Elektronische Langzeitarchivierung: Probleme und Lösungsansätze

Lothar Schmitz

Digital documents last forever -
or five years, whichever comes first.

Jeff Rothenberg: Avoiding Technological Quicksand,1999

Inhalt

1. Einführung: Probleme
2. Organisatorisches Umfeld
3. Lösungsansätze im Überblick
4. Emulation vertieft
5. Perspektive

Quellen

1. Einführung: Probleme

Für Langzeitarchivierung verantwortlich sind **Bibliotheken** und **Archive**.

für Wissenschaft

=> komfortabler Zugriff

Gedächtnis & Beweismittel des Trägers

=> große Archive:

- NASA Satellitenbilder
- NARA Verwaltungsnotizen (email)
- Australiens Ölvorkommen
- Wetter- / Versicherungsdaten

Erfahrungs-austausch

Elektronische Archivierung: Die Lösung ...

- Elektronische Kataloge und Ausleihsysteme (OPAC)
- Schriftgut entsteht meist digital ***(„born digital“)***
- Multimediale Dokumente auf CD mit ISBN
- Digitale ***Konservierung*** alter Bestände

Vorzüge der digitalen Form:

- ✓ *Perfekte, kompakte Kopien*
- ✓ *Zugriff per Internet weltweit*
- ✓ *Volltextsuche*
- ✓ *Weiterverarbeitung*

... oder das Problem?

- geringe Lebensdauer digitaler *Medien*
- rasch veraltende *Gerätetypen*
- *Software:* noch größere Vielfalt und Neuerungsrate

☞ ***Negativfolgen von Wettbewerb und Fortschritt***

Inhalt der Dokumente
den menschlichen Sinnen nicht zugänglich!
(vgl. Höhlenmalerei, Hieroglyphen, ...)

☞ ***Kulturzeugnisse proaktiv retten, sonst Verlust!***

Beispiele (aus RLG Task Force Bericht 1996)

1. **US Volkszählungsdaten von 1960** (Census Bureau) auf UNIVAC type II-A Bändern. Restauration 1976-79. Verlust ca. 10000 Datensätze (von ca 1,5 Mio).
2. **Erste email** (Tomlinson, 1974, in Boston) verloren.
3. **Satellitenbilder des brasilianischen Regenwalds `70** (wichtig für Klimaforschung) nicht mehr brauchbar.
4. **Land Use and Natural Resources Project** (Cornell und NY, Ende der 60er Jahre), jetzt NY Staatsarchiv. Digitale Daten unbrauchbar => bei Bedarf Hardcopies neu erfassen.

***UK Public Record Office** bearbeitet laut Gesetz 50 Jahre alte Regierungsdaten => Was passiert 2015?*

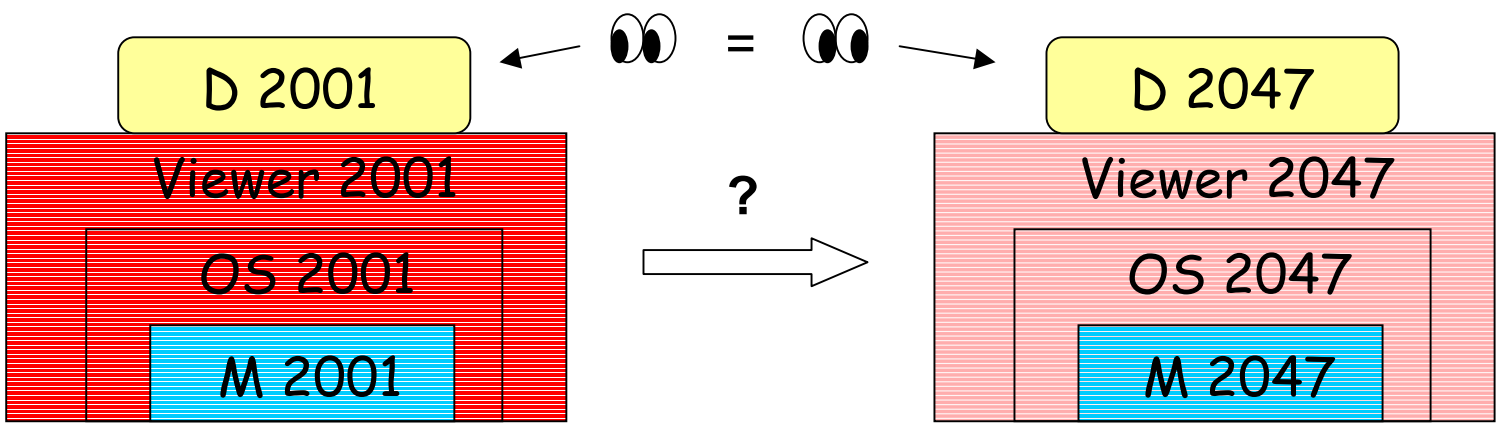
(Michael Lesk, Bellcore, 1995)

Das technische Langzeitarchivierungsproblem

besteht aus zwei Teilproblemen:

Im Prinzip gelöst !

1. **Erhalt des digitalen Dokuments,**
   d.h. des (vom Datenträger unabhängigen) ***Bitstroms***.
2. **Erhalt des Zugangs zum digitalen Dokument D,**
   d.h. einer ***„Abspielvorrichtung“***



Task Force Definition des Problems

„Migration is the periodic transfer of digital materials from one hardware / software configuration to another, or from one generation of computer technology to a subsequent generation.

The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display and otherwise use them in the face of constantly changing technology.“

Nicht-technische Probleme

- **Juristische Fragen:**
  Rechte am Dokument?
  (Änderungen, Weitergabe?)
  Gesetzliche Ablieferverpflichtung?
  Rechte an Abspielkomponenten?

- **Gesellschaftliche Fragen:**
  Welche Dokumente sollen erhalten werden?
  (Kriterien?)
  Wer trägt die Kosten?
  Für die ***sichere und unterbrechungsfreie Erhaltung*** zuständige Institution?

Internationale Anstrengungen

- **R**esearch **L**ibraries **G**roup (USA)
  Task Force Report: Preserving Digital Information, 1996
- **CCSDS** (NASA, DLR, ...) Reference Model for
  Open **A**rchival **I**nformation **S**ystems (**OAIS**), 1999
- **C**onsortium of **U**niversity **R**esearch **L**ibraries (UK)
  CURL **E**xemplars for **D**igital **AR**chive**S** (CEDARS)
- **N**etworked **E**uropean **D**eposit **Lib**rary (NEDLib)
  OAIS => **DSEP** und Emulation ***(RAND, IBM Almaden)***
- **C**ouncil on **L**ibrary and **I**nformation **R**esources (USA)
  Migrationsstudie & Workbook (***Cornell University)***
- ...

2. Organisatorisches Umfeld

Allgemein akzeptierter Standard:

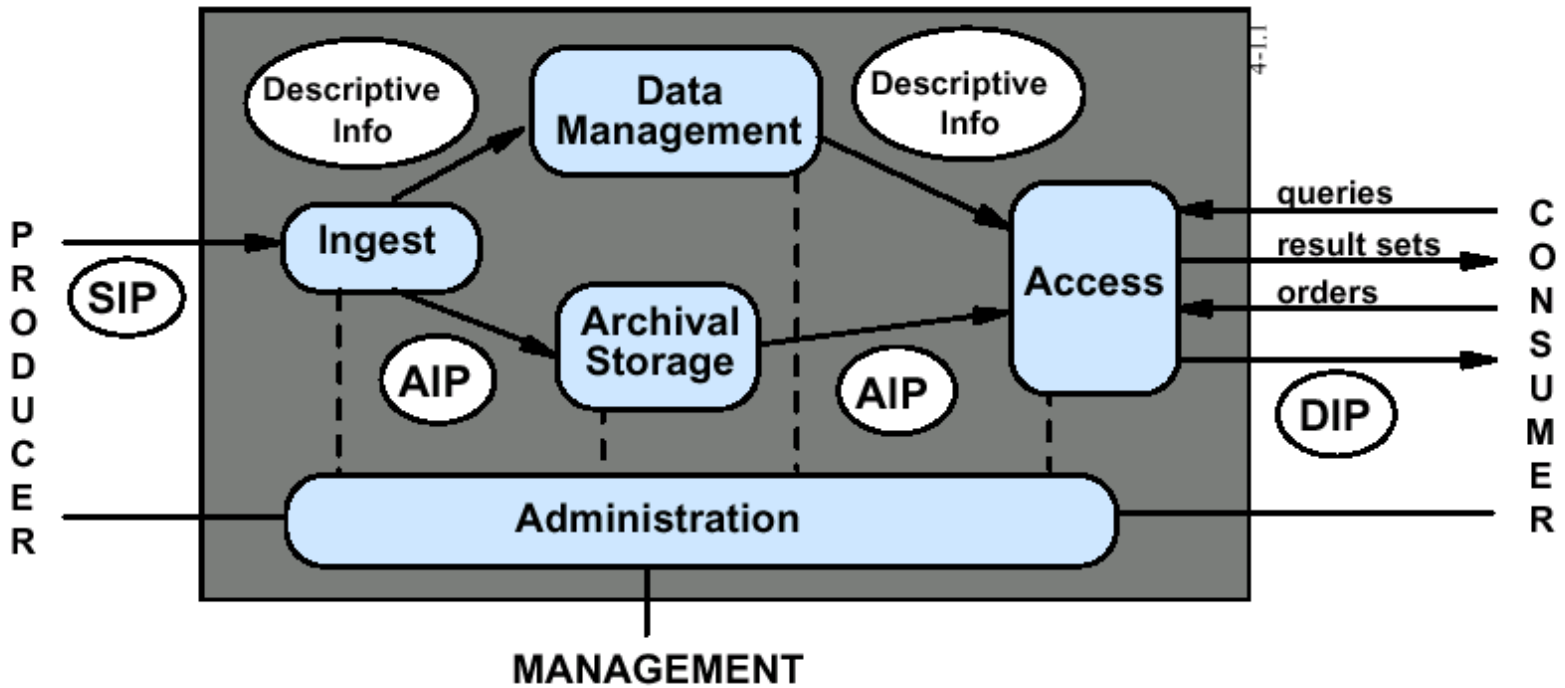
OAIS Reference Model

- ***Prozessmodell***
  „Functional Model“
- ***Datenmodell***
  „Information Model“

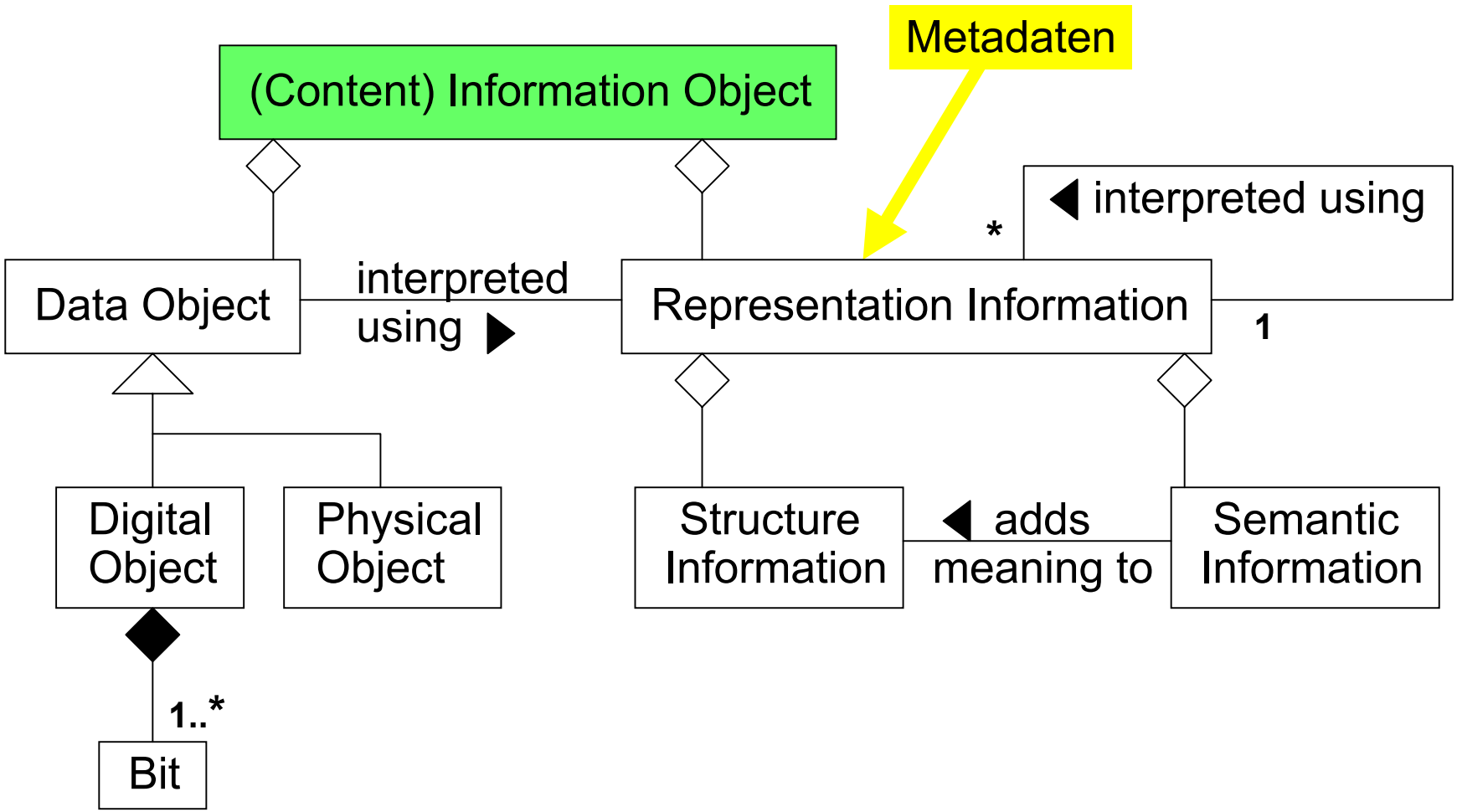
NEDLib
DSEP Modell

- für ***Bibliotheken***
- spezialisierte Prozesse

Das „OAIS Functional Model“



Information Model des OAIS (I) - Content



Information Model des OAIS (II) - AIP

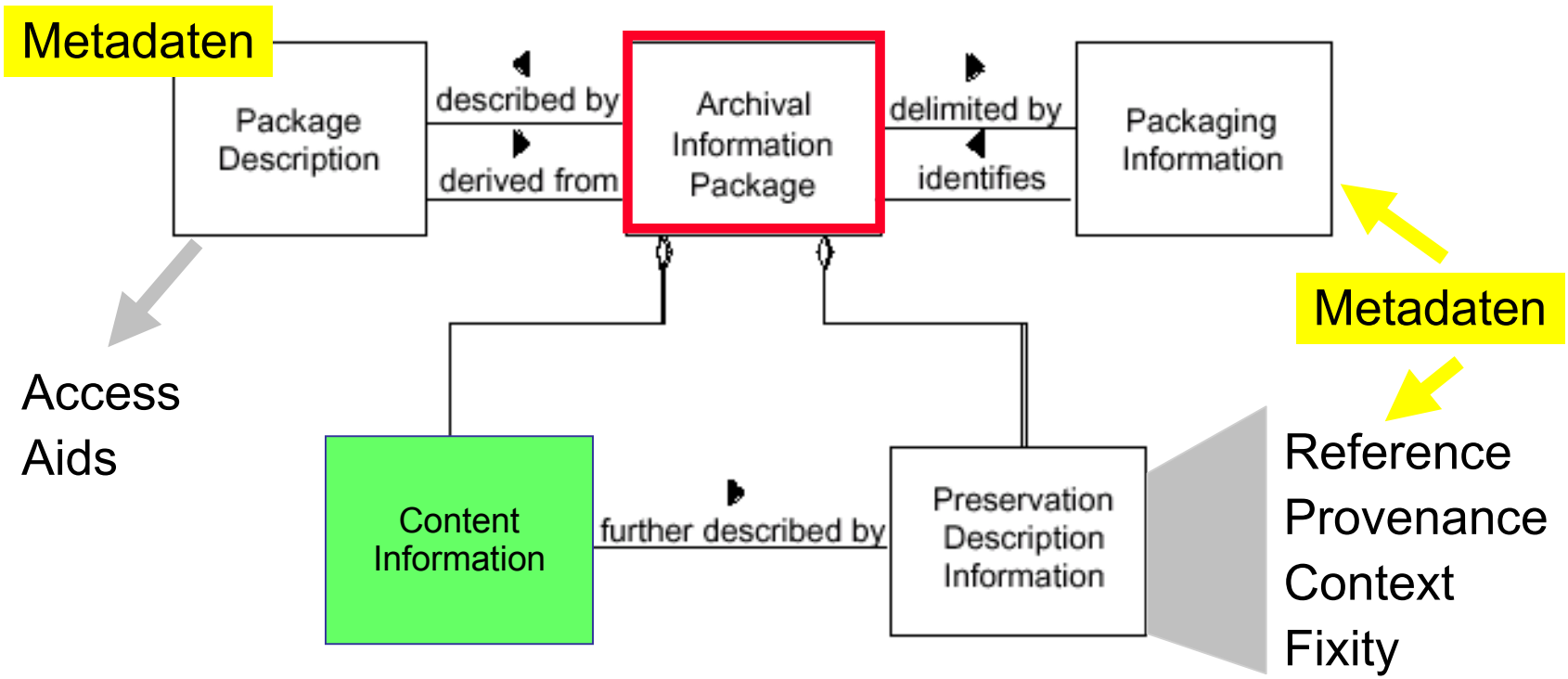


Figure 4-16: Archival Information Package (AIP)

Arbeitsvoraussetzungen beim Erhalt

- ***stetig wachsende Datenvolumina***
- kostengünstig => ***durch automatisierte Prozesse***
- ***langfristig*** (auch > 100 Jahre) => unterbrechungsfrei
- ***zugänglich*** *(evtl. weiterverarbeitbar)* ***erhalten***
- ***bei größtmöglicher Authentizität***
- ***ohne Rückgriff auf Autor*** und dessen Kontextwissen

jedenfalls keine „heroischen Maßnahmen“ !!

3. Lösungsansätze im Überblick

Notlösung:

*Ausweichen auf **nicht-digitale Formate***
wie Papier, Mikrofiche, Metallfolie, Lochstreifen, ...

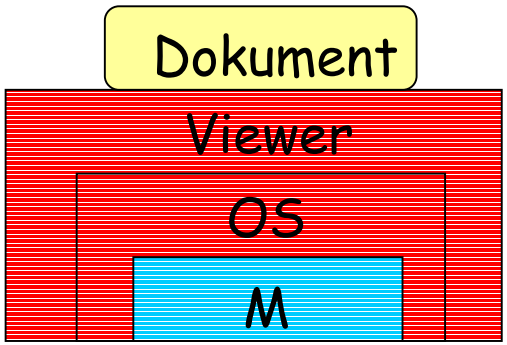
\+ dauerhafter

\+ teils dem Auge direkt zugänglich

\- Vorteile digitaler Dokumente gehen verloren

\- nur einfache Text- und Bilddokumente möglich, keine Multimedia, Links, Spreadsheets, CAD, ...

Teilproblem „Erhalt des Zugangs“



Drei Basisstrategien:

1) **Dokument an neue Umgebung anpassen**
   => ***Transformation*** *des Dokuments*
2) **Umgebung für Originaldokument erhalten**
   => *anstelle des Dokuments die Umgebung anpassen*
3) **Umgebungsunabhängigen Inhalt aufbewahren**
   => *Form und damit Abspielumgebung unwichtig*

Varianten der Basisstrategien

1) **Dokument an neue Umgebung anpassen**
   1.1) reversible **Transformationen**
   1.2) irreversible **Transformationen**
2) **Umgebung für Originaldokument erhalten**
   2.1) Emulation
   2.2) Portierung von Viewern
   2.3) Museumsansatz
   2.4) Digitale Archäologie („Rosetta Stone“)
3) **Umgebungsunabhängigen Inhalt aufbewahren**
   z.B. in selbstbeschreibenden, erweiterbaren XML-Formaten

1.1) reversible Transformationen

mit unterschiedlicher Zielsetzung:

a) **Wechsel der Codierung:**
   z.B. *ASCII* <=> *Unicode-Subset*

b) **OS-spezifische Textdateiformate:**
   z.B. *Unix* <=> *Windows* <=> *Mac*

c) **Datenkompression:**
   z.B. *zip / unzip*

d) **Verschlüsselung:**
   z.B. *PGP*

=> kein Informationsverlust wenn ...

1.2) irreversible Transformationen

„Migration im engeren Sinn“

z.B. Formate von Gleitpunktzahlen

a) Direkte Konversion:

- **aufwärtskompatible Programmversionen**
  z.B. *verschiedene Word-Versionen*
- **Import-Filter oder Konversions-Tools**
  z.B. *WordPerfect -> Word*

=> *typisch zwischen proprietären Formaten*

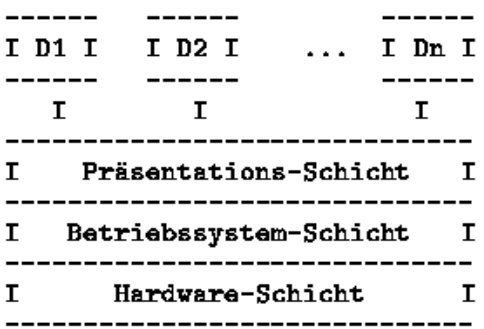
b) Konversion über Zwischenformat:

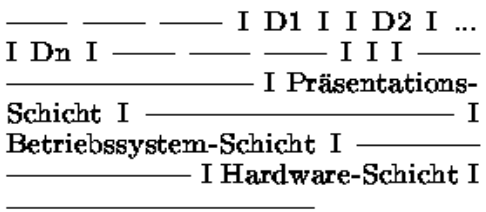
z.B. über ASCII zwischen Word und *TeX*

=> *Gefahr von Informationsverlusten !*

Implizite Information als Migrationshindernis

- ASCII-Textgrafik => *LaTeX*





- *TeX*-Formel nach MathML: $f(b+c)$

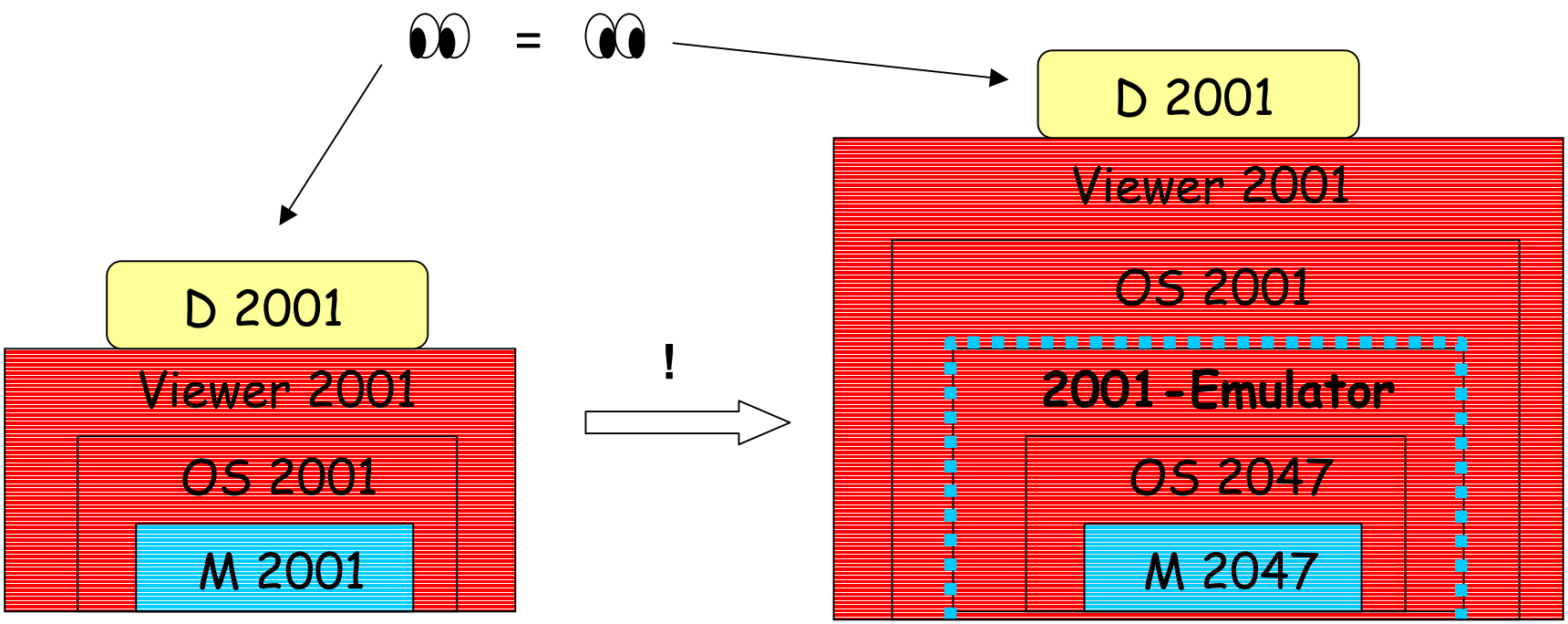
Funktionsanwendung oder **Produkt** ?

- Zeitung: `Fortsetzung nächste Seite, Spalte 3 unten.`

2.1) Emulation

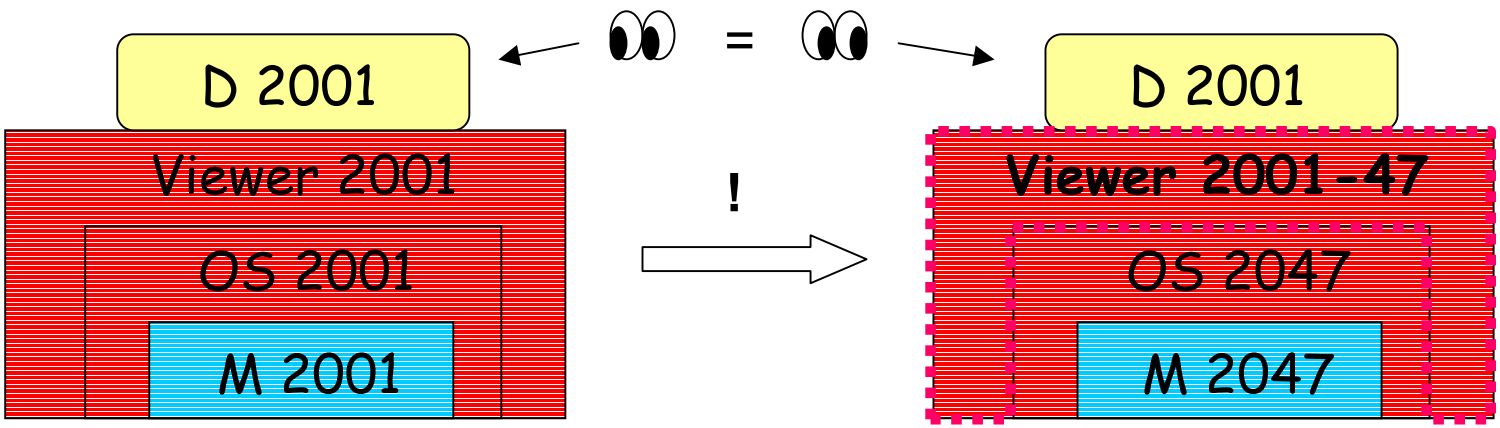
Idee: 1. Alten Computer auf neuem emulieren

2. Darauf Originalsoftware und -dokument laden



2.2) Portierung von Viewern

Idee: 1. Viewer-Software auf neuen Computer portieren
2. Darauf Originaldokument laden



- Quelle des Viewers und 2047er Compiler benötigt.
- Für alle Viewer durchzuführen.

Gegenüberstellung von Strategien

	Migration	**Emulation**
Authentizität	gefährdet	hoch
Erhaltungs-aufwand	je Typ: hoch je Dokument: hoch	je Plattform: hoch je Dokument: wenig
Technisches Risiko	nur kurzfristig abschätzbar	derzeit hoch anzusehen
geeignet für Langzeitarchivierung	***noch nicht erwiesen***	***noch nicht erwiesen***

In jedem Fall nützlich:

- **Standardisierung:**
  Beschränkung auf wenige Plattformen und Formate (z.B. ASCII, PDF, TIFF, SVG, SMIL, ...)
  => Reduzierung des Aufwands
- **Erweiterbare Formate:**
  Kein Stillstand der Entwicklung abzusehen
  => Stabilität nur durch Änderbarkeit
- **Offene, nicht proprietäre Formate (z.B TeX):**
  Rechte, Formate und Quellen als Gemeingut
  => Basis für Gemeinschaftsentwicklung, nicht durch kommerzielle Interessen eingeschränkt

4. Emulation vertieft

Emulation

= „Migration der Abspielumgebung“

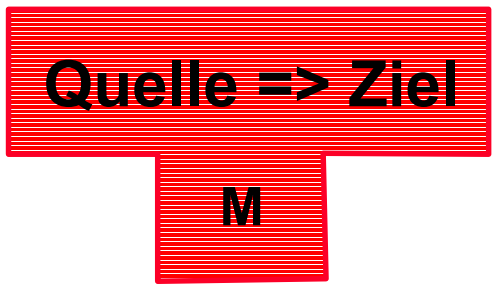
⊆ Migration von Software

Portierung

Standardtechniken
- Compiler
- Virtuelle Maschinen

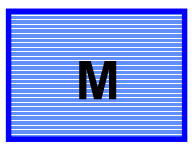
Diagrammbausteine nach N. Wirth

- **T-Diagramm**
  (**T**ranslator, ***Compiler***)



- **I-Diagramm**
  (**I**nterpreter,
  auch ***Viewer,***
  Emulator)

- **M-Diagramm**
  (**M**aschine)

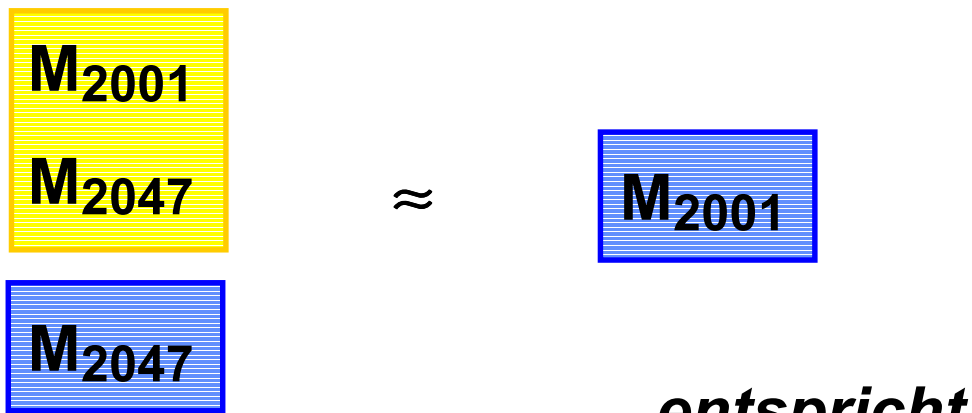


ähnlich wie Dominosteine aneinanderzulegen !

Verwendung eines Emulators

M_{2001}-Emulator

auf M_{2047} …

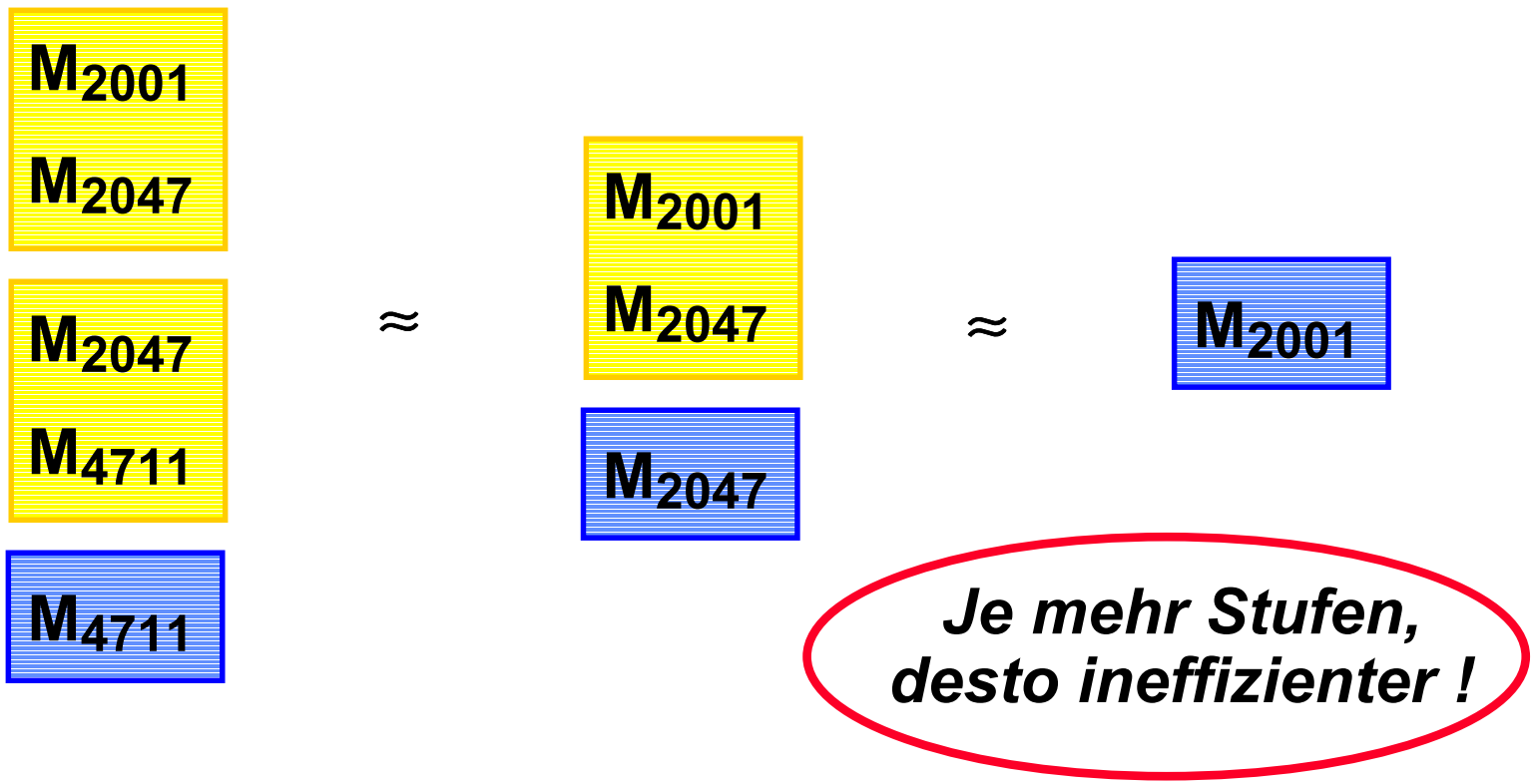


… entspricht

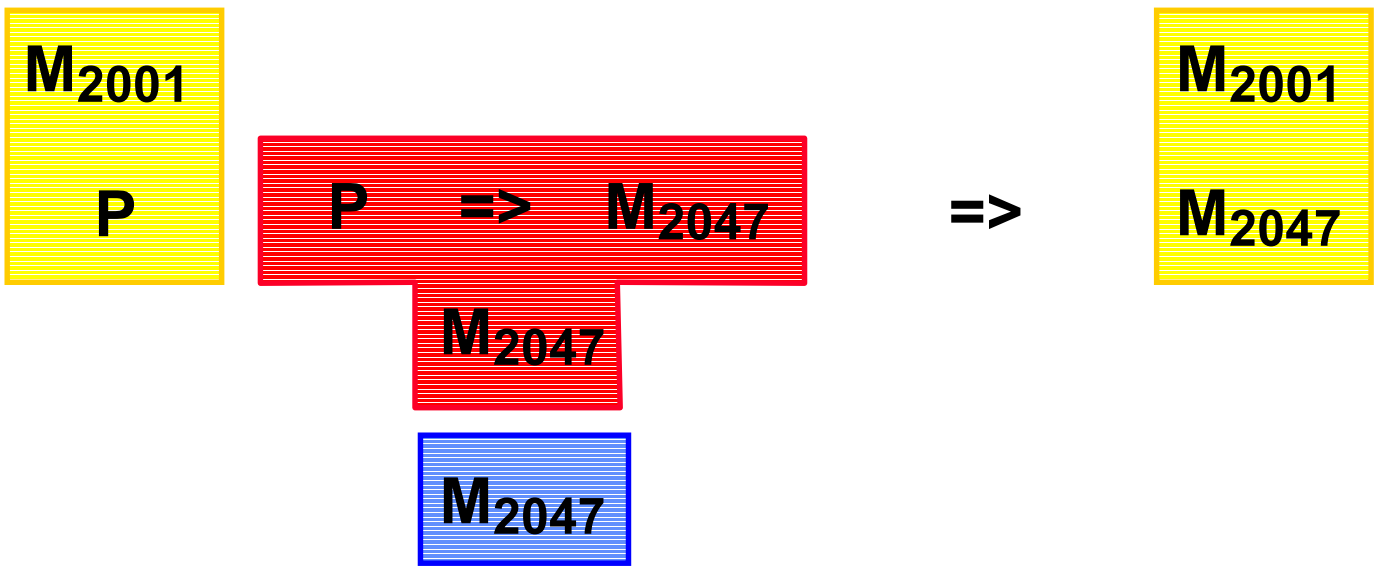
M_{2001}

Mehrstufige („layered“) Emulation

hier **zweistufig:**



Erzeugung / Portierung eines Emulators

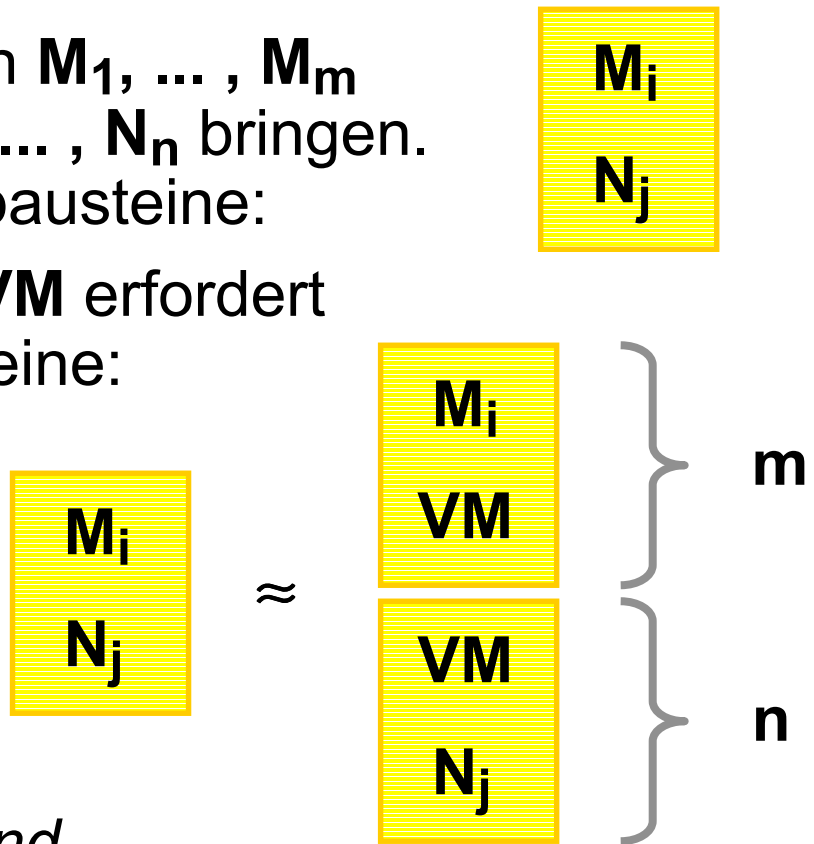


mit höherer Programmiersprache P und Compiler

(hier für M_{2047} ; genauso für M_{4711} , sofern ...)

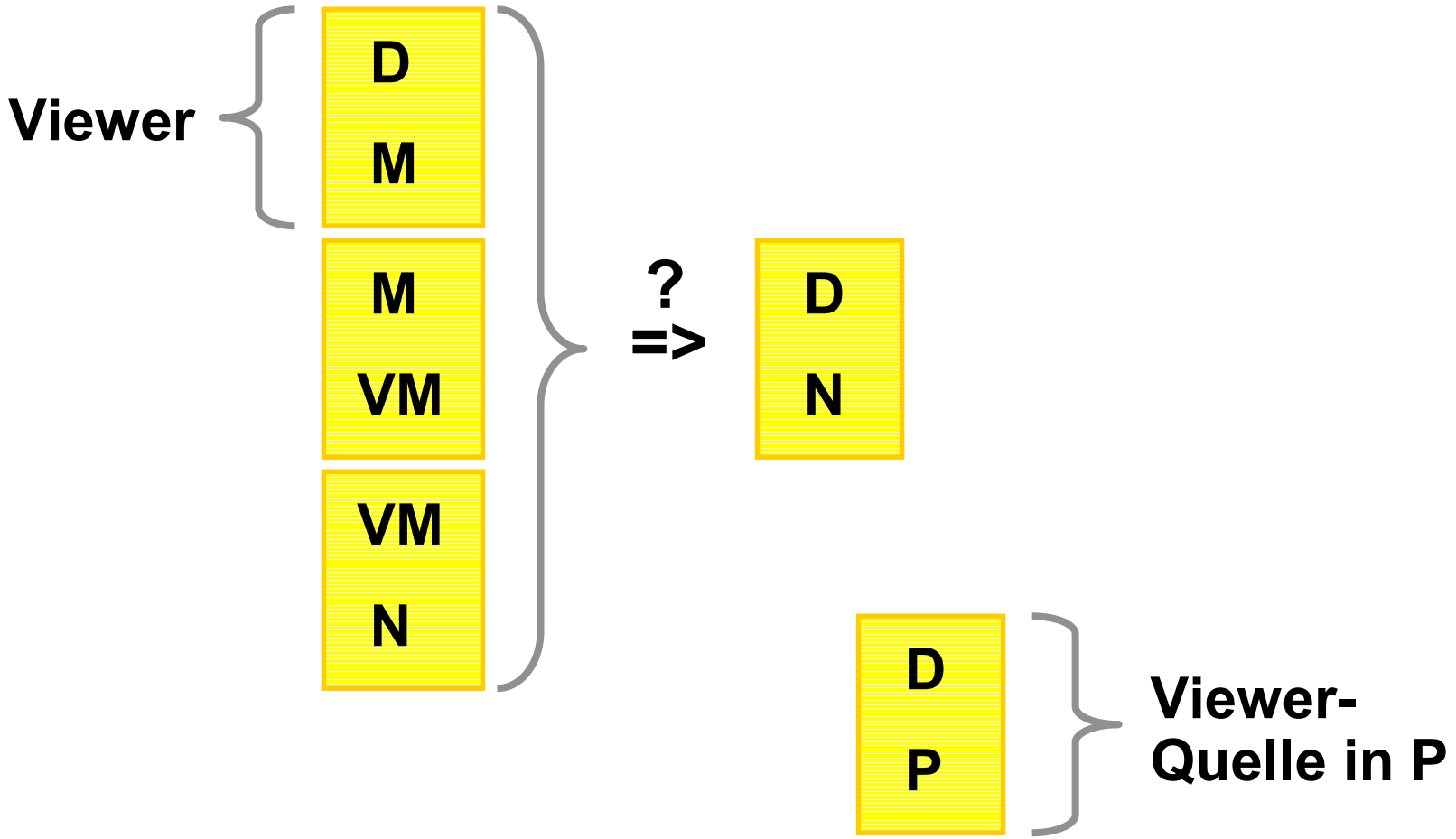
Emulation mit Virtueller Maschine VM

- **Aufgabe:** alte Plattformen $\mathbf{M_1, \ldots, M_m}$ auf neue Plattformen $\mathbf{N_1, \ldots, N_n}$ bringen. Erfordert **m*n** Emulatorbausteine:
- Die gleiche Aufgabe mit **VM** erfordert nur **m+n** Emulatorbausteine:

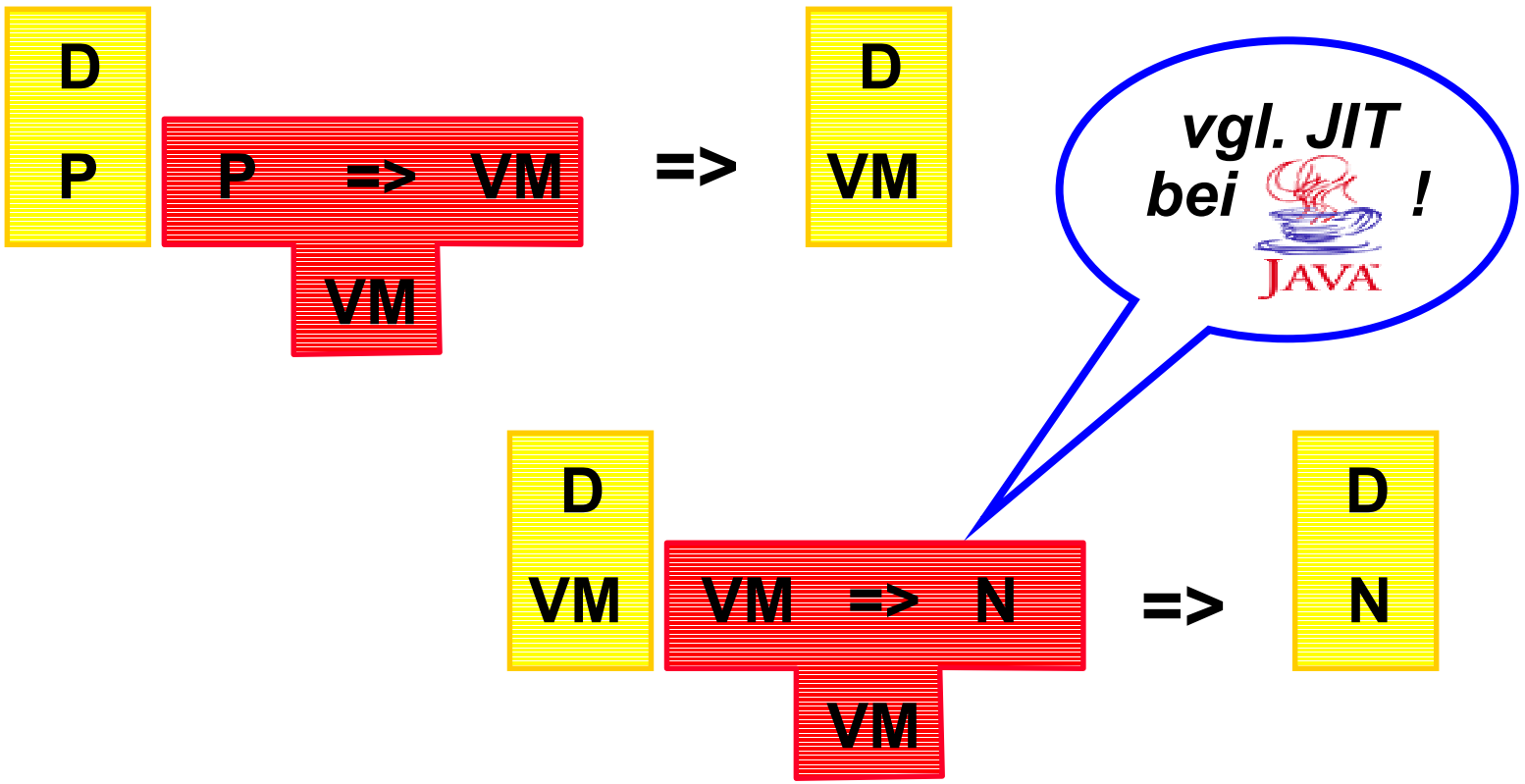


- **V**irtuelle **M**aschine
  - ➢ *schafft Stabilität*
  - ➢ *kontinuierlicher Aufwand*

Ineffizienz vermeiden ...



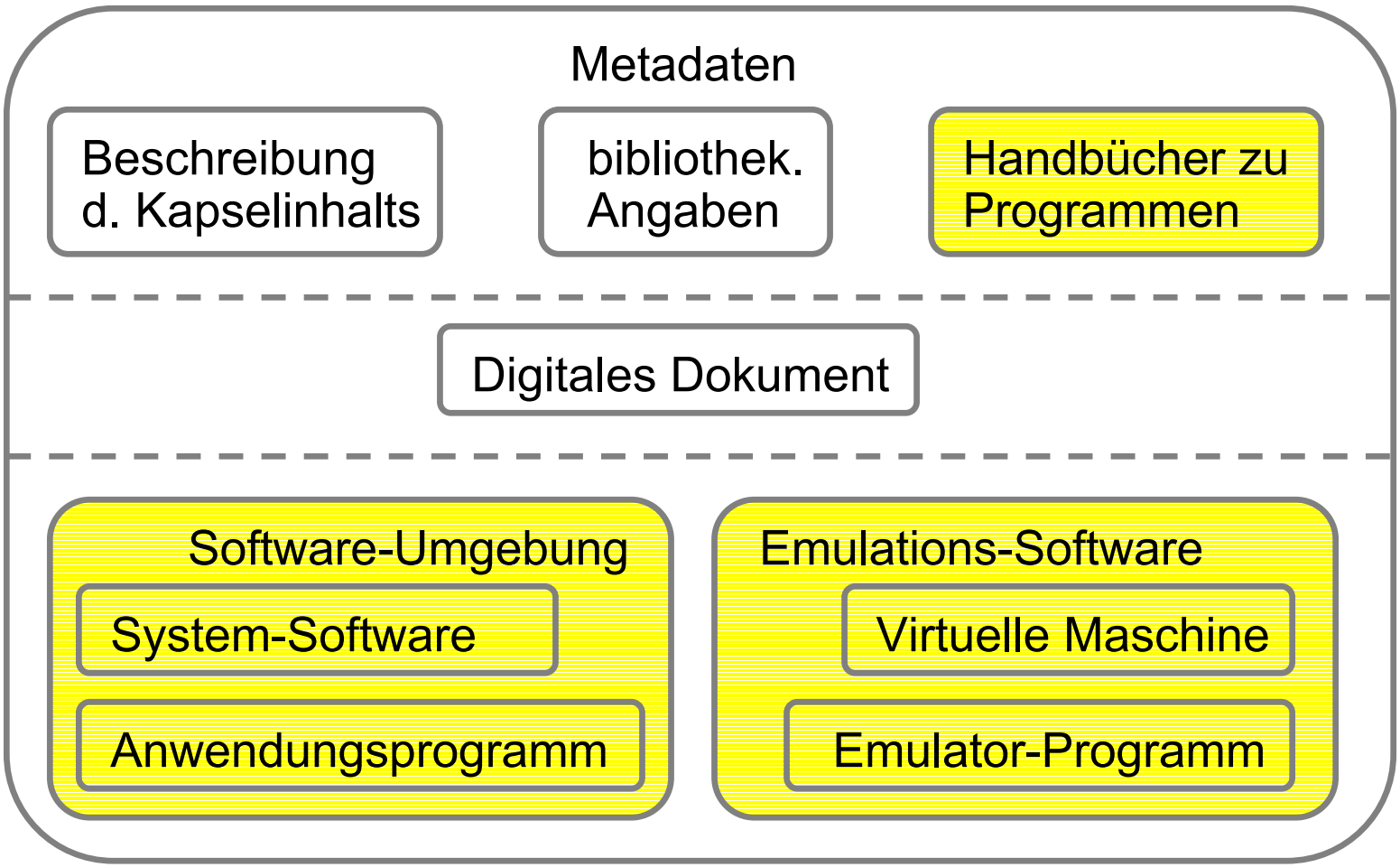
... durch zwei Compilationsschritte



Funktioniert Emulation?

- Emulatoren bei der **Entwicklung neuer Computer**.
  - ❖ Motorola 68000-Emulator von **Apple** beim Übergang zum PowerPC-Prozessor verwendet.
- Emulatoren für populäre **Videospiel-Plattformen** im Netz verfügbar.
- Anspruchsvolle Emulatoren (z.B. **MMIX von D. Knuth**) in der Informatik-Ausbildung im Einsatz.
  - ❖ **Experimente von Rothenberg** mit kommerziellem Windows-Emulator auf Macintosh (MM-CDs).
- **ICL1900-Emulator von Holdsworth** implementiert (in C auf Win32, Solaris, Linux), darauf Timesharing-OS George3 und Algol68-Compiler.

AIP-Speicherkapsel für Emulation



Kritik und Klärungsbedarf

- OAIS Red Book
  - ***E/A- und Zeit-Abhängigkeiten zu wenig beachtet.***
  - ***Fehlende Unterstützung (Werkzeuge) im Markt.***
- D. Bearman („Reality and Chimeras ...“)
  - ***Undefiniert, welche Dokumentattribute wichtig sind.***
  - ***Nicht Software, sondern Belege wesentlich (=>DB)!***
- J. Rothenberg („Using Emulation ...“)
  - ***Entwicklung eines Formalismus für Emulatorspezifikationen.***
  - ***Sichere Datenkapselungstechniken (für komplexe AIPs).***
  - ***Selbstbeschreibung durch Mensch-lesbare Annotationen.***
  - ***Hilfe-System für künftige Nutzer.***

Aktuelle Entwicklung

- Anstelle von Jeff Rothenberg, RAND-Europe, jetzt **Raymond Lorie**, IBM Almaden, Partner von NEDLib.
- Bestandteile seines Konzepts:
  1. Unterscheide zwischen Erhalt von Dokumenten und Programmen (beides nötig, aber verschieden).
  2. Speichere Dokumente „XML-ähnlich“ und beschreibe Dokumentenstrukturen „XML-DTD-ähnlich“.
  3. Verwende einen **U**niversal **V**irtual **C**omputer (**UVC**).

     Dokumenttypdefinierer: **UVC**-Viewer

     Computerhersteller: Computer <=> **UVC**

     (je zwei Emulatoren)

5. Perspektive

Dringend zu klärende Fragen:

- Emulation
  - *geeignete, umfassende Spezifikationssprache?*
  - *Bedienbarkeit der Programme?*
- Migration
  - *akkumulierende Verfälschungen vermeiden?*
  - *Aufwand dauerhaft tragbar?*
- Inhaltsorientierte Archivierung
  - *authentisch genug?*
- Rechtliche und gesellschaftliche Probleme

Komplementäre Nutzung der Strategien

- ***Originalfassung*** für „Dokumenten-Historiker“: Originaldokument und Emulatoren.
- Von häufig gebrauchten Dokumenten daraus ***aktuelle Fassungen*** durch Migration ableiten.
- Für ***automatische Auswertung*** von Inhalten:
  - Erschlossene Informationen in einem offenen **XML**-Format halten.
  - Metadaten in Form von **RDF**-Statements (Resource Description Framework) basierend auf **Dublin Core**-Begriffen.

Links und Quellen

Diese Folien sind zugänglich unter:

http://ist.unibw-muenchen.de/People/lothar/DigBib.htm

Guter Startpunkt für eigene Recherchen:

http://www.nla.gov.au/padi/

Literatur

Oft zitiert

- *Preserving Digital Information,* Report of the Task Force on Archiving of Digital Information, Research Libraries Group, Mai 1996.
- CCSDS Draft Recommendation CCSDS 650.0-R-1: *Reference Model for an Open Archival Information System (OAIS),* Red Book, Mai 1999.

Fallstudie zur Migration

- G.W.Lawrence, W.R.Kehoe, O.Y.Rieger, A.R.Kenney: *Risk Management of Digital Information: A File Format Investigation,* CLIR, Jun. 2000.

Jeff Rothenberg zu Emulation

- *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation,* RAND-Europe, Jan.1999.
- Gemeinsam mit Tora Bikson: *Carrying Authentic, Understandable and Usable Digital Records Through Time,* RAND-Europe, Aug. 1999.
- *An Experiment in Using Emulation to Preserve Digital Publications,* NEDLib Report 1, Apr. 2000.
- *Using Emulation to Preserve Digital Documents,* RAND-Europe and Koninklijke Bibliotheek, Jul. 2000.

Weitere Arbeiten zum Emulationsansatz

- D.Bearman: *Reality and Chimeras in the Preservation of Electronic Records,* D-Lib Magazine, Apr. 1999.
- S.Gilheany: *Preserving Information Forever and a Call for Emulators,* Digital Libraries Conf., Singapur, 1998.
- D.Holdsworth: *C-ing ahead for digital longevity,* CAMiLEON Project, University of Leeds, ??.
- R.A.Lorie: *Long-Term Archiving of Digital Information,* IBM Almaden Research Report RJ 10185, Mai 2000.
- R.A.Lorie: *Long Term Preservation of Digital Information,* IBM Almaden Research, Okt. 2000.