# Challenges of Longterm Preservation of Digital Data

*DAAD / UFPR, Dept. Computer Science*

*Research Seminar*

*15th March 2011, Curitiba*

Dr. Dirk von Suchodoletz

Faculty of Engineering, University Freiburg, Germany

# Overview of the Things to Come

- My Background
- Threats on digital objects
- Developments in this field
  - International initiatives and Projects
  - Existing components and ideas
- Preservation action and Emulation
  - Formalizing the rendering requirements
  - Software archiving
- Workflow integration and automation

# My Background

- Lecturer/researcher at the professorship of communication systems
  - Lectures, seminars and student projects on computer networking and communication systems
  - PhD on “Longterm preservation of dynamic digital objects” in 2008
  - Participating in large scale EU integration project PLANETS
  - Research at National Archives of New Zealand beginning this year

# My Background

- Chair in Communication Systems
  - Rather small entity with a small number of scientific assistants/lecturers working in different fields
  - Offers lectures and seminars in Internet Working, Telecommunication Systems, Network Technology; in cooperation with Max-Planck-Institute for Foreign and international Crimial Law seminars on Internet&Law
  - Got into domain of DP via PhD theses on Emulation and PLANETS project – Preservation and Longterm Access to NETworked Services

# The Professorship

- Digital Preservation involvements
  - EU PLANETS Project (finished)
  - Founding member of the *Open Planets Foundation*
  - Cooperation with the National Library and National Archives in The Netherlands
  - Member of the German *nestor* initiative and founding member of its *Emulation WG*
  - Long cooperation with the *Computer Games Museum* opened last weekend in Berlin
  - Supervision of a number of Bachelor- and Master theses in this field

# Short Risk Analysis of Digital Objects

- Threatened on several layers
  - Physical
  - Technological
  - Intellectual

# Physical Risks

- Decay of media
  - More risky are removable, because often “uncovered”, unprotected media
    - Optical media like CDROM, DVD, BlueRay
    - DVD is pretty risky even for brand new pressed disks (bacteria eating up the glue layer)
    - CD for several years, no much known on BlueRay (general not a good idea)
  - Less problematic for hard-drives – but number of mechanical problems here

# Physical Risks

- Deprecation of connectors and standards
  - Anyone knows the MFM/RLL controller?
  - Last disk of this standard might be produced just 20 years ago
  - SCSI – pretty old standard – anyone succeeded in connecting the old SCSI-disk to a modern SCSI-320 controller?
  - Same for IDE → SATA

# Technological Risks

- Even if the bit-stream recovery was successful (disk completely copied to a modern, accessible medium)
- Rapid changes of the typical work-desks
  - Changes of (G)UI concepts (IBM 286 and Apple iPad just 25 years apart)
  - Different hardware architectures
  - Different operating systems
  - What to do with the old file formats (WordStar, AmiPro, Lotus-1-2-3, WordPerfect, old MS-Office)?

# Intellectual Risks

- Extremely relevant for archival data
- Changes in contextual knowledge
  - Missing or incomplete documentation
  - Lost context of single objects or groups of objects, linked objects
  - Ambiguous data formats and descriptions
  - Changes in terminology and basic assumptions

# Additional Risks

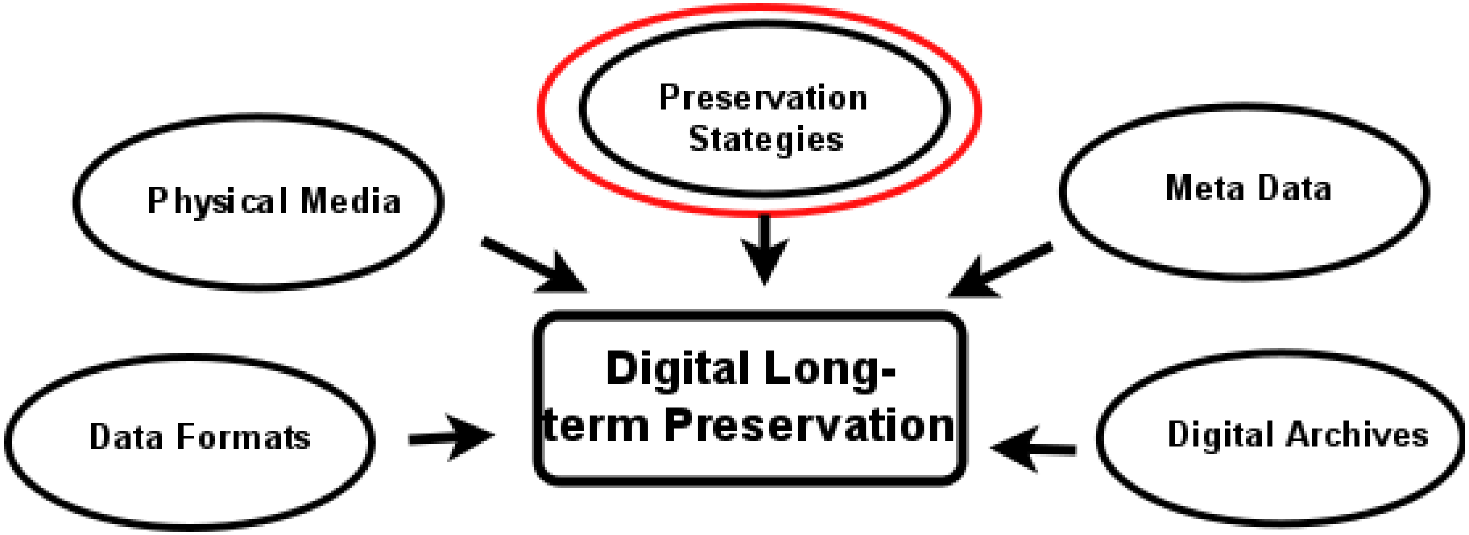
- Human errors
- Technical, machine breakdowns (e.g. a hot summer maneuvered the Freiburg universities computer center to nearly shutdown of most of the server machines in the “air-conditioned” hall)
- Catastrophes
- Security flaws, forgery of data, sabotage ...

# Additional Risks

- Typical risks in many memory institutions
  - Under-financed IT services
  - Understaffed, ill-qualified personnel
  - Missing rights management, authorization, …
  - Missing Know-How

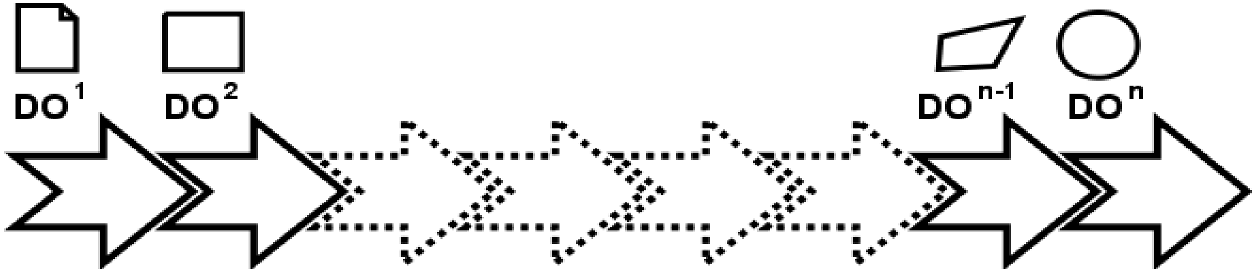
# From Risks to Solutions

- Research Domain of Digital Preservation pretty young but differentiates into sub domains
- Our specialization is preservation action: Emulation



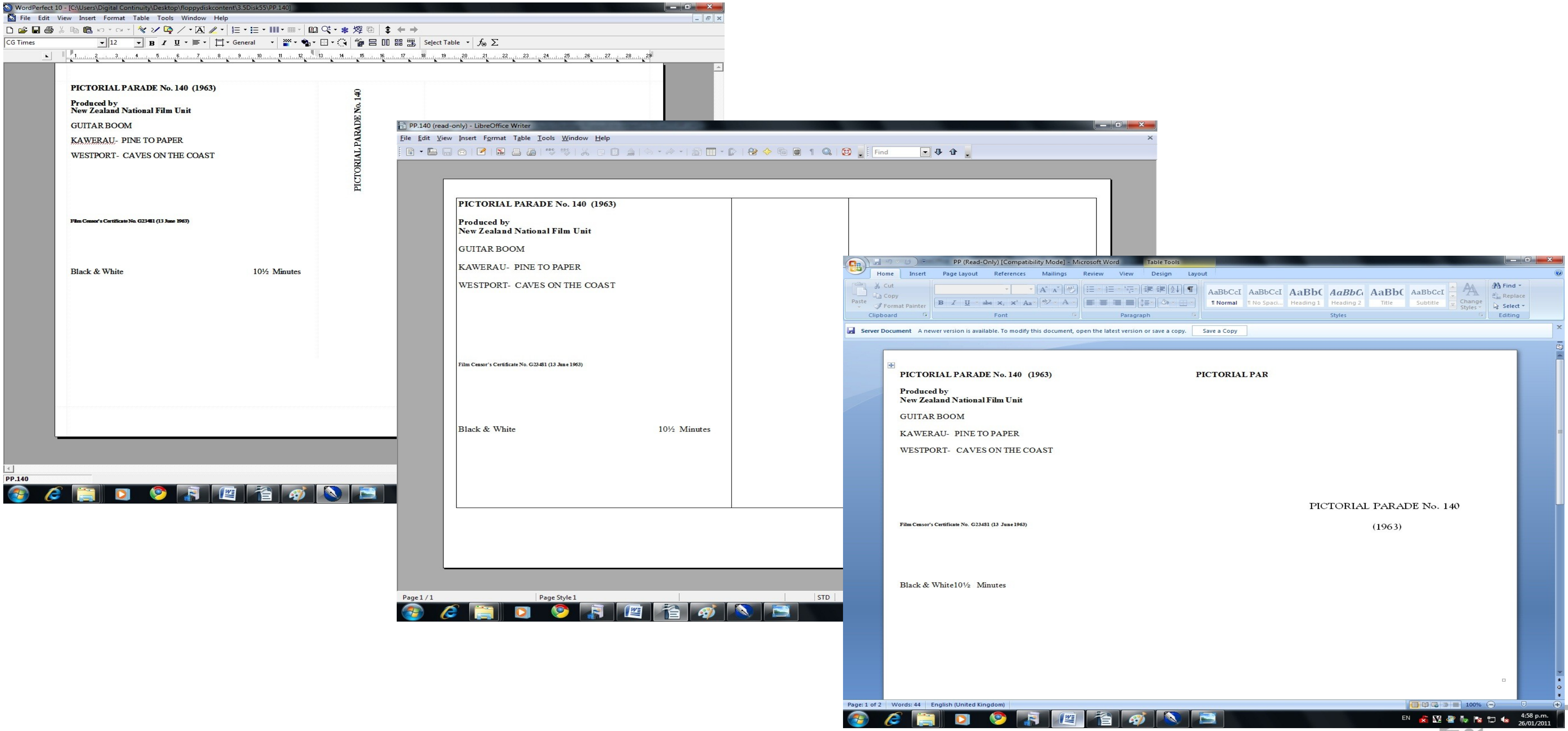
# Dynamic Objects & Authenticity

- Preservation challenges
  - Digital objects require software / hardware environments to be accessed
  - Environments change over the time and obsolete most of digital material
  - Mainline strategy: *Migration*
  - Risky to rely on it exclusivly
  - Not suitable for all object types



# Authenticity Tests / Experiments

Problems easily spotted: Rendered a test corpus in different original applications/environments


WordPerfect 10 - [C:\Users\Digital Continuity\Desktop\floppydiskcontent\3.5Disk55\PP.140]
File Edit View Insert Format Table Tools Window Help
CG Times
12
General
Select Table
PP.140

**PICTORIAL PARADE No. 140 (1963)**

**Produced by**
**New Zealand National Film Unit**

GUITAR BOOM

KAWERAU- PINE TO PAPER

WESTPORT- CAVES ON THE COAST

PICTORIAL PARADE No. 140

Film Censor's Certificate No. G23481 (13 June 1963)

Black & White 10½ Minutes

PP.140 (read-only) - LibreOffice Writer
File Edit View Insert Format Table Tools Window Help
Find
Page 1 / 1
Page Style 1
STD

**PICTORIAL PARADE No. 140 (1963)**

**Produced by**
**New Zealand National Film Unit**

GUITAR BOOM

KAWERAU- PINE TO PAPER

WESTPORT- CAVES ON THE COAST

Film Censor's Certificate No. G23481 (13 June 1963)

Black & White 10½ Minutes

**Server Document** A newer version is available. To modify this document, open the latest version or save a copy.

**PICTORIAL PARADE No. 140 (1963)** **PICTORIAL PAR**

**Produced by**
**New Zealand National Film Unit**

GUITAR BOOM

KAWERAU- PINE TO PAPER

WESTPORT- CAVES ON THE COAST

PICTORIAL PARADE No. 140

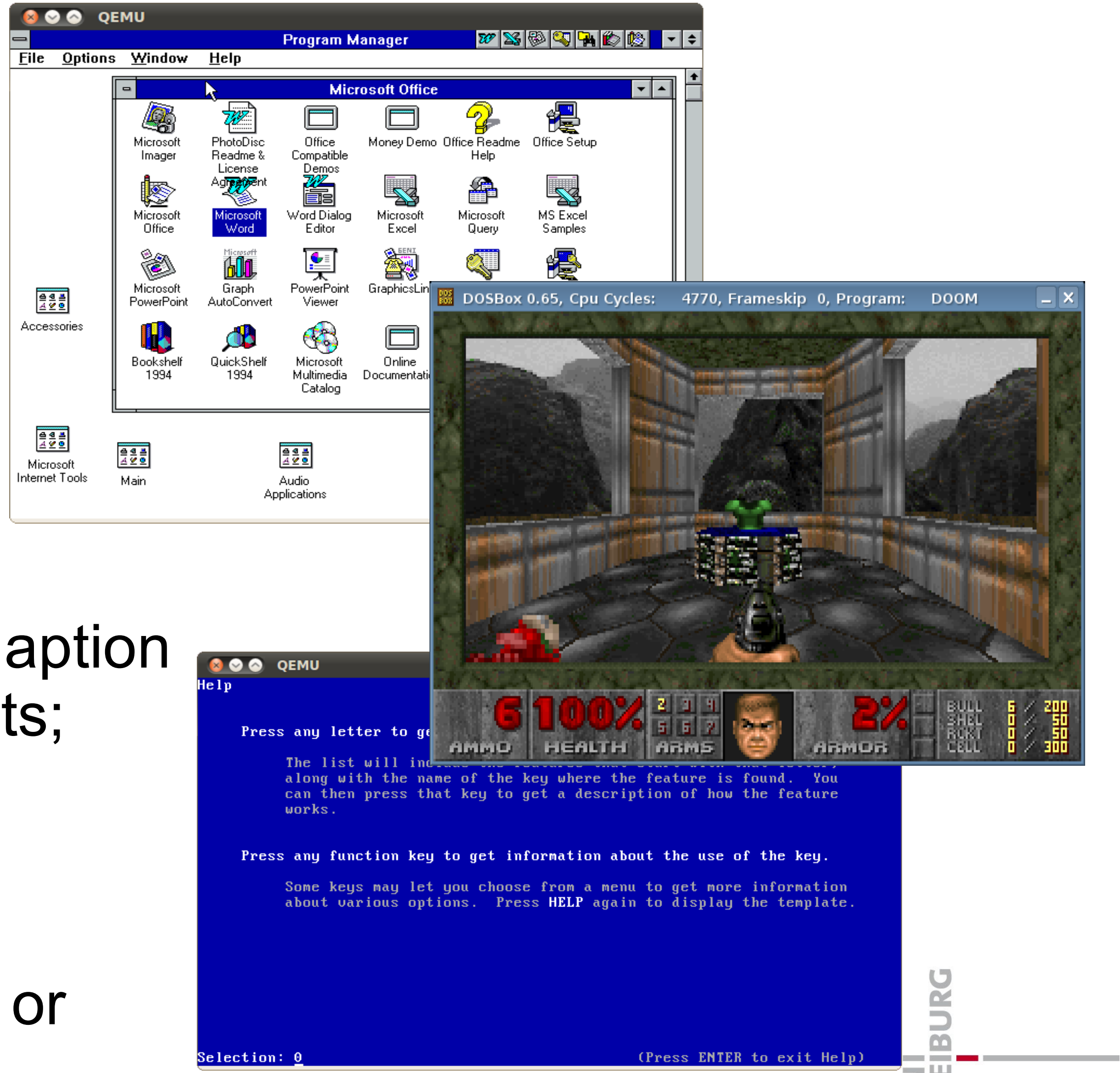
(1963)

Film Censor's Certificate No. G23481 (13 June 1963)

Black & White10½ Minutes

# Dynamic Digital Objects

- Objects like
  - Applications
  - Operating systems
  - Databases
- Non-linear, user interaction, multiple views
- No real option:
  - Printing of source, adaption to recent environments; even **if** source code available
  - Video-recording, screenshots of game or application session



Help

Press any letter to ge

The list will in
along with the name of the key where the feature is found. You can then press that key to get a description of how the feature works.

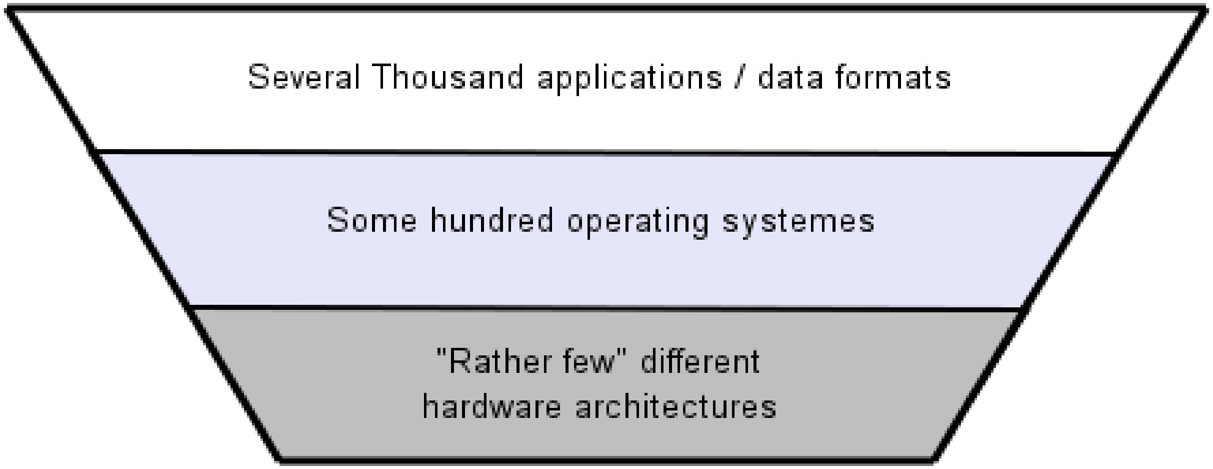
Press any function key to get information about the use of the key.

Some keys may let you choose from a menu to get more information about various options. Press **HELP** again to display the template.

Selection: 0 (Press ENTER to exit Help)

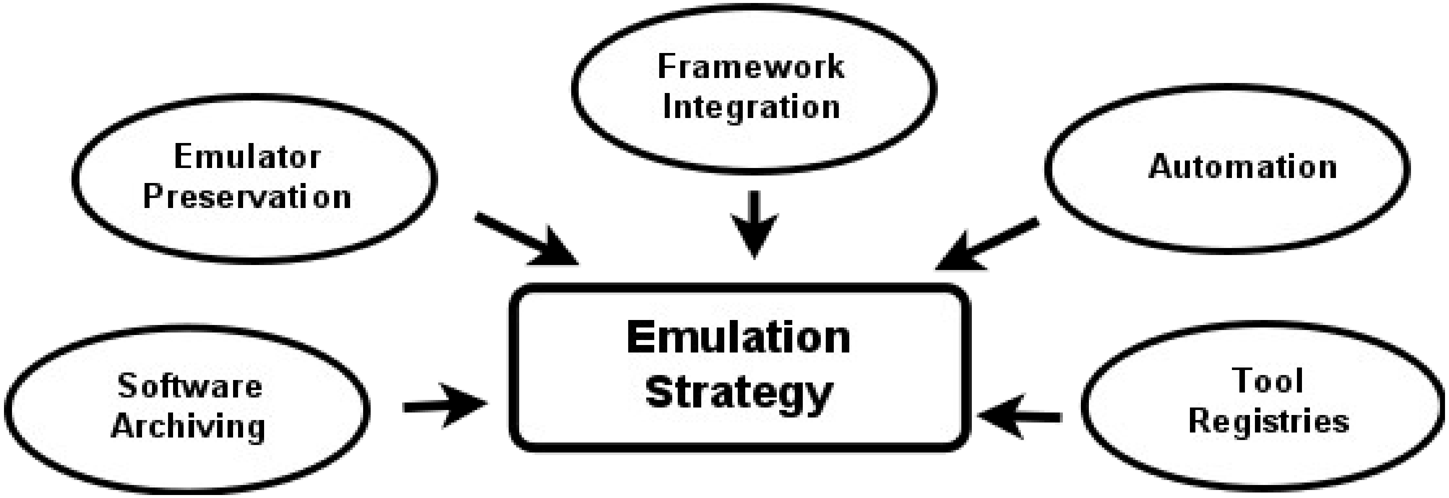
# Different Approach – Emulation

- No changes on object, but reproduction of original environment
  - Emulators around for quite a while, supplemented by virtualization
  - Can operate on different layers of software/hardware stack
  - Number of objects to cover differs significantly; thus hardware layer seems very attractive to focus on


Several Thousand applications / data formats
Some hundred operating systemes
"Rather few" different hardware architectures

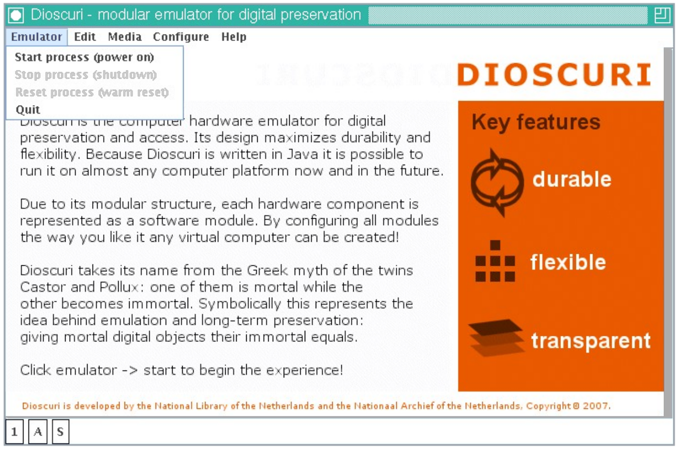
# Emulation Strategy

- More complex approach involving larger number of additional software components, complexities
  - Standalone emulation does not help much
  - Different sectors of ongoing research
  - Number of open issues



# Emulator Examples

- Dioscuri X86 emulator recreating an 286, 386 PC of the early 1990th
  - Java programming language, modular approach – components like disk, floppy, VGA, CPU, RAM put together to form the machine
  - Running DOS and Windows 3.0
- QEMU – using popular C program-ming language multi architecture emulator for X86, PPC, Sparc, ...
- Both Open Source – no vendor lock-in, adaptable



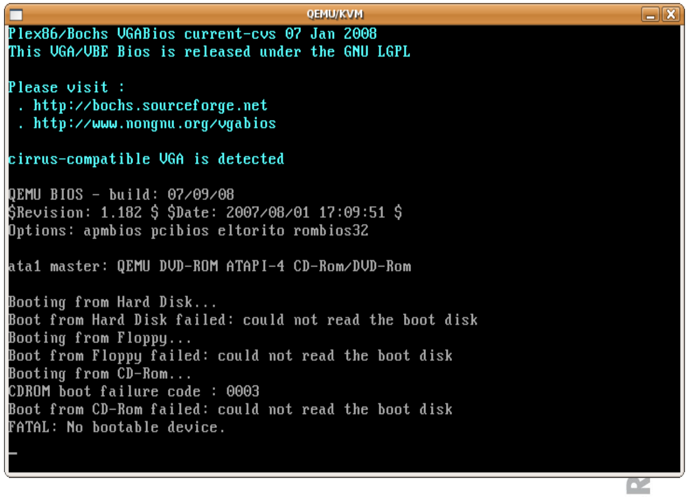
Dioscuri is the computer hardware emulator for digital preservation and access. Its design maximizes durability and flexibility. Because Dioscuri is written in Java it is possible to run it on almost any computer platform now and in the future.

Due to its modular structure, each hardware component is represented as a software module. By configuring all modules the way you like it any virtual computer can be created!

Dioscuri takes its name from the Greek myth of the twins Castor and Pollux: one of them is mortal while the other becomes immortal. Symbolically this represents the idea behind emulation and long-term preservation: giving mortal digital objects their immortal equals.

Click emulator -> start to begin the experience!

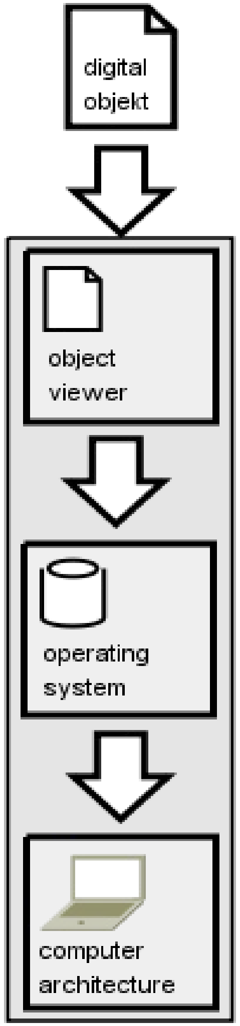
Dioscuri is developed by the National Library of the Netherlands and the Nationaal Archief of the Netherlands. Copyright © 2007.



```
Plex86/Bochs VGABios current-cvs 07 Jan 2008
This VGA/VBE Bios is released under the GNU LGPL

Please visit :
 . http://bochs.sourceforge.net
 . http://www.nongnu.org/vgabios

cirrus-compatible VGA is detected

QEMU BIOS - build: 07/09/08
$Revision: 1.182 $ $Date: 2007/08/01 17:09:51 $
Options: apmbios pcibios eltorito rombios32

ata1 master: QEMU DVD-ROM ATAPI-4 CD-Rom/DVD-Rom

Booting from Hard Disk...
Boot from Hard Disk failed: could not read the boot disk
Booting from Floppy...
Boot from Floppy failed: could not read the boot disk
Booting from CD-Rom...
CDROM boot failure code : 0003
Boot from CD-Rom failed: could not read the boot disk
FATAL: No bootable device.

_
```

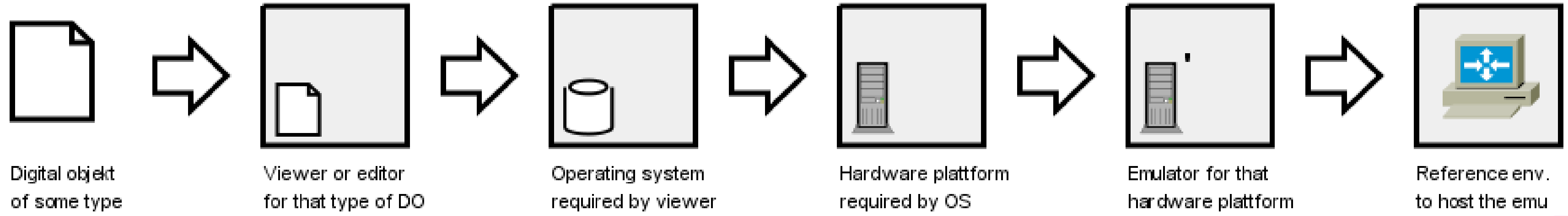
# Requirements for Emulation

- Emulation not working just on its own – additional software is required
- Emulation approach requires recreation of ancient hardware / software environments for access / execution
  - E.g. spreadsheet document requires the proper spreadsheet application for interpretation and displaying
  - Spreadsheet software is dependent on an operation system
  - Operating system was programmed for a very specific or a range of hardware architectures
  - Additional components like fonts might be needed for range of documents, especially for non-latin typesets



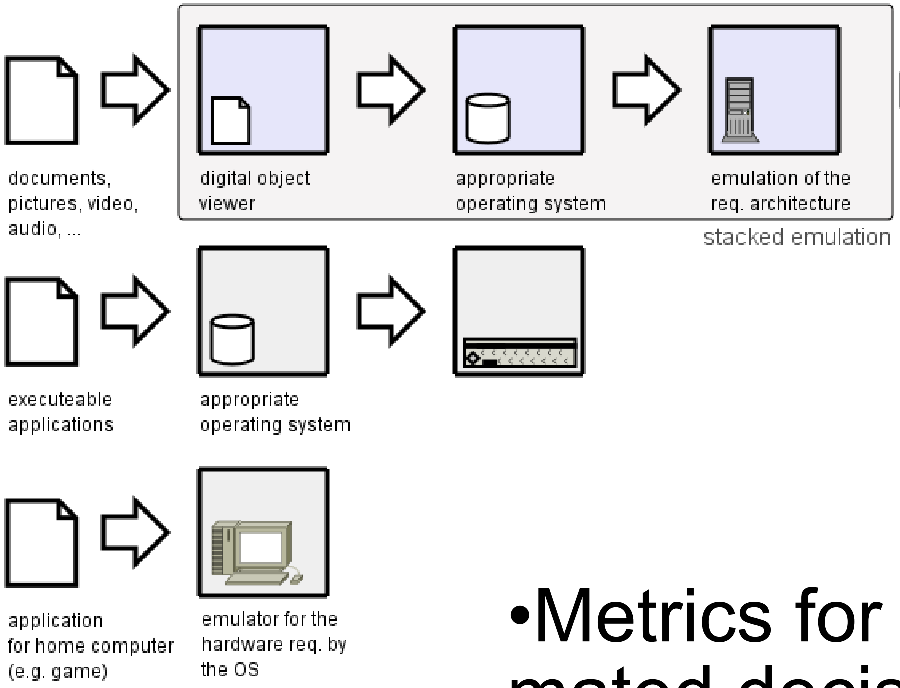
# Formalization of Requirements

- View Path – pathway from object to specific environment
- Formalization needed – view path as the requirements to be followed to actually access, display the object of interest
- Reference environment – specifically defined software hardware combination for object access, rendering
- Concept to describe dependencies between objects

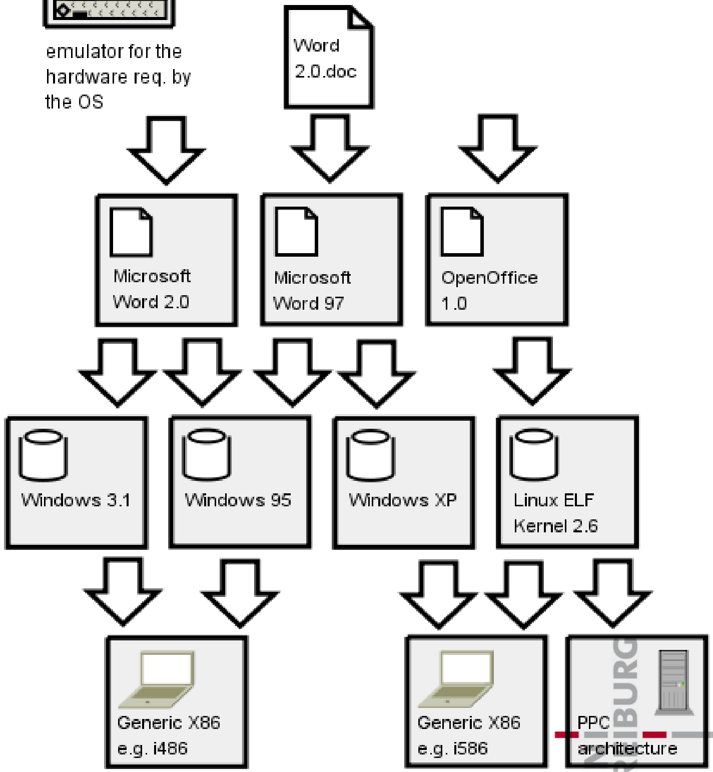

Digital objekt of some type
Viewer or editor for that type of DO
Operating system required by viewer
Hardware plattform required by OS
Emulator for that hardware plattform
Reference env. to host the emu

# Formalization of Requirements

- View Path dynamic – depend on regarded object, actual working environment, emulator preservation strategy; often multiple options

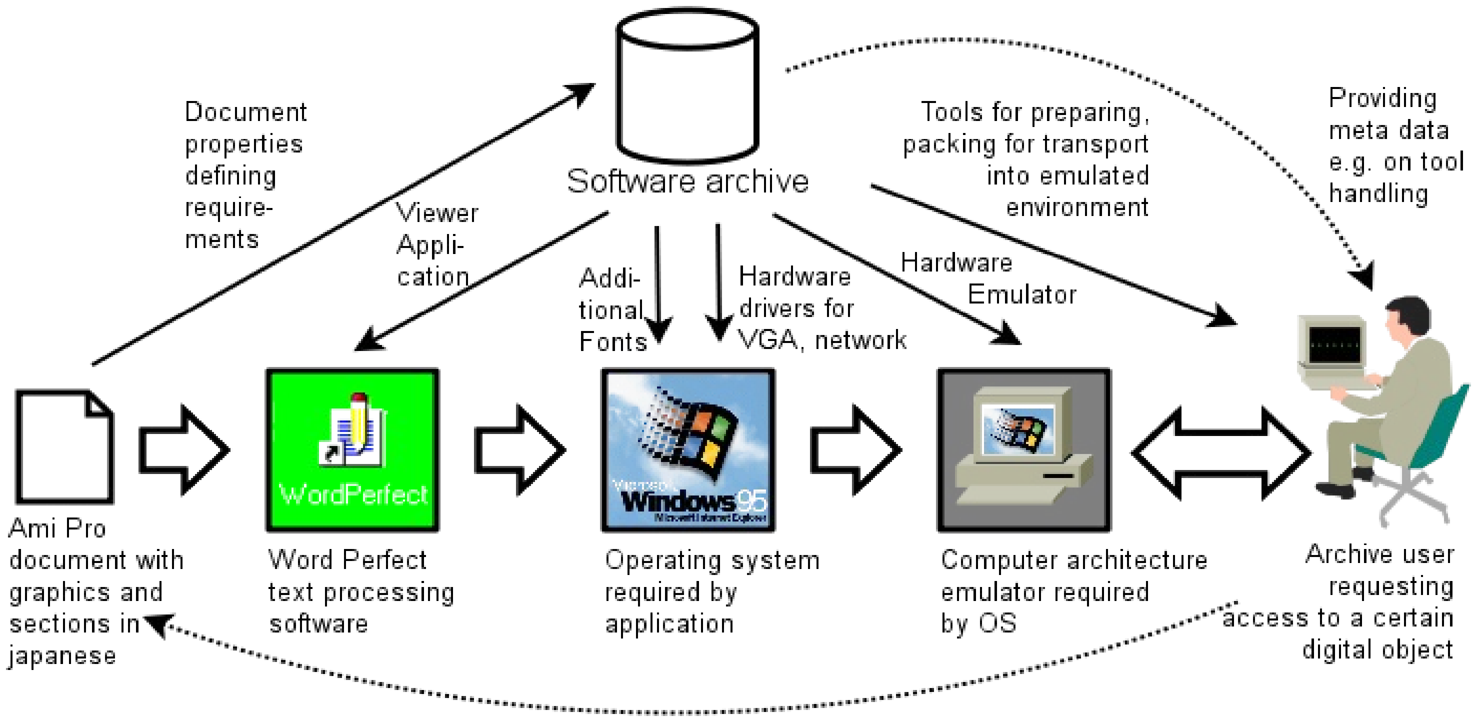

documents, pictures, video, audio, ...
digital object viewer
appropriate operating system
emulation of the req. architecture
stacked emulation
appropriate operating system
emulator for the hardware req. by the OS
executeable applications
appropriate operating system
application for home computer (e.g. game)
emulator for the hardware req. by the OS

- Metrics for automated decision making



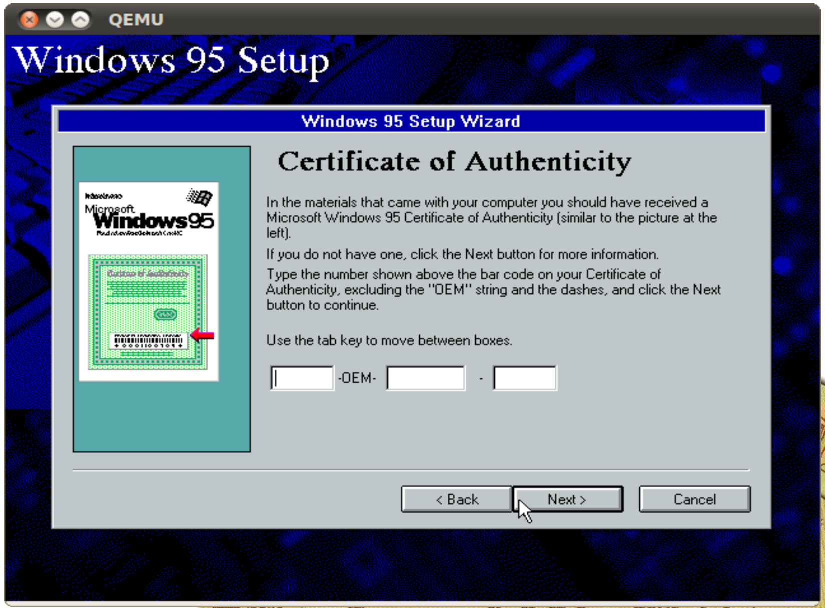
# Software Archiving

- Software archive containing all necessary additional single objects or for production systems prefabricated view-paths


Software archive
Document properties defining require-ments
Viewer Appli-cation
Addi-tional Fonts
Hardware drivers for VGA, network
Hardware Emulator
Tools for preparing, packing for transport into emulated environment
Providing meta data e.g. on tool handling
WordPerfect
Windows95
Ami Pro document with graphics and sections in japanese
Word Perfect text processing software
Operating system required by application
Computer architecture emulator required by OS
Archive user requesting access to a certain digital object

# Additional Information

- Additional information and metadata needed in software archive
  - Application handbooks
  - Howtos and trouble shooting guides
  - Application update packages
  - License keys, access codes
- Depending on object
  - Fonts for documents
  - Codecs for video, audio
  - Software extensions like DirectX, OpenGL libraries



Windows 95 Setup

Windows 95 Setup Wizard

Certificate of Authenticity

In the materials that came with your computer you should have received a Microsoft Windows 95 Certificate of Authenticity (similar to the picture at the left).

If you do not have one, click the Next button for more information.

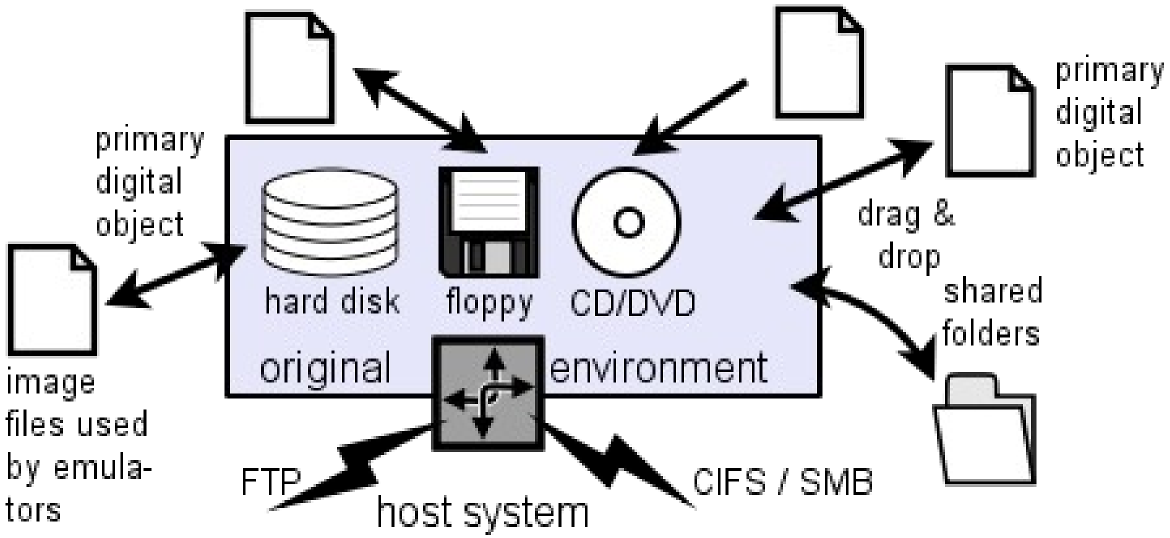
Type the number shown above the bar code on your Certificate of Authenticity, excluding the "OEM" string and the dashes, and click the Next button to continue.

Use the tab key to move between boxes.

-OEM- -

# Data Exchange with Emulators

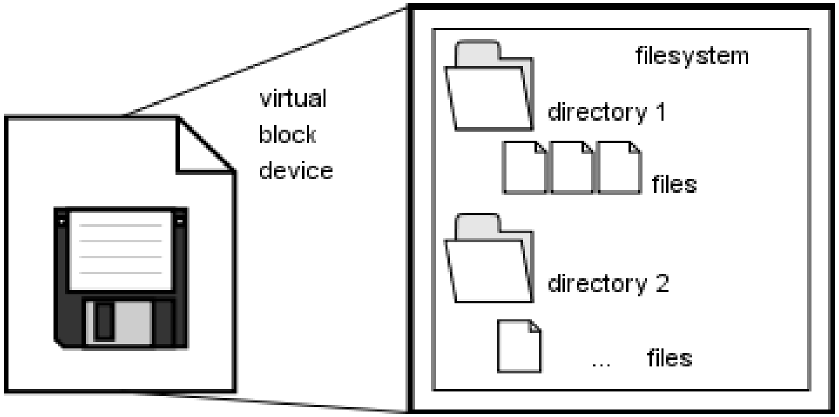
- Object is to be transported into emulation environment
  - Different ways: After or during enviroment setup

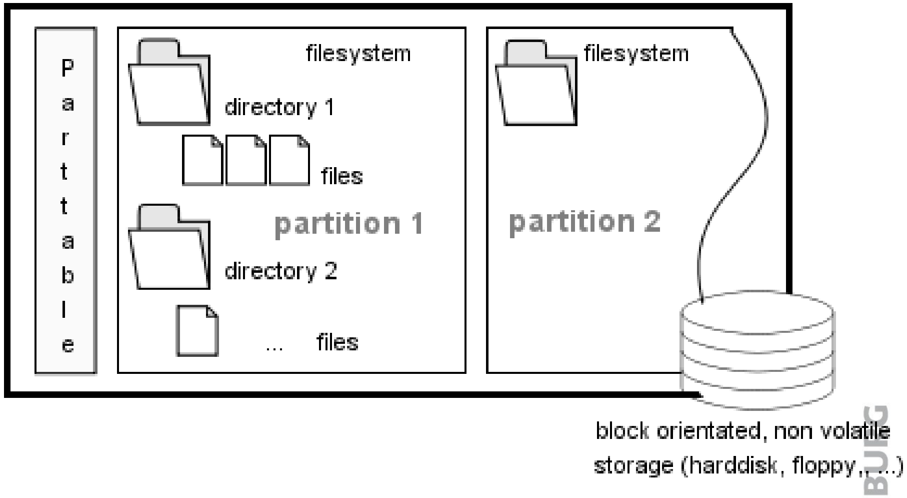


- Means of object transport
  - Virtual optical (ISO) or floppy disks as images
  - Disk container files
  - Network connections like FTP, SMB/CIFS
  - “Shared Folders” (as e.g. found in VMware or VirtualBox)
  - Copy&Paste (e.g. text areas in Dioscuri)

# Transport Containers

- Data transport requires fromats understood by the target environment, e.g.
  - Floppy disks, ubiquious in for many platforms for a rather long period
  - Images easy to create and store
  - Optical disks: ISO images well understood by many emus
- More complex
  - Container files of the several emulators
  - Creator tools required
  - Adding objects to disk container files before emulators started





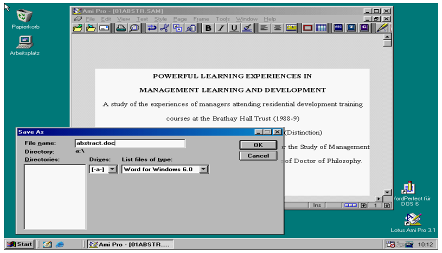
# Challenge: Access to Emulation

- Groundworks laid – how to make emulation accessible?
- Emulation environments
  - Often deal with outdated concepts of software interaction
  - Typically complex and require specific knowledge
  - Require depending on the digital object to be rendered or executed a bunch of additional software components which may need prior installation

# Enabling Access to Emulation

- Major goal is to allow non-technical users access to those services an easy to use, abstract interface is required
- During PLANETS project a prototype for emulation wrapping created – GRATE
- Different emulators like Dioscuri, MESS, QEMU, Hatari and others put into a single networked application

BITTE DRÜCKEN SIE DIE ENTER-TASTE
UM DAS SPIEL ZU BEGINNEN



POWERFUL LEARNING EXPERIENCES IN

MANAGEMENT LEARNING AND DEVELOPMENT

A study of the experiences of managers attending residential development training courses at the Brathay Hall Trust (1988-9)

(Distinction)

r the Study of Management

of Doctor of Philosophy.

File name: abstract.doc

Directory: a:\

Directories:

Drives: [-a-]

List files of type: Word for Windows 6.0

# Knowledge and Automation

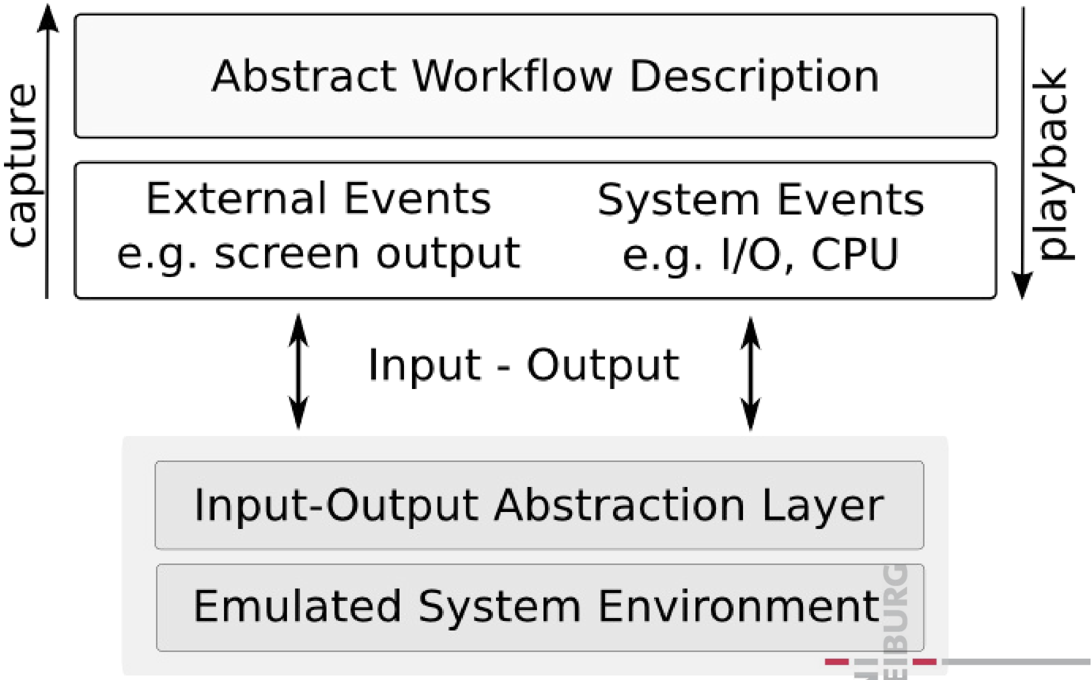
- GRATE focuses on traditional human interaction model, but
  - Requires certain knowledge getting more and more uncommon for todays users
  - Taking system images of emulated environments for granted
  - Handling only limited, prefabricated VPs
- Unsuitable for integration into non-interactive large scale preservation work-flows

# Knowledge and Automation

- Typical applications most digital objects produced with are interactive
- Standard migration work-flows like opening a document and save it in a different format require a human user to type or point&click
- Such manual procedure sub-optimal for e.g. mass migration scenarios
- Next step: A method to replace the human-interaction in GRATE with generic recording and monitoring

# Automation of Interaction

- Define an interactive work-flow as ordered list of interactive events passed on (e.g. mouse and keyboard events)
- Each event is linked with a precondition and an expected outcome
- Built the solution on top of the VNC-Play tool, which offers visual synchronization points



# Framework Integration

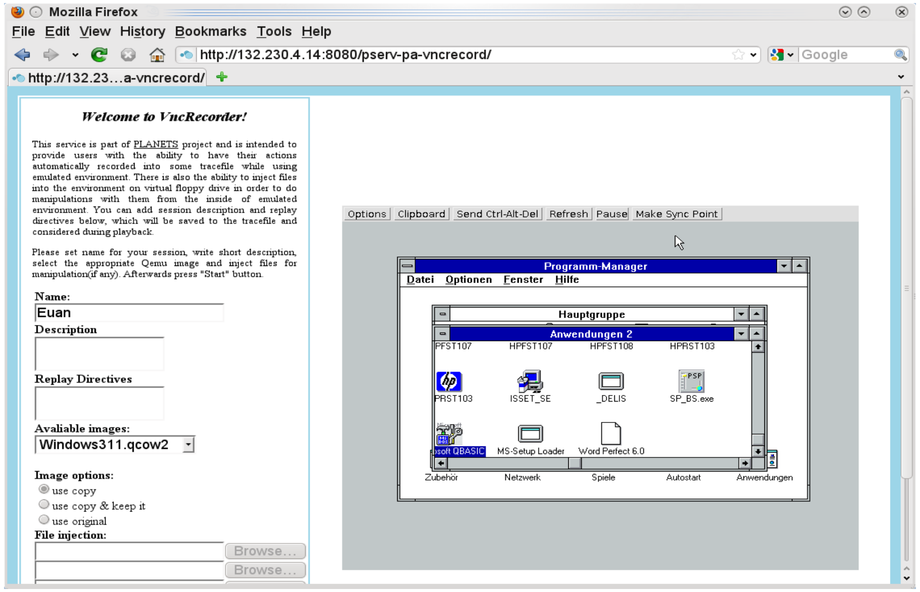
- PLANETS – Preservation and Longterm Access to NETworked Services
  - Offers a set of standardized Web services like Characterization, View, Validation, Comparison, Migration, ...
- Defines a set of APIs Web services need to conform to

# Goals

- Emulation services should allow
  - Occasional users to view digital objects and compare digital objects in their original environment
  - Occasional users to experience ancient (graphical) interactive user environments
  - Documentation and preservation of user interactions and interactive processes in ancient user environments
  - Automated migration of files using the original application in emulation

# Required Services

- After reviewing these goals
  - View service to allow traditional interactive access to objects
  - Automated migration by emulation service



*Welcome to VncRecorder!*

This service is part of PLANETS project and is intended to provide users with the ability to have their actions automatically recorded into some tracefile while using emulated environment. There is also the ability to inject files into the environment on virtual floppy drive in order to do manipulations with them from the inside of emulated environment. You can add session description and replay directives below, which will be saved to the tracefile and considered during playback.

Please set name for your session, write short description, select the appropriate Qemu image and inject files for manipulation(if any). Afterwards press "Start" button.

**Name:** Euan

**Description**

**Replay Directives**

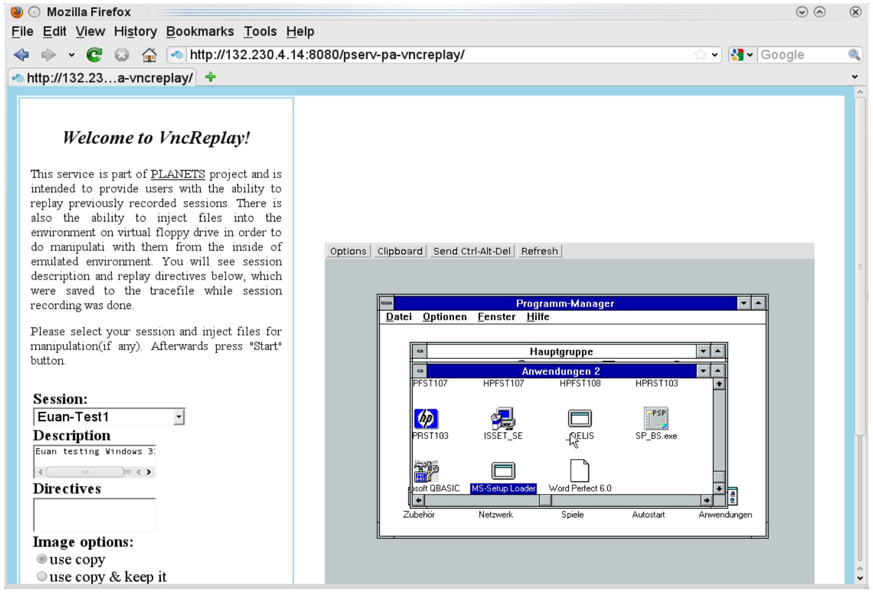
**Avaliable images:** Windows311.qcow2

**Image options:**

● use copy
○ use copy & keep it
○ use original

**File injection:**

Browse…
Browse…



*Welcome to VncReplay!*

This service is part of PLANETS project and is intended to provide users with the ability to replay previously recorded sessions. There is also the ability to inject files into the environment on virtual floppy drive in order to do manipulati with them from the inside of emulated environment. You will see session description and replay directives below, which were saved to the tracefile while session recording was done.

Please select your session and inject files for manipulation(if any). Afterwards press "Start" button.

**Session:** Euan-Test1

**Description**

Euan testing Windows 3:

**Directives**

**Image options:**

● use copy
○ use copy & keep it

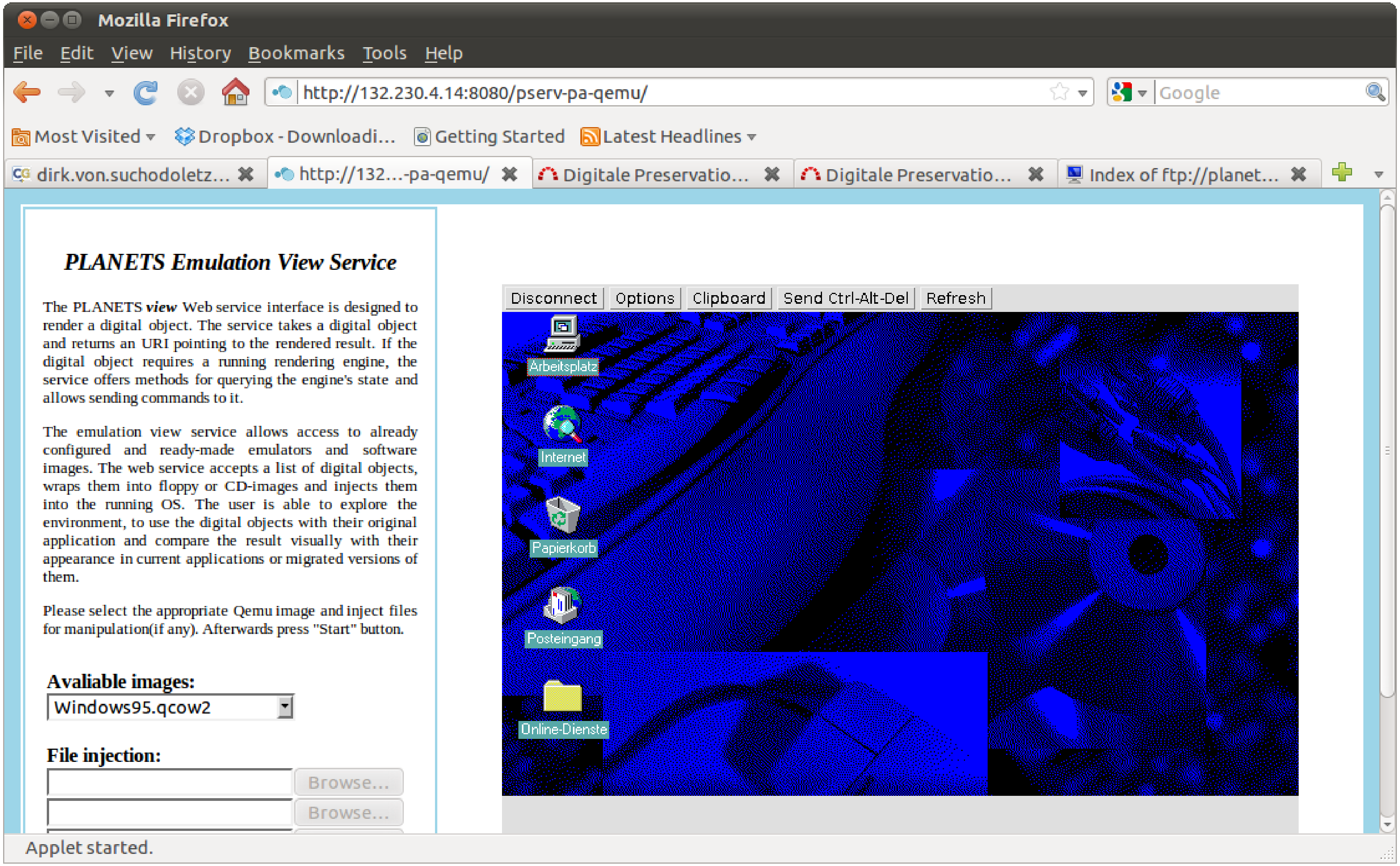
# Emulation View Service

- A generic PLANETS view service takes a digital object and returns an URI pointing to the rendered result
- If the digital object requires a running rendering engine the service offers methods for querying and sending
- View service developed allows access to already configured and ready-made software environments

# Emulation View Service

- Implemented as a PLANETS Web service
  - Accepts a list of digital objects
  - Wraps them into a CD image
  - Makes them available for running operating system
- User is able
  - Explore the original environment
  - Use within the original application
  - Allows visual comparison for migrated objects
  - Do manual migrations by saving or printing
- Process can be generalized, recorded

# Emulation View Service



## *PLANETS Emulation View Service*

The PLANETS ***view*** Web service interface is designed to render a digital object. The service takes a digital object and returns an URI pointing to the rendered result. If the digital object requires a running rendering engine, the service offers methods for querying the engine's state and allows sending commands to it.

The emulation view service allows access to already configured and ready-made emulators and software images. The web service accepts a list of digital objects, wraps them into floppy or CD-images and injects them into the running OS. The user is able to explore the environment, to use the digital objects with their original application and compare the result visually with their appearance in current applications or migrated versions of them.

Please select the appropriate Qemu image and inject files for manipulation(if any). Afterwards press "Start" button.

**Avaliable images:**

Windows95.qcow2

**File injection:**

# Migration by Emulation Service

- Good for viewing but not for large scale preservation tasks
- Second important service for PLANETS using emulation
- Interface expects a digital object as input, a designated output format (PUID) and an optional list of service parameters
- Outcome will be a successfully transformed object or an error message

# Actual State of Workflows

- Got some promising results for simple migrations like loading an AMI Pro document and send it to virtual PDF printer
- At the moment pretty expensive regarding time and compute resources:
  - For every object complete cycle from mounting, loading, system execution and shutdown required
  - In future: start system once and handle multiple objects in succession

# Black Box Emulation Environment

- Pretty much “black box” at the moment
  - Not yet reliable
  - Migration processes just started difficult to monitor
  - Unknown execution time
  - Unreliable behavior – system might take infinitely or might crash – mostly unobserved

# Measurement and Evaluation

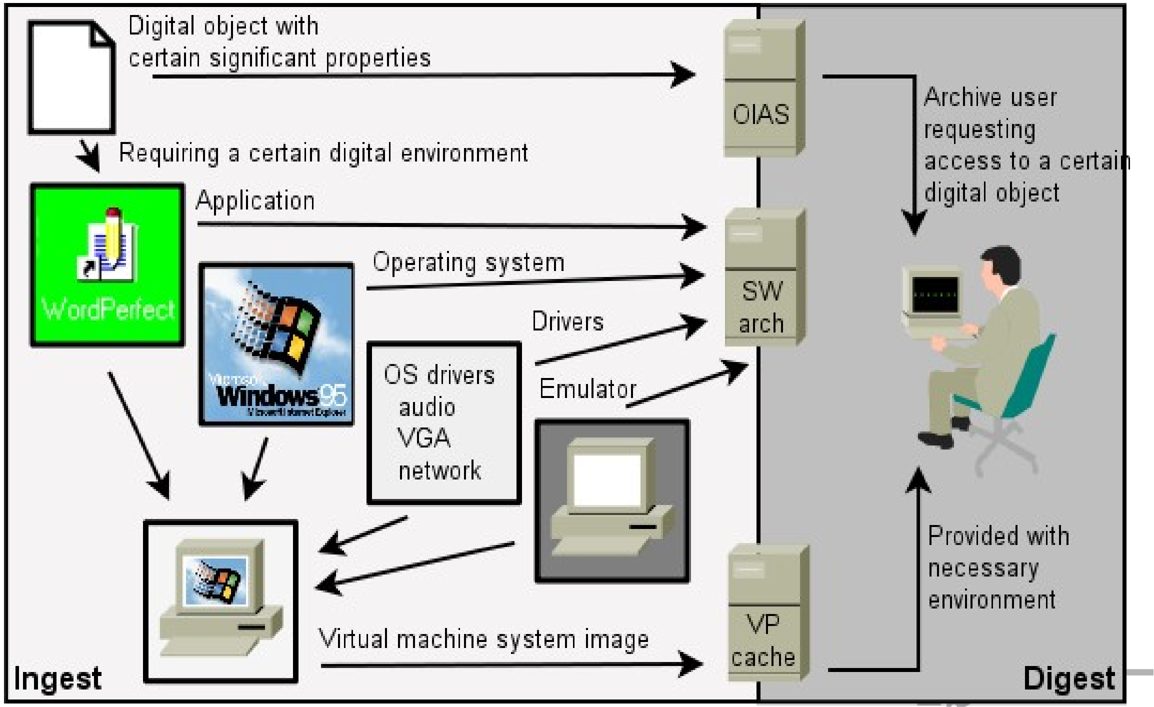
- Very scarce measurement and evaluation options in todays emulators
  - Difficult to calculate runtime, effort, host system resources
  - Tedious to observe file operation e.g. when the object is finished to be processed and completely written to disk again
- Without measurements and metrics for significant characteristics comparisons of workflows and different emulators pretty impossible

# Major Challenges

- To make migration via emulation workflows comparable
  - Monitoring and evaluation framework needed
  - Have metrics for certain emulation characteristics
  - Test or prove completeness of emulation

# Outlook: General Integration

- Integrate software archive into preservation work-flows
  - Check software list on object ingest
  - Store single software components
  - Documentation
- Preserve knowledge by storing workflow recordings and complete emulation environments



# Ongoing Research / Theses

- Map view path to software archiving / library to preservation workflows
- Large scale workflow automation
- Creating sample reading room workstation for object access through emulation (services)
- Automated emulator testing
- Define future emulator requirements
    - Control APIs (input, automation, monitoring, …)
    - Stable presentation towards original environments
- General: Long-term stable software platforms

# Thank you for your Attention!

Questions / Comments?

**Dirk von Suchodoletz**

Chair on Communication Systems
Faculty of Engineering
University of Freiburg / Germany
*dirk.von.suchodoletz@rz.uni-freiburg.de*