PERICLES - Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics

Mark Hedges, Simon Waddington[1]
King's College London
{mark.hedges, simon.waddington}@kcl.ac.uk

Jens Ludwig, Philipp Wieder[5]
Georg-August-Universität Göttingen
ludwig@sub.uni-goettingen.de, philipp.wieder@gwdg.de

Rob Baxter[9]
Edinburgh Parallel Computing Centre
r.baxter@epcc.ed.ac.uk

Sándor Darányi, Elena Maceviciute, Tom Wilson[2]
University of Borås
{sandor.daranyi, elena.maceviciute}@hb.se, wilsontd@gmail.com

Paul Watry, Adil Hasan, Fabio Corubolo[6]
University of Liverpool
P.B.Watry@liverpool.ac.uk, {adilhasan2, corubolo}@gmail.com

Pip Laurenson[10]
Tate
pip.laurenson@tate.org.uk

Yiannis Kompatsiaris, Stamatia Dasiopoulou[3]
Centre for Research and Technology Hellas
{ikom, dasiop }@iti.gr

Rani Pinchuk[7]
Space Applications Services NV
rani.pinchuk@spaceapplications.com

Christian Muller[11]
B.USOC
christian.muller@busoc.be

Odysseas Spyroglou[4]
Dotsoft
ospyroglou@dotsoft.gr

Jean-Pierre Chanod, Jean-Yves Vion-Dury[8]
Xerox
{jean-pierre.chanod, jean-yves.vion-dury}@xrce.xerox.com

ABSTRACT
This poster paper describes the objectives, approach and use cases of the EC FP7 Integrated Project PERICLES. The project began on 1st February 2013 and runs for four years. The aim is to research and prototype solutions for digital preservation in continually evolving environments including changes in context, semantics and practices. The project addresses use cases focusing on digital art, media and science.

Categories and Subject Descriptors
Information Systems [**Information Systems Applications**]: Digital Libraries and Archives

General Terms
Theory.

Keywords
Preservation models, lifecycle, data analytics, semantics, policies.

1. INTRODUCTION
This poster paper describes the objectives, approach, use cases and proposed deliverables of the EC FP7 Integrated Project

PERICLES: http://www-pericles-project.eu. The PERICLES project was funded through the FP7 ICT Call 9 Digital Preservation. The project involves partners of a range of complementary types, including six academic partners, one multinational corporation, two SMEs and two non-academic public sector organisations.

2. PROBLEM DESCRIPTION
As digital content and its related metadata are generated and used across different phases of the information lifecycle, and in a continually evolving environment, the concept of a fixed and stable 'final' version that needs to be preserved becomes less appropriate. As well as dealing with technological change and obsolescence, long-term sustainability requires us to address changes in context, such as changes in semantics - for example, the 'semantic drift' that arises from changes in language and meaning - or disciplinary and societal changes that affect the practices, attitudes and interests of the 'stakeholders', whether these be curators, artists, scientists, or indeed a broader public, such as visitors to exhibitions.

Such a changing environment necessitates a corresponding evolution of the strategies and approaches for preservation if stakeholder communities are to be able to continue to use and interpret content appropriately. A key issue is the provision of sufficient contextual information to enable both lifecycle management and preservation on the one hand, and re-use or re-

interpretation of content on the other, as well as the facility to model and describe preservation processes, policies and infrastructures as they themselves evolve. Capturing and maintaining this information throughout the lifecycle, together with the complex relationships between the components of the preservation ecosystem as a whole, is key to an approach based on 'preservation by design', through models that capture intents and interpretative contexts associated with digital content, and enable content to remain relevant to new communities of users.

The project will address these preservation challenges in relation to digital content from two quite different domains: on the one hand, digital artworks, such as interactive software-based installations, and other digital media from Tate's collections and archives; on the other hand, experimental scientific data originating from the European Space Agency and International Space Station.

3. AIMS AND OBJECTIVES

The project has three main objectives.

Objective 1: To enable trusted access to digital content that is complex, heterogeneous, highly-interconnected, and subject to change, and to facilitate continued understanding and reuse of those objects across all phases of the lifecycle. This will be achieved by:

- Developing a model based on a linked data paradigm for describing the resources in preservation environment - including content, metadata, processes, users, and policies - and for managing their dependencies and consistency as the environments evolve.
- Adapting and extending preservation and lifecycle models to address the evolution of digital ecosystems and their dependencies, and developing an associated framework and tools.
- Developing a range of analytical methods for identifying and capturing preservation-related information - semantics, users, interpretative contexts - from digital content and its environment.

Objective 2: To evaluate our approaches, processes and tools against requirements and user communities in different application domains. This will be achieved by:

- Developing case studies to evaluate the approaches taken by PERICLES against the requirements of the user communities targeted by the project.
- Assessing the potential for deploying project outputs in operational environments.

Objective 3: To facilitate sustainability and exploitation of project outputs by disseminating the knowledge created by the project, and in particular by:

- Building communities of practice around a number of topics addressed by PERICLES.
- Engaging with standardisation activities regarding contribution to relevant standards.
- Engaging with commercial organisations to facilitate the take-up of project outputs by industry.

4. APPROACH AND METHODOLOGY

4.1 Case studies and evaluation

The research carried out by PERICLES will be driven by and evaluated against two distinct groups of case studies focused around different application domains and communities in media and science. While on the surface very different, these two areas have in common an environment that evolves continually, not only in terms of the technologies used, but also as regards meaning, and the practices, attitudes and interests of stakeholders, whether these be curators, artists, scientists, engineers, or indeed a broader public, such as visitors to exhibitions. By addressing the preservation challenges raised by digital material from two quite different domains we aim to ensure that our results are of broad applicability.

Rather than a single system, PERICLES will produce a variety of components (models, tools, policies, architectural approaches etc.) that can be used independently in different combinations to support a range of preservation requirements. We will also implement two prototypes that integrate the various technologies developed by the project so as to meet the requirements of the two broad communities involved in the case studies. These prototypes will serve as test beds for the evaluation of the project against the two case studies, and as demonstrators of the project technologies to a wider audience.

4.2 Core research activities

The research in PERICLES is focused around three core research work packages:

- WP3 will develop a conceptual framework and unified model, based on linked data principles, for representing dynamic preservation ecosystems composed of distributed interdependent resources, together with a language and tools describing and managing change in such ecosystems.
- WP4 will investigate and develop a range of analytical techniques and tools for identifying, extracting, analysing and encapsulating information about digital objects and their environments of relevance to their preservation, appraisal and reuse, such as representation information, provenance, contextual information, semantic content descriptions, and metadata more generally.
- WP5 will extend existing lifecycle and preservation models, which focus on technological change, to address the broader evolution of preservation ecosystems, including changes in semantics and user communities, as well in the policies, processes and systems of the preservation infrastructure itself. It will also develop processes and tools that support the management of preservation ecosystems in accordance with these models, and in particular for appraisal processes.

These three work packages are closely interlinked, as illustrated in Figure 1.

- The processual models - and corresponding processes and policies - developed in WP5 will describe and influence the evolution of the more generic models developed in WP3. At the same time, components, such as processes and policies, created by WP5 will themselves be part of the ecosystem and will thus be represented in the model.
- WP5 will produce concrete processes and policies that will be composed from a variety of smaller components, including many of the tools and services for extracting,

analysing and encapsulating information developed by WP4. At the same time, these components will provide relevant metadata to control preservation management and appraisal processes developed by WP5.

- Finally, the information captured and extracted by the tools developed in WP4 will serve to instantiate and populate the linked resources ecosystem model developed by WP3.

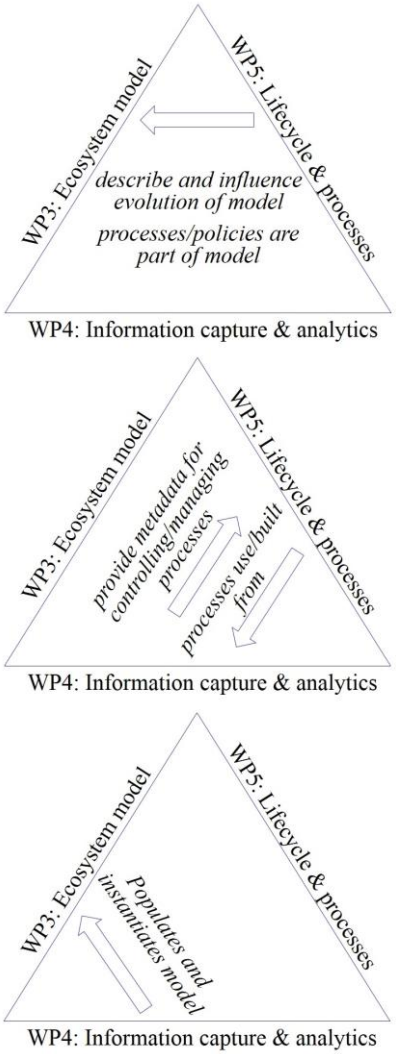

Figure 1. Interactions between research work packages

4.3 Dissemination, communities of practice and exploitation

As well as the two specific communities addressed in the Case Studies, PERICLES will undertake dissemination and engagement activities within a number of communities of practice. Some of these communities will correspond with broad application domains, and will represent an opening out of the domains addressed by the case studies (science and engineering; media and art; archives and other memory institutions); others will correspond to "enabler" stakeholder groups, orthogonal to the domain-based communities and cutting across disciplines (facilities and operations centres; data infrastructure technology; policies and standards; business and sustainability).

The principle behind these communities of practice is that they will function as coordination points for promoting the findings of the project, for seeking input and feedback, and for extending collaborations to new communities. Each community will carry out a range of activities, such as workshops and liaison with external organisations, of relevance to the area it addresses.

A key objective of PERICLES is to set up pathways for the take-up and exploitation of project outputs in production environments, both commercially and in the public sector. While this objective cuts across the communities of practice, because of its importance this will be organised through a dedicated work package (WP10) that will be tasked with identifying exploitation opportunities and developing an exploitation strategy.

5. USE CASES

5.1 Science case study

Preserving space science data is critical for the wider research community. Collecting data in space is extremely expensive - the design of the payload (the experiment device) for operating in orbit is very complex, the cost of launching them to orbit is very high, and operating the payload is very demanding. Moreover, observational data (e.g. sun or weather observations) are simply impossible to replicate.

The science case is based on data from space operations created at B.USOC. B.USOC is one of the European distributed operation centres of the International Space Station (ISS) payloads and is, amongst other functions, Facility Responsible Centre for the SOLAR package. SOLAR is a set of instruments measuring the variations of the energy output of the sun in spectral ranges going from the far ultraviolet to the near infrared. This experiment package has been running on the International Space Station since 2008 but has a longer history. The instruments actually belong to a series extending to first designs in 1976. Thus the current data series span most of the space age and constitute an important source of reference on the impact of solar variations on earth's climate and environment for three solar cycles.

The data from space experiments are collected only during the mission, after the payload is launched to space. However, related metadata (e.g. design documentation) is collected from the start of the project, sometimes many years before the launch – during the mission analysis, feasibility, definition, qualification and production phases. For the SOLAR payload, for example, there are over 900 documents from these preparation phases. These documents include specification and design documents, acceptance data packages, test plans and reports, interface control documents, assessment reports, safety data packages, operation manuals, certifications, thermal analysis etc.

Valuable metadata is collected also during missions – such metadata describes how the experiment devices were operated and can explain for example different anomalies in the results. This metadata includes the different operation interface procedures, flight rules and payload regulations, payload data files, minutes of meetings, console logs, flight notes, checklists, daily operation reports, science planning, command schedules etc.

The space science data itself includes telemetry sent from the payload and telecommands sent to the payload. The telemetry includes housekeeping data – this consists of engineering measurements about the state of the payload (e.g. temperature, voltage and current reading), health and status data which include measurements from sensors outside the payload, and science data.

The raw science data must be calibrated. In SOLAR for example, the detector is put in front of a black body in certain temperature, and since the spectrum of a black body in certain temperature is known, the detector can be calibrated. Another example is that during the mission the detector must be further calibrated as it decays. The SOLSPEC detector is calibrated once in 24 hours during sun's visibility window. The calibration is done using a calibration lamp that its spectrum is known. The South Atlantic Anomaly Disturbances is another example. This anomaly causes an increased flux of energetic particles in the south Atlantic region and exposes orbiting satellites to higher than usual levels of radiation. Apparently, these disturbances have to be taken into account when calibrating the results. The raw science data is processed to include the information from the calibration curves, and may later be further processed to include other calculations of the scientists.

The telecommands are structured data sent to payloads during the operations. They may contain control structures for shutting up or starting various modules, as well as uploads of data and scripts.

In addition to the above, auxiliary data is also collected. Most of auxiliary data comes from public sources. For instance, current B.USOC operations related to the SOLAR payload heavily depend on TLE (two-line elements) to predict the position of the ISS and on the ISS attitude timeline (ATL) to predict the orientation of ISS towards the sun. The two external data sources are combined in order to create a full prediction of the upcoming month allowing clear scientific planning and an optimal operations support plan to be created.

5.2 Media case study

The Media case study, led by Tate, will reflect the activities and responsibilities of distinct areas of the organisation.

- Digital art from the main fine art collection includes software-based art, video and audio content, and file-based material such as vector graphics.
- Born-digital material from artists' estates and from institutional records, for example from the archives of galleries.
- Audio Arts collection. This is a rich archive of digitised audio material generated as an audio magazine that was produced and distributed from 1973 to 2006.
- Tate Media Productions. This content comprises unedited footage, edited programmes, and other digital assets generated in the creation of these programmes.

Each of these categories contains digital assets of very high value, which are moreover associated with a great deal of interrelated contextual information, including (for example) information generated during and (implicitly) documenting the process of creation, as well as content generated in social media sites (e.g. blogs, Facebook, Twitter) once an artefact has been exhibited. The lifecycle of an object thus involves a number of high-level processes that may be represented as formal workflows, which will be used to define the use case.

To take a simple example, in the case of digital art, this description will map out the creation and acquisition of these works into the collection, and the subsequent cycles of display, maintenance and preservation or recovery associated with their life within the museum. The PERICLES project will map not only the digital assets that constitute the components of the artworks themselves, but also the rich information that surrounds them and that describes the context in which they exist as their lifecycle progresses.

The use cases will also describe the existing systems and storage infrastructure used to manage and curate the material, and their part in the broader digital and preservation ecosystem within the institution. Importantly, the use cases will also identify the "pressure points" in the existing workflows, and highlight areas where currently there are in place no practices that implement appropriate preservation strategies for these digital objects.

6. CONCLUSIONS

This paper describes the PERICLES project, its objectives and approach, and details its case studies. The conference poster will describe some early findings of the requirements gathering in both the science and media use cases. These will be used to provide some more concrete examples of the research problems being addressed in the project. PERICLES aims to develop several prototypes to validate the concepts and models being defined, and some initial ideas on these will be discussed.

7. ACKNOWLEDGMENTS

This work was supported by the European Commission Seventh Framework Programme under Grant Agreement Number FP7-601138 PERICLES.

8. ADDRESSES OF AUTHORS

1. Centre for e-Research, King's College London, 26-29 Drury Lane, London WC2B 5RL, UK..
2. University of Borås, Swedish School of Library and Information Science, Borås, Sweden.
3. Information Technologies Institute, Centre for Research and Technology Hellas, Thermi-Thessaloniki, Greece.
4. Dotsoft, Kountourioti, 54625 Thessaloniki, Greece.
5. Georg-August-Universität Göttingen, Wilhelmsplatz, 37073 Göttingen, Germany.
6. University of Liverpool, Brownlow Hill, Foundation Building, Liverpool L69 7ZX, UK.
7. Space Applications Services NV, Leuvensesteenweg, 1932 Zaventem, Belgium.
8. XEROX Research Centre Europe, Avenue du President Wilson, Immeuble Le Jade, 93200, France.
9. Edinburgh Parallel Computing Centre, Old College, South Bridge, Edinburgh EH8 9YL, UK.
10. The Board of Trustees of Tate Gallery, Millbank, SWIP 4RG London, UK.
11. B.USOC (Belgian User Support and Operations Centre), Avenue Circulaire 3, 1180 Brussels, Belgium.