FP6 CALL 5: Digital Preservation Projects Meeting

Hilton Glasgow Hotel
November 23rd, 2006

Introduction

This meeting provided a chance for partners in the FP6 DPE, PLANETS and CASPAR projects to reflect on project achievements and to share operational, technical and methodological practice. The event was also a conduit for discussion surrounding a movement towards consensus for the FP7 preservation roadmap.

Session 1: Introduction to FP6 Projects

Professor Seamus Ross, co-director of the DCC at HATII, welcomed partners to the meeting and introduced the chair of the morning session, Andrew McHugh, also of HATII. Seamus gave an overview of the activities of DPE, setting the scene for the day's proceedings by addressing current issues facing the European preservation community.

David Giaretta of the CCLRC and Director of CASPAR discussed CASPAR activities to date, and Adam Farquhar of the British Library, PLANETS Director, delivered a complementary session for PLANETS. Esther Conway of the CCLRC contributed an account of the CASPAR Testbed, and PLANETS Testbed development was discussed by Max Kaiser of the Austrian National Library.

A round table session chaired by Seamus Ross then precipitated animated discussion surrounding what kind of experiments the community would encourage within a Testbed environment. Discussion opened up to include a wealth of additional topics including differing definitions of Testbeds, issues surrounding participation of the designated community, as well as cost benefit models and knowledge management.

Session 2: Project Presentations

Lunch was followed by a session on preservation planning in PLANETS by Hans Hofman of the Nationaal Archief van Nederland and Andreas Rauber of the Vienna University of Technology. Manfred Thaller from the University of Koln supplemented this with an overview of characterisation in PLANETS. David Giaretta and Luigi Briguglio of Engineering Ingegneria Informatica then gave an account of the use of the Enterprise Architecture for CASPAR and Yannis Tzitzikas of FORTH-ICS delivered a presentation on the role of knowledge management in the CASPAR environment.

Another round table session followed, chaired by Reagan Moore of the SDSC, centring on the establishment of a preservation research roadmap. Though this discussion focussed largely on technical aspects of the needs of the designated community, considerable debate did take place with regard to the ever increasing need for shared practice and collaboration.

Presentations and roundtable discussions

Some of the day's presentations are available on the CASPAR wiki (http://dev.dcc.ac.uk/caspar/bin/view/Main/MeetingCasparPlanetsDpe20061123) and summaries of the pertinent points of the roundtable discussions are as follows.

Roundtable session 1

Role of Gathering Empirical Evidence in Digital Preservation

Overview

The purpose of this discussion was to elicit from FP6 project partners what kind of experiments they might envisage conducting as part of the CASPAR and PLANETS Testbeds. The session was chaired by Professor Seamus Ross, co-director of the DCC at HATII. This report summarises some of the main thematic areas discussed.

Testbeds

Defining the Testbeds

The discussion was preceded by a request from Adam Farquhar of the British Library, Director of the PLANETS project, for the group to establish clear definitions for what each of the projects meant by the term Testbed. Max Kaiser of the Austrian National Library reiterated the PLANETS definition from his Testbed presentation. The PLANETS project defines a preservation Testbed as a controlled environment for conducting and evaluating preservation experiments in a laboratory setting, consisting of hardware and software, as well as benchmarking procedures, metrics and content.

David Giaretta of the CCLRC, Director of the CASPAR project, explained that the CASPAR Testbed could be defined as a set of validation procedures underpinned by semantics, usability and understanding of a dataset and which is driven by the OAIS Reference Model. The Designated Community is core to defining all aspects of the Testbed, based on its knowledge and understanding of the information to be preserved.

Hans Hofman of the Nationaal Archief van Nederland commented that the preservation planning activities carried out within the PLANETS testbed includes a clear model of the Designated Community for whom the preservation is being carried out, the content that they work with, and the goals of the preserving institution.

Adam Farquhar observed that the main roles for the PLANETS Testbed were to:

1) validate the PLANETS preservation planning approach;
2) evaluate preservation plans for other bodies, and;
3) evaluate third party tools (e.g., to assess a tool that purports to migrate an Oracle database into another database architecture).

Terminology convergence/ divergence

Chris Rusbridge, co-director of the DCC, suggested that it was arguably not necessary to have a convergence of definitions of the PLANETS and CASPAR Testbeds and that there was potentially more value in keeping each of these efforts discrete. David Giaretta added that the Testbeds are intended for fundamentally different purposes, with the CASPAR Testbed testing the effectiveness of the tools and processes developed by the project through involving the Designated Community in establishing usability and understandability.

Though the central concepts of the two Testbeds were agreed to be not so very different, one fundamental difference is that within PLANETS the domain could be bounded; that the tools and processes are intended for use within a single environment and that the results may be recorded and compared to other tests. Whilst the PLANETS Testbed may be ‘rented’ as a usability lab, the CASPAR Testbed is more specifically geared towards usability studies and focussed on user testing. The CASPAR Testbed questions whether its tools and techniques are adequate to provide additional Representation Information and metadata, and is concerned with usability and abstractions and how these can be reflected in user systems.

Cost benefit models

Financial implications of the Testbeds were addressed in terms of cost-benefit models and what could feasibly and affordably be achieved. David Giaretta explained that the CASPAR Testbed is not explicitly geared towards cost modelling, though this is an important consideration, in that it is essential to understand Representation Information and the complexity of the capture process in order to identify costs of a specific format.

The Designated Community

Role of the Designated Community

In discussing the role of the Designated Community in defining validation criteria and authenticity/ audit procedures, David Giaretta pointed out that the Designated Community is meant to be defined by the preservers of the digitally encoded information, and provide the means by which those who claim to be able to preserve the understandability and usability of such digital content may be tested. Omitting consideration of the Designated Community leaves us with only the equivalent of interoperability testing. Reagan Moore of the San Diego Super Computer Centre considered aspects of the RLG/NARA checklist for establishing trusted digital repositories. It was suggested that a successful Testbed might be thought of as one in which migration and retrieval could take place without any loss of integrity and authenticity (and for which a second Testbed may be needed). Characterisation of data was agreed to be dependent on who does the characterisation and with what stake in the Designated Community. In addition, the role/ purpose of a document as well as its nature, and whether simply the content of a file must be preserved or whether information about the structure is more crucial to maintain, were all discussed as impacting significantly on perception and approach to preservation. The preservation of data, as distinct from documents, raised the issue of the need to take care of the semantic content embodied in that data.

Vassilis Christophides of FORTH-ICS posed the question of how much government the Designated Community should have over metrics. As metrics are strongly focussed on user groups, and those user groups only particularly interested in certain types of data, there is an associated challenge in identifying a more generic approach to metrics in order to be of more widespread relevance to a larger user community. An additional key danger is that metrics are focussed on individual institutional needs.

Defining the Designated Community

Dialogue between members of the Designated Community in order to establish what aspects of datasets are important is a vital aspect of creating a Testbed environment, and assists in aims to expand those encapsulated by the definition of the Designated Community.

David Giaretta re-iterated that it is up to the data preservers to define their Designated Community. This level of detail is needed for a) stakeholders to judge the validity of this Designated Community and b) to allow future users to understand the knowledge base of the Designated Community in the Testbed build process.

Ontologies vs tacit knowledge

Ross King of the ARC commented that as a tool, the PLANETS Testbed would allow designated communities to review and define not only all kinds of metrics, but also help to make common knowledge within the Designated Community explicit, and therefore play a role in institutional knowledge management. However, a Testbed may be of use in preserving community ontologies but cannot be relied on for the preservation of semantic content.

Alain Bonardi of IRCAM introduced the concept of artistic interpretation, and how the tacit knowledge of a Designated Community is more difficult to preserve than information relating to agreed or shared information. For example, in the re-performance of a piece of 17th century music although the score could be accessed, information on appropriate instrumentation, tempo, timbre, etc, may be less easily sourced.

Experiments

Time as a factor in Testbed experiments

In considering different types of Testbed experiments, the topics of temporality in experimentation, how the factor of time could be included in a Testbed and the concept of the future emulated or tested were raised. It was mooted that this could be achieved by examining how past objects were used or shared in a simulation, that testing could be done in an environment replicated to reflect the technology available at a specified time, for example, in 1995 when OAIS first began to be discussed.

According to a specific subject, there may be a disparity in which changes more rapidly – technology or the community, and this may have some influence when considering timing in Testbed experiments. Vassilis Christophodes gave the example of the biological sciences, where technology may not change at the same pace as the Designated Community’s understanding of biological concepts. In such a scenario, the Designated Community itself is potentially diverse and could include members of the biological science community, computing science as well as medicine, all of whom have varying demands, understanding and definitions of curation. This will in turn impact upon how the experiment would be carried out and time perceived.

Testing the Testbed and Designated Community

Esther Conway, also of the CCLRC, commented on the possibility of creating a detailed workflow of the CASPAR Testbed by way of experimentation, to go through the preservation process component by component. This could be used to identify areas which are vital to the Testbed, as Chris Rusbridge described it 'desecated information', and a bare minimum approach to preservation. This linked in with Helen Hockx-Yu's suggestion of testing the knowledge of the Designated Community by running experiments incorporating users with a different knowledge base, and ascertaining if the same information that began the preservation process was retrieved unchanged at the end. David Giaretta confirmed that Helen's suggestion was an essential part of the CASPAR Testbeds.

Planning

Funding implications

Helen Hockx-Yu raised the issue of funding and what types of scenarios might result from the cessation of funding. Adam Farquhar defined 3 potential scenarios for PLANETS:

1. that the Testbed would be dismantled and become irrelevant, being no more persistent that the outputs of the empirical research;
2. that the Testbed software become a component in a suite of preservation tools picked up by a vendor for sale and distribution or;
3. that the Testbed is continually managed to provide emulations or further tests.

David Giaretta recognised an element of overlap with these comments though the CASPAR Testbed is not as confined as a single piece of software. Validation procedures are of more significance to the CAPSAR tools, and if these processes are successful, then the Testbed itself becomes irrelevant.

Dissemination and communication

A key point arising throughout the morning's discussion was the care required in articulating both the definition of and the work undertaken by the Testbeds to the wider community, as if such little consensus exists in the community of practice there is considerable scope for misinterpretation or ambiguity externally.

EC perspective

Arian Labat was invited to comment on the perspective of the European Comission, and underscored the continual need to evaluate the needs of Testbeds and their users. Arian agreed that Testbeds should not necessarily be creating tools for specific communities but aiming towards more distributed solutions, as an enormous diversity of designated communities have very different preservation needs.

Summary

It was concluded that there is considerable need for persistent research in the Testbed area, and that there is no universal, one-size-fits all approach to their development of effectiveness.

There was agreement that there was a need for both the PLANETS and the CASPAR approaches to Testbed design, development, implementation and experimentation, and that there would be a need for other preservation and curation testbeds built on other models.

Roundtable session 2

Constructing an Achievable Preservation Research Roadmap in the Context of FP7

Introduction

Reagan Moore of the San Diego Super Computer Centre began proceedings by outlining his observations on the creation of a preservation research roadmap.

Reagan presented the idea of preservation in the context of safeguarding the integrity and authenticity of digital records in their entirety in a changing external world. The preservation environment also maximises the reciprocal relationships with this external world to allow and enhance access to digital resources. Interpreting digital data is fundamentally built on characterizations of structures within the digital data and characterizations of relationships between the structures. These characterizations correspond to representation information in the OAIS model. The preservation of representation information is key in maintaining the integrity, authenticity and trustworthiness of records. The representation information also provides a way to exchange records between projects with differing management and operational strategies.

Preservation environments

Preservation environments attempt to generate a generic infrastructure that will allow a high level of abstraction for managing the authenticity and integrity properties required by a designated community. Such preservation environments should ideally also provide support for digital libraries, sensor systems, workflow provenance, cyberinfrastructure, etc, in order to minimise the risk that the infrastructure becomes obsolete.

Scientific data format virtualisation aims to characterise properties of a digital entity independently of the creation application. By separating the parsing of digital entities from the manipulation of the structures within the digital entity, it should be possible to apply modern display mechanisms to obsolete data formats, as well as the restricted set of manipulations that were originally available. Manipulation of structures within a digital entity is confined to the relationships present between the structures, and includes logical (including semantic labelling), structural (such as mapping to co-ordinates of a system), spatial (mapping of the coordinate system to a geometry), temporal, and functional relationships. Bit streams themselves could be abstracted as relationships on structures, which could in turn facilitate the development of a system to manipulate them based on the representation information, independently of the creation application.

In managing persistent objects, separate knowledge is required to parse bits than is needed in controlling the manipulation of behaviours. Reagan commented that for office products, Multivalent is an effective method of capturing structure and relationships, and thus implementing behaviours that are defined by these interactions. Since the Multivalent technology is written in Java, it provides an alternative to the UVC technology for implementing generic data parsing technology.

RLG/ NARA Assessment criteria

The designated community plays the pivotal role of defining standard semantics, encoding formats and standard services (using standard formats). Reagan discussed the digital preservation environment policy framework model as employed in the RLG/ NARA assessment criteria, and its role in facilitating automated validation mechanisms for trustworthy repositories. The RLG/NARA assessment criteria encompass management policies. Preservation capabilities have been defined for the NARA Electronic Records Archives. A current NARA research goal that is being pursued at SDSC is the mapping of assessment criteria to management policies, then the expression of the management policies based on the preservation capabilities, then the development of rules to automate the application of the management policies. The rules control the execution of services that implement the required preservation capabilities. Persistent state information is managed to record the outcomes of applying the management policies and to track whether the assessment criteria are met.

iRODS

iRODS, is intended to automate this rule implementation. Six logical name spaces are required to manage preservation environments: records, persons, storage resources, rules, micro services and persistent state information. Given control of these name spaces, the properties of preserved records can be defined independently of the choice of storage technology. Based on analysis of the ERA capabilities list, around 600 operations should be supported by the data management system and controlled by about 174 generic rules that manage the micro services that manipulate the data. The ability to automate management policies is imperative for large collections.

Essentially there are two types of rules: those which manage micro services (that replicate, validate, synchronise and allow comparison of outcomes) and those that manage structured information (that parse information from a disposition and submission agreements and format information for inclusion in dissemination or archival packages).

Reagan’s future thoughts focussed on the evolution of higher levels of granularity (more sophisticated structures) for the organization and management of policies, micro services and assertions of trustworthiness.

Discussion

When discussion opened up, the first comments focussed on a collaborative approach to the development of the roadmap – although this might precipitate a greater amount of meetings. Greater discussion is required amongst the community itself in order to facilitate developments.

Practical measures to reduce the causes of digital obsolescence were also raised, and it was agreed that more interaction with the creators of file formats is needed to reduce the likelihood of new versions being released year on year with new and varied functionalities.

Issues surrounding handling metadata and Representation Information with respect to particular formats were also discussed, and the example of a data dictionary to accompany a FITS file by way of explanation was given. Adam Farquhar of the British Library commented that it was important to encourage format designers to consider long-term preservation issues. For example, leaving the data dictionary out of the FITS file makes it slightly more compact, but much more difficult to understand later on. He suggested that communities may be more willing to consider such recommendations if they had a tangible scientific basis ratified by the community of practice.

Reagan Moore then raised the issue of desktop access to repositories and commented that with the increasing proliferation of data grids and computational leverage, this could easily become a reality within the next five years.

Reagan went on to discuss that such a system could provide end-to-end encryption and error recovery mechanisms such that any document preserved would not be decrypted until it leaves the repository, and that recovery from single-bit errors could be a property of the data format. Discussion within the room then turned to address the diversity of proprietary file formats and how these are designed by communities who wish to share data and the problems this creates in providing a standard service. Chris Rusbridge of the DCC commented that he thought this may be an idealistic idea, and that members of the community would be eager to create new file formats purely from a research and development and creation of new science point of view.

Reagan Moore returned to the topic of encryption and the use of description formats such as EAST and DFDL. David Giaretta of the CCLRC added that no description language will be able to describe everything – different techniques would be needed for different types of encodings. He went on to talk about the creation of extra abstractions to make files more easily usable by software in the future and thereby keep costs down.

Reagan then spoke briefly about a project involving the Thomas Collection within the Library of Congress, in which temporal and semantic relationships present within the material were explicitly characterized. It was then possible to characterize all involved senator's names (identified by party), and consequently permitted the chronological ordering of bills according to amendment date. In this case documents could only be parsed according to the relationships dependent on when they were created.

Reagan also mentioned compound documents which contain different areas with different embedded relationships.

Summary

Though various technical considerations of the preservation roadmap were discussed, time did not permit the further exploration of a number of issues surrounding the urgency of shared practice in informing the development of FP7 projects. Ariane Labat of the

European Commission mentioned the establishment of 'Centres of Competence' in the upcoming FP7 call, thus outlining the need for shared expertise based on a hub-and-spoke model.

A collaborative approach and greater interaction were identified as being key tasks within FP7 and are central to the upcoming call. A diversity of operational aspects of FP6 projects such as PLANETS and CASPAR will be used to inform the methodology and logistical interactions of the community of practice in undertaking future research.

Event summary

Overall, the meeting allowed partners to learn about individual subprojects, highlighting both similarities and differences between operation approaches and methodologies. This kind of communication should encourage further project collaboration and enhance future shared effort, which will be key in addressing the centrality of communication in the upcoming call for FP7 projects.

Victoria Boyd (CASPAR/HATII) & Kellie Snow (PLANETS/HATII)
27 November 2006