Home Archiving:
Moving Digital Preservation Capabilities from Large Institutions to SMEs and Home Users

Andreas Rauber

Department of Software Technology and Interactive Systems

Vienna University of Technology

www.ifs.tuwien.ac.at/~andi

Motivation

- Digital Preservation R&D predominantly for large, professional institutions (libraries, archives, museums)
- Need for Digital Preservation (DP) solutions in
  - smaller institutions
  - SMEs
  - SOHO: Small Office / Home Office
  - individuals
- Currently hardly any solutions available
- CMS, no real preservation support apart from back-up
- Goal:
  A solution automating preservation activities so that they can be deployed in non-expert settings

Requirements

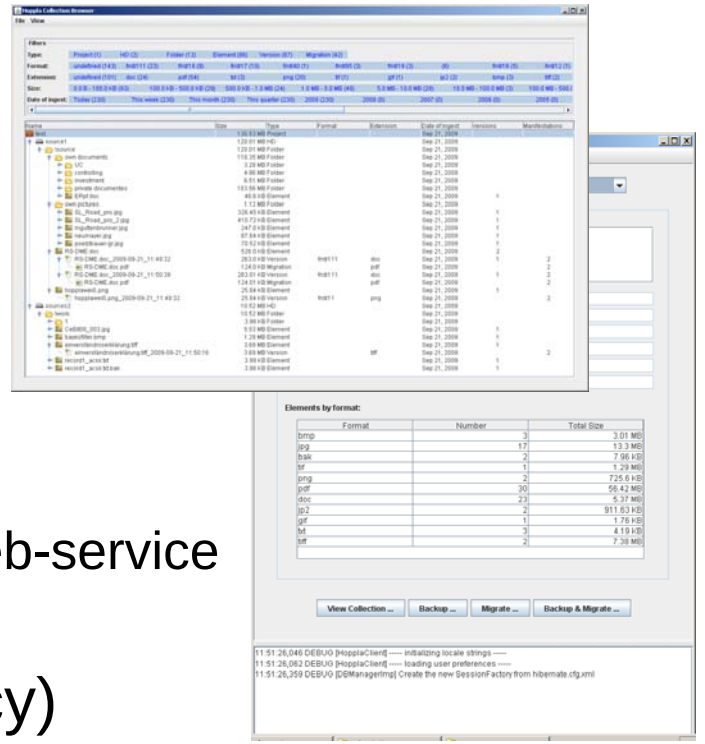
- Operate with little or no DP expertise
- Operate with little or no IT expertise
- Effortless operations: not core business, infrastructure support
- Highly automated
- High level of robustness
  - non-professional interaction with system
  - recovery on complete loss of core management system
- Support for common (also non-archival) media for storage
- “Easier” challenges, more focused collections
- Probably weaker requirements on quality levels

HOPPLA

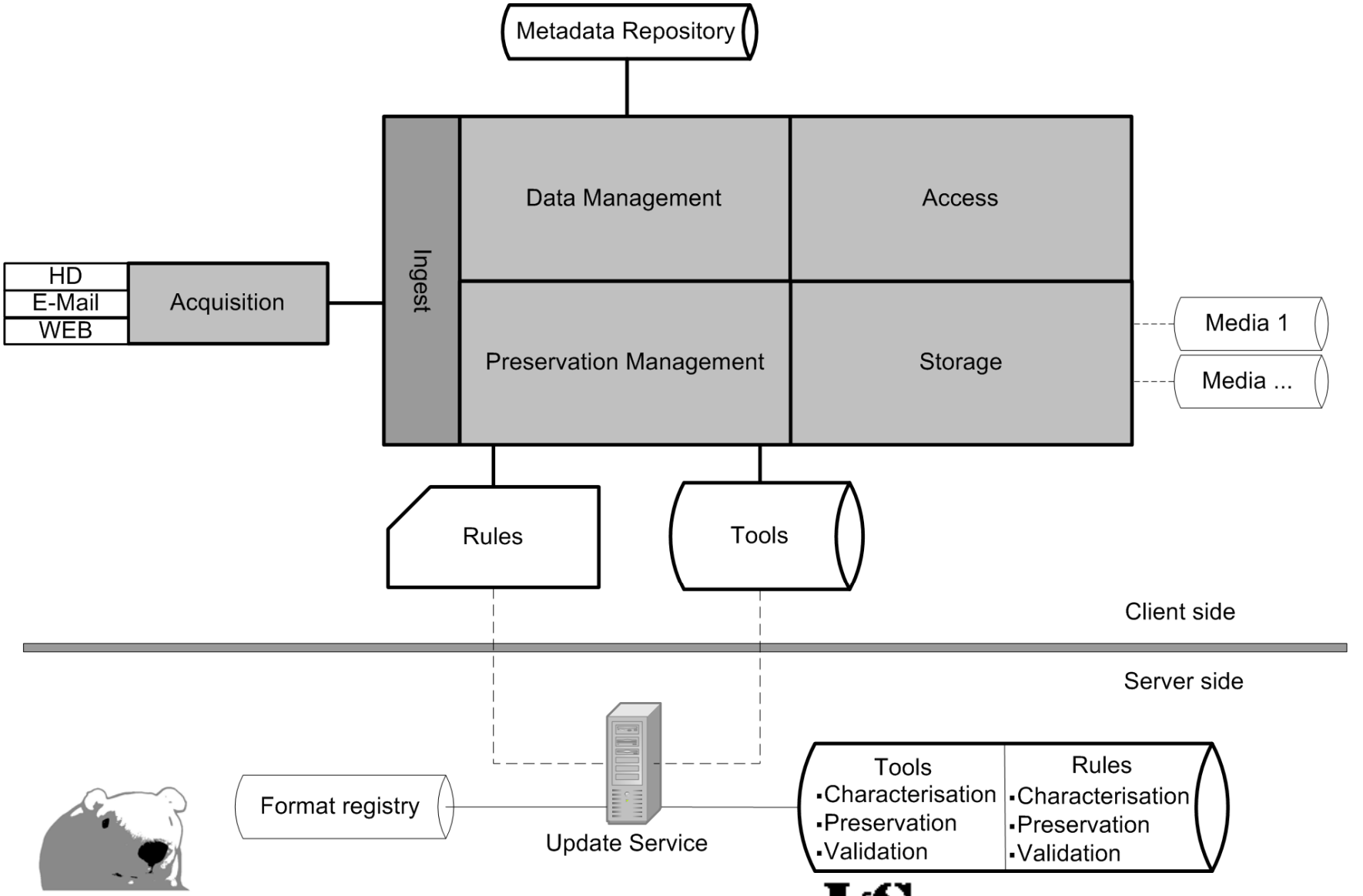
- **HOPPLA:**
  Home Office Painless Persistent Long-term Archiving
- Client-server system
- Inspired by: antivirus-SW and software firewall solutions
  - data resides with client
  - server provides DP expertise and solutions
- Support for
  - ingest of data from different sources (home, e-mail, on-line)
  - multiple back-ups on (also low-end) storage (DVD, ext. HDD, RAID systems, on-line storage)
  - recovery on loss of system data
- Focus on robustness and automation
- Meet requirements of audit and certification initiatives

HOPPLA

- Developed in Java (plattform-independent)
- Combination of back-up and DP
- Outsourcing of expertise
- Flexible client – server architecture
  - rules
  - tools (mostly plattform-dependent, plus web-service based solutions on request)
- Data remain only on client side (privacy)
- Metadata provided by external experts and automated tools/services



HOPPLA

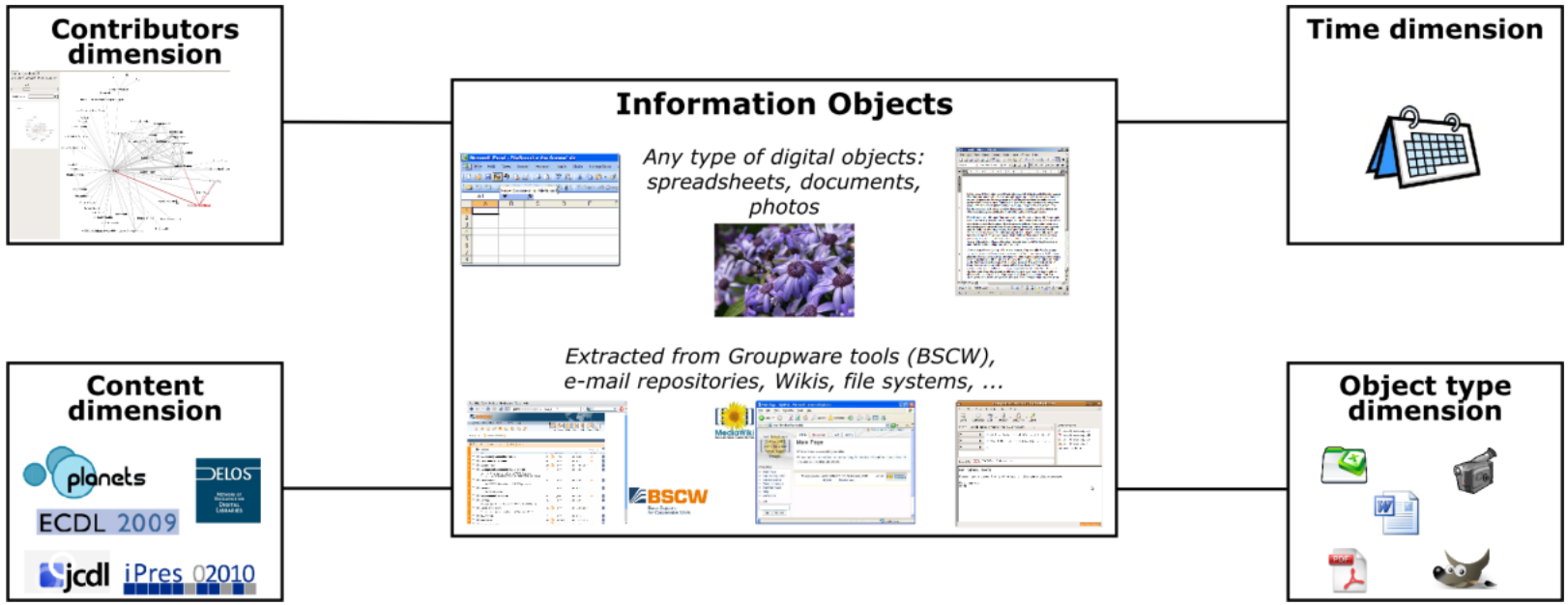

Metadata Repository
Ingest
Data Management
Access
Preservation Management
Storage
HD
E-Mail
WEB
Acquisition
Media 1
Media ...
Rules
Tools
Client side
Server side
Format registry
Update Service
Tools
▪Characterisation
▪Preservation
▪Validation
Rules
▪Characterisation
▪Preservation
▪Validation

Core Workflow

- Ingest from different source media
  - Creation of collection profile (technical metadata)
  - Extract / estimate contextual information from source info
- Collection profile is sent to server
  (adjustable level of detail – not implemented yet)
  - Experts on server side provide registry with preservation plans (Plato)
  - Appropriate preservation plan is chosen according to user profile (data volumes, risk level, cost/benefit settings)
  - Preservation action plan sent to client
- Client performs migration activities
  - Data stored redundantly
  - Media refresh

HOPPLA

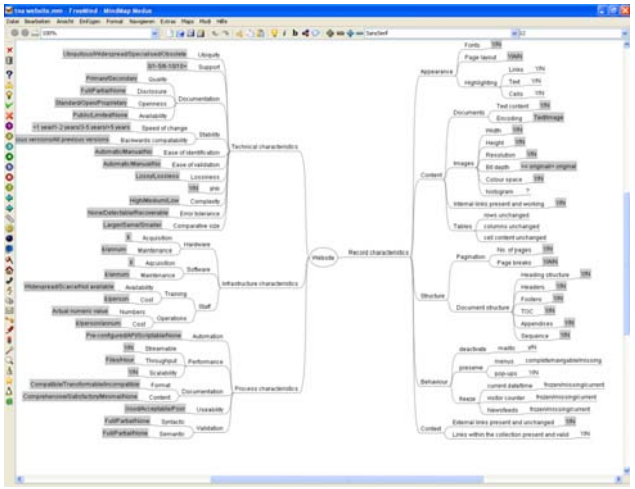
- Technical Metadata: JHOVE, Pronom, XCDL, …
- Semantic metadata: utilising context of objects
  - Extract context using IR, IE and NLP techniques
  - Organise objects along multiple dimensions (DWH-inspired)
  - Finding groups of related objects (semi-automatic)

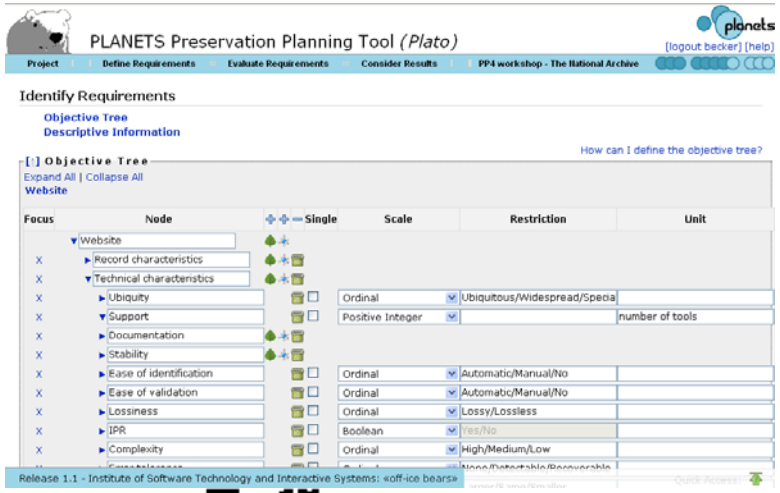


Preservation Planning

Server-side: Preservation Planning

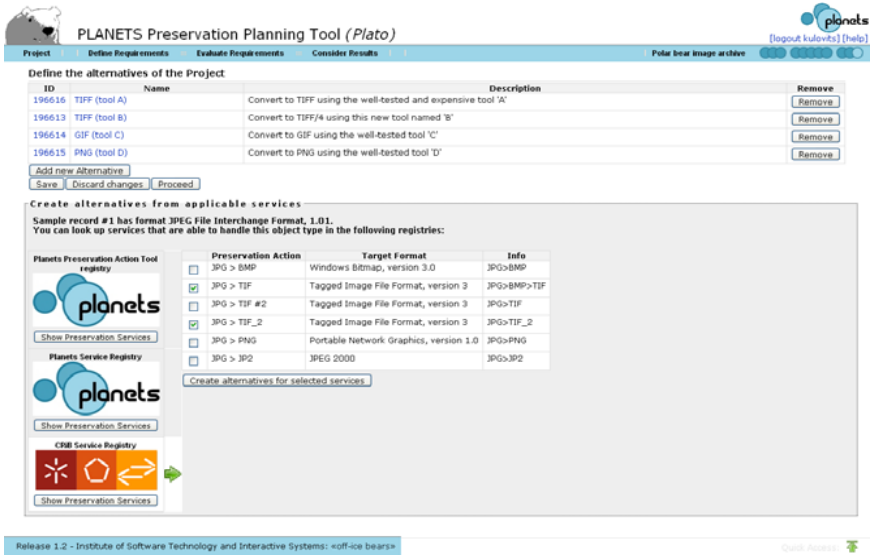
- Plato Preservation Planning Tool
- Implements Planets Preservation Planning Workflow
- Allows creation of objective tree
  - within application or via import of mindmaps
- Allows the selection of Preservation action tools to be evaluated

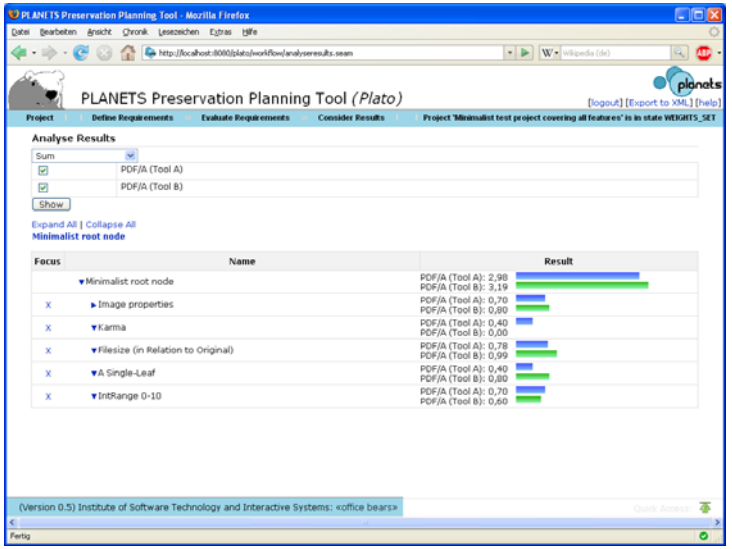




Preservation Planning

- Runs preservation action experiments, documents results
- Allows definition of transformation rules, weightings
- Performs evaluation, sensitivity analysis,
- Provides recommendation (ranks solutions)
- Stores/exports preservation plan with evidence, triggers,…





Preservation Planning

- Preservation Planning to ensure “optimal” preservation
- Operated by experts on server side
- A simple, methodologically sound model to specify and document requirements
- Repeatable and documented evaluation
- Basis for well-informed, accountable decisions
- Concretization of OAIS model
- Follows recommendations of TRAC and nestor
- Plato:
  - tool support to perform solid, well-documented analyses
  - creates core preservation plan
- Rules to match preservation requirements with plans

Preservation Planning

- Server side has risk profile for object types
- Client side has preferences with
  - user / institution implicit preferences (e.g. obj. importance, …)
  - degree of risk avoidance based on object type, file size,…
  - preferences in terms of storage space availability / cost
- Objects identified at some risk level are matched with preservation plans
- List of potential preservation actions:
  - tools installed on client side
  - tools wrapped as plug-ins
  - external tools potentially to be installed at client side (license / cost / willingness to install)
  - external webservice (if client is willing to send data)
- Identify most suitable solution

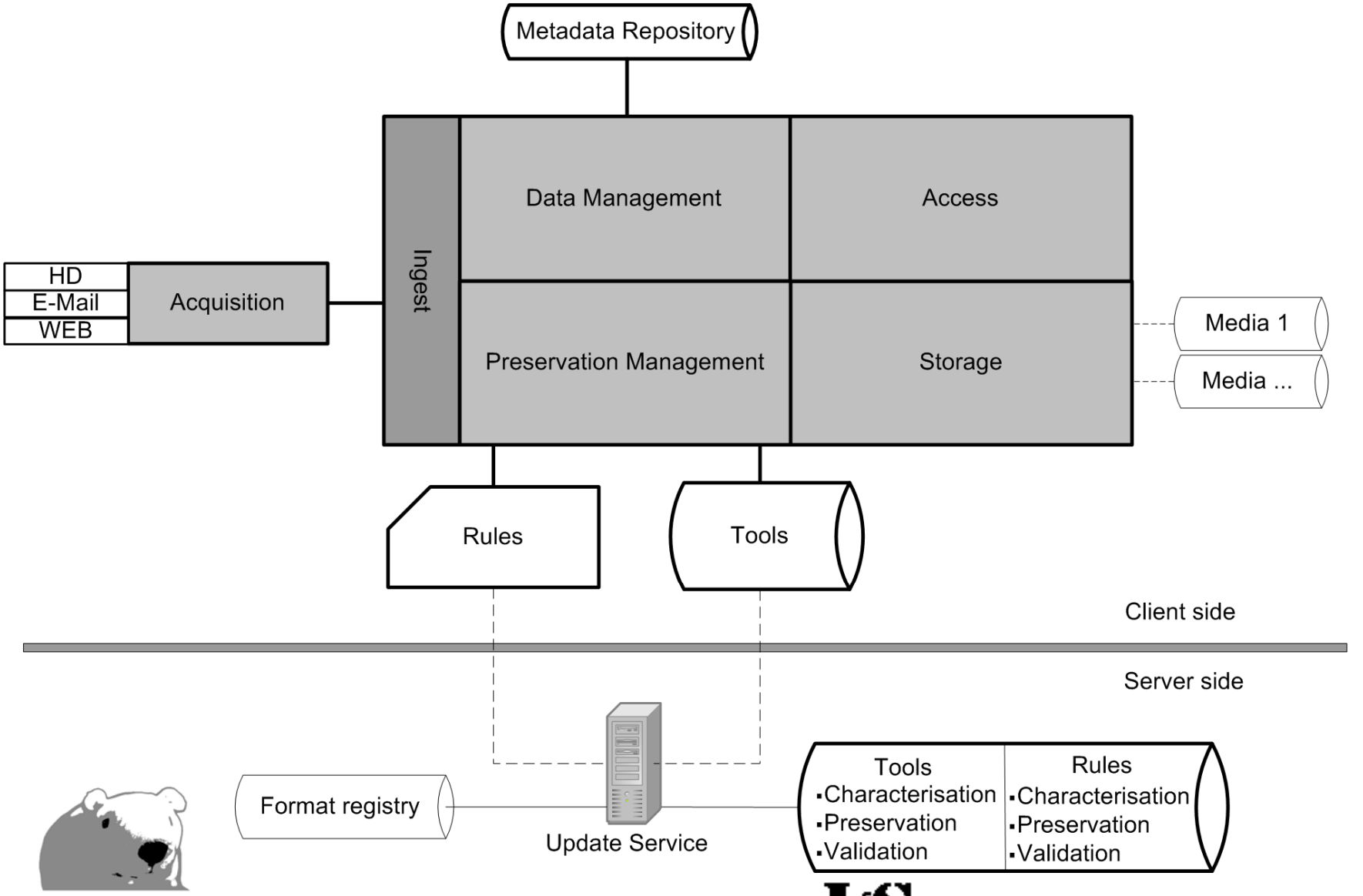
HOPPLA

- Server receives form client
  - collection profile (technical metadata)
  - preferences settings
  - list of locally available tools
  - level of detail will be configurable to meet privacy requirements
- Server provides to client
  - Preservation action plan: recommended preservation actions
  - Preservation plan for documentation / evidence
  - wrapped migration tools or install packages (potentially also emulators)
- Client
  - performs migration actions and redundant storage

Preservation Action Plan

```
[…]
<migrationToolId>1</migrationToolId>
<constraint>
  <metadataConstraints><metadataConstraint>
  <characterisationRuleId>1</characterisationRuleId>
  <comparisonMode>Smaller</comparisonMode>
  <metadataId>height</metadataId>
  <constraintValue>10000</constraintValue>
  </metadataConstraint><metadataConstraint>
 […]
  <minSize>3</minSize>
  <maxSize>10</maxSize>
[…]
<riskScore>bestpractice</ riskScore>
<estChangeInSize>50%</estChangeInSize>
<estDurationPerMb>5</estDurationPerMb>
[…]
```

HOPPLA


Metadata Repository
Ingest
Data Management
Access
Preservation Management
Storage
HD
E-Mail
WEB
Acquisition
Media 1
Media ...
Rules
Tools
Client side
Server side
Format registry
Update Service
Tools
▪Characterisation
▪Preservation
▪Validation
Rules
▪Characterisation
▪Preservation
▪Validation

Storage

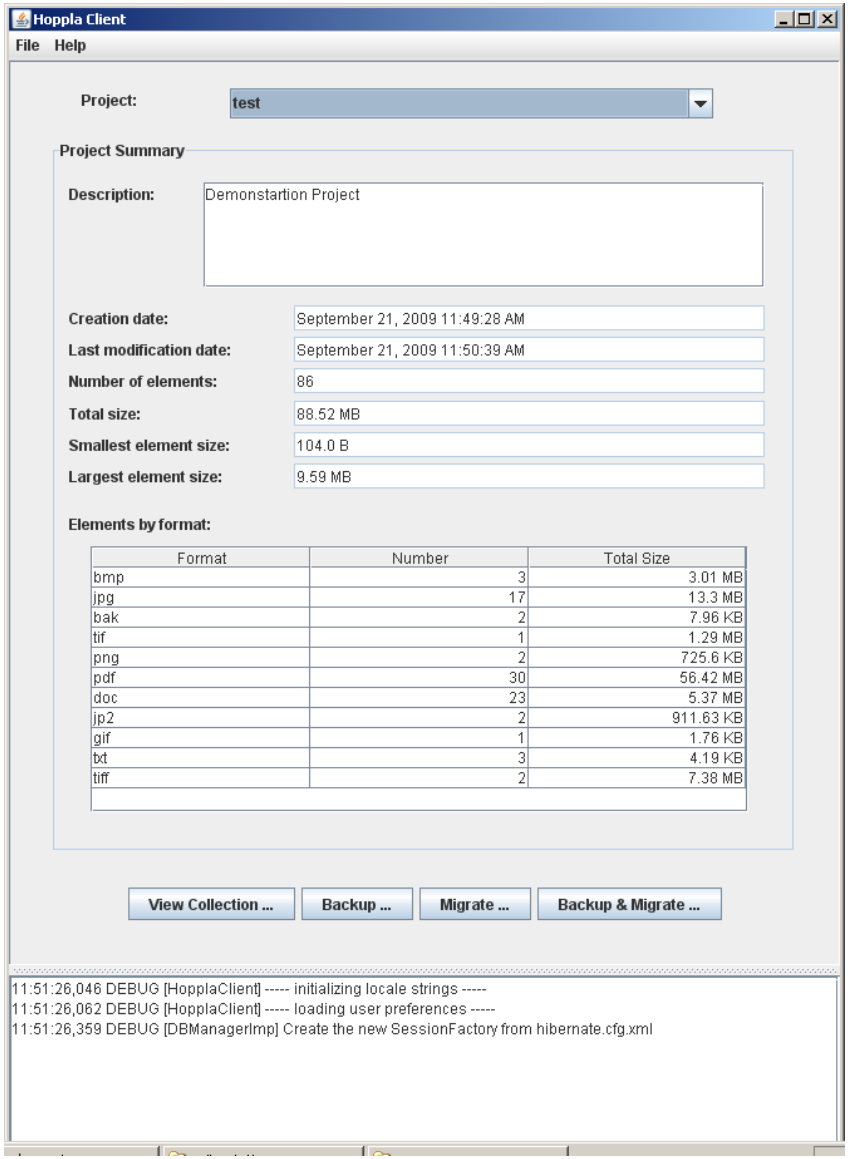
- Re-create source hierarchy
  - directory structure
  - mail folder structure
  - complex objects with sub-directories
- Locate each object in respective location
- Embed some metadata in filename (version, timestamp)
- Add XML file per directory with all object metadata in data management
- Bit-level preservation:
  configure degree of redundancy and media types
- media refreshment via reminders / automatic
- Can be used directly (even if somewhat limited) without an system

HOPPLA

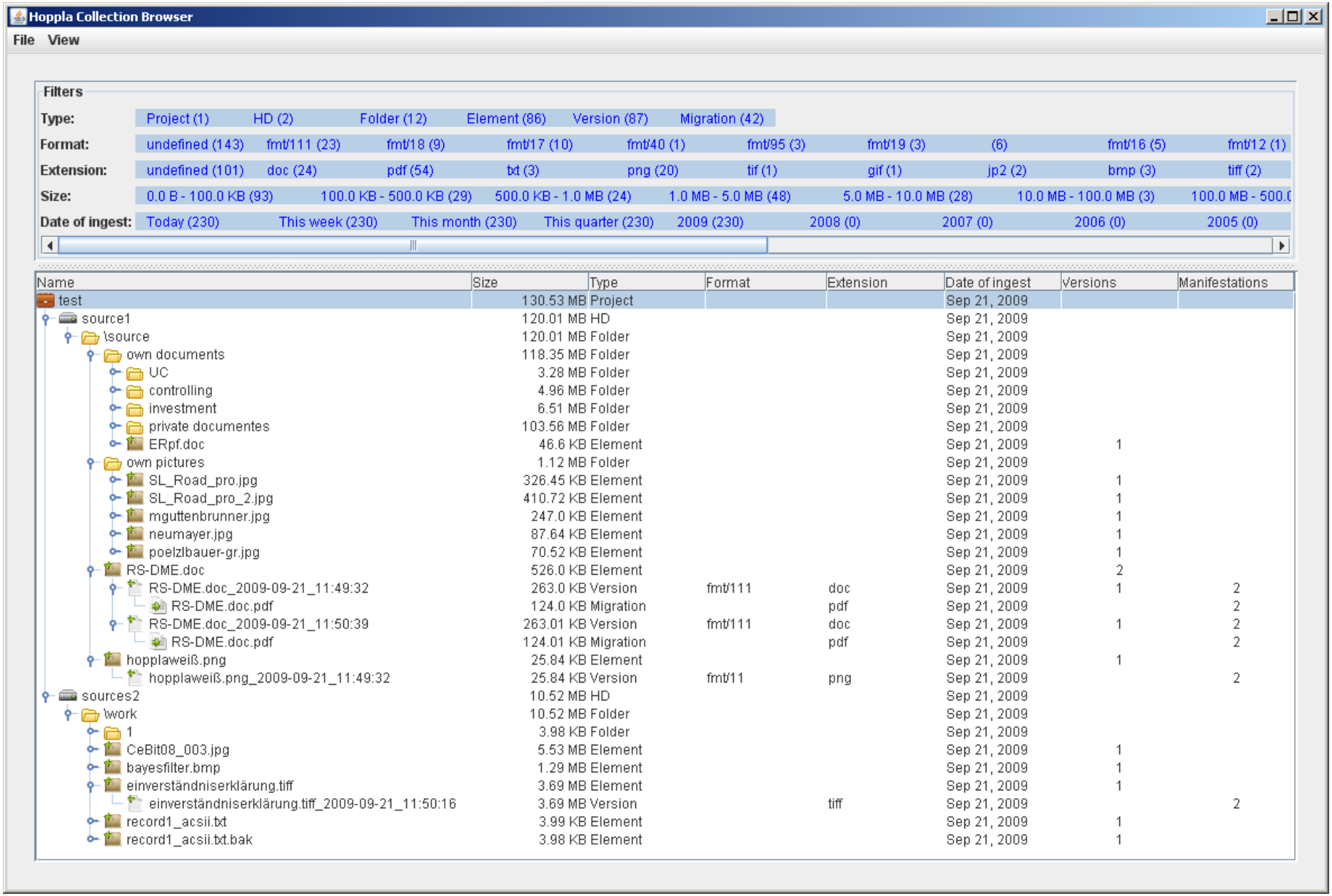
Access

- Via collection browser
  - representation of directory structure
  - archived versions and time-stamps
  - migrated versions and time stamps
  - operates via data management, faceted browser
- Can be used directly (even if somewhat limited) without any HOPPLA system
  - needs to be able to mount file system (FAT, ISO9660, on-line)
  - use physical directory structure
  - limited set of metadata in filenames
  - detailed information (basis for recovery) in XML file
- Preservation Plans as regular objects in dedicated directory

Hoppla - Software Prototype

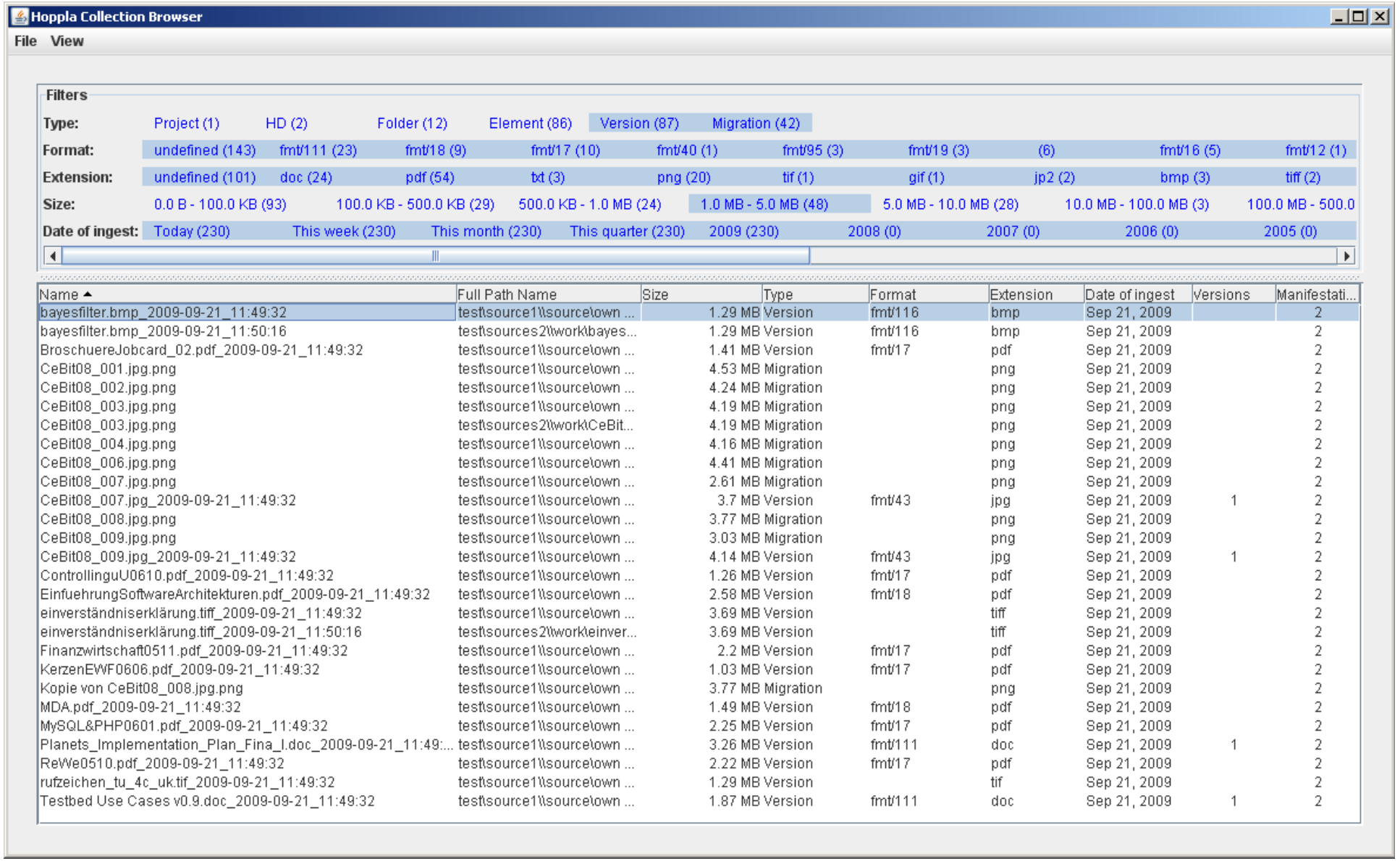


Hoppla - Software Prototype



Name	Size	Type	Format	Extension	Date of ingest	Versions	Manifestations
test	130.53 MB	Project			Sep 21, 2009		
source1	120.01 MB	HD			Sep 21, 2009		
\source	120.01 MB	Folder			Sep 21, 2009		
own documents	118.35 MB	Folder			Sep 21, 2009		
UC	3.28 MB	Folder			Sep 21, 2009		
controlling	4.96 MB	Folder			Sep 21, 2009		
investment	6.51 MB	Folder			Sep 21, 2009		
private documentes	103.56 MB	Folder			Sep 21, 2009		
ERpf.doc	46.6 KB	Element			Sep 21, 2009	1	
own pictures	1.12 MB	Folder			Sep 21, 2009		
SL_Road_pro.jpg	326.45 KB	Element			Sep 21, 2009	1	
SL_Road_pro_2.jpg	410.72 KB	Element			Sep 21, 2009	1	
mguttenbrunner.jpg	247.0 KB	Element			Sep 21, 2009	1	
neumayer.jpg	87.64 KB	Element			Sep 21, 2009	1	
poelzlbauer-gr.jpg	70.52 KB	Element			Sep 21, 2009	1	
RS-DME.doc	526.0 KB	Element			Sep 21, 2009	2	
RS-DME.doc_2009-09-21_11:49:32	263.0 KB	Version	fmt/111	doc	Sep 21, 2009	1	2
RS-DME.doc.pdf	124.0 KB	Migration		pdf	Sep 21, 2009		2
RS-DME.doc_2009-09-21_11:50:39	263.01 KB	Version	fmt/111	doc	Sep 21, 2009	1	2
RS-DME.doc.pdf	124.01 KB	Migration		pdf	Sep 21, 2009		2
hopplaweiß.png	25.84 KB	Element			Sep 21, 2009	1	
hopplaweiß.png_2009-09-21_11:49:32	25.84 KB	Version	fmt/11	png	Sep 21, 2009		2
sources2	10.52 MB	HD			Sep 21, 2009		
\work	10.52 MB	Folder			Sep 21, 2009		
1	3.98 KB	Folder			Sep 21, 2009		
CeBit08_003.jpg	5.53 MB	Element			Sep 21, 2009	1	
bayesfilter.bmp	1.29 MB	Element			Sep 21, 2009	1	
einverständniserklärung.tiff	3.69 MB	Element			Sep 21, 2009	1	
einverständniserklärung.tiff_2009-09-21_11:50:16	3.69 MB	Version		tiff	Sep 21, 2009		2
record1_acsii.txt	3.99 KB	Element			Sep 21, 2009	1	
record1_acsii.txt.bak	3.98 KB	Element			Sep 21, 2009	1	

Hoppla - Software Prototype



Name	Full Path Name	Size	Type	Format	Extension	Date of ingest	Versions	Manifestati...
bayesfilter.bmp_2009-09-21_11:49:32	test\source1\\source\own ...	1.29 MB	Version	fmt/116	bmp	Sep 21, 2009		2
bayesfilter.bmp_2009-09-21_11:50:16	test\sources2\\work\bayes...	1.29 MB	Version	fmt/116	bmp	Sep 21, 2009		2
BroschuereJobcard_02.pdf_2009-09-21_11:49:32	test\source1\\source\own ...	1.41 MB	Version	fmt/17	pdf	Sep 21, 2009		2
CeBit08_001.jpg.png	test\source1\\source\own ...	4.53 MB	Migration		png	Sep 21, 2009		2
CeBit08_002.jpg.png	test\source1\\source\own ...	4.24 MB	Migration		png	Sep 21, 2009		2
CeBit08_003.jpg.png	test\source1\\source\own ...	4.19 MB	Migration		png	Sep 21, 2009		2
CeBit08_003.jpg.png	test\sources2\\work\CeBit...	4.19 MB	Migration		png	Sep 21, 2009		2
CeBit08_004.jpg.png	test\source1\\source\own ...	4.16 MB	Migration		png	Sep 21, 2009		2
CeBit08_006.jpg.png	test\source1\\source\own ...	4.41 MB	Migration		png	Sep 21, 2009		2
CeBit08_007.jpg.png	test\source1\\source\own ...	2.61 MB	Migration		png	Sep 21, 2009		2
CeBit08_007.jpg_2009-09-21_11:49:32	test\source1\\source\own ...	3.7 MB	Version	fmt/43	jpg	Sep 21, 2009	1	2
CeBit08_008.jpg.png	test\source1\\source\own ...	3.77 MB	Migration		png	Sep 21, 2009		2
CeBit08_009.jpg.png	test\source1\\source\own ...	3.03 MB	Migration		png	Sep 21, 2009		2
CeBit08_009.jpg_2009-09-21_11:49:32	test\source1\\source\own ...	4.14 MB	Version	fmt/43	jpg	Sep 21, 2009	1	2
ControllinguU0610.pdf_2009-09-21_11:49:32	test\source1\\source\own ...	1.26 MB	Version	fmt/17	pdf	Sep 21, 2009		2
EinfuehrungSoftwareArchitekturen.pdf_2009-09-21_11:49:32	test\source1\\source\own ...	2.58 MB	Version	fmt/18	pdf	Sep 21, 2009		2
einverständniserklärung.tiff_2009-09-21_11:49:32	test\source1\\source\own ...	3.69 MB	Version		tiff	Sep 21, 2009		2
einverständniserklärung.tiff_2009-09-21_11:50:16	test\sources2\\work\einver...	3.69 MB	Version		tiff	Sep 21, 2009		2
Finanzwirtschaft0511.pdf_2009-09-21_11:49:32	test\source1\\source\own ...	2.2 MB	Version	fmt/17	pdf	Sep 21, 2009		2
KerzenEWF0606.pdf_2009-09-21_11:49:32	test\source1\\source\own ...	1.03 MB	Version	fmt/17	pdf	Sep 21, 2009		2
Kopie von CeBit08_008.jpg.png	test\source1\\source\own ...	3.77 MB	Migration		png	Sep 21, 2009		2
MDA.pdf_2009-09-21_11:49:32	test\source1\\source\own ...	1.49 MB	Version	fmt/18	pdf	Sep 21, 2009		2
MySQL&PHP0601.pdf_2009-09-21_11:49:32	test\source1\\source\own ...	2.25 MB	Version	fmt/17	pdf	Sep 21, 2009		2
Planets_Implementation_Plan_Fina_l.doc_2009-09-21_11:49:...	test\source1\\source\own ...	3.26 MB	Version	fmt/111	doc	Sep 21, 2009	1	2
ReWe0510.pdf_2009-09-21_11:49:32	test\source1\\source\own ...	2.22 MB	Version	fmt/17	pdf	Sep 21, 2009		2
rufzeichen_tu_4c_uk.tif_2009-09-21_11:49:32	test\source1\\source\own ...	1.29 MB	Version		tif	Sep 21, 2009		2
Testbed Use Cases v0.9.doc_2009-09-21_11:49:32	test\source1\\source\own ...	1.87 MB	Version	fmt/111	doc	Sep 21, 2009	1	2

HOPPLA

Next steps

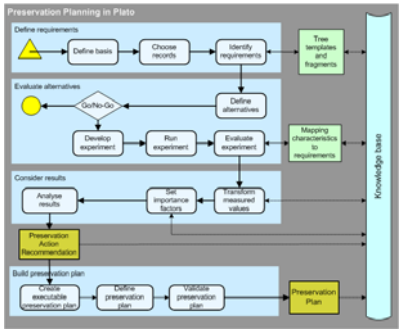
- Prototype development via DME-funded project
- Co-funding for adapting and integrating prototype in partner systems
- First functional prototype winter 2009 (internal, partner)
- Tighter coupling of preservation actions on client side and preservation planning on server side
- Design adaptations to allow more flexible integration into different systems
  - repository solutions
  - back-up systems
- Eventually better support for audit trails

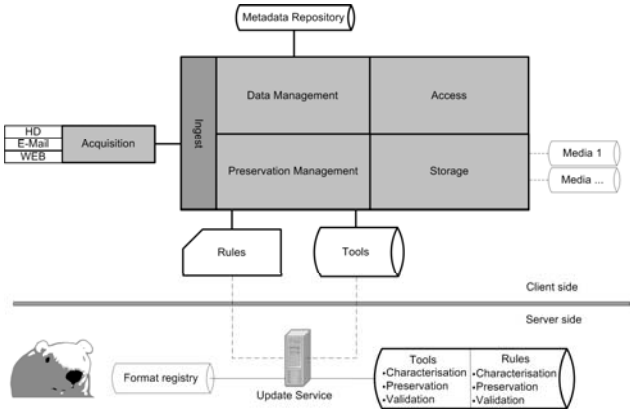
Conclusions

- Digital Preservation is a challenge for everybody
- Lack of solutions for small institutions / individuals
- Digital Preservation as a service
- Automation: metadata creation, preservation actions
- Server side:
  - Preservation planning
  - Tool provisioning
- Covering bit preservation and logical preservation
- Flexible adaption to needs via rule-based mappings
- Outsourcing of expertise

http://www.ifs.tuwien.ac.at/dp/hoppla

Thank you!





http://www.ifs.tuwien.ac.at/hoppla

http://www.ifs.tuwien.ac.at/dp/plato