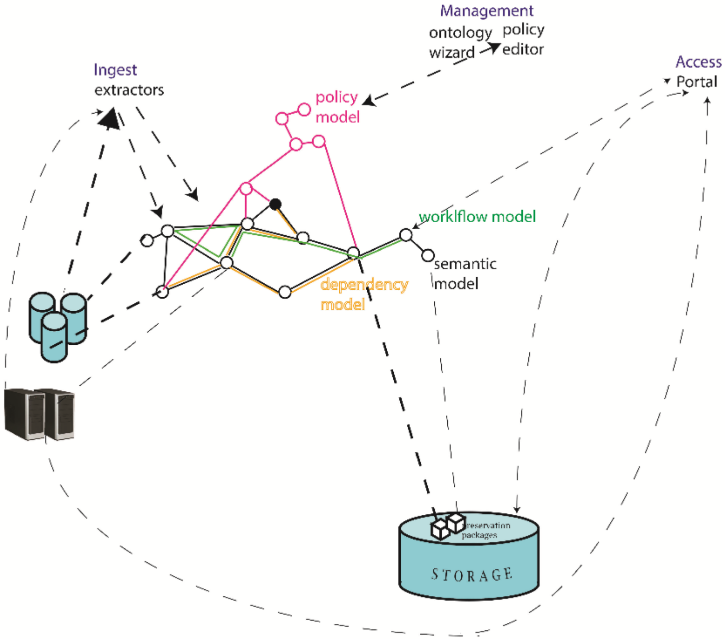

Fig 4: schematic reuse of an experiment data to access a stored experiment through a dedicated portal. The different models indicated can found at : http://www.pericles-project.eu/deliverables.

The formalism used departs from the standard OAIS (Open Archival Information System, CCSDS, 2012) to use the LRM (Linked resources Model), it is too complex to be described in this abstract but has been published in detail by Lagos et al. (2015) and Vion-Dury et al (2015).
Several related tools have been developed and are outlined in Waddington et al (2016). The only one to have been tested on a mirror of the actual SOLAR data was the PET (Processor Extraction Tool), an anomaly detector was also tested on the actual data flow but was not selected as a final PERICLES product. While developing the different tools, it was discovered that archiving tools could play a role in the operational chain well before data was transmitted, in particular, the appraisal tool developed by King's College, London (http://www.preserveware.com/tools/pericles-technical-appraisal-tool/) could select the data to be sent in the chain in the case of an insufficient bandwidth due to poor transmission conditions.

7 Final disposition of SOLAR data and conclusions.

For B.USOC, the PERICLES project at proposal level intended to produce an archive of the entire SOLAR mission data so that the mission could be replayed in a more advanced technological future and so that future scientists not related to the present teams could reuse them with new paradigms unknown to us. The proposal itself was much less ambitious, B.USOC should deliver use cases and analyse them with tools supplied by the IT partners. The first stage of the B.USOC action was to produce the data survey and related ontologies for the SOLAR case (PERICLES, 2015). The number of data sets and related metadata came as a surprise as well as the length of the description, showing the extent of the task necessary to preserve all of them for a longer period than the ten years specified in the ESA contract. Moreover, the ESA data policy requirement on manipulating the data only inside the B.USOC secure network limited the access to other partners than B.USOC itself and its contractor Space Application Services. B.USOC decided than to offer publicly accessible EUMETSAT data to the PERICLES partners to validate the different PERICLES softwares.

The effort made in PERICLES convinced however ESA HSO that long term data preservation had to be envisaged for space observations made from the ISS and the final decision was to archive the SOLAR data after approval by the science teams in the heliophysics data base of the ESA science directorate in ESAC (European Space Astronomy Centre, Villafranca, Spain). The ISS SolACES data are now already publicly accessible there (http://isssolac.esac.esa.int/iss-solaces/index.html) In the meanwhile, detailed scientific publications describing the SOLAR series results began to appear (Bolsée et al, 2017, Meftah et al, 2016, 2017).

The SOLAR programme was however a good example of a data set which needs to be preserved as its series contain information that will never be reproduced by new observations because the sun is constantly changing. It benefited from the increase of consciousness in the need of long term data preservation and its data is now better preserved than most manned space experimental data obtained since 1961. This partial success can be credited to the different groups of scientists using solar irradiance data in synergetic programmes and to related studies like PERICLES.

Acknowledgments

The authors wish to thank the European Space Agency (ESA,Directorate of Human Spaceflight and Operations) and the Belgian Science Policy Office (BELSPO) (ESA Prodex and other programmes) for SOLAR funding and support of the ground based solar UV network. PERICLES is funded under FP7 ICT Call 9, and addresses work programme objective ICT-2011.4.3 Digital Preservation and we are grateful to the European Commission for its support.

References

Bolsee D, N. Pereira, D. Gillotay D, P. Pandey, G.Cessateur, T.Foujols, S.Bekki, A. Hauchecorne, M. Meftah, L.Dame, M.Herse, A.Michel, C. Jacobs A Sela , SOLAR/SOLSPEC mission on ISS: In-flight performance for SSI measurements in the UV. Astronomy and Astrophysics. 2017 April 1; 600: A21. DOI: 10.1051/0004-6361/201628234.

CCSDS - Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2, 2012

Dewitte, S., D. Crommelynck, S. Mekaoui, and A. Joukoff,: Measurement and uncertainty of the long-term total solar irradiance trend. Solar Physics, 224, 209-216, 2004

Frohlich, C., 2006: Solar irradiance variability since 1978 - Revision of the PMOD composite during solar cycle 21. Space Science Reviews, 125, 53-65.

Jacobs,C., S. Klai, M. Schmitt, The YAMCS Notification Add-on: an automated notification tool for operations in human space flight, AIAA-2016-2307, SpaceOps 2016 Conference, Deajon, Korea, 2016

Klaï S., E. Sevinc., B.Fontaine, C.Jacobs C. and C.Muller, CUBIST: Semantic Business Intelligence Supporting Payload Operations, AIAA 2012-1279934, SpaceOps 2012 Conference, Stockholm, Sweden, 2012

Kopp, G., and J. Lean,: A new, lower value of total solar irradiance: Evidence and climate significance. Geophysical Research Letters, 38, L01706.,2011

Lagos, N., S.Waddington and J.Y. Vion-Dury, : On the preservation of evolving digital content - the continuum approach and relevant metadata models. In E. Garoufallou, R. J. Hartley, & P. Gaitanou (Eds.), Metadata and semantics research (Vol. 544, pp. 15–26). Cham: Springer International Publishing., 2015

Meftah M, D. Bolsee, L.Dame, A.Hauchecorne, N.Pereira, A. Irbah, S. Bekki G. Cessateur, T.Foujols, R. Thieblemont : Solar irradiance from 165 to 400 nm in 2008 and UV variations in three spectral bands during solar cycle 24. Solar Physics.; 291(12): 3527-3547. DOI: 10.1007/s11207-016-0997-8., 2016.

Meftah M, L.Dame, D.Bolsee, A.Hauchecorne, N.Pereira, D.Sluse, G.Cessateur, A. Irbah, J.Bureau,M. Weber, K.Bramstedt K, T.Hilbig, R. Thieblemont, M.Marchand, F.Lefevre, A.Sarkissian, S.Bekki : SOLAR-ISS: A new reference spectrum based on SOLAR/SOLSPEC observations. Astronomy and Astrophysics 26; epub: 13 pp. DOI: 10.1051/0004-6361/201731316, 2017

S. Mekaoui , S. Dewitte ,C. Conscience, A. Chevalier, Total solar irradiance absolute level from DIARAD/SOVIM on the International Space Station, Advances in Space Research , 45 1393–1406, 2010.

PERICLES Consortium. Deliverable D2.3.2: Data survey and domain ontologies for case studies. Available at: http://www.pericles-project.eu/deliverables/48, 2015

Sela, A. , Mihalache, M and Moreau, D, YAMCS - A Mission Control System, SpaceOps2012, American Institute of Astronautics, 2012.

Schmidtke,G., C. Fröhlich and G. Thuillier, ISS-SOLAR: Total (TSI) and spectral (SSI) irradiance measurements, Advances in Space Research , 37, 255-264, 2006.

Schmidtke, G., R.Brunner, D.Eberhard, B.Halford, U.Klocke, M. Knothe, W. Konz, W.J. Riedel, H. Wolf, SOL–ACES: Auto-calibrating EUV/UV spectrometers for measurements onboard the International Space Station., Advances in Space Research ,37, 273-282, 2006.

Thuillier,G., T. Foujols, D. Bolsée, D. Gillotay, M. Hersé, W.Peetermans, W.Decuyper, H. Mandel, , P. Sperfeld P., S.Pape et al, SOLAR/SOLSPEC: Scientific Objectives, Instrument Performance and Its Absolute Calibration Using a Blackbody as Primary Standard Source, Solar Physics 157, 185-213, 2010

Vion-Dury, J.-Y., N. Lagos, E. Kontopoulos, M. Riga,P. Mitzias, G. Meditskos, S. Waddington, P.Laurenson, and I. Kompatsiaris,: I. Designing for Inconsistency – The Dependency-based PERICLES Approach. In T. Morzy, P. Valduriez, L. Bellatreche (Ed.), New Trends in Databases and Information Systems, 539, (pp. 458-467). Springer Berlin, 2015

Waddington, S., M. Hedges, M. Riga, P. Mitzias, E. Kontopoulos, , I. Kompatsiaris,, J.Y. Vion-Dury, N. Lagos, S. Darányi, F. Corubolo, C. Muller and J. McNeill, PERICLES – Digital Preservation through Management of Change in Evolving Ecosystems The Success of European Projects using New Information and Communication Technologies, Chapter: 4, Publisher: SCITEPRESS Digital Library, Editors: Sofiane Hamrioui, pp.59-82, 2016

Willson, R.C. and A.V. Mordvinov. Secular total solar irradiance trend during solar cycles 21-23. Geophysical Research Letters, 30(5):1199, doi:10.1029/2002GL016038, 2003