



7th DELOS Summer School 2006
Digital Preservation in Digital Libraries:
Emerging Approaches

Convento de Cappuccini, San Miniato, Italy
4th-10th June, 2006

Session Plans & Biographical Sketches of Lecturers

- ♦ **DELOS: A Network of Excellence on Digital Libraries**
Instrument: Network of Excellence
Thematic Priority: IST-2002-2.3.1.12
Technology-enhanced Learning and Access to Cultural Heritage
European Commission Project no. 507618

Co-Sponsors:

- ♦ **Fondazione Giangiacomo Feltrinelli (Milan)**
- ♦ **Fondazione Rinascimento Digitale (Florence)**

Seamus Ross, Maria Guercio, & Hans Hofman
Academic Directors of the DELOS 2006 Summer School
On behalf of the Digital Preservation Cluster of DELOS
<http://www.dpc.delos.info>

Lecturers and Session Presenters

- *Birte Christensen-Dalsgaard*, Head of Development at the State and University Library in Aarhus.
- *Michael Day*, Research Officer, UKOLN
- *Wendy Duff*, Professor in Faculty of Information Studies the University of Toronto.
- *David Giaretta*, Associate Director for Development at the UK Digital Curation Centre.
- *Mariella Guercio*, Professor in Archival Science and Electronic Record Management, ISTBAL, Università degli Studi di Urbino.
- *Ross Harvey*, inaugural Professor of Library and Information Management at Charles Sturt University's School of Information Studies.
- *Hans Hofman*, co-director of ERPANET and senior advisor at the Nationaal Archief (National Archives) of the Netherlands.
- *Yunhyong Kim*, Digital Curation Centre Researcher, HATII University of Glasgow
- *Andrew McHugh*, Audit and Certification Manager, Digital Curation Centre (DCC), HATII University of Glasgow
- *Heike Neuroth*, Research & Development & DINI secretary Göttingen State and University Library (SUB)
- *Seamus Ross*, Professor of Humanities Informatics and Digital Curation, and Director HATII University of Glasgow.
- *Andreas Rauber*, Professor at the Department of Software Technology and Interactive Systems at the Vienna University of Technology.
- *Manfred Thaller*, Professor of Historisch-Kulturwissenschaftliche Informationsverarbeitung, University of Köln.
- *Andrew Wilson*, Arts and Humanities Data Service

Course Timetable

Sunday, 4 June 2006

- 18:30 Overview of the Summer School and Introduction to the Lecturers
- 19:30 Opening Reception at San Miniato

Monday, 5 June 2006

- 9:00 - 12:45 Digital Curation in Digital Libraries: Issues, Obstacles, and Possibilities (*Hieke Neuroth*, Research & Development & DINI secretary Goettingen State and University Library (SUB))
- 14:00 - 17:30: Unpacking The OAIS Model (*David Giaretta*, Associate Director for Development at the UK Digital Curation Centre)

Tuesday, 6 June 2006

- 9:00 - 12:45 Creating and Using Metadata and Registries (*Wendy Duff*, Professor in Faculty of Information Studies at the University of Toronto)
- 14:00 - 17:30: Identifying, Evaluating and Selecting Preservation Methods: An Introduction to the DELOS Testbed and Utility Analysis (*Hans Hofman*, senior advisor at the Nationaal Archief of the Netherlands and *Andreas Rauber*, Professor at the Department of Software Technology and Interactive Systems at the Vienna University of Technology)

Wednesday, 7 June 2006

- 9:00 - 12:45 Approaches to Preservation (*Michael Day*, UKOLN and *Andrew Wilson*, Arts and Humanities Data Service)
- 14:00 - 17:30: Methodologies of Selection and Appraisal (*Ross Harvey* is the inaugural Professor of Library and Information Management at Charles Sturt University's School of Information Studies)

Thursday 8 June 2006

- 9:00 - 12:45 Managing Ingest: Handling, Documenting, and Automating (*Birte Christensen-Dalsgaard*, Head of Development at the State and University Library in Aarhus)
- 14:00 - 17:30 Active Ingest: Metadata Extraction, Creation, and Workflow (*Yunhyong Kim*, Digital Curation Centre Researcher, HATII University of Glasgow & *Seamus Ross*, Professor of Humanities Informatics and Digital Curation, and Director HATII)

Friday, 9 June 2006

- 9:00 - 12:45: Preserving for 2016, 2116, 3016. Or: "Is there a life for an object outside a digital library?" (*Manfred Thaller* Professor of Historisch-Kulturwissenschaftliche Informationsverarbeitung, University of Köln)
- 14:00 - 17:30: Audit and Certification of Preservation Processes and Repositories (*Andrew McHugh*, Audit and Certification Manager, Digital Curation Centre (DCC)).

Session:

Overview of the Summer School & Introduction to the Lecturers

Abstract:

The DELOS Network of Excellence is a four-year project funded by the European Commission under the Sixth Framework Programme to synergize and foster technology for the next-generation of Digital Libraries (www.delos.info). This delivery of this event is co-sponsored by the Digital Curation Centre (DCC) and . The aim of the summer school is to assist participants in understanding how to address digital preservation challenges in the context of the digital library. The workshop also provides a networking opportunity for participants to meet with other students and researchers, international experts, and practitioners across disciplinary and national boundaries.

During this 6-day summer school, internationally established lecturers will each lead half-day sessions and most will be available through the week for further discussion. By the end of the course, participants should have:

- gained an appreciation of the issues surrounding digital preservation within the context of digital library development and management;
- developed a grasp of the core research in the area of digital curation and preservation;
- developed a coherent and practical understanding of activities surrounding digital preservation;
- gained experience with issues surrounding workflow modelling, metadata definition, and ingest process management;
- acquired an appreciation of the different approaches to selecting and appraising potential digital acquisitions;
- become familiar with the OAIS model and gained a knowledge of the approaches to repository design and deployment;
- a working knowledge of the issues surrounding audit and certification of digital repositories;
- a working knowledge of the techniques and practices that underlie digital curation;
- develop a firm understanding of the issues of authenticity, integrity, and reliability in relation to digital libraries; and
- considered how digital curation and preservation requirements can be integrated into approaches to digital library development

Session

Digital Preservation in Digital Libraries: Issues, Obstacles, and Possibilities

Lecturer and Session Leader:

Heike Neuroth, Head Research & Development Department at the Goettingen State and University Library, Germany

Abstract:

This session provides a general introduction in digital preservation and explains the key issues in this area. The focus lies on the big challenges in this relatively new area like organisational aspects or strategic approaches for preservation planning. The key questions addressed in this session include:

- What is digital preservation
- Why is digital preservation important for digital libraries
- Which is the relevant material, documents and data for digital preservation in digital libraries
- What are the big challenges digital libraries have to deal with (organisational, strategic, technical issues etc.)?
- How could a preservation strategy be developed?
- What are the relevant standards, initiatives, programs?

Aims and learning outcomes of this session:

This session aims to:

- explain the needs and challenges of digital preservation in general
- provide an overview of the role of digital preservation in digital libraries (why, what)
- describes the various aspects of digital preservation concentrating on technical, organisational, strategic issues as well as considering copy-right questions and relevant standards
- give some examples of the heterogeneity of digital objects which might be under responsibility of digital libraries
- offer an outlook of ongoing activities

By the end of this session students will have:

- developed a feel for digital preservation needs in digital libraries
- become familiar with the challenges digital libraries might face when dealing with the digital preservation and the variety of digital objects
- understood the heterogeneous aspects which should be considered by developing a digital preservation strategy

Session outline:

09:00 – 09:45 *Lecture* (45 minutes): Introduction and key issues of digital preservation in digital libraries

09:45 – 10:00 *Discussion* (15 minutes) and short break

10:00 – 10:45 *Activity* (45 minutes): Development of specific action items, e.g. considering technical aspects, organisational issues etc. in groups

10:45 – 11:30 *Discussion with embedded break* (45 minutes): The groups present their initial results, followed by discussion

11:30 – 12:00 *Activity* (30 minutes): The groups develop a digital preservation strategy

12:00 – 12:30 *Summary discussion* (30 minutes)

Session

Unpacking the OAIS Model

Lecturer and Session Leader:

Dr David Giaretta, Associate Director for Development at the UK Digital Curation Centre

Abstract:

This session aims to elucidate the Open Archival Information Systems (OAIS) Reference Model, its origins, its purpose, the reasons behind its structure, the several models within it and the follow-on standards planned or in progress. OAIS is one of the key standards concerned with the long term preservation of digital information.

Aims and learning outcomes of this session:

This session aims to:

- provide a clear understanding of the origins and role of the OAIS Reference Model
- explain the real issues being addressed by OAIS
- describe the various components of the model
- give examples of mapping existing repositories to OAIS
- indicate how the OAIS Reference Model fits into the wider context of Digital Libraries

By the end of this session students will have:

- developed an appreciation of the significance of the OAIS Reference Model
- understood the roles and interrelations between the components of the reference model
- become familiar with the limitations of the OAIS Reference Model
- grasped some of the ways in which the reference model can be applied.

Session outline

- *Lecture* (60 minutes): The key ideas behind the OAIS Reference Model, and the role of its various components
- *Discussion* (participants, facilitated by the presenter, 20 minutes): Given a stream of bits, what is needed to produce usable information? How will this change over time?
- *Seminar* (30 minutes): discussion of two or three key readings (details of these to be advised)
- *Lecture* (25 minutes): Applicability of the OAIS Reference Model.
- *Activity* (30 minutes): Mapping OAIS to existing repositories.
- *Summary* (15 minutes)

Session:

Creating and Using Metadata and Registries

Lecture and Session Leader:

Wendy Duff Professor in Faculty of Information Studies the University of Toronto

Abstract:

Metadata can be defined simply as structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage resources. In recent years, there has been a growing awareness of the role of metadata in supporting the preservation and reuse of digital resources. For example, in 2003 the US National Science Foundation's Blue-Ribbon Advisory Panel on Cyberinfrastructure argued that the creation and maintenance of metadata was essential for the ongoing stewardship and curation of research data. To date, much metadata development has been fragmented, typically focused on particular implementations, domains or resource types. More recently, however, groups like the PREMIS (Preservation Metadata: Implementation Strategies) working group have begun to identify the core functional requirements of preservation metadata, basing their work partly on the increasingly influential Reference Model for an Open Archival Information System (OAIS).

Aims and learning outcomes of this session:

The session aims to:

- Introduce students to the important role of metadata in supporting digital preservation strategies and highlight some key initiatives
- Identify generic preservation metadata requirements, i.e. the functions that metadata might support within a preservation system
- Highlight the difficulty of defining a single schema that might be suitable for all resource types and all preservation strategies

By the end of the session students will have:

- Gained an appreciation of why the creation, capture and management of metadata should be a key part of most digital preservation strategies
- Developed a feel for the range of metadata initiatives and standards currently in use (or under development)
- Appreciated some of the practical reasons why generic preservation metadata schemas are difficult to develop and implement

Session outline:

- *Lecture* (30 minutes): Welcome and session outline, defining metadata, outlining its importance in supporting preservation strategies (including examples). The primary focus will not be specific schemas or implementations, but generic metadata requirements (e.g. those identified by the OAIS information model) and how these might interact with different resource types and preservation strategies.
- *Discussion* (30 minutes): Questions will include: What are the most important types of metadata needed to support preservation? What may be missing from the OAIS information model categories? How much of this metadata might exist in other forms already, or may be able to be captured automatically?

- *Lecture* (30 minutes): Introducing selected schemas (looking in detail at the results of the PREMIS activity), implementation issues, and managing metadata.
- *Activity* (1 hour, 15 minutes; 45 minutes discussion in groups, 30 minutes reporting back time and discussion): participants (including other summer school tutors) divided into teams, each looking at a different resource type (e.g., simple dataset, online e-journal, TIFF image, etc.). Using a template, teams will need to select an appropriate preservation strategy, identify what types of metadata would be necessary to support this, consider what metadata exists already (and whether this can be captured), then mapping this on to the OCLC/RLG framework (or the PREMIS Core Metadata model, if available) The aim is to see whether such models are useful (or sufficient), and get students to think about the issues surrounding the generation of metadata in a practical way.
- *Summary and discussion* (15 minutes): Summing up and a final opportunity for questions

Session

Identifying, Evaluating and Selecting Preservation Methods: An Introduction to the DELOS Testbed and Utility Analysis

Lecturers and Session Leaders:

Andreas Rauber Associate Professor at the Department of Software Technology and Interactive Systems at the Vienna University of Technology.

&

Hans Hofman co-director of ERPANET and senior advisor at the Nationaal Archief (National Archives) of the Netherlands

Abstract:

During the last couple of years, several preservation strategies have been developed, tested, rated and implemented, but until now, none of them is clearly better than all the others. Thus, many inexperienced users, but also experts are not sure which solution to apply for which data collection. Unfortunately, the decision depends not only on the types of records in a collection, but also by many other qualities, such as specific requirements, user satisfaction or costs.

Choosing the optimal preservation strategy for a given setting constitutes a complex decision, depending on a large range of possibly conflicting requirements. During the first year of DELOS the Vienna University of Technology and the Dutch National Archive developed two evaluation environments, which were joint to one common DELOS Testbed. This DELOS Testbed is a powerful tool to describe the requirements of an institution and to clearly rank possible alternatives for long-term preservation.

In this session the process of the DELOS Testbed is described in detail both theoretically and with practical examples. The process of setting requirements for long-term preservation is described as well as the setting of alternatives, the weighting of preservation goals within an organization and the evaluation and ranking of alternatives. The focus is set on the workflow, but also the foundations necessary to understand the steps of the testbed are described.

The key questions that are addressed in this session include:

- What are the requirements, which a preservation solution must fulfill?
- What are the criteria, which can be considered, when choosing one or another preservation approach?
- Which alternatives are viable and how to detect inappropriate preservation solutions?
- How can different alternatives be compared and how can a suggestion for one or another of them can be made?

Aims and learning outcomes of this session:

This session aims to:

- Introduce students to the concepts of the DELOS Testbed
- Provide a detailed description of the testbed and evaluation process
- Demonstrate the usability with the help of a practical example

By the end of this session students will have:

- An improved sensibility on requirements for long-term preservation
- The knowledge to apply the DELOS Testbed in their home-institution
- The ability to make informed and accountable decisions selecting one or another preservation solution for their individual requirement.

Session outline:

09:00-09:30 Lecture (30 minutes): Theoretical introduction. Detailed description of the single steps

09:30-09:45 Discussion (participants, facilitated by the presenter, 15 minutes) and short break

09:45-10:30 Activity (45 minutes): Creation of individual objective trees in groups supported by an instructor. Definition of possible preservation alternatives.

10:30-11.15 Discussion with embedded break (45 minutes): The groups present their initial results, discussion of these ideas, short break.

11:15-12:00 Activity (45 minutes): The groups apply the other steps of the DELOS Testbed and so choose the optimal preservation solution for their individual preservation environment.

12:00-12:30 Summary discussion (30 minutes)

Session

Selection and appraisal for preservation in digital libraries:

Why? What? How?

Lecturer & Session Leader:

Ross Harvey is the inaugural Professor of Library and Information Management at Charles Sturt University's School of Information Studies

Abstract:

Librarians and recordkeepers have long acknowledged their responsibility for preserving documents for future use, and have developed criteria and processes for identifying the documents to which they will devote resources to ensure their preservation. However, these criteria and processes have been developed and applied to documents in what have been primarily paper-based collections – they do not automatically translate to digital material. They need to be revisited and modified to ensure that they can be applied effectively to digital materials. The basis of selection is defining value – and here the challenges begin. How is value defined? Value to whom? Does value change as time passes? Will users in the future define value in the same way that we do? How do we address the preservation needs of valuable material in the way most appropriate to that value? For digital materials selection decisions are ‘not a choice made once and for all near the end of an item’s life cycle, but rather ... an ongoing process intimately connected to the active use of the digital files’ (Paul Conway). The ‘digital mortgage’ that is the consequence of the selection decisions also need to be considered: ‘Program costs don’t cease when the Web site disappears’ (Diane Vogt-O’Connor).

This session notes selection criteria traditionally applied in library practice and appraisal criteria traditionally used by archives, then indicates why these selection criteria, developed for physical artefacts, do not transfer well to digital materials. It considers additional factors to be considered when developing effective selection policies and practices for digital materials, such as considering the role of intellectual property ownership, the importance of preserving context, the place of stakeholder input, and risk management approaches. Emerging frameworks for selecting digital documents are discussed. The key questions that are addressed in this session include:

- Are selection and appraisal necessary? Why can’t we preserve everything?
- What is the mandate for collecting? Who is the audience (institutional missions, users, clientele)? Are there other stakeholders we should engage?
- What do we select for preservation? What criteria might we use?
- What techniques do we have available to assist in selection and appraisal?
- How might we make selection and appraisal a routine activity in developing and managing digital libraries?

Aims and learning outcomes of this session:

This session aims to:

- Introduce students to the key challenges of selection and appraisal of digital material for long-term preservation

- Provide students with a basic understanding of selection and appraisal principles and practices for the purposes of digital preservation
- Indicate how selection and appraisal principles and practices can be applied to digital library development and management.

By the end of this session students will have:

- An appreciation of why selection and appraisal is a necessity for digital libraries
- Basic understanding of selection and appraisal principles and practice as they might be applied to the preservation of digital materials
- Familiarity with techniques that might apply to assist selection and appraisal
- Awareness of how the issues are differently articulated in different digital library environments.

Session outline:

- *Lecture* (20 minutes): Why is selection necessary? What are the key issues?
- *Discussion* (participants, facilitated by the presenter, 40 minutes): Discussion of two or three key readings (details of these to be advised), using focus questions: Who are the users of digital libraries? What do they require now? What will they require in the future?
- *Lecture* (20 minutes): What selection criteria might we use? What techniques do we have available to assist in selection and appraisal? How might we make selection and appraisal a routine activity in developing and managing digital libraries?
- *Examples* (30 minutes): presentation of examples from practice and class discussion of them
- *Activity* (30 minutes): Scenarios to explore the different requirements of selection in different digital library contexts. Based on a template that participants fill in, according to their backgrounds and/or their organizations, which asks for responses to questions about the users (now? in ten years time? in 100 years time?), their requirements (now? in ten years time? in 100 years time?), and the basis on which selection to meet user needs might be made
- *Group reporting of scenarios* (30 minutes)
- *Summary* (10 minutes)

Session

Approaches to Preservation

Lecturer & Session Leader:

Dr Andrew Wilson, Preservation Services and Projects Manager, Arts & Humanities Data Service (AHDS), UK. & *Michael Day*, UKOLN

Abstract:

Unlike most paper based materials, just keeping digital objects as they are is insufficient for preservation. Digital curators will need to take active steps with digital resources to ensure that they survive and are useable into the future. This introduces what has been called ‘the paradox of preservation’: the desire to maintain digital resources intact but also to keep them useable across time and space. It has been observed that “our digital environment has fundamentally changed our concept of preservation requirements. If we hold on to digital information without modifications, accessing the information will be come increasingly difficult, if not impossible”.¹

After a basic introduction to the ‘problem of preservation’, this session will explore the whole question of what makes a preserved object authentic. If we must change digital objects in order to preserve them over time what of the source digital object do we need to retain to ensure the object is still considered ‘authentic’ over time? Is it ‘look and feel’, content, or structure? Do we need to maintain all of these or some combination of them? The notion of significant properties will be introduced and analysed. The session will continue with a discussion of the methods that are currently being used for preserving digital resources. The focus of this part of the session will be on the various approaches to preservation, specifically, technology retention, migration, emulation (including the UVC concept), and the digital archaeology approach. Examination of each approach will use practical examples of real world implementations where possible, and will discuss the issues and problems associated with each approach.

Aims and learning outcomes of this session:

This session aims to:

- introduce participants to the problems of digital preservation
- discuss authenticity issues and examine the notion of ‘significant properties’
- provide a detailed description of different approaches to preserving digital objects
- investigate examples of current implementations of the different approaches

By the end of the session participants will have:

- gained an understanding of the principles underlying digital preservation
- understood the notion of ‘significant properties’

¹ Su-Shing Chen, “The Paradox of Digital Preservation”, *Computer*, March 2001, 2-6.

- reached an awareness of answers to the question: ‘what do we preserve?’
- developed an appreciation of role of bitstream preservation
- become familiar with emulation, migration, and digital archaeology as preservation approaches
- developed an appreciation of the strengths and weaknesses of each approach

Session outline:

- *Lecture (30 minutes):* The Digital Preservation Problem and authenticity
- *Discussion/Activity (40 minutes):* A breakout session discussing the issues around the question: ‘what makes a preserved digital object authentic?’
- *Presentation of breakout group conclusions (20 minutes)*
- *Lecture (60 minutes):* Approaches to Digital Preservation – emulation, migration, digital archaeology
- *Discussion (30 minutes):* Advantages and disadvantages of each approach

Session

Managing ingest – issues relevant for creating an archive.

Lecturer and Session Leader:

Birte Christensen-Dalsgaard, State and University Library in Aarhus

Abstract:

The ingest phase involve many channelling problems, both in the selection process on what to include and in the inclusion phase, where the methodology needs to be decided. The nature of the problem is linked to the legal situation for the institution where in some cases (e.g. national archives) the institution can control the formats accepted whereas in other (e.g. deposit libraries) all material and all formats will have to be accepted.

The problem of actual collecting the material have different approaches and as a case the lecture will address the situation in Denmark, where the legal deposit law has been changed and from July 1 all material published on the net will be archived and all radio- and television broadcasts will be digitised and stored in a digital archive.

Objectives:

Lecture: There will be a one hour lecture around the problems addressed – both in their generalised form and in the more specific form related to the web archiving project in Denmark. The lecture will also address the negotiations between the national library and the Danish Broadcasting service on how to optimise the use of taxpayer's money and at the same time optimise the quality of the digital archive.

Discussion: Discussion around some of the issues raised.

Activity: The last part of the afternoon session will be around case studies, where participants will discuss the best approach for the negotiation phase and the choice of strategies for collecting and ingesting the material to the digital archive. The work will start in groups and end in a plenum discussion.

Session:

Active Ingest: Metadata Extraction and Creation, and Workflow

Lecturer & Session Leader:

Yunhyong Kim, Research Associate, HATII & *Seamus Ross*, Professor of Humanities Informatics and Digital Curation, and Director HATII (Humanities Advanced Technology and Information Institute).

Abstract:

The useful role of metadata (i.e. information describing a given piece of data) has been recognised by the digital library community in facilitating the management of digital materials. In particular, metadata helps to quickly catalogue objects for ingest, retrieval and appraisal without having to read, view or analyse the content. There is a continuing discussion on the appropriate standards of metadata which would provide sufficiently detailed and purposeful description of digital objects to make this process efficient by accurately identifying the content, structure, technical format, provenance and function of digital material. Even assuming such an appropriate standard can be established, the manual collection of metadata is becoming increasingly difficult as digital materials pour in at an exponential rate from different conventions and formats of publishing and organisations at all levels (from individuals to well-established publishing companies) via a number of varied media.

In this session we will discuss possibilities of automating the extraction/creation of metadata; the distinction between extraction and creation, what methods can be employed, what limitations might exist, how far automation can be taken and how this could fit in with the general ingest, preservation, and appraisal frameworks discussed in other sessions. The session will introduce some basic level of information technology, including information retrieval, language technology and image processing as a relevant component of preservation and suggest the first few steps necessary to place automatic metadata extraction in the work flow of a preservation model of digital libraries. It will attempt to describe how far we have come and what direction we might take to go further in automatic extraction and creation of metadata.

Aims and learning objectives:

This session aims to

- introduce the concept of automation in the area a metadata collections
- introduce possible methods of automating metadata extraction
- show how automation might fit in with the work flow of other established frameworks of digital preservation

After this session students should have:

- an understanding of the benefits and necessity of automatic metadata extraction/creation.
- grasped a sense of available methods which can be utilized for automatic collection of metadata
- developed an understanding of the difficulties involved in automatic metadata extraction
- the ability to visualize how to fit automatic metadata extraction into standard preservation processes

Session outline:

- Lecture -- 45 minutes -- Introduction to automation of metadata extraction and creation

- Exercise and discussion -- 30 minutes -- details of the exercise will be given at the event
- Lecture -- 45 minutes -- Integration into the preservation framework
- Activity and discussion -- 45 minutes -- details of the activity will be given at the event
- Summing up & questions -- 15 minutes

Session

Preserving for 2016, 2116, 3016. Or: “Is there a life for an object outside a digital library?”

Lecturer & Session Leader:

Manfred Thaller Professor for *Historisch-Kulturwissenschaftliche Informationsverarbeitung*

Abstract:

Almost all current planning for digital longevity assumes, that the institutional (as well as the technical) context for digital collections will not change fundamentally along time. Therefore most preservation planning takes a highly developed framework for preservation as granted. We will look successively at three target dates for preservation, check what a comparison with similar dates in the past lets us assume about the future and see, what follows for preservation policies, and how far these assumptions can be converted into technical recommendations.

For 2016 we can presumably expect a relatively unchanged situational context: It is somewhat improbable, that major parts of civilization break down until then and the major technical building blocks will not really have changed. (We will discuss why.) We also can assume, that current institutional frameworks will be either intact or their responsibilities will have been taken over by purpose built successors.

When we look further into the future we will try to estimate, how changes in these assumptions influence the viability of survival for digital material preserved today. For longer periods we will try to enhance this by separating what we preserve as far as possible from the conditions we have to postulate for our attempts to remain sustainable. Out of this we will derive the notion of an “autonomous digital object” which should be fit for long term survival.

Turning these themes into itemized questions:

- What are the current tacit assumptions about the future in preservation?
- How long can they be expected to survive?
- What will a failure of these assumptions mean for the preservation of digital material?
- What degrees of object autonomy can be expected?
- How to we work towards it?

Aims and learning outcomes of this session

This session aims to:

- Create an understanding for some of the consequences of recent developments in distributed systems for wide area repositories.
- Create an understanding for some of the areas, where current assumptions about the further development of digital technology may be too conservative.

- Provide them with a starting point for the planning of preservation policies which react to these changes.

By the end of this session students will:

- Be aware of some of the implications of "distributed technologies".
- Be able to react more flexible to the resulting challenges.
- Be able to think a bit further ahead.

Session outline

- Lecture (30 minutes): What are the current assumptions in preservation for 2016?
- Group work (30 minutes): What do participants think, could have gone wrong with these in 2116?
- Lecture (30 minutes): What does the lecturer think, could have gone wrong with the initial assumptions in 2116?
- Group work (30 minutes): What do participants think, could have gone wrong with the assumptions in 3016?
- Lecture (30 minutes): What does the lecturer think, could have gone wrong with the initial assumptions in 3016?
- Lecture (15 minutes): Do we have a chance to plan that far ahead?
- Final discussion and summary (15 minutes)

Session

Audit and Certification

Lecturer and Session Leader:

Andrew McHugh, Digital Curation Centre and HATII

Abstract:

It is widely acknowledged that the purpose of audit is to check that organisations, whether public or private, meet commonly agreed or expected standards or values, and that they are doing what they are supposed to do and performing their activities in an appropriate and acceptable manner. Audits provide insight into such factors as whether resources are used effectively and efficiently; they play a critical role in ensuring good governance and accountability. Certification should provide an independent mark as to how well the organization has done in meeting these standards.

Audit aims to provide a level of certainty and confidence that we can trust organizations with which we deal. That is why audit reports are mostly made public and organizations that fail audits frequently clearly identified. While audits can provide valuable external assessments of an organizations, many want to know how they perform against their mission, targets, and processes and procedures. An audit report can provide an overview of strong and weak points and of the effectiveness and efficiency of the organization. In the area of digital preservation audits can be used to assure the public that the information that is provided by information providers, including digital repositories, still is authentic, reliable, secure, and has been maintained to the highest standards. This is critical in a digital environment where information is no longer physical or tangible and for its survival very much dependent on the fast developing information technology.

Audits require clearly defined goals, standards and procedures against which processes, services, and products can be measured. These can either be defined externally or they can be defined internally by the organization being audited. Certification, if it follows from audit, usually requires that certain that the audit produce information which enables auditors to assess its quality and completeness and to take a view as to its accuracy. In some contexts audit may be a process that happens without warning, or may be something that is only initiated by the audited organization as a way of ensuring its readiness for external spot audits.

Depositing our digital assets with external repositories requires that we are able to trust the ability of the repository to secure our information over the long term. Audit and certification can be used to reduce the unknowns associated with our use of repositories whether as depositors or consumers of information held in the repositories. Depositors and users of resources may have legitimate concerns about repositories that do not subject themselves to audits or have failed to achieve certain levels of certification. Repositories may even decide that their business cases makes it possible for them to meet certain types of certification and not others. Return on investment may be a critical factor in deciding whether or not audit and certification are paths that repositories wish to pursue.

Aims and learning outcomes of this session:

This session aims to:

- Develop an participants understanding of why we use audit and certification mechanisms.
- Provide a perspective on what audits should achieve and what we should be auditing.
- Examine whether audit frameworks support different levels of compliance (or should they?)
- Create an appreciation of the steps involved in the audit process.
- Investigate the relationship between audit and certification.
- Define the consuming communities of audit and certification services in the econtent sector.
- Cover the ways to develop a framework that can serve as guidance in an organisation with respect to creating and managing or preserving digital entities.
- Demonstrate the objectives of certification.
- Study whether audit should be part of a process to achieve certification of a digital repository.
- Offer participants a chance to understand whether developing sustainable business models for digital repositories depends upon the establishment of a viable audit and certification environment
- Establish how audit and certification services might be provided, what can be automated and what must be done manually, and how often audits and certifications may be required.
- Sketch how organizations can prepared for audits and how they should respond to a failed audit.

Session outline:

- *Lecture:* Introduction to audit and certification services.
- *Discussion:* Focused on the RLG/OCLC study on *Attributes of a Trusted Digital Repository*, 2002, <http://www.rlg.org/longterm/repositories.pdf>
- *Breakout Session:* Defining approaches to audit and certification.

Biographical Sketches of the DELOS Summer School Team—

Birte Christensen-Dalsgaard is head of development at the State and University Library in Aarhus and is also responsible for the groups working on digital archives, the user oriented webservices and the usability work. Her work over the last five years, after she joined the Library sector coming from a computer centre working on educational material and research networks, has focussed on areas such as general library services, IT architecture for digital libraries and on preservation issues. In connection with the latter, the work on changing the Danish Legal deposit law has been quite dominating. She was in charge of the initial projects to demonstrate the relevant activities and relevant strategies in connection with archiving the web. She was member of a national committee working on recommendations to parliament on relevant measures to secure the electronic cultural heritage. Over the last year a group consisting of members from the two national libraries has worked to develop the administrative system to initiate webarchiving

Michael Day is a member of the research and development team at UKOLN, based at the University of Bath, United Kingdom. His research interests largely relate to metadata development, particularly as it relates to resource discovery, interoperability and management for long-term access and preservation. He also has a general interest in digital preservation and curation issues, for example, with regard to institutional repositories or resources made available through the World Wide Web. He has published widely in these fields. Since joining UKOLN, Michael Day has worked on a large number of metadata or preservation-related projects funded by the European Commission, the Joint Information Systems Committee, the Wellcome Trust, and others. Currently, most of his activities are based on the advisory services and research work of the UK Digital Curation Centre - of which UKOLN is a partner - and the preservation cluster of the DELOS Network of Excellence on Digital Libraries.

Wendy Duff is an associate professor at the University of Toronto, Faculty of Information Studies. She received her PhD from the University of Pittsburgh. While doing her doctoral work she was the project co-ordinator for the University of Pittsburgh Electronic Recordkeeping Project. She has served as a member of the ICA Adhoc Commission on Descriptive Standards, the Canadian-US Task Force on Archival Description (CUSTARD) Steering Committee, the Encoded Archival Description Working Group, the Encoded Archival Context Working Group, the Working Group on Subject Indexing and The Planning Committee on Descriptive Standards. She was chair of the Canadian Committee on Archival Description. Her primary research interests are user studies, archival description, and electronic records. Her current research focuses on the usability of a text analysis portal, and the development of generic user-based evaluation tools.

David Giaretta is the Associate Director for Development at the UK Digital Curation Centre. He is one of the co-authors of the OAIS Reference Model. His background is Astronomical research, in particular using data from astronomical satellites. He has led many archive and software projects and is currently based at the Rutherford Appleton Laboratory in the UK.

Mariella Guercio, ISTBAL, Università degli Studi di Urbino, is a full professor in archival

science and electronic record management at the University of Urbino where she entered in 1998. For twenty years (1978-1998) she worked as State Archivist for the Ministry of Cultural Heritage where she cooperated with the Authority for information technology in the public administration to define the Italian legislation related to the electronic recordkeeping systems. She chaired the ICA Committee on current records (1990-1992) and part of the Committee on program management of the ICA itself. She chaired the Italian team for the international project InterPARES 1 (1999-2001). She is a co-director of the European project ERPANET, a network for digital preservation. She runs many Italian projects funded by the Ministry for Research in the area of digital preservation. She takes part of the digital preservation cluster of the DELOS network of excellence. Since 2002, she is the director of the journal *Archivi & Computer*. She has published widely.

Ross Harvey is the inaugural Professor of Library and Information Management at Charles Sturt University's School of Information Studies. He has held academic positions at Curtin University of Technology and Monash University in Australia, Nanyang Technological University in Singapore, and the New Zealand Library School, and has been a Visiting Professor at the University of California Los Angeles. His current research and teaching interests include the preservation of library and archival material, especially in digital form. He has published widely in the fields of bibliographic organisation, library education, and the preservation of library and archival material, a recent book being *Preserving Digital Materials* (K.G. Saur, 2005). Ross Harvey was based at the National Library of Australia from April to June 2003 as a National Library Fellow, where he researched issues associated with the preservation of digital information. A full CV for Ross Harvey is available at <http://www.elibank.net>

Hans Hofman is co-director of ERPANET and senior advisor at the Nationaal Archief (National Archives) of the Netherlands. In his position at the Nationaal Archief he is involved in e-government projects and initiatives throughout Dutch government with respect to access and management of digital records and information in general. On the international scene he is co-investigator and representative of the Nationaal Archief in the Inter Pares 2 research project, and representative of the Netherlands in the ISO TC46/SC11 on Records Management. Within this committee he is chair of the Working Group on Records Management metadata. Finally, he is a researcher in the Delos 2 project on digital libraries with respect to digital preservation. He has given numerous presentations and written many articles on topics like digital preservation, recordkeeping metadata and electronic records management.

Yunhyong Kim is a Postdoctoral Researcher at the Digital Curation Centre based in the Department of Humanities Advanced Technology and Information Institute, University of Glasgow. She specialises in automatic extraction of metadata from digital objects to aid automated ingest and appraisal in digital repositories. She has a Ph.D. in Mathematics from University of Cambridge and a M.Sc. in Speech and Language Processing from the University of Edinburgh.

Andrew McHugh is responsible for leading the UK's Digital Curation Centre (DCC) audit and certification activities. After graduating from Glasgow University with Honours in Scots Law and a Masters in Information Technology he took a post as Resource Development Officer responsible for Music. He revolutionised the Department of Music's information infrastructure, performing a diverse range of roles, with responsibilities ranging from database and server administration to web programming, application

development and desktop cluster design and management. He joined the DCC project in late 2004, in the position of Advisory Services manager, coordinating a world-class team of digital curation practitioners in offering leading edge expertise and insight in a wide variety of issues to a mainly Higher and Further Education based audience. He has recently moved within the project, and performs the central role in the DCC's efforts to deliver a means of achieving effective trusted repositories for digital materials in the UK HE/FE communities and beyond. In addition to his DCC duties he also administers and teaches on the Department of Computing Science's MSc in IT Multimedia Systems course. Andrew's main interests include open source and web development.

Heike Neuroth holds a Ph.D in Geology and is the head of the department Research and Development at Göttingen State and University Library (SUB), Germany. She is engaged in several national and international working groups dealing with digital libraries. She is also the secretary of DINI (German Initiative for Network Information) which was originated from the Coalition of Network Information (CNI) in the USA.

Seamus Ross, Professor of Humanities Informatics and Digital Curation, and Director of Humanities Computing and Information Management at the University of Glasgow, runs HATII (Humanities Advanced Technology and Information Institute) (<http://www.hatii.arts.gla.ac.uk>) of which he is the founding director. Currently he is also Principal Director of ERPANET (Electronic Resource Preservation and Network) (IST-2001-32706) a European Commission activity to enhance the preservation of cultural heritage and scientific digital objects (<http://www.erpanet.org>). He is an Associate Director of the Digital Curation Centre in the UK (<http://www.dcc.ac.uk>) and a co-principal investigator in the DELOS Digital Libraries Network of Excellence. He is a lead partner in The Digital Culture Forum (DigiCULT Forum, IST-2001-34898), which works to improve the take-up of cutting edge research and technology by the cultural heritage sector (<http://www.digicult.info>). Before joining the University of Glasgow he was Head of ICT at the British Academy and a technologist at a company specialising in knowledge engineering. He earned a doctorate from the University of Oxford.

Andreas Rauber is Associate Professor at the Department of Software Technology and Interactive Systems at the Vienna University of Technology. He furthermore is head of the iSpaces research group at the Electronic Commerce Competence Center (EC3). He received his MSc and PhD in Computer Science from the Vienna University of Technology in 1997 and 2000, respectively. In 2001 he joined the National Research Council of Italy in Pisa as an ERCIM Research Fellow, followed by an ERCIM Research position at the French National Institute for Research in Computer Science and Control, at Rocquencourt, France, in 2002. He actively participates in research projects in the field of Digital Libraries, focusing on the organization and exploration of large information spaces, as well as Web archiving, and digital preservation. His research interests cover the broad scope of digital libraries, including specifically text and music information retrieval and organization, information visualization, and data analysis and neural computation. For further information on Andreas Rauber see <http://www.ifs.tuwien.ac.at/~andi>

Manfred Thaller holds a chair for "Historisch-Kulturwissenschaftliche Informationsverarbeitung" – roughly: Computer Science applied to Humanities and Cultural Heritage – at the University at Cologne. He has previously worked at the Max-Planck-Institute for History, Göttingen, Germany and is a former director of the Humanities Information Technology Program at the University of Bergen, Norway. He has been visiting

professor at the Hebrew University, Jerusalem, Queen Mary and Westfield College, London, and the European University Institute, Firenze. His research interests have been derived from a general definition of the requirements of Humanities' research in computing. This has in more recent years led to his involvement in a large number of digital library projects in the cultural heritage area. Manfred Thaller is also a member of the library and information systems committee of the German National Research Council (DFG). A full CV for Manfred Thaller is available at <http://www.hki.uni-koeln.de/people/thaller/mt.html>

Andrew Wilson currently Preservation Services and Projects Manager at the Arts and Humanities Data Service (AHDS), where I am involved in day-to-day operational work as well as participating in a number of preservation-related research projects. Before moving to the AHDS, I worked for the National Archives of Australia for about 10 years. A lot of my time there was spent developing and implementing a Dublin Core-based metadata element set, AGLS, now an Australian Standard: AS5044. I was also heavily involved in development and deployment of recordkeeping metadata and represented the Archives on a sub-committee of Standards Australia developing a recordkeeping metadata standard. I later managed a Research & Development project developing and implementing a long-term digital records preservation approach. I am currently a member of the Dublin Core Usage Board and Advisory Board, and chair of the DC-Agents Working Group. Educated as an Archivist, Andrew holds a doctorate in Ancient History.