

TU VIENNA

Digital Preservation

Andreas Rauber
 Department of Software Technology and
 Interactive Systems
 Vienna University of Technology
<http://www.ifs.tuwien.ac.at/~andi>

1 **I/S** FACULTY OF **INFORMATICS**

TU VIENNA

Overview

Part 1: Introduction

- What is Digital Preservation?
- What is the OAIS Reference model?
- How do we build a preservation plan?

2 **I/S** FACULTY OF **INFORMATICS**

TU VIENNA **Why do we need Digital Preservation?**

Questions / discussion:

- What is *Digital Preservation*?

3 **I/S** FACULTY OF **INFORMATICS**

TU VIENNA **Why do we need Digital Preservation?**

4 **I/S** FACULTY OF **INFORMATICS**

TU VIENNA **Why do we need Digital Preservation?**

5 **I/S** FACULTY OF **INFORMATICS**

TU VIENNA **Why do we need Digital Preservation?**

- Digital Objects require specific environment to be accessible :
 - Files need specific programs
 - Programs need specific operating systems (-versions)
 - Operating systems need specific hardware components
- SW/HW environment is not stable:
 - Files cannot be opened anymore
 - Embedded objects are no longer accessible/linked
 - Programs won't run
 - Information in digital form is lost (usually total loss, no degradation)
- Digital Preservation aims at maintaining digital objects authentically usable and accessible for long time periods.

6 **I/S** FACULTY OF **INFORMATICS**


 **Why do we need Digital Preservation**

- The goal of Digital Preservation is to **maintain digital objects accessible and usable in an authentic manner for a long term** into the future.

..... 7  FACULTY OF **INFORMATICS**

 **Why do we need Digital Preservation?**

- Essential for all digital objects
 - Office documents, accounting, emails, ...
 - Scientific datasets, sensor data, metadata, ...
 - Applications, simulations, business processes, ...
- All application domains
 - Cultural heritage data
 - eGovernment, public administration
 - Science / Research
 - Industry
 - Health, pharmaceutical industry
 - Aviation, control systems, construction, ...
 - Private data
 - ...


..... 8  FACULTY OF **INFORMATICS**

 **Why do we need Digital Preservation?**


Questions / discussion:


- What is *digital data*?
- What is *digital storage*?
- What do we mean by
 - *accessible*?
 - *authentic*?
 - *long-term*?

..... 9  FACULTY OF **INFORMATICS**


 **Why do we need Digital Preservation**


- 3 levels of threat / preservation
 1. Bit rot – physical preservation / bit preservation
 2. Object formats – logical preservation
 3. Authenticity, interpretability – semantic preservation

..... 10  FACULTY OF **INFORMATICS**

 **Bit-level preservation**


- Maintain bit-sequence
- Redundant storage:
 - Lockss: lots of copies keeps stuff safe
 - Cloud
- Distributed storage – physically separated
- Different technologies / platforms / production batches
- Controlled storage conditions
- Regular maintenance: type rewinding, disc spinning, ...
- Maintain devices for accessing storage!
- Trade-off capacity, energy, effort

..... 11  FACULTY OF **INFORMATICS**

 **Bit-level preservation**

Questions / discussion:

- How long do tapes / CDs / DVDs / HDDs / SSD last?
- What are the costs of bit-level preservation?
- What are the logistic challenges?
- Is a DVD that last for 200 years a solution?
- Distribution and Trust?
- Are we allowed to store redundantly?
 - Copyright
 - Copy protection

..... 12  FACULTY OF **INFORMATICS**

! TU VIENNA **Why do we need Digital Preservation**

- 3 levels of threat / preservation
 1. Bit rot – physical preservation / bit preservation
 2. Object formats – logical preservation
 3. Authenticity, interpretability – semantic preservation

..... 13 **I/S** FACULTY OF **INFORMATICS**

! TU VIENNA **Logical Preservation**

Deja vue:

- Digital Objects require specific environment to be accessible :
 - Files need specific programs
 - Programs need specific operating systems (-versions)
 - Operating systems need specific hardware components
- SW/HW environment is not stable:
 - Files cannot be opened anymore
 - Embedded objects are no longer accessible/linked
 - Programs won't run
 - Information in digital form is lost (usually total loss, no degradation)
- Digital Preservation aims at maintaining digital objects authentically usable and accessible for long time periods.

..... 14 **I/S** FACULTY OF **INFORMATICS**

! TU VIENNA **Logical Preservation**

- Identify objects at risk
- File format no longer supported
- HW / OS platform changes

..... 15 **I/S** FACULTY OF **INFORMATICS**

! TU VIENNA **Strategies for Logical Preservation**

Strategies
 (grouped according to Companion Document to UNESCO Charter validity of grouping is questionable - maintained only for structuring reasons)
<http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>

- Investment strategies:
 - Standardization, Data extraction, Encapsulation, Format limitations
- Short-term approaches:
 - Museum, Backwards-compatibility, Version-migration, Reengineering
- Medium- / long-term approaches:
 - Migration, Viewer, Emulation
- Alternative approaches:
 - Non-digital Approaches, Data-Archeology
- No single optimal solution for all objects

..... 16 **I/S** FACULTY OF **INFORMATICS**

! TU VIENNA **Standardisierung**


- Verwendung von offenen oder de-facto Standards
- + Vereinfacht Preservation Prozess
- + Zahlreiche Tools verfügbar
- + Tools können in Zukunft leichter erstellt werden
- Aufwand zur Standardisierung
- Verlust bei Konvertierung
- Manche Objekte nicht standardisierbar

..... 17 **I/S** FACULTY OF **INFORMATICS**


! TU VIENNA **Datenextraktion und Strukturierung**


- Schaffung abstrakter Repräsentation der Daten (z.B. Datenbanken in XML)
- + Unabhängig von spezifischer Infrastruktur
- + Zahlreiche Tools verfügbar
- + Tools können in Zukunft leichter erstellt werden
- Hoher Entwicklungsaufwands für Tools zur Abstraktion und Interpretation
- Interpretationstools limitieren Funktionalität
- Manche Objekte nicht abstrahierbar

..... 18 **I/S** FACULTY OF **INFORMATICS**


 **Encapsulation**

- Daten werden mit Metadaten und Software gekapselt („Zwiebelprinzip“)
- + Vereinfacht das Auffinden oder die Entwicklung einer Preservation-Lösung bei Bedarf
- + Guter Ansatz, der die Anwendung anderer Strategien jederzeit erlaubt
- Löst nicht die Problematik
- Selbst mit gekapselter Information kann die Schaffung einer Lösung unmöglich sein

..... 19  FACULTY OF **INFORMATICS**

 **Formatbeschränkung**

- Einschränkung der zur Preservation akzeptierten Dateiformate
- + Reduziert die Problemstellung auf überschaubare Anzahl von Formaten
- + Verfeinerung des Standardisierungsansatzes
- Löst nicht die Problematik
- Schränkt die Art von Material ein, die akzeptiert werden kann
- Konvertierung in Formate kann Datenverlust bedeuten
- Erfordert strikte Kontrolle der Formate

..... 20  FACULTY OF **INFORMATICS**

 **Universal Virtual Computer (UVC)**

- Generelle Virtuelle Maschine, Zwischenplattform, die auf jeder Plattform leicht implementiert werden kann
- Zum Preservation-Zeitpunkt wird logische Beschreibung der Daten plus Dekodierungsprogramm für UVC geschaffen
- + Kann sowohl für Dokumente als auch Software funktionieren
- + Einheitliche Zwischen-Plattform reduziert Entwicklungsaufwand
- + Entwickelte Programme bei Erstellung testbar
- Sehr komplex, noch in Entwicklung
- Entwicklungsaufwand zum Archivierungszeitpunkt
- Erfordert Kooperation von Produzenten für Software-Objekte
- Bei notwendiger Abstraktion der Daten Informationsverlust

..... 21  FACULTY OF **INFORMATICS**

 **Strategien zur Langzeitbewahrung**


- **Investment Strategien:**
Standardisierung, Datenextraktion, Encapsulation, Formatbeschränkung, UVC
- **Kurzfristige Ansätze:**
Museum, Rückwärtskompatibilität, Versionsmigration, Migration, Reengineering
- **Mittel-/langfristig Ansätze:**
(Migration), Viewer, Emulation, (UVC)
- **Alternative Ansätze:**
Nicht-digitale Ansätze, Daten-Archäologie

..... 22  FACULTY OF **INFORMATICS**


 **Museum - Technologiebewahrung**


- Bewahrung der Hardware (Laufwerke, Rechner,...)
- + Voller Funktionsumfang wird bewahrt
- + Gewinnt Zeit zur Entwicklung alternativer Strategien
- + Erforderliche Dokumentation der HW und SW erlaubt besseres Verständnis für Objekte
- + Einzige Strategie für manche Objekte
- Langzeitverfügbarkeit von Ersatzteilen nicht finanzierbar
- Erfordert umfangreiches „Museum“
- Erfordert umfangreiche technische Expertise

..... 23  FACULTY OF **INFORMATICS**


 **Rückwärtskompatibilität und Versionsmigration**


- Aktuelle Software liest ältere Formate und migriert
- + Meist verfügbar
- + Gewinnt Zeit bis zu umfangreicheren Transformationen
- + Oft gleiche (oder bessere) Funktionalität
- Langzeitverfügbarkeit unwahrscheinlich
- Versionsänderung kann unerwünschte Änderungen bewirken, unmittelbar oder über mehrere Migrationsschritte
- Nicht für alle Objekte anwendbar
- Keine Garantie von Seiten der Produzenten

..... 24  FACULTY OF **INFORMATICS**


 **Migration**

- Transformation in anderes Format, kontinuierlich oder on-demand (Viewer)
- + Verbreitet im Einsatz
- + Vergleich mit un-migriertem Objekt zum Zeitpunkt der Migration
- + Jederzeit zugreifbar
- Unerwünschte Änderungen unmittelbar oder über mehrere Migrationsschritte
- Nicht für alle Objekte anwendbar
- Kontinuierliche Migration erforderlich

..... 25  FACULTY OF **INFORMATICS**

 **Re-Engineering**


- Software Source wird auf neuer Plattform angepasst und kompiliert, Neuentwicklung, oder Reverse-Engineering
- + Anwendbar für Software-Objekte
- Source Code oft nicht verfügbar
- Portierung auf neue Umgebungen oft sehr schwer
- Erheblicher Zeit- und Ressourcenaufwand
- Reverse-Engineering oft illegal

..... 26  FACULTY OF **INFORMATICS**


 **Strategien zur Langzeitbewahrung**


- **Investment Strategien:**
Standardisierung, Datenextraktion, Encapsulation, Formatbeschränkung, UVC
- **Kurzfristige Ansätze:**
Museum, Rückwärtskompatibilität, Versionsmigration, Migration, Reengineering
- **Mittel-/langfristig Ansätze:**
(Migration), Viewer, Emulation, (UVC)
- **Alternative Ansätze:**
Nicht-digitale Ansätze, Daten-Archäologie

..... 27  FACULTY OF **INFORMATICS**


 **Viewer**

- Migration bei Bedarf, Interpretation durch Viewer-Software
- + Original-Datenstrom wird interpretiert
- + Keine kontinuierliche Migration
- + Keine kumulativen Fehler
- Viewer können nicht alle Teilobjekte darstellen
- Technologische Spanne bei Viewer-Erstellung
- Viewer müssen mit Technologie mitgezogen werden
- Schwer zu evaluieren, ob Viewer das Objekt korrekt wiedergeben

..... 28  FACULTY OF **INFORMATICS**

 **Emulation**


- Emulation von Hardware oder Software (Betriebssystem, Applikation)
- + Prinzipien weit verbreitet
- + Zahlreiche Emulatoren verfügbar
- + Potentiell vollständige Funktionalitätsbewahrung
- + *Dokument bleibt unverändert*
- *Dokument bleibt unverändert*
- Komplexe Technologie, Forschungsaufwand
- Erfordert penible Dokumentation des Systems
- Erfordert in Zukunft Erfahrung im Umgang mit derzeitigen Systemen
- Emulatoren müssen ebenfalls migriert/emuliert werden
- Emulatoren potentiell fehlerhaft (Komplexität)

..... 29  FACULTY OF **INFORMATICS**

 **Strategien zur Langzeitbewahrung**


- **Investment Strategien:**
Standardisierung, Datenextraktion, Encapsulation, Formatbeschränkung, UVC
- **Kurzfristige Ansätze:**
Museum, Rückwärtskompatibilität, Versionsmigration, Migration, Reengineering
- **Mittel-/langfristig Ansätze:**
(Migration), Viewer, Emulation, (UVC)
- **Alternative Ansätze:**
Nicht-digitale Ansätze, Daten-Archäologie

..... 30  FACULTY OF **INFORMATICS**

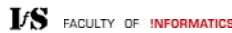
 **Nicht-digitale Strategien**


- „Ausdruck“ auf Papier, Mikrofilm, Nickel-Platten
- + Erfordert Transformation in lesbare Form – technologieunabhängig
- + Codierung für digitale Daten möglich
- + Erfahrung im Umgang mit analogen Objekten
- + Hohe Stabilität -> Bit-stream Preservation
- Funktionsverlust, Verlust der Vorteile digitaler Technologien
- Nicht für alle Objekte anwendbar
- Kosten für Erhaltung mancher analogen Trägermaterialien sehr hoch

..... 31 

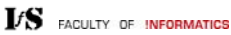
 **Data-Recovery, Archäologie**


- Analyse des bit-streams um Daten zu interpretieren
- + Potentiell einziger Ansatz um andernfalls verlorenes Material wieder zugreifbar zu machen
- Keine Garantie
- Ohne Dokumentation of nur „raten“
- Extrem hohe Kosten pro Objekt
- Nicht abschätzbar ob erfolgreich anwendbar auf bestimmtes Objekt

..... 32 

 **Logical Preservation**

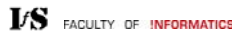
- Changing object, environment
- Loss upon migration / emulation
- Decision of what to preserve → **Significant Properties!**
- Range of strategies available, none is perfect
- Combination of strategies
- No solution forever -> DP is a process!


..... 33 

 **Logical Preservation**

Questions / Discussion:

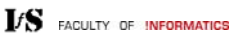
- What are the problems of logical preservation?
- What is the optimal strategy?
- What is the optimal strategy for a specific object?
- What is a good format / platform (e.g. to migrate to)?
- What are characteristics of good formats/platforms/... ?
- What is the difference between emulation and migration? Are they different? Are they not different?
- What are the significant properties of an object / process?
- “I want to preserve everything” – (how) can we do this?
- How can we find out what we loose with a strategy?


..... 34 

 **Logical Preservation**

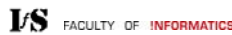
Questions / Discussion (2):


- What is the “original object”?
- Is XML the solution to DP?
- What is the complexity of each strategy? Costs? Effort?
- What know-how do we need to decide on a strategy?
- Which objects are most at risk?
- Which objects are most difficult to preserve?
- If we loose significant properties with a strategy, what is the impact on authenticity? Can we use a “changed” object?

..... 35 

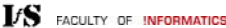
 **Why do we need Digital Preservation**


- 3 levels of threat / preservation
 1. Bit rot – physical preservation / bit preservation
 2. Object formats – logical preservation
 3. Authenticity, interpretability – semantic preservation

..... 36 

 **Semantic preservation**

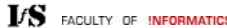
- Threats at semantic level
 - meaning of terms change: city names, ...
 - measurement scales, sensor sensitivity, ...change
 - interpretation of facts change: alcohol levels, ...
- Rather long-term, but subtle to notice
- Consider context of objects
 - purpose, setting, limitations, cultural context, related objects, ...
- Approaches / solutions:
 - semantic enrichment
 - metadata
 - migration at semantic level
 - documentation of context
 - tracing of metadata


..... 37 

 **Semantic preservation**

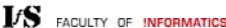
Questions / discussion:


- How do we identify need for action?
- What is the risk of missing timely action?
- How do we solidly identify and document context?
- How can we implement semantic enrichment / semantic migration, ...?
- Are differences in the communication protocol at an API level a problem of logical or semantic preservation?

..... 38 


 **Digital Preservation - Summary**


- Is a complex task
- Requires a concise understanding of the objects, their intellectual characteristics, the way they were created and used and how they will most likely be used in the future
- Requires a continuous commitment to preserve objects to avoid the „digital dark hole“
- Requires a solid, trusted infrastructure and workflows to ensure digital objects are not lost
- Is essential to maintain electronic publications & data accessible
- Will become more complex as digital objects become more complex
- Needs to be defined in a preservation plan

..... 39 

 **Digital Preservation**

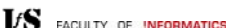
- Reference Models
 - Records Management, ISO 15489:2000
 - OAIS: Open Archival Information System, ISO 14721:2003
- Audit & Certification Initiatives
 - RLG- National Archives and Records Administration Digital Repository Certification Task Force: Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)
 - NESTOR: Catalogue of Criteria of Trusted Digital Repositories
 - DCC/DPE: DRAMBORA: Digital Repository Audit Method Based on Risk Assessment


..... 40 

 **Digital Preservation**

Questions / Discussion:

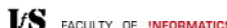
- At what levels are digital objects threatened?
- What are the time intervals at each level?
- How can we identify objects at risk?
- What can we do to mitigate the risk?
- How can we recover if mitigation fails / is missed?
- How do we organize DP for an organization?
- What competences do we need?
- How would a training/education program look like?
- How do we know if somebody is doing a good job at DP?

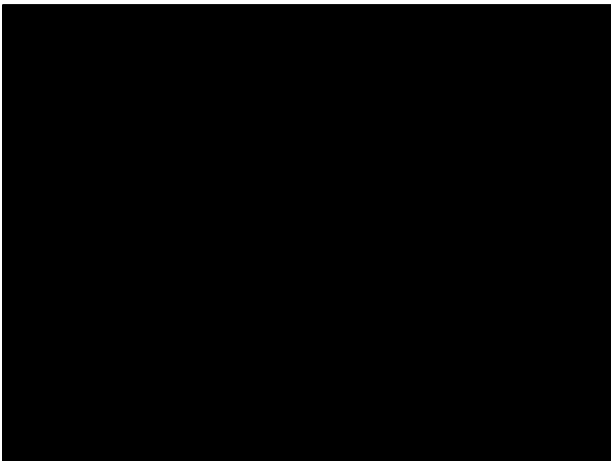
..... 41 


 **Overview**

Part 1: Introduction

- What is Digital Preservation?
- **Break? - Video?**
- What is the OAIS Reference model?
- How do we build a preservation plan?


..... 42 




 **Overview**


Part 1: Introduction


- What is Digital Preservation?
- What is the OAIS Reference model?
- How do we build a preservation plan?

..... 44 

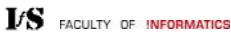
 **OAIS**


- NASA: National Space Science Data Center
 - NASA's first digital archive
 - Experienced many technological changes since 1966
- Consultative Committee for Space Data Systems
 - International group of space agencies
 - Developed range of discipline-independent standards
 - Evolved into ISO TC 20/ SC 13 working group around 1990
 - TC20: Aircraft and Space Vehicles
 - SC13: Space Data and Information Transfer Systems

..... 45 


 **OAIS**


- Reference Model for an Open Archival Information System (OAIS), Blue Book, CCSDS 650.0-B-1, January 2002
- ISO 14721:2003
- slides based on Blue Book and:
 - Don Sawyer, Lou Reich: ISO Reference Model for an Open Archival Information System (OAIS) Tutorial Presentation, LOC, June 13 2003
- <http://ssdoo.gsfc.nasa.gov/nost/isoas/overview.html>

..... 46 


 **OAIS**

- Framework for understanding and applying concepts needed for long-term digital information preservation
 - Long-term: long enough to be concerned about changing technologies
 - Starting point for model addressing non-digital information
- Provides set of minimal responsibilities to distinguish an OAIS from other uses of 'archive'^{alcohol levels}
- Framework for comparing architectures and operations of existing and future archives
- Addresses a full range of archival functions
- Applicable to all long-term archives and those organizations and individuals dealing with information that may need long-term preservation
- Does NOT specify an implementation

..... 47 

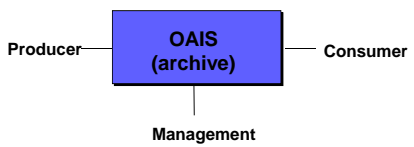
 **OAIS**

- OAIS helps understanding / structuring DP
- Is not "perfect"
 - Conflicting models, different views
- Does NOT specify an implementation model !!!
- Difficult balance between high-level structure and detailed guidelines, not consistently solved
- Has to be understood wrt. its time of origin and purpose
- Standards create their own dynamics

..... 48 

! TU VIENNA

OAIS



Producer — **OAIS (archive)** — Consumer
 |
 Management

- Producer is the role played by those persons, or client systems, who provide the information to be preserved
- Management is the role played by those who set overall OAIS policy as one component in a broader policy domain
- Consumer is the role played by those persons, or client systems, who interact with OAIS services to find and acquire preserved information of interest

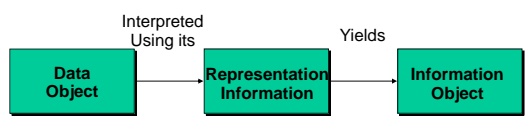
49 **I/S** FACULTY OF **INFORMATICS**

! TU VIENNA

OAIS

OAIS Information Definition

- Information is always expressed (i.e., represented) by some type of data
- Data interpreted using its Representation Information yields Information
- Information Object preservation requires clear identification and understanding of the Data Object and its associated Representation Information

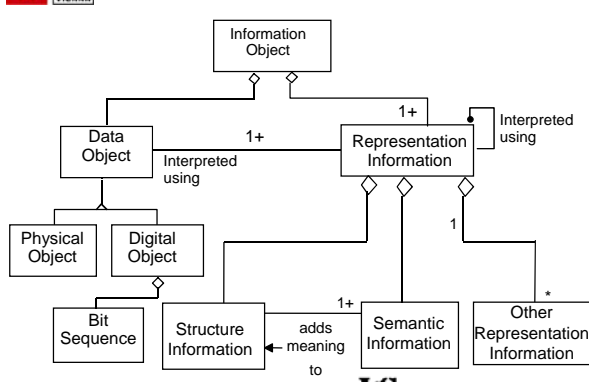


Data Object — Interpreted Using its —> Representation Information — Yields —> Information Object

50 **I/S** FACULTY OF **INFORMATICS**

! TU VIENNA

OAIS



Information Object
 |
 Data Object (1+) — Interpreted using —> Representation Information (1+)
 |
 Physical Object | Digital Object
 |
 Bit Sequence | Structure Information
 |
 Semantic Information (1+) — adds meaning to —> Structure Information
 |
 Other Representation Information (*)

51 **I/S** FACULTY OF **INFORMATICS**

! TU VIENNA

OAIS

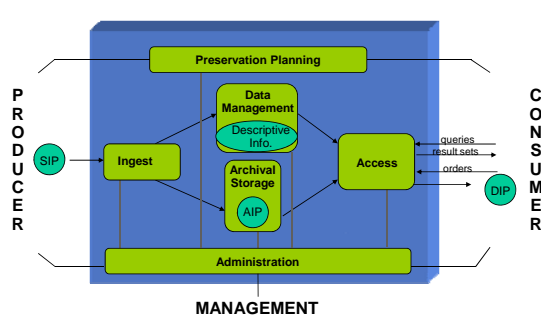
Information Package Variants

- **SIP**: Submission Information Package
 - Negotiated between Producer and OAIS
 - Sent to OAIS by a Producer
- **AIP**: Archival Information Package
 - Information Package used for preservation
 - Includes complete set of Preservation Description Information (PDI) for the Content Information
- **DIP**: Dissemination Information Package
 - Includes part or all of one or more Archival Information Packages
 - Sent to a Consumer by the OAIS

52 **I/S** FACULTY OF **INFORMATICS**

! TU VIENNA

OAIS



PRO D U C E R —> **OAIS (archive)** —> C O N S U M E R
 |
 M A N A G E M E N T

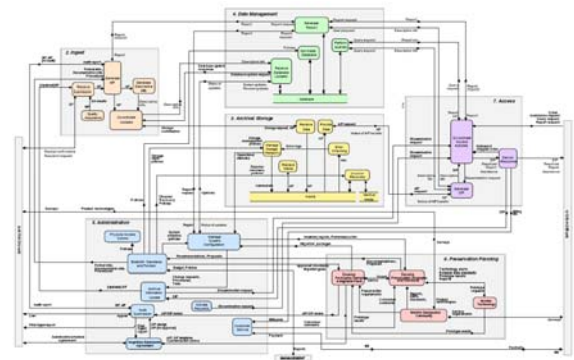
Preservation Planning
 Ingest —> Data Management (Descriptive Info, Archival Storage) —> Access
 Administration

SIP = Submission Information Package
 AIP = Archival Information Package
 DIP = Dissemination Information Package

53 **I/S** FACULTY OF **INFORMATICS**

! TU VIENNA

OAIS



54 **I/S** FACULTY OF **INFORMATICS**

TU VIENNA **Overview**

Part 1: Introduction

- What is Digital Preservation?
- What is the OAIS Reference model?
- How do we build a preservation plan?

55 **I/S** FACULTY OF **INFORMATICS**

TU VIENNA **Preservation Planning**

Why Preservation Planning?

- Several preservation strategies developed
 - For each strategy: several tools available
 - For each tool: several parameter settings available
- How do you know which one is most suitable?
- What are the needs of your users? Now? In the future?
- Which aspects of an object do you want to preserve?
- What are the requirements?
- How to prove in 10, 20, 50, 100 years, that the decision was correct / acceptable at the time it was made?

56 **I/S** FACULTY OF **INFORMATICS**

TU VIENNA **Preservation Planning**

What is Preservation Planning?

- Consistent workflow leading to a preservation plan
- Analyses, which solution to adopt
- Considers
 - preservation policies
 - legal obligations
 - organisational and technical constraints
 - user requirements and preservation goals
- Describes the
 - preservation context
 - evaluated preservation strategies
 - resulting decision including the reasoning
- Repeatable, solid evidence

57 **I/S** FACULTY OF **INFORMATICS**

TU VIENNA **Digital Preservation**

What is a preservation plan?

- 10 Sections
 - Identification
 - Status
 - Description of Institutional Setting
 - Description of Collection
 - Requirements for Preservation
 - Evidence for Preservation Strategy
 - Cost
 - Trigger for Re-evaluation
 - Roles and Responsibilities
 - Preservation Action Plan

[Preservation Plan Template](#)

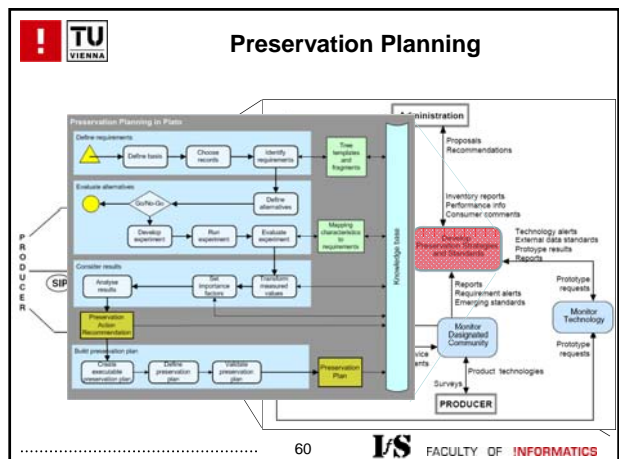
58 **I/S** FACULTY OF **INFORMATICS**

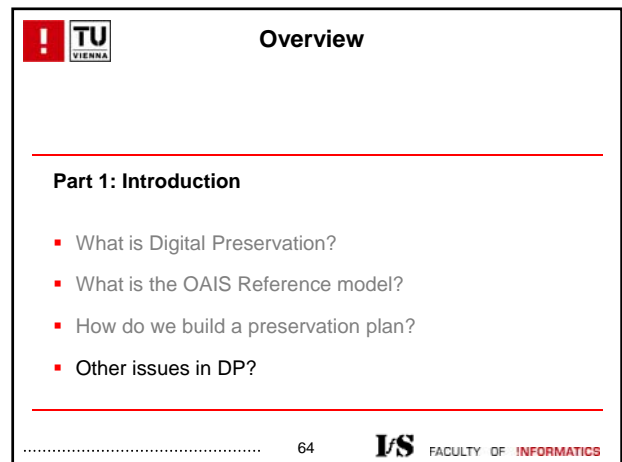
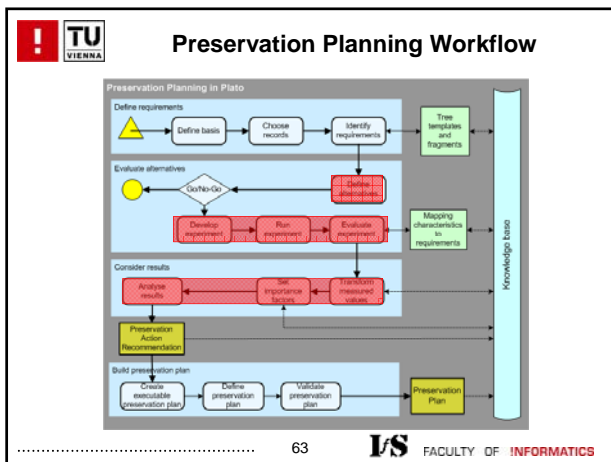
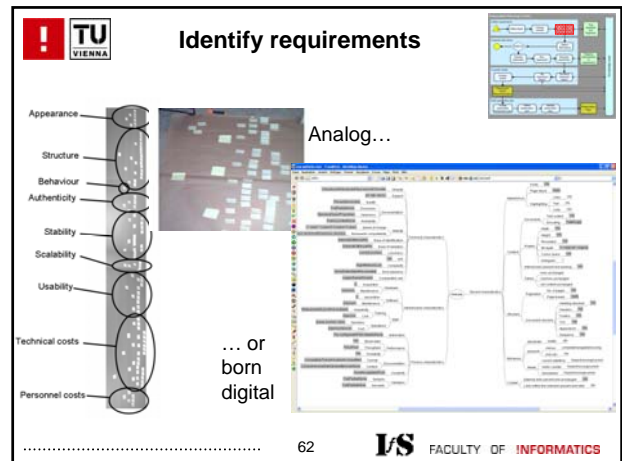
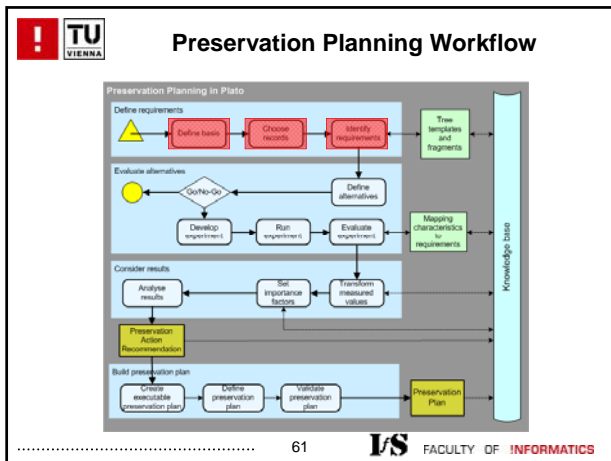
TU VIENNA **Preservation Planning**

Preservation Planning Workflow

- Originally developed within the DELOS DP Cluster now refined and integrated within PLANETS
- Based on
 - Preservation Planning approach based on Utility Analysis, developed at TU Vienna
 - Testbed/lab for evaluation developed at Nationalarchief, The Netherlands
- Follows the OAIS model
- Consistent with requirements specified by ORLC/TRAC and Nestor criteria catalogue

59 **I/S** FACULTY OF **INFORMATICS**





- ### Current Issues
- Personal & SOHO Archiving
 - Web Archiving: DP, IR & Ethics
 - Interactive Content
 - Emulation of Computer Games, Multimedia Art
 - Context of objects
 - Long-term storage: holography, solid-state,...
 - Database Preservation
 - Security: Signatures, encryption, audit, certification
 - Atomic file formats, stability of file formats
 - Scalability, Semantics
 - Digital forgetting
 - Preservation of programs, processes
- 65 I/S FACULTY OF INFORMATICS

Thank you!

<http://www.ifs.tuwien.ac.at/dp>

66 I/S FACULTY OF INFORMATICS