A pragmatic approach to significant environment information collection to support object reuse

Fabio Corubolo
IPHS, University of Liverpool
Waterhouse Building, Brownlow St.,
Liverpool L69 3GL, UK
corubolo@gmail.com

Anna Grit Eggers
Göttingen State and University
Library, Georg-August-Universität
37070 Göttingen, Germany
eggers@sub.uni-goettingen.de

Adil Hasan
IPHS, University of Liverpool
Waterhouse Building, Brownlow St.,
Liverpool L69 3GL, UK
adilhasan2@gmail.com

Mark Hedges
Department of Digital Humanities,
King's College London
Strand, London WC2R 2LS, UK
mark.hedges@kcl.ac.uk

Simon Waddington
Department of Digital Humanities,
King's College London
Strand, London WC2R 2LS, UK
simon.waddington@kcl.ac.uk

Jens Ludwig
Göttingen State and University
Library, Georg-August-Universität
37070 Göttingen, Germany
ludwig@sub.uni-goettingen.de

ABSTRACT

When aiming to ensure the long-term usage of digital objects, it is important to carefully select what information to keep, considering also what lives outside of them. In the PERICLES project we start by analysing how such information has been described in related work, considering common definitions of metadata, context, significant properties and environment, and we come to the conclusion that we need to consider the broadest set of information, which we term environment information. Building on previous definitions, we introduce the concept of Significant Environment Information (SEI) that takes into account the dependencies of the digital object on external information for specific purposes and significance weights that express the importance of such dependencies for the specific purpose. From there we expand the definition in time considering the importance of collecting SEI during any phase of the digital object lifecycle, following the sheer curation perspective. Examples of SEI are illustrated in the very diverse use cases considered in the project, that include diverse data types from the Art domain and data from space observations in the Science domain. Finally we introduce our PERICLES Extraction Tool, that we developed to capture SEI, and present methods to extract SEI with experimental results supporting the approach. The PET tool automates the novel techniques we describe, supports sheer curation, as a continuous transparent collection process that otherwise the user (e.g. scientist, artist in our use cases) would have to find time to perform manually.

General Terms

Infrastructure, communities, preservation strategies and workflows, specialist content types, case studies and best practice.

Keywords

Digital preservation, significant properties, significant environment information, environment information, dependency graph, sheer curation, significance weight, dependency extraction.

1. INTRODUCTION

The PERICLES project (http://www.pericles-project.eu/) is an EU-funded Integrated Project focused on the problem of digital preservation. One of the areas of study is the investigation of what could constitute Environment Information (EI), in its broadest sense, in order to be able to select and capture the relevant part of that information that will sustain the use and reuse of the Digital Objects (DOs). One of the principles we have adopted is to try to explore the information based on the purpose that users have when interacting with a DO.

In Section 2, we explore relevant work and definitions of the information for the interpretation of a DO and describe our view on the subject.

In Section 3, we define Significant Environment Information (SEI) of a DO in a way that takes into account the purposes and the measure of significance of the purpose. This relates to, and in a way extends the definitions of Significant Properties (SP) of a DO. We also introduce the importance of gathering such use information in the user environment, in the sheer curation context, and describe methods to measure significance.

In Section 4 we look at examples of SEI that can be captured in the context of PERICLES case studies, in the art domain, for Software Based Art (SBA), and in the Science domain, in the scope of SOLAR experiment observations[1].

Section 5 introduces the PERICLES Extraction Tool (PET), the software tool we designed to capture SEI, and illustrate some of the techniques for environment information collection and how these can easily be adapted to different domains of use. The focus of the tool is on the context of unstructured workflows, as in many use cases users do not adopt workflow systems that can be used to analyse their flow of work.

Section 6 will describe some detailed experiments, and evaluate the results obtained using the PET tool.

Section 7 will draw conclusions and describe future work.

2. DIGITAL OBJECT INFORMATION: PREVIOUS WORK

In this section, we examine previous work on identifying and representing the information for a digital object that is relevant to support the reuse of that object, both in the long term, and across different user communities and for different purposes. We structure this examination by beginning with information that comes from the DO itself, then moving beyond the DO with the aim of identifying a broader set of information that needs to be taken into account to better support DO reuse, as illustrated in Figure 1. We recognise that this classification is one among many;

[1] http://en.wikipedia.org/wiki/SOLAR_(ISS)

the aim is however to show one thread that leads us to the topic of significant environment information, introduced in Section 3.

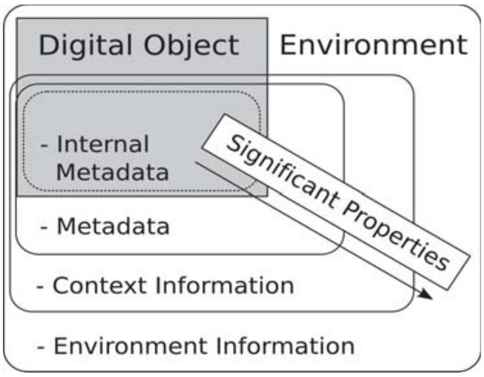


Figure 1. Our view on Digital Object and related information, from the narrowest to the broadest.

2.1 Metadata

Metadata can be defined as the information necessary to find and use a DO during its lifetime [1]. This definition covers a wide variety of information, and the Consultative Committee for Space Data Systems further refined it in their reference model for an Open Archival Information System (OAIS) [2]. This refinement covered the information necessary for the long-term storage of DOs, and they identified a number of high-level metadata categories, as follows. *Descriptive Information* (DI) consists of information necessary to understand the DO, for example its name, a description of its intended use, when and where it was created, etc. The *Preservation Description Information* (PDI) consists of all the information necessary to ensure that the DO can be preserved, including fixity (e.g. a checksum), access rights, unique identifier, context information (described in more detail in the following subsection) and provenance, which describes how the object was created. The final category arises from the fact that the OAIS manages not the DO itself, but information packages which consist of the DO as well as the DI, PDI and information required to interpret the contents of the DO (which is described by the *Representation Information (RI)*). The *Packaging Information* (PI) category describes how the information package is arranged such that individual elements can be accessed.

Standard file formats have standard structural metadata (e.g. MPEG21)[2], and *de facto* standards (e.g. the Text Encoding Initiative)[3] exist for popular formats. The situation on standardisation for the descriptive part of the RI is more complex due to the different needs of different communities, although many approaches contain the Dublin Core metadata element set [3] as a core. A catalogue of metadata standards for different communities can be found on the Digital Curation Centre website[4].

Metadata may be held internally in a DO, e.g. in the header of a structured file, or externally, e.g. in a database. Metadata may be treated as a separate entity, as it can be accessed without accessing the DO, but lack of metadata adversely affects the access to or reuse of the DO. While such information is essential for the reuse of the DO it is not in general sufficient; information concerning the external relationships of a DO, whether to other DOs, stakeholder communities, or other aspects of the environment within which a DO is created or curated, also need to be taken into account to ensure that the DO can be used fully and appropriately. This is addressed in the following sections.

2.2 Significant Properties

The concept of significant properties (SP) has been much discussed in Digital Preservation (DP) over the past decade, in particular in the context of maintaining authenticity under format migrations, given that some characteristics are bound to change as formats are migrated. The issue here was to identify which properties of an object are significant for maintaining its authenticity.

Early work in this direction may be found in [4], where SP are introduced as a "canonical form for a class of digital objects that, to some extent, captures the *essential characteristics* of that type of object in a highly determined fashion". Later work [5] investigated ways of classifying the properties: "Significant Properties, also referred to as "significant characteristics" or "essence", are essential attributes of a digital object which affect its appearance, behaviour, quality and usability. They can be grouped into categories such as content, context, appearance (e.g. layout, colour), behaviour (e.g. interaction, functionality) and structure (e.g. pagination, sections)." The concept has been adopted by standards such as [6], which describes SP as "Characteristics of a particular object subjectively determined to be important to maintain through preservation actions." Such characteristics may be specific to an individual DO, but can also be associated with categories of DOs.

An important aspect of SP is that significance is not absolute; a property is significant only relative to an intended purpose [7], or a stakeholder [8], or some other way of identifying a viewpoint. This intuition is also highly relevant to the work described in this paper.

While the concept of SP is useful for digital preservation, in its application it has usually been restricted to internal properties of a DO, for example the size and colour space of an image, or the formatting of text documents, rather than the potentially valuable information that is external to the object itself. There have been some indications of a broader conception: [5] identifies context as a category of SP, [9] refers to the need to preserve properties of the environment in which a DO is rendered, and [8] introduces the notion of characteristics of the environment. The latter associates environments with functions or purposes; this differs from what we are aiming at, which is to describe the significance of information from a DO's environment in relation to the purpose the user is following (such as editing the object, processing the object, etc.). We thus see the purpose as qualifying the significance, not the environment – a piece of information is significant for a specific purpose, but not for some other purpose.

2.3 Context

Context is a term with many definitions, a basic dictionary definition being "the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood[5]". This clearly relates context to the purpose of *understanding* information, and this is a key feature of context in relation to digital objects. Context encompasses a broader range of information than metadata; it describes the setting that enables an *understanding* of a DO [10], including for example other DOs, metadata, significant properties, relationships, and policies governing the curation or use of the DO. In [11] context is defined even more broadly as 'all those things which are not an inherent

[2] MPEG21 http://mpeg.chiariglione.org/standards/mpeg-21

[3] TEI http://www.tei-c.org/index.xml

[4] http://www.dcc.ac.uk/resources/metadata-standards

[5] http://www.oxforddictionaries.com/definition/english/context

part of information phenomena, but which nevertheless bear some relation to these', where the nature of the 'relation' is left unspecified.

The OAIS model views context as the relationship between a DO (equivalent to the Content Information in OAIS terms) and its environment. In this view, the environment is considered to be necessary for using the DO, although it does not take into account two factors that we consider essential for our purposes: firstly, the possible variety of different uses to which a DO may be put, which will in general differ in the demands of 'necessity' they make on the environment; secondly, the variable strengths of the relationship with different aspects of the environment. These factors will be described and supported with examples in the following sections.

In TIMBUS [12] context is explored from the point of view of supporting business processes in the long term, describing a meta-model based on enterprise modelling frameworks. The context parameters cover a wide set of parameters, from the legal, business to the system, and technological ones, with the aim of supporting the execution of processes in the long term.

[13] presents a broad notion of context, close to our definition of environment, and recognises the importance of relationships between DOs in a collection. It further proposes a framework for contextual information that takes into account the different phases where DO context information should be gathered, and a general taxonomy of contextual entities. Automated methods for context population are also presented, but those are not overlapping the ones we are investigating in this paper, more focused on the semantic aspects of context.

2.4 Environment information

The widest set of information is the environment, which we define as consisting of all the entities (DOs, metadata, policies, rights, services, etc.) useful to correctly access, render and use the DO. The definition supports the use of unrelated DOs and conforms to the definition for environment used by PREMIS [6].

PREMIS builds on the OAIS reference model and defines a core set of metadata semantic units that are necessary for preserving DOs. The set is a restricted subset of all the potential metadata and only consists of metadata common to all types of DO. The current, published set (PREMIS2) defines a data model consisting of four entities (see Figure 2): the *Object* entity allows information about the DO's environment to be recorded amongst other information. The *Rights* entity covers the information on rights and permissions for the DO. The *Events* entity covers actions that alter the object whilst in the repository. The *Agents* covers the people, organizations or software services relevant that may have roles in the series of events that alter the DO or in the rights statements. The *Intellectual Entity* allows a collection of digital objects to be treated as a single unit.

The PREMIS working group undertook an investigation of the environment information metadata based on feedback from their user-groups that found the existing support to be difficult to use. The group reported in [14] their findings which entailed promoting the environment information to a first-class entity and not a subordinate element of the DO for the next version of PREMIS (PREMIS 3). They advocate the use of the *Object* entity to describe the environment, which allows relationships between different environment entities. This approach neatly supports the PERICLES view of the environment although PERICLES makes a distinction between the general environment and the environment significant for a particular set of purposes (termed the Significant Environment Information for a DO), which is described in the following section.

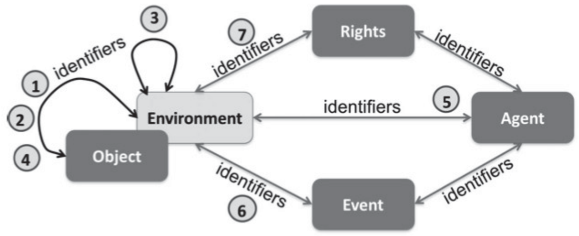


Figure 2. From [14], proposed changes for PREMIS 3 to make environment a first class object (light grey)

We consider the environment information for a DO to be the widest set of entities that a DO has the potential to be related to. This would include by definition all other DO, information, services and other information that can relate to the DO, but also other information from the environment that is not necessarily directly related to the DO but is useful for its uses (and depending on a specific use).

We consider this a wider set, although related, to the one described in the OAIS as *Representation Information*, and we take a different focus than that defined by PREMIS3. Another important distinction is that in general, we look at the environment as something defined from a DO upwards, so its definition will be related to a DO (although of course we will make sure to avoid redundancy by making good use of a linked model). In PERICLES, there is another separate view that is that of the ecosystem, that takes the view from the institution downwards. Furthermore, we consider that a part of the environment information will only be observable in the live environment of the user, so it's important to observe it in a sheer curation context as described later.

While looking at the DO environment, we consider the user an important part of it, and for that we want to observe the interaction between users and their communities, the DO, and the rest of the environment. We think that this perspective will allow us to capture the information based on the pragmatic, sometimes neglected aspects of the real requirements for making use of DO. This will also help us in the task of inferring dependencies that are not explicit and determined relevant information based on real use information.

3. SIGNIFICANT ENVIRONMENT INFORMATION

Based on the definition of environment from the previous section, we now introduce our definition of Significant Environment Information (SEI).

We define **Environment Information** for a source DO, based on the broad definition of environment in section 2.4, to be the information about the set of relationships between the source DO and any related objects from its environment.

We further define **purpose** as one specific use or activity applied to the source DO, by a given user or community. It is possible to imagine a hierarchy of purposes where a higher level purpose (as for example, 'render DO with faithful appearance' purpose) will lead to a set of detailed purposes (such as, accurate colour reproduction, accurate font reproduction etc.).

We further define **significance weight,** with respect to a purpose, as a continuous value (e.g. in the range 0-1, we have recently started refining the weights semantics) expressing the importance of each environment information relationship for that particular purpose. The significance weight will be a property of each relationship between the source DO and the DOs constituting its environment information.

Finally, we define '**Significant Environment Information**' (**SEI**) for a DO, with respect to a given purpose(s) as the set of environment information relationships qualified with significance weights. This will include both the dependency relationship (with purpose and weights) and the information that is target of the dependency.

Once SEI is determined for a collection of DOs, the different relationships can form a graph structure, where DOs in the collection could have relationship between each other (when a DO in the collection will depend on another DO in the collection for a specific purpose) or other DOs representing extracted information. This graph can be the basis to appraise (e. g. by selecting by threshold based on the weight) the set of DO that, together with their SEI, constitute the information to be preserved in order to support the selected uses of the DOs in the future.

In a less formal way, what we are aiming at is to determine "more or less all you need" when interacting with a DO for a specific purpose, and the significance of each of these information units.

Comparing this definition to that of SP to the environment, as for example described in [6], we note that the former is aimed at the collection of SEI for a DO to support the different purposes a user can have with respect to a DO, while the latter is defining the significant properties of an environment in itself. The information we aim to collect is defined by qualified relationships to other DOs, as opposed to properties of the environment.

As we mentioned in the introduction, the perspective we take is that of observing the current use of a DO before it enters a Digital Preservation system, in the systems where the DO is used, as this will allow better determination of what is significant for its use. We consider that knowing the significant information necessary to support current purposes will allow us to cover or at least know more precisely the needs also for the long term, as long as we try to support different user communities. This is because different user communities will have different purposes and different requirements, so this is a good approximation of knowing the needs of future communities (that we cannot know in advance).

3.1 Measuring significance

At this time, we are focusing on collecting a wide array of environment information, based on the relationships it can have to the DO and its estimated relevance. We are also trying to infer object dependencies that have an implied significance, by looking at use data, as described later. Still, a very relevant part of what we need is measuring the significance of the collected data. Although we don't have experimental results for now (those will come at a later stage of the project), we have clear ideas on how to define and collect it. It's in answering the question 'what for – for what purpose?' that should help us define what is significant.

Collections of data often have more than one use. Determining what information is significant depends on the use of the data. For example, the calibration of the solar measurement instrument will require calibration data, which may be a subset of the complete collection of data as well as applications necessary to read and analyse the calibration data. For a given collection not all of its environment information may be necessary for every potential use. To represent this we propose to assign weights to each relation between the collection and the environment information. The weights are based on the number of times the information is necessary for a given use. Weights will vary between 0 and 1. A weight of 1 indicates the information is essential for all intended uses of the data. Monitoring the access of information as well as regular review of the information required for each use would provide the opportunity to update the weights and could also accommodate new uses of the data.

Other factors can also be included in the weight, to express value, such as cost in time and money to collect the information as well as whether the information is proprietary (which may limit the accessibility to the information). There may also be constraints from licensing which restrict from where the data can be accessed. Any factor which influences access to the information may contribute to its weight.

Significance is also useful in the long term preservation perspective, for example to support critical analysis of the science data, as it will be a useful representation of the point of view and importance of the information for the stakeholders. It can also provide a key to understand the information.

3.2 SEI in the digital object lifecycle

In recent years, there have been various efforts within the digital curation community to establish new methods of carrying out curation activities from the outset of the digital lifecycle. A major constraint that mitigates against this is that data creators (such as researchers) typically have time only to meet their own short-term goals, and – even when willing – may have insufficient resources, whether in terms of time, expertise or infrastructure, to spend making their datasets preservable, or reusable by others (e.g. [15]). Moreover, the very volume of information that may be useful can preclude this as a practical approach, and in any case the researcher may be unaware of the utility, or even the existence, of much of this information

One approach to this challenge has been termed sheer curation (by Alistair Miles of the Science and Technology Facilities Council, UK), and describes a situation in which curation activities are integrated into the workflow of the researchers creating or capturing data. The word sheer here is used to describe the 'lightweight and virtually transparent'[6] way in which these curation activities are integrated, with minimal disruption.

Sheer curation is based on the principle that effective data management at the point of creation and initial use lays a firm foundation for subsequent data publication, sharing, reuse, curation and preservation activities, and it may be contrasted with post-hoc curation, which takes place only after the period during which the digital objects are created and primarily used.

The sheer curation model has not been extensively discussed in the scientific literature. The term has sometimes been interpreted as motivating the performance of curatorial tasks by data creators and initial users of data by promoting the use of tools and good practice that add immediate value to the data. This is, in particular, the take of [16], which discusses the role of such an approach to the distributed, community-based curation of business data.

However, this interpretation does not really address the challenges outlined above, and a more common understanding of sheer curation depends on data capture being embedded within the data creators' working practices in such a way that it is automatic and invisible to them. For example, the SCARP project[7], during which

[6] http://alimanfoo.wordpress.com/2007/06/27/zoological-case-studies-in-digital-curation-dcc-scarp-imagestore/

[7] http://www.dcc.ac.uk/projects/scarp

the term sheer curation was coined, carried out a number of case studies in which digital curators engaged with researchers in a range of disciplines, with the aim of improving data curation through a close understanding of the researchers' practice [17] [18].

In [19] the concept of sheer curation is extended further to take account of process and provenance as well as the data itself. The work examined a number of use cases in which scientists processed data through various stages using different tools in turn; however, as this processing was not carried out in any formally controlled way (e.g. by a workflow management system), it would have been impossible for a generic preservation environment to understand the significance of the various digital objects produced from the information available, as the story of the experiment was represented implicitly in a variety of opaque sources of information, such as the location of files in the directory hierarchy, metadata embedded in binary files, filenames, and log files. This was addressed by capturing information about changes on the file system as these changes occurred, when a variety of contextual information was still available, and the provenance graph was constructed from this dynamically using software that embedded the knowledge and expertise of the scientists.

The most effective way to capture SEI is through observation in the environment of creation and use of the object. We look at the interaction between the DO, the environment and the user, with time dimension. This allows us to infer dependencies that are not explicit and determine relevant information useful for use and reuse of the DO.

4. SEI IN PERICLES CASE STUDIES

The concept of SEI is now illustrated by examples in the area of digital art and space science, which constitute the main areas of interest of the use case partners of the PERICLES project.

4.1 Software Based Artworks

The following use example illustrates the SEI investigation inspired by the Software Based Art scenario from the Tate gallery. In this example a Software Based Artwork (SBA) should be migrated to a new computer system for the purpose of an exhibition. The software component of the SBA causes a strong dependency on the computer system environment. A description of SBA and an extensive study on their SP can be found at [20] and [21].

We assume there is a computer system with a validated SBA installation, which should be preserved to be able to configure and emulate the computer system environment as closely as possible for future exhibitions. The problem cannot be solved by preserving only the SBA as a DO, as the original appearance and behaviour of the software cannot be reconstructed based only on the metadata that belongs to the DO. In the context of executing the SBA's software for the exhibition are for example other dependencies such as external libraries and applications, and data dependencies (data used at run-time by the SBA). However, we have to look further at the whole environment to conceive all information that could be important for this scenario, as for example context-external running processes can affect the availability of resources, or external network dependencies. The determination, extraction and preservation of SEI are essential to solve the problem of enabling a future faithful emulation of the original system. An investigation of the environment information influence on the SP of the DO helps to identify the SEI for this use case. An example of SEI influencing the SP is when software changes the execution speed, based on the system resources, since program procedures can adapt their execution speed to the available resources depending on the programming style. This will make information about system resources SEI for the "maintaining the speed of execution" purpose. Information about display settings, as colour profile and resolution, used fonts, the graphic card and its driver is SEI that can affect the SBA appearance ('render DO with faithful appearance' purpose). Changes of programming language-related software can result in execution bugs or different speed of execution. The user interaction experience with the SBA can be affected by the peripheral driver or setting or response times that are dependent on the execution speed.

In order to determine its SEI, each SBA has to be individually analysed, regarding the use purpose and based on the properties of the artwork and the artist's beliefs regarding the SP of his artwork. Typical SEI to emulate the environment for a SBA is: information about computer system specifications, available resources, required resources, installed software and software dependencies. Other relevant dependencies to capture can be for example all the files that are used during the SBA execution, and peripheral dependencies, which can be identified by analysing the peripheral calls of the SBA. System resource requirements can be estimated on the basis of resource usage. Another example of SEI purpose, with a different set of significance weights, is when the SBA has to be recompiled because of a migration to another platform or to fix malfunctions. Here the SBA behaviour has to be validated by the comparison of behaviour patterns measured at the original system continuously in a sheer curation setting. Examples for such measurements are processing timings, log outputs, operating system calls, calls of libraries and dependent external software, peripheral calls and commands, resource usage, user interaction, video and audio recordings. The last two can be used to validate also the appearance of the artwork. If the SBA has a component of randomness, it is more difficult to evaluate its behaviour based on the measured patterns. Furthermore information about the original development environment can be useful for a recompilation, and to identify the source of a malfunction.

4.2 Space science scenario

As one of the two main use cases, the PERICLES project is considering capture and preservation of information relating to measurements of the solar spectrum being carried out by the SOLAR payload[8] of the International Space Station. The information includes operational data concerning the planning and execution of experiments, engineering documentation relating to the payload and ground systems, calibration data, as well as scientific measurements performed by solar scientists. The ultimate aim of SOLAR is to produce a fully calibrated set of solar observations, together with appropriate metadata.

We now consider three examples to illustrate the capture and use of Significant Environmental Information.

In order to validate the experimental observations of the SOLAR instrument, it is necessary to understand the impact of many complex extraneous factors on the instruments. For example, vehicles visiting the ISS can affect the trajectory of the ISS itself

[8] http://www.esa.int/Our_Activities/Human_Spaceflight/Columbus/SOLAR

and cause pollution and temperature changes. Such effects are often only uncovered by a long term analysis of the data by the scientists. Hence there is a need to capture as much of the environment as possible at the time the observations are made to enable such analysis. This includes the capture of a wide range of complex environment information relating to the instruments, the operational data, the payload sensors and events on the ISS itself. In this case, the purpose of SEI is to enable critical analysis of the solar observations by the scientists. The significance weights reflect the influence the DO captured have on the critical analysis task. These weights can change over time as additional environmental factors may be uncovered that have an impact on the scientific data. The SEI (at a given time) will therefore reflect the DOs that are relevant to critical analysis with an appropriate weighting.

In order to validate the solar measurements made by the SOLAR instrument, frequent comparisons are made with data collected independently by other scientific teams. Often the techniques and instruments are different, which provides a good way to ensure the results are not subject to unwanted effects caused by the experimental methods used. The data from other teams and the comparisons that have been made that are a valuable part of the environment metadata for the SOLAR data. The capture of the validation experiments themselves can be captured by the PERICLES PET tool and appropriate metadata created. This would include validation scripts and dependencies between subsets of the data, and would constitute (part of) the environment information. The purpose associated to the SEI is the validation of the scientific data by the science community. The significance weights reflect the value of specific data objects in the validation of the SOLAR dataset. The SEI can assist scientists in assessing the quality and reliability of the data produced.

A third example relating to the PERICLES science case study relates to the operational data for the SOLAR experiment, which is primarily created and managed by the mission operators, who operate the experiments on ISS remotely from the ground station. The operations data includes the planning, telemetry and operations logs. Given the huge complexity and volume of the space mission information, a major issue for the operators is information overload. An important task for the operators is to resolve anomalies. Anomalies occur when the normal operational parameters of the instrument are exceeded, such as overheating. Identifying and resolving anomalies often requires extensive research in the archived operations data and documentation. In this case, the digital object to be preserved is the catalogue of known anomalies and the environment information is the aggregation of all operations data. The purpose for the SEI is the identification of a specific anomaly. In this case, the significance weights indicate the relevance of a specific DO, such as a piece of documentation for the instrument or an excerpt from the archived telemetry to the particular anomaly. Thus the SEI provides a way to indicate all the environment information relevant to identifying and debugging a specific anomaly.

5. SEI EXTRACTION AND THE PERICLES EXTRACTION TOOL

Based on these premises, we are building a tool to help capture and record the environment information from the systems where the DOs are used. While different projects looked at sheer curation for very specific domains and use cases [16],[18],[19], we have built a generic, modular framework that can be adapted to support different use cases and domains with specific modules and configuration profiles. Our tool is focused on information extraction, while others target different aspects of information curation. We have also addressed the context of unstructured workflows, where the user is not adopting any workflow system, making it important to observe the flow of events in an agnostic framework.

5.1 General scenario for SEI capture

We briefly describe here a general scenario for the information capture that we are aiming at with our PERICLES Extraction Tool (PET). This should make the tool description more clear. In this scenario we observe and collect environment information from a user's computer as he interacts with DOs for different purposes. The tool is installed with the agreement and under the full control of the user. We want to look individually at the environment changes as the user e.g. calibrates some data, runs unstructured analysis workflows, creates new DOs and in different ways makes use of the data by access, interaction, and transformation.

We have different objectives that we want to accomplish, where each depends on the previous one:

1. Use the PET tool to collect environment information when the DOs are used, based on specific profiles;
2. Analyse the information to infer new relationships;
3. Assign values to the dependencies based on the purpose and significance (significance weights).

The current development status that covers mostly the first objective and starts to address the second.

5.2 The PERICLES Extraction Tool

The PET is a framework for the extraction of SEI, soon to be open sourced[9]. It can be used in a sheer curation scenario, where it runs in the system background and reacts to events related to the creation and alteration of DOs and the information accessed by processes, to extract environment information with regard to these events. All changes and successive extractions are stored locally on the curated machine for further analysis. It supports also an environment information snapshot mode, which is intended for the extraction of information that doesn’t change frequently.

The tool aims to be generic, as it is not created with a single user community or use case in mind, but can be specialized with domain specific modules and configuration. PET provides several methods for the extraction of SEI, implemented as extraction modules as displayed in Figure 3. Once PET has been configured for a particular scenario, it runs in a way that doesn't interfere with system activities and follows the sheer curation principles. An automated selection of SEI based on the use of the DOs and following the ideas outlined in paragraph 3.1 is going to be developed in a future phase.

Two different types of environment information can be distinguished: one is information directly related to a specific file, such as the location of the file at the system, or information related to the modification of the file. The other type is independent of any DO and specific to the environment, as for example general system specifications. Monitoring daemons, also displayed in Figure 3, observe the environment continuously, and trigger customised extractions based on events, as for example modifications of observed files, the creation of new files in observed directories, processing events as file openings by applications, and specific system calls. So extraction in PET sheer

[9] Apache licensed, final approval pending; source code will be available at: https://github.com/pericles-project/pet

curation mode is always related to environment changes, to avoid redundancies.

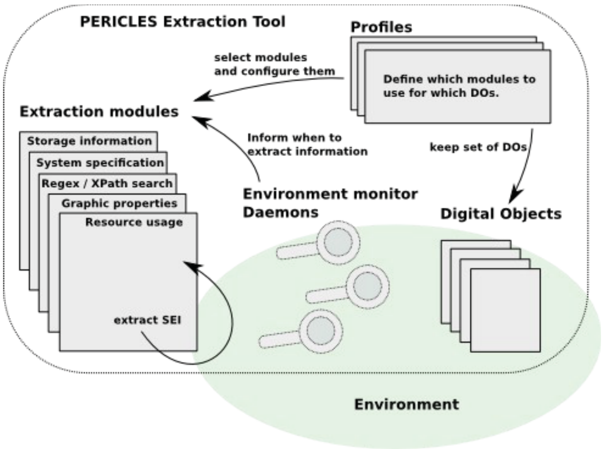


Figure 3. SEI extraction with the PERICLES Extraction Tool.

As a principle we have used existing libraries and tools, where possible, to reduce the module development times. Currently implemented information extraction modules include, among others, modules to extract:

- Available and used system resources;
- Information in files with the help of configurable regular expressions or XPath expressions;
- File format identification and checksums;
- Currently running processes;
- Event information (file and network) from processes;
- Graphic configuration information;
- MS Office and PDF font dependencies.

Furthermore we implemented generic configurable modules to execute native system commands configurable for specific needs.

With PET it is possible to create extraction profiles for different purposes to manage the information diversity. The profiles contain a set of investigated files, belonging to DOs, and a set of configured extraction modules to fit for the purpose. Future developments will include significance evaluation, as described in section 3.1, for the creation and selection of extraction profiles. Daemon modules for process and file monitoring allow the inference of process and file dependencies as described in the next section.

It is important to note that the major aim of the tool has been to enable the collection of the relevant information from the live environment, and in response to relevant events. The raw data collected will be further analysed in the tool in later tasks in the PERICLES project to conclude higher level SEI. We are also investigating techniques to encapsulate the extracted SEI together with the related DOs, to avoid data loss. These techniques will be implemented in a further PERICLES tool which will interact directly with the PET.

5.3 Extracting SEI by monitoring software

A promising technique to extract relevant information from a DO environment that we have started to develop is to look at the software currently executing on the observed system, and based on that to perform an analysis of the system calls and files used by an application. Based on a configurable set of parameters, it allows a more accurate examination of the system, and to infer dependencies between observed files based on the system activity. This will allow a reasonable amount of general information to be gathered all the time, while going in depth with the analysis of the activities when an interesting set of parameters will indicate the likelihood of a particular activity being executed.

We first describe here a simple scenario that will allow us to illustrate how such SEI collection should happen:

A scientist is calibrating data, using some specific scripts, as described in section 4.2. The PET tool is running on the scientist's machine, monitoring the environment for events that can have importance for the information collection.

The execution of a specific script triggers the event: data calibration, indicating that the user is calibrating this set of data using this script with these parameters;

Based on the event information and the state of the system the tool will first start examining the system in more detail, for example by starting a more detailed examination of the parameters and input data for the script, or observing other target applications such as office software;

A series of events and environment information is collected; this will be used to infer the activity being executed (user's purpose), and the dependencies between DOs (by looking at patterns of use, and co-occurring use of different DOs from specific software processes, dependencies based on the script, its input and output parameters; or based on other heuristics).

- By using a larger series of this collected data, we may be able to assign a significance value to the dependencies (for example by looking at how often DOs of type X is used in conjunction with DOs of type Y).
- The collected data could also be stored and kept for improving the analysis, for example by using more complex and time-consuming algorithms.

These dependencies can be mapped, automatically when possible, into a graph structure, where the edges are weighted to illustrate the importance of each dependency for the execution of an activity. The most important dependencies can then be identified defining the environment information to be extracted, on the base of the dependency graph, which helps to determine the SEI to be extracted for similar scenarios.

5.4 Provenance and other related work

Provenance information is a type of metadata that is used to represent the history of the transformations and events for a DO.

As part of our scenario, some of the data collected will be in the form of provenance information. We are exploring how such processing history of the DO can help us to infer dependency relationships, as described in more details in paragraph 5.3.

Our tool's final aim is to collect relationships between the original DO and the significant information for a specific purpose, in contrast to provenance that addresses how the DO had been created. Such dependencies are not related to single events, and are not reports of what has happened, as in the case of provenance information. It will be still possible to use our tool to collect useful provenance information, although it is not our main focus. In the development of PET, we have considered different provenance collection tools, to see if they could be helpful for our use cases.

One such example, SPADE[10], is a cross platform tool for the collection of provenance information. Its architecture [22] is similar in some ways to the one we independently designed for PET, with reporters that have a role similar to that of our modules. Spade and its modules are focused on collecting provenance information, and do not cover the variety of information we are addressing with the PET tool. We are also trying to limit the amount of information to the portion that is useful to determine SEI, and we discovered there is not a good match with the existing modules (although some of the techniques used have similarities).

The TIMBUS project [12] investigates the preservation of business processes. Although the environment information for this purpose marginally overlaps with one we are considering in this paper, TIMBUS aims at the context of the business processes, whereas we focus on assessing which environment information is, or could become, important for different uses and reuses, taking into account various purposes and informal user context while looking at the interaction with the DOs. Both our PET and TIMBUS context population tools[11] could benefit from each other by supporting the other's information extraction techniques. For example, [23] presents a different scenario: a scientist is running experiments using a formal workflow system (Taverna[12]), and the aim is that of preserving the process used by the experiment. While this is of course a worthy approach, it differs from our intent as it is based on a scenario where the user defines formally the workflow and other relevant information. In our case, we attempt to capture the process in an existing environment, where a formal workflow may not be defined.

The National Software Reference Library (NSRL) Diskprints[13] "is an attempt to comprehensively describe the changes in a computer system as a result of the influence of a software package", and could be used as an extension module to investigate in detail the state of the environment's software in its lifecycle.

6. EXPERIMENTAL RESULTS AND EVALUATION

We here describe the experiments we have set up to validate the functionality and important aspects of the framework. In all these experiments, the common steps are:

1. The PET tool is installed, configured and started on the machine where the DOs are used
2. The user interacts with the system while PET collects EI in the background
3. The environment information, DO events and changes are stored for future use and analysis

6.1 Space science

6.1.1 Operations: anomaly related information

As described in the third example of paragraph 4.2, operators dealing with anomalies usually find their solution searching through a multitude of documents. This can include for example solutions from previous anomalies, telemetry, console logs, meeting notes, emails, etc. Such data, although present in the storage, requires experience and its selection is a task that requires specific knowledge that is usually passed from operator to operator. For this reason we are addressing the collection of such dependencies between anomalies and mission documentation, in order to preserve useful information that is otherwise not captured.

In more detail, when an anomaly occurs, the issue is recorded on the 'handover sheet'. Different procedures are executed to solve the issue, and the operator's need to access the relevant documentation. We have set up a simplified experiment to show what significant environment information can be collected in this scenario. In order to support this scenario, we set up a specific PET profile that tracks the use of relevant software on specific files, using the PET software monitor; this enables us to have a trace of the documents that have been used at a given moment in time, as illustrated in Figure 4.

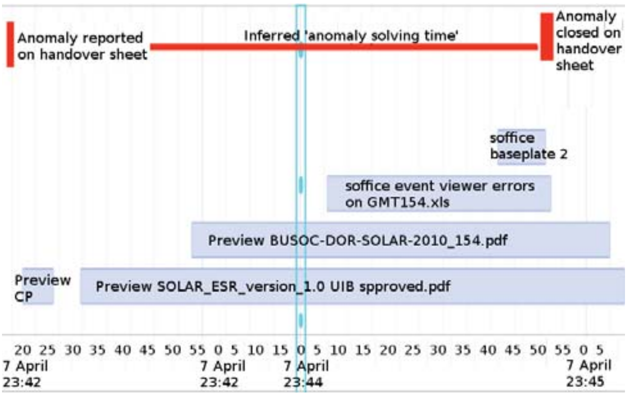
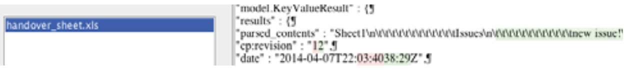


Figure 4. Trace of document use (based on open and close times) collected from process monitoring (blue) with anomaly solving time (red) collected using file change monitoring.

At the same time, it is possible to observe the 'handover sheet' and track the reporting of an anomaly start and end times (as shown in Figure 5 where a new issue is written in the document).



```
"model.KeyValueResult" : {¶
"results" : {¶
"parsed_contents" : "Sheet1\n\t\t\t\t\t\t\t\t\t\tIssues\n\t\t\t\t\t\t\t\t\t\tnew issue!"
"cp:revision" : "12",¶
"date" : "2014-04-07T22:03:4038:29Z",¶
```

Figure 5. Screenshot showing changes in the 'handover sheet' tracked by the PET tool, used to determine anomaly time

The connection between the documentation track and the 'handover sheet' tracking can allow us to infer the 'anomaly solving time span' (indicated with a red line in Figure 4) and assume there is a dependency between the solution to the anomaly and the documentation that was used between the start and end of the anomaly.

In future work we will consider more complex issues that we have ignored in this simplified example, such as the 'noise' that can be reported by the event tracking. This 'noise' can be for example due to the fact that users often multitask, so there can be unrelated documentation that was used but not relevant to the anomaly solution, or documentation that was quickly opened and closed may also indicate in some cases that the document was not relevant. We will explore also ways to obtain a fine-grained tracking, as for example to include what pages have been consulted in a document. We are planning to dedicate effort to a more careful analysis of the collected data in the next phases.

[10] https://code.google.com/p/data-provenance/

[11] https://opensourceprojects.eu/p/timbus/context-population/wiki/Home/

[12] http://www.taverna.org.uk/

[13] http://www.nsrl.nist.gov/dskprt/diskprints.html

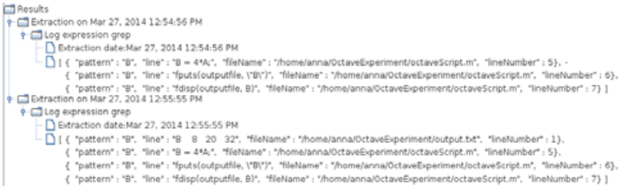
6.1.2 Extracting results of scientific calculations

The following experiment illustrates how SEI extraction can be useful for examining scientific calculations. This experiment uses an extraction module that extracts whole lines from files. It is configured to monitor an output directory of the open source tool GNU Octave[14] and to extract calculation results with the aid of a regular expression. The extraction module is originally intended for the extraction of particular log messages from a log directory.

The scientist uses PET to track the resulting development of an Octave-script execution over time and in relation to the script lines that are relevant for the result. This enables the possibility to understand the resulting changes in relation to script formula changes. First the user configures the module by specifying the output directory and the regular expression to search the result line, which is, similar to the name of the result variable, just "*B*" at this example. Then the sheer curation mode of PET is started, to monitor the directory, which triggers an initial extraction. At the time of this first extraction the script wasn't executed. We used the following script for this example:

```
#script for octave example
1;
outputfile = fopen('output.txt', 'w');
A = [ 2, 5, 8];
B = 4*A;
fputs(outputfile, "B")
fdisp(outputfile, B)
fclose(outputfile)
```

Then the scientist starts his normal octave workflow and executes the script. The PET detects the file changes in the configured output directory and triggers a new extraction of the selected module. The following screenshot shown in Figure 6 displays the results of the first and second extraction:



```
Results
  Extraction on Mar 27, 2014 12:54:56 PM
    Log expression grep
      Extraction date:Mar 27, 2014 12:54:56 PM
      [ { "pattern" : "B", "line" : "B = 4*A;", "fileName" : "/home/anna/OctaveExperiment/octaveScript.m", "lineNumber" : 5}, -
        { "pattern" : "B", "line" : "fputs(outputfile, \"B\")", "fileName" : "/home/anna/OctaveExperiment/octaveScript.m", "lineNumber" : 6},
        { "pattern" : "B", "line" : "fdisp(outputfile, B)", "fileName" : "/home/anna/OctaveExperiment/octaveScript.m", "lineNumber" : 7} ]
  Extraction on Mar 27, 2014 12:55:55 PM
    Log expression grep
      Extraction date:Mar 27, 2014 12:55:55 PM
      [ { "pattern" : "B", "line" : "B   8   20   32", "fileName" : "/home/anna/OctaveExperiment/output.txt", "lineNumber" : 1},
        { "pattern" : "B", "line" : "B = 4*A;", "fileName" : "/home/anna/OctaveExperiment/octaveScript.m", "lineNumber" : 5},
        { "pattern" : "B", "line" : "fputs(outputfile, \"B\")", "fileName" : "/home/anna/OctaveExperiment/octaveScript.m", "lineNumber" : 6},
        { "pattern" : "B", "line" : "fdisp(outputfile, B)", "fileName" : "/home/anna/OctaveExperiment/octaveScript.m", "lineNumber" : 7} ]
```

Figure 6. Screenshot of the PET showing a calculation result extraction

The result of the first extraction shows the locations of the scripts result variable *B* in the not yet executed script, which also lies in the observed output directory. At the result of the second extraction the line of the output file with the result variable *B* and its line number can be seen. This is the extracted result of the scientific calculation.

Since also the location of the result variable at the original script was extracted, an easier understanding of the dependencies between results and locations at the script is enabled. A continuous extraction over long periods of time makes an observation of result changes in relation to changes of the script formulas possible. The PET indeed needs highly customised configurations for the example, but these enable it to adapt to specialised scenarios.

6.2 SBA: system information and dependencies

This experiment is about collecting dependencies from a SBA, as described in paragraph 4.1. The PET tool is executed on the SBA and will extract a series of information useful for the understanding and future use of the SBA.

6.2.1 System information snapshot

Various information pertinent to the scenario of emulating an environment of a SBA, as described in section 4.1 can be extracted by the PET with a snapshot extraction. To these belongs mainly information that doesn't change continuously, as the systems hardware specification or installed graphic drivers.

Table 1 shows a portion of the result of a snapshot extraction executed by the PET. To the significant information belong system hardware specifications, the CPU, system graphic settings as the installed fonts and display information, and information about the operating system and development toolkits used to program the SBA's software are listed here.

Table 1. SEI snapshot extraction with the PET

Extraction Module: CPU specification snapshot	
model	Intel Core(TM)i5-3470CPU@ 3.20GHz
totalCores	4
Extraction Module: Graphic System properties snapshot	
font_family_names	Bitstream Charter", "Cantarell",...
displayInformation	isDefaultDisplay=true, refreshRate=60 ..
Extraction Module: Operating System properties	
user_language	En
os_name	Linux
Extraction Module: Java installation information	
java_home	/opt/java/jre
java_vendor	Oracle Corporation
java_version	1.7.0_15

In order to capture the type of information that changes constantly (in the SBA scenario this is mainly the use of system resources) it's possible to use PET's continuous extraction mode. A measurement of resource usage values over a long period of time can be analysed to identify behaviour patterns, which can be used to validate the correct behaviour of a new software installation. Other examples of measurements are those of CPU usages, executed by PET's *"CPU usage monitoring"*-module, whereby the changes over time can be traced.

Another example of such runtime information that can be collected and be useful for assessing the dependencies of a SBA is the file-system and network usage information (all the files and network connections used during the execution of the SBA) that can be collected by the PET tool with a specific extraction profile.

The extraction results enable the configuration and emulation of a new environment for a SBA, as described in section 4.1.

6.2.2 Extracting font dependencies

The PDF format gives the ability to embed the font types used in a document, to guarantee faithful reproduction of the document even when the DO is moved to an environment does not include them. It still is the case that PDF documents are created without the inclusion of at some necessary fonts (for user choice or application blacklisting). To recognize such missing external font dependencies, that are particularly relevant in the case of a PDF file used in a SBA, we created a module that will analyse PDF files and extract a list of used but not embedded fonts. This list determines dependencies between the DO and the listed fonts, relevant for accurate rendering.

[14] GNU Octave https://www.gnu.org/software/octave/

7. CONCLUSIONS, FUTURE WORK

In this paper we presented our work on determining what information is significant to collect, from the widest set of the Environment Information. We presented a definition of Significant Environment Information that takes into account the purpose of use of a DO, and can apply to relationships with significance weights. We also presented ways to determine significance weights and their relations to the DO lifecycle.

Finally, we presented the tool we are developing to collect such information, together with its methods of extraction, and showed experimental results to support the importance of such information. We believe the importance of the contribution also lies in the way that the information is collected, that is domain agnostic and aims at collection in the context of spontaneous workflows, with minimal input from the user and very limited assumption on the system structure and its infrastructure.

We plan to continue our work on exploring new methods of automated information collection, and improving the filtering and inference of dependencies. We also plan to explore and implement the methods for determining significance described in section 3.1, and look at the aspects of dependency graphs based on the purpose and significance weights that the tool will allow to infer.

8. ACKNOWLEDGMENTS

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

9. REFERENCES

[1] Press, N. I. S. O. 2004. National Information Standards Organization. Understanding Metadata.

[2] CCSDS, J. 2012. *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-M-2, Magenta Book.

[3] Dublin Core Metadata Initiative. 2008. Dublin core metadata element set, version 1.1.

[4] Lynch, C. 1999. Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information. *D-Lib Magazine*, 5, 9 (Sept. 1999)

[5] Hedstrom, M., and Lee, C. A. 2002. Significant properties of digital objects: definitions, applications, implications. In *Proceedings of the DLM-Forum* 200, (May 2002), 218-27.

[6] PREMIS Editorial Committee. 2008. PREMIS data dictionary for preservation metadata, version 2.0.

[7] Knight, G. 2008. Deciding factors: Issues that influence decision-making on significant properties. InSPECT project report. *Arts and Humanities Data Service/The National Archives. At http://www.significantproperties.org.uk/ deciding-factors.html*

[8] Dappert, A., and Farquhar, A. 2009. Significance is in the eye of the stakeholder. In *Research and Advanced Technology for Digital Libraries*. 297-308. Springer Berlin Heidelberg.

[9] Knight, G. 2010. Significant Properties Data Dictionary. InSPECT project report. *Arts and Humanities Data Service/The National Archives. At http://www.significantproperties.org.uk/sigprop-dictionary.pdf*

[10] Chowdhury, G. 2010. From digital libraries to digital preservation research: the importance of users and context. *Journal of documentation*, 66,2, 207-223.

[11] Kari, J., and Savolainen, R. 2007. Relationships between information seeking and context: A qualitative study of Internet searching and the goals of personal development. *Library & Information Science Research*, 29, 1, 47-69.

[12] The TIMBUS EU project, http://timbusproject.net/

[13] Lee, Christopher A. 2011. A Framework for Contextual Information in Digital Collections. *Journal of Documentation* 67,1, 95-143.

[14] Dappert, A., Peyrard, S., Chou, C. C., and Delve, J. 2013. Describing and Preserving Digital Object Environments. *New Review of Information Networking*, 18, 2, 106-173.

[15] Perspectives, K. 2010. Data dimensions: disciplinary differences in research data sharing, reuse and long term viability: A comparative review based on sixteen case studies. *Digital Curation Centre, UK, available at: http://www. dcc. ac.uk/sites/default/files/documents/ publications/SCARP-Synthesis.pdf.*

[16] Curry, E., Freitas, A., and O'Riáin, S. 2010. The role of community-driven data curation for enterprises. In *Linking enterprise data*. 25-47. Springer US.

[17] Lyon, L., Rusbridge, and C., Neilson C., Whyte, A. 2009 Disciplinary Approaches to Sharing, Curation, Reuse and Preservation, *Digital Curation Centre, UK, available at: http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf*

[18] Whyte, A., Job, D., Giles, S., and Lawrie, S. 2008. Meeting curation challenges in a neuroimaging group. *International Journal of Digital Curation*, 3, 1, 171-181.

[19] Hedges, M., and Blanke, T. 2013. Digital Libraries for Experimental Data: Capturing Process through Sheer Curation. In *Research and Advanced Technology for Digital Libraries.* 108-119. Springer Berlin Heidelberg.

[20] Falcão, P. 2010. Developing a Risk Assessment Tool for the conservation of software-based artworks. MA thesis, BFH, Hochschule der Künste Bern (HKB).

[21] Laurenson, P. 2014. *Old media, new media? Significant difference and the conservation of software-based art.* In Graham, B. *New Collecting: Exhibiting and Audiences after New Media Art.* Chapter 3. University of Sunderland, UK.

[22] Gehani, A., and Tariq, D. 2012. SPADE: Support for provenance auditing in distributed environments. In *Proceedings of the 13th International Middleware Conference,* (2012 Dec.), 101-120. Springer-Verlag New York, Inc..

[23] Strodl, S., Mayer, R., Rauber, A., & Draws, A. 2013. Digital Preservation of a Process and its Application to e-Science Experiments. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (IPRES 2013)* 1. Springer.