DELOS
The Digital Preservation Cluster
Progress in JPA2

Prof Seamus Ross
Director, HATII (University of Glasgow)
http://www.dpc.delos.info

DP Cluster Objectives

- To contribute to the elimination of the duplication of effort of research activities by researchers at different institutions and to enable identification, collection, and sharing of knowledge and expertise;
- To examine core issues that will deliver essential guidelines, methods, and tools to enable the construction of preservation functionality within digital library activities and deliverables are created;
- The establishing of testbeds and validation metrics;
- To relate the digital preservation research agenda more directly to the development of exploitable product opportunities and to develop links with the industrial sectors;
- To catalyse the research and funding environment to enable delivery of the DELOS/NSF research agenda for Digital Preservation and Archiving; and,
- To raise the profile of digital preservation issues within the Digital Library Community.

JPA 1 High-Level Objectives

- To develop cross institutional research potential.
- Create the right mix of research activity and delivery in theoretical, methodological and practical domains.
- To establish a research team balance between eContent owning and creating institutions and research institutes.
- To create the appropriate level of shared domain ownership to enable the cluster's progression from synthetic and evaluative activity to experimental and analytical research.
- to lay the foundation for testbeds and necessary metrics and tools for assessing preservation strategies
- to collaborate with other international bodies to ensure consistencies of digital repository standards.

DELOS
NETWORK OF EXCELLENCE ON DIGITAL LIBRARIES

DP Cluster JPA2 Objectives

- establish a digital preservation testbed;
- make progress on the automation (or at least semi-automation) of the processes of ingesting material into preservation environments
- complete work on defining of attributes and functionalities that need to be represented;
- foster approaches to integrate preservation analysis and design approaches into application design methodologies;
- promote the adoption of preservation technologies in digital library development designs;
- raise the profile of digital preservation issues within the Digital Library Community; and
- increase our collaboration with other international researchers conducting research within the digital libraries and preservation communities

Information Society

WHO IS IN OUR SANDBOX

- University of Glasgow
- Technical University of Vienna
- University of Köln
- UKOLN
- National Archives of The Netherlands
- Phonogrammarchiv
- CNR
- Göttingen State and University Library (Germany),
- Austrian National Library (Austria)

Digital Preservation Testbed Forum *JPA1WP6Task*1 -- Achievements

- Primary Objective
  - Establish a framework of a digital preservation testbed environment.
- Achievements
  - Defined on paper a framework for a digital preservation testbed environment and produce metrics for testing and validating digital preservation strategies supporting
  - Design Digital Preservation Testbed Environment for conducting repeatable and measurable experiments for identifying appropriate preservation strategies and methods,
  - Ensuring consistency in the framing of research questions,
  - Delineating the experiment process, describes the Testbed environment, and presents the products of the research
  - Will provide an add-on to OAIS which lacks preservation system

Designing, Deploying, & Managing Digital Repositories *JPA1WP6Task2* -Achievements

- Primary Objective
  - Evaluate the current and emerging systems digital repositories models.
- Achievements
  - Examined the current generation of storage models for digital repositories (e.g. DSpace, LOCKSS)
  - Agreed criteria for evaluating digital repository models
  - Agreed areas where research is needed in digital repositories. Special focus is on the points of ingest and access within the repository environment.
  - Integrated political, economic, and organizational issues as well as research into the technical models—repository models
  - Laid groundwork to compare approaches on scalability and ingest being conducted using HATII's SAN

File Formats, Classification, and Typology *JPA1WP6Task3* - Achievements

- Primary Objective
  - Contribute to the development of file format registries and the mechanisms for their use and define the relationship between file format types and preservation methods
- Achievements
  - Fundamental file format criteria identified and discussed
  - Analysis of existing literature showed inconsistencies
  - Digital preservation strategies for the file format management are presented
  - Defined the concept of transformability as
  - file format categorization and the identification of the existent classification criteria. Defining the relationship between file format types and preservation methods to be done in WP6-T3 to do by feeding into WP6-T4
  - Defined further research that is needed but the research questions are outside the funding scope.

Documentation of Functionality and Behaviour Metrics *JPA1WP6Task4 -- Achievements*

- Primary Objective
  - Define framework for documenting behaviour and functionality through building the attributes of functionality and behaviour described and layout mechanisms for representing them.
  - Investigate the viability of automating the process of functionality and behaviour verification
- Achievements
  - A workflow for the elicitation of these requirements was designed, resulting in a tree of objectives.
  - First approach to make preservation approaches comparable and measurable
  - the application of preservation actions evaluated and different preservation strategies examined
  - a model was put into first field trials with partners of the preservation cluster, particularly focusing on the audio collection of the OEAW-Phonogrammarchiv.
  - Supports transparent and informed decision making

Enabling the Integration of Digital Preservation Architectures *JPA1WP6Task5* -- Achievements

- Primary Objective
  - Develop the requirements for a preservation functionality modelling tool.
  - Develop a models of preservation functionality.
- Achievements
  - a comparative evaluation of diverse design methodologies (completed)
  - Examination as to how preservation aspects can be integrated into design methodology (completed)
  - The effort is currently directed, therefore, to an attempt to identify as many possible spots within the existing UML language constructs to which "preservation extensions", based primarily upon the state diagrams of UML, to which preservation aspects can easily be attached.

JPA2WP6Task5 : Modelling Preservation Functionality

- Modelled in UML two frameworks of digital libraries—a pragmatic and a formal one
- Added persistency functionality modules to the design process of any DL
- Demonstrated that modules fit both of our base models
- The work completed includes UML model for preservation components of *any* digital library.

JPA2WP6Task5 : NEXT STEPS

- **A draft version of the report of *JPA2WP6Task5 available.***
- **Workshop in February to validate the applicability of the approach through analysis of digital library systems and applications**
- **A revision will be released towards the end of March 2006 following the workshop**

JPA2WP6Task6

- Builds on results of *JPA1WP6Task1* (testbed development), J*PA1WP6Task3* (Format analysis), and *JPA1WP6Task4* (utility analysis)
- Software tool support was developed to assist in the processing of the Utility Analysis to inform the selection of preservation methodologies.
- Conceptual integration of the Austrian Utility Analysis Tool with of the Dutch testbed procedures (DELOS Digital Preservation Testbed) has been completed
- Use Case Study has been done focusing on Audio records of the Austrian Phonogram Archive, Video records of the Austrian Phonogram Archive, and Document records of the Dutch National Archive

JPA2WP6Task6 : Next Steps

- Software support integration for the functionality offered by the Dutch testbed
- Further use(r) case studies, specifically for database preservation (in cooperation with CNR), collections of thesis documents (in cooperation with the Austrian National Library) and a special collection of Göttingen.
- Validation of tools and their refinement

JPA2WP6Task7

- look at ways of automating the semantic metadata extraction process and create a prototype tool, integrate this tool with other metadata extraction tools and ingest processes for automatic population of document repositories.
- Using linguistic and layout analysis techniques to automate this process of metadata extraction and creation.
- The research within this task can be divided into six activities:
  - (a) selecting metadata to be extracted and that can be extracted or created (*done),
  - (b) integrating previous and current related research (*done),
  - (c) designing a prototype metadata extraction and creation tool (*completed but not validated) [focusing on genre classification (e.g., 59 Genre such as research article vs email)
  - (d) implementing a prototype metadata extraction tool (*prototyped,
  - (e) establishing a well-designed corpus of documents to validate the effectiveness of the prototype, and
  - (f) testing and refining the prototype

JPA2WP6Task7: Next Steps

- completion of the prototype tool and its testing on a narrow dataset,
- work on constructing a million document corpus reflecting document types, domain and distribution of size (10k to 20MB)
- work towards construction of the testbed environment,
- Validation testing and refining of the prototype tool, and
- extrapolation of the model to enable it to be used.

DPC and JPA3

- Options –
  - – Continue all three current tasks – lots still to do
  - – Incorporate results of *JPA2WP6Task5 in overall DL Model JPA2WP1Task4 and continue with JPA2WP6Task6* and *JPA2WP6Task7*
  - – *Pass the outcomes of JPA2WP6Task6* and *JPA2WP6Task7 to other projects and begin some new preliminary work in selection and appraisal building on this earlier work*

DELOS DPC JPA3

- ensure that the WP works comprehensively together :
- we will continue work in the area of the automated extraction of metadata and the evaluation of the results for their capability to improve the preservation process (Task 6.7);
- we will integrate the current extraction tool with outcomes from Task 6.6;
- we will promote the extrapolation of our work to ensure that other tools are developed; and,
- we will aim to proceed with automating the process of appraisal (and re-appraisal).

- Primary goal of DELOS DPC is to ensure that we provide fundamental building blocks to ensure that preservation is adequately represented in digital library framework and activity cycle and that we can created new funded research opportunities.

DELOS

Network of Excellence on Digital Libraries

DPC JPA3 What we want to do

- Build on work to automate extraction and creation of information about digital objects (either content or context or both) (Tasks 6.6 and 6.7);
  - support for automated ingest of objects in a repository, specifically with a focus on the automated extraction of metadata and the subsequent evaluation of the objects with respect to preservation issues. (suitability and stability of the file format, complexity)
  - support for automating preservation action evaluation using file format characteristics information and object metadata.
  - evaluate the potential for automated preservation action advice for existing repositories based on object metadata
- analyse this information based upon appraisal criteria being established by the Urbino/Naneth team—this will provide us with a selection and appraisal model;
- attribute appropriate appraisal/selection/disposal metadata to the digital object (GU/Cologne/Vienna)—result should be ingest or dispose trigger;
- Implement a tool representing the appraisal decision making process and integrated it with the utility analysis tools, and the metadata extraction and creation tools.

- **What we want to investigate is whether we can control the ingest pipeline and reduce costs of acquisition of content**

Information Society

DELOS WP6: Coordination -- 2005

- Summer School and Sophia Antipolis Cluster Meeting
- Cluster Conference Calls
- Den Haag Mtg (October 05)
- Sub-Cluster mtgs
- Exchanges
- DELOS DPC Website – http://www.dpc.delos.info

DELOS DPC Summer School

- Summer School on Digital Preservation in the context of Digital Libraries in Sophia Antipolis (FR) in June 2005
- Planning a second in Cortona (IT) in June 2006

Cortona Summer School

- Monday
  - (Morning) Digital Curation in Digital Libraries: Issues, Obstacles, and Possibilities (*Anne R. Kenney)*
  - (Afternoon) Unpacking The OAIS Model (TBC)
- *Tuesday,*
  - (Morning) Creating and Using Metadata and Registries (*Wendy Duff),*
  - (Afternoon) Identifying, Evaluating and Selecting Preservation Methods: An Introduction to the DELOS Testbed and Utility Analysis (*Hans Hofman* and *Andreas Rauber*)
- *Wednesday,*
  - (Morning) Methodologies of Selection and Appraisal (*Ross Harvey)*
  - *(Afternoon)*Functional Requirements for Preservation Systems, (TBC)
- *Thursday*
  - (Morning) Managing Ingest: Handling, Documenting, and Automating (*Birte Christensen-Dalsgaard*)
  - (Afternoon) Active Ingest: Metadata Extraction, Creation, and Workflow (Y Kim & *S Ross)*
- *Friday*
  - (Morning) Digital Libraries as Persistent Collections of Autonomous Objects (*Manfred Thaller)*
  - (Afternoon) Audit and Certification of Preservation Processes and Repositories
    - Approaches to Preservation *(Bill Underwood……)*
    - Audit and Certification of Preservation and Repositories (TBC)