# Trustworthy Preservation Planning

## Skalierbare vertrauenswürdige Planung: Entscheidungen in der digitalen Langzeitarchivierung

30. Mai 2011

Dagstuhl

**Christoph Becker**

Institut für Softwaretechnik und Interaktive Systeme
Technische Universität Wien

http://www.ifs.tuwien.ac.at/~becker

FACULTY OF !NFORMATICS

# Why do we need Digital Preservation?

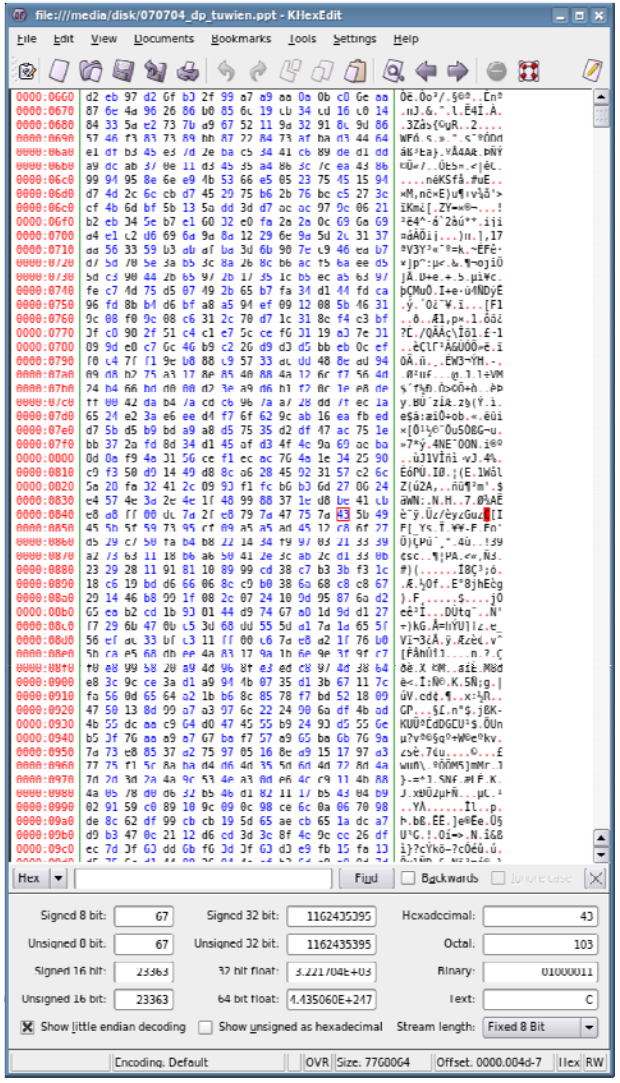
- .
  - .
  - .
  - .
- . .
  - .
  - .
  - rograms won‘t
  - .
- .

## 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.

# Why do we need Digital Preservation?





# Digital preservation and DP actions

- 1,2 zettabyte of digital information by 2010
  - From camera raw files to blogs to HEP experiment data
- Given a set of digital artifacts ...
  - ...each artifact needs a certain environment to fulfil their purpose
  - Threat levels: Physical, **Logical,** Semantic
- The mission of Digital Preservation
  - Address risks to authenticity and understandability
  - Provide understandable authentic content with optimal efficiency
- DP is a young field and of concern in different scenarios
  - Cultural heritage, space industry...
  - Repositories: BL, DNB,...
  - CERN, MSRC, SAP, IBM, ESA, Airbus, Helmholtz,...
  - ... almost everyone

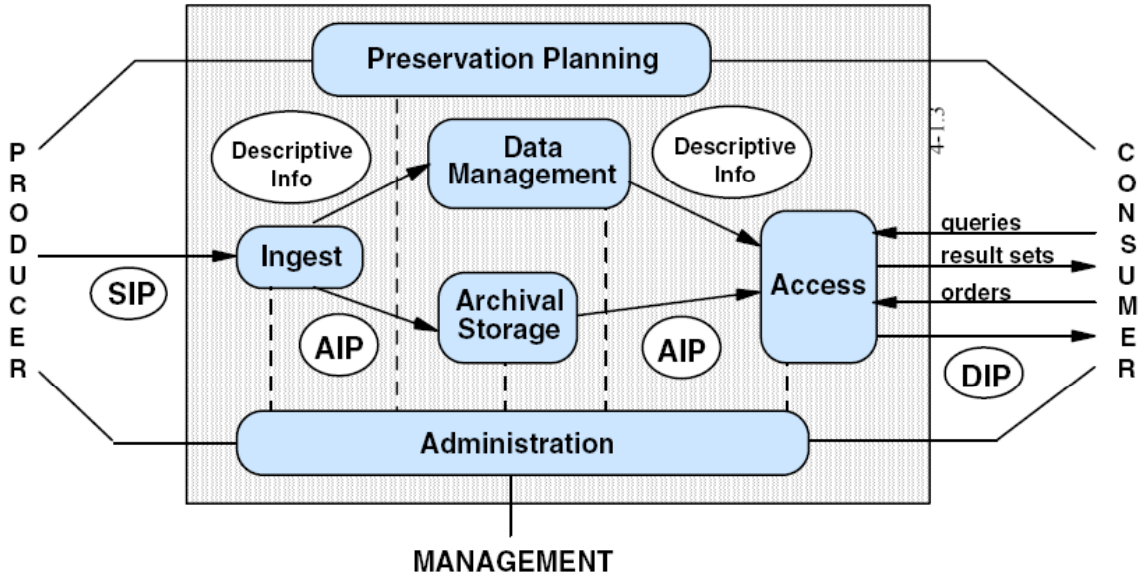
# Evaluating preservation actions

- Several actions available (migration, emulation,...)
- Challenges:
  - Quality varies across tools
  - Properties vary across content
  - Usage varies across communities
  - Requirements vary across scenarios
  - Risk tolerance varies across collections
  - Preferences and constraints vary across organisations
  - Cost structures and compatibility varies across environments
  - Constraints, priorities and requirements shift constantly
- Multi-criteria decision analysis problem
  - competing objectives
  - comparable to component selection

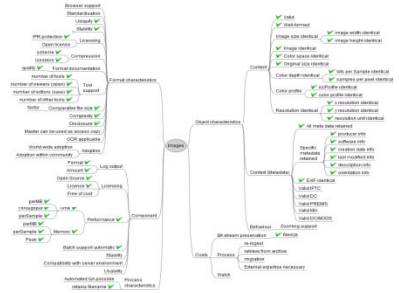
# Research Questions

1. How to select the optimal preservation action for a given scenario?
2. How to ensure trustworthy preservation planning?
3. How to enable decision processes to scale up?

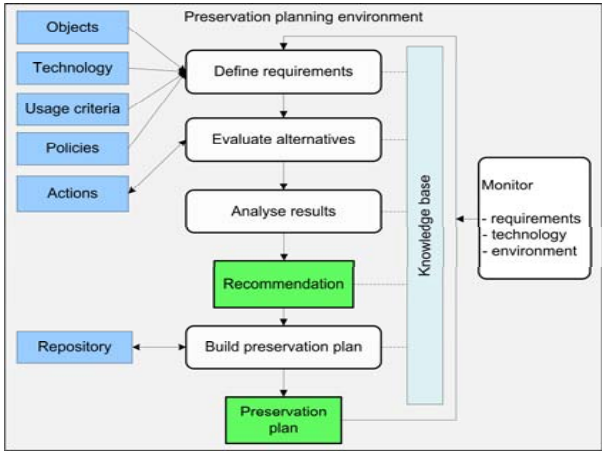
- Context: Trustworthy repositories
  - Legal mandates of national institutions
  - Open Archival Information Systems model (OAIS)
  - Trustworthy repositories criteria (ISO RAC, nestor)
  - What is a *preservation plan?*

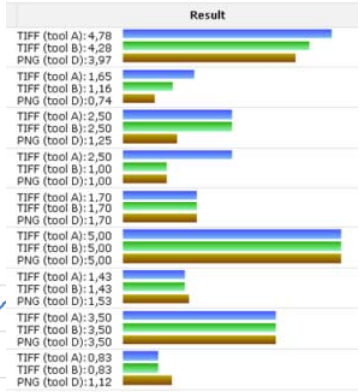


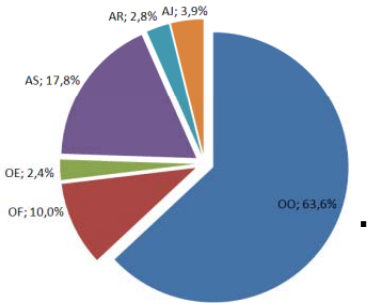
# Contributions

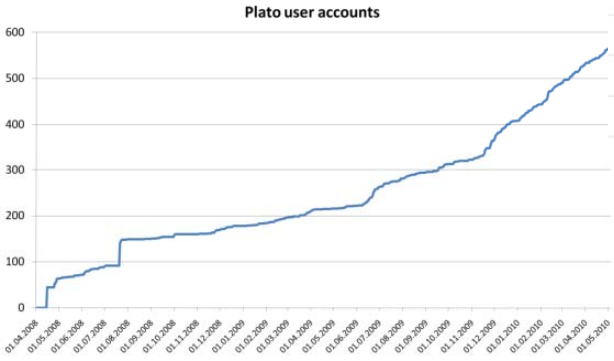


- Preservation planning
  - Definition, method, workflow and continuous monitoring [IJDL]
  - Component selection [SAC RE'09]
  - Case studies [ICADL'07, RCDL'07, iPRES'08, IJDC]
  - Critical assessment and gap analysis
- Planning tool Plato [JCDL'08]
  - Integration of planning and actions [ECDL'08]
  - Characterisation in planning [SAC DE'08, JUCS]
  - Controlled experimentation [INFSOF]
  - Quality-aware migration [ICWE'09, ECDL'09]
- Decision criteria and measurements
  - Measurement framework and coverage analysis [JASIST]
  - Decision making in DP [JCDL'11]



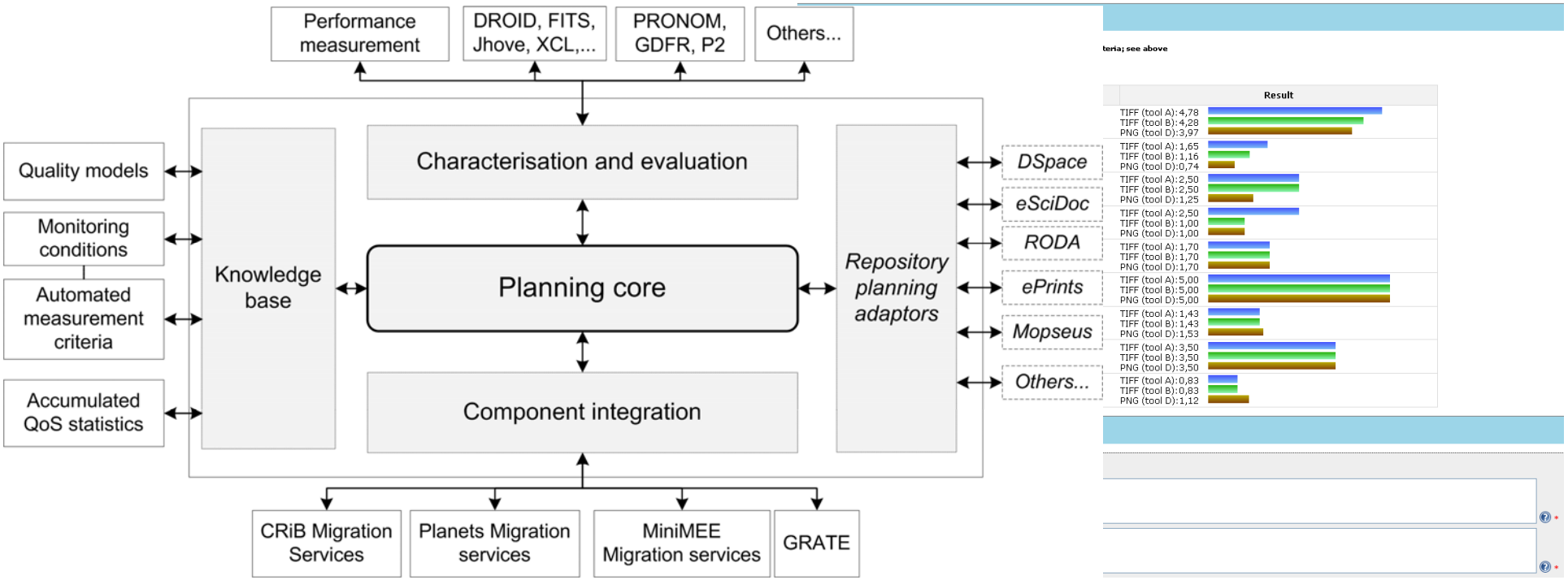



# Repeatable evidence for objective decisions

- Standardised workflow and automated documentation
  - Evaluate potential actions objectively against scenario-specific requirements in a repeatable way
- Objective tree: key element of requirements definition and assessment
  - weighted hierarchy of objectives leading to measurable criteria
  - *Criteria* are incommensurable → *utility function* for each criterion specifying the organisation's assessment for the range of possible values
  - Wide range of influence factors are relevant
  - What are the essential properties of a digital object, intellectually and technically?
- Controlled experimentation on sample content
- A framework for automated measurement

# The planning tool PLATO

- Plato implements the planning workflow
  - 14 steps in four phases
  - Knowledge base contains quality models and template trees
  - Automated experiments, action discovery and invocation
  - Visual analysis of results, plan specification, traceable documentation
- http://www.ifs.tuwien.ac.at/dp/plato

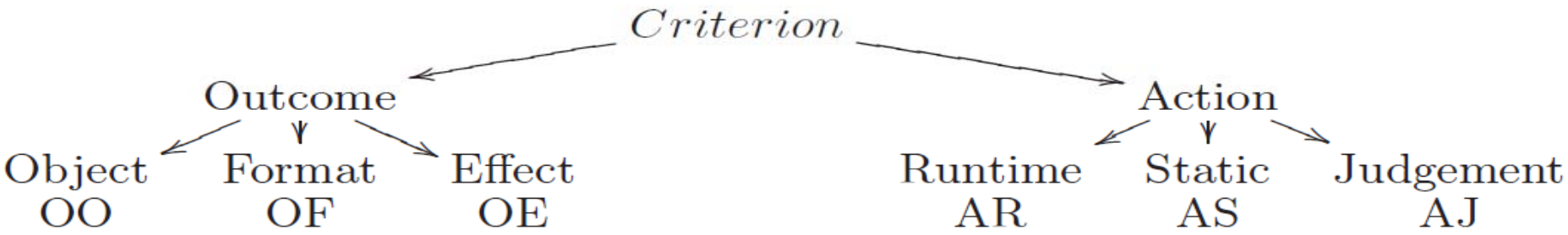


# Case studies

- Variety of case studies
  - Electronic theses
  - Web pages
  - Relational databases
  - Interactive art and computer games
- Scanned images in national libraries
  - British Library, 80TB TIFF5
  - Bavarian State Library, 72TB TIFF6
  - Royal Library of Denmark, ~10.000 aerial photographs,TIFF6
- Validated the decision making framework
  - Trustworthy? Yes. Scalable? No… Open issue: automation
  - Manual effort for evaluation is prohibitive
- Analysis of ~600 decision criteria from case studies →

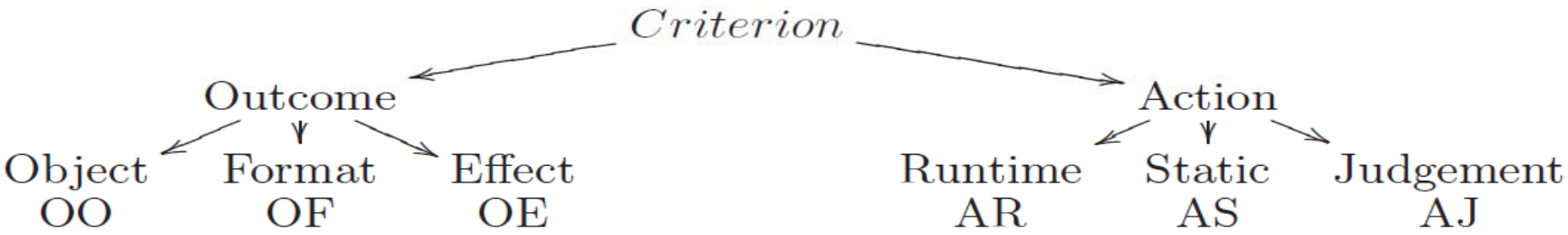
# Decision criteria

- Each criterion concerns either the action or its outcome
- Outcome
  - Object (authenticity, editability, …)
  - Format (licensing, standardisation, complexity…)
  - Effect (Costs…)
- Action
  - Runtime properties (performance, stability, logging…)
  - Static (price, license…)
  - Judgement (configuration interface usability…)



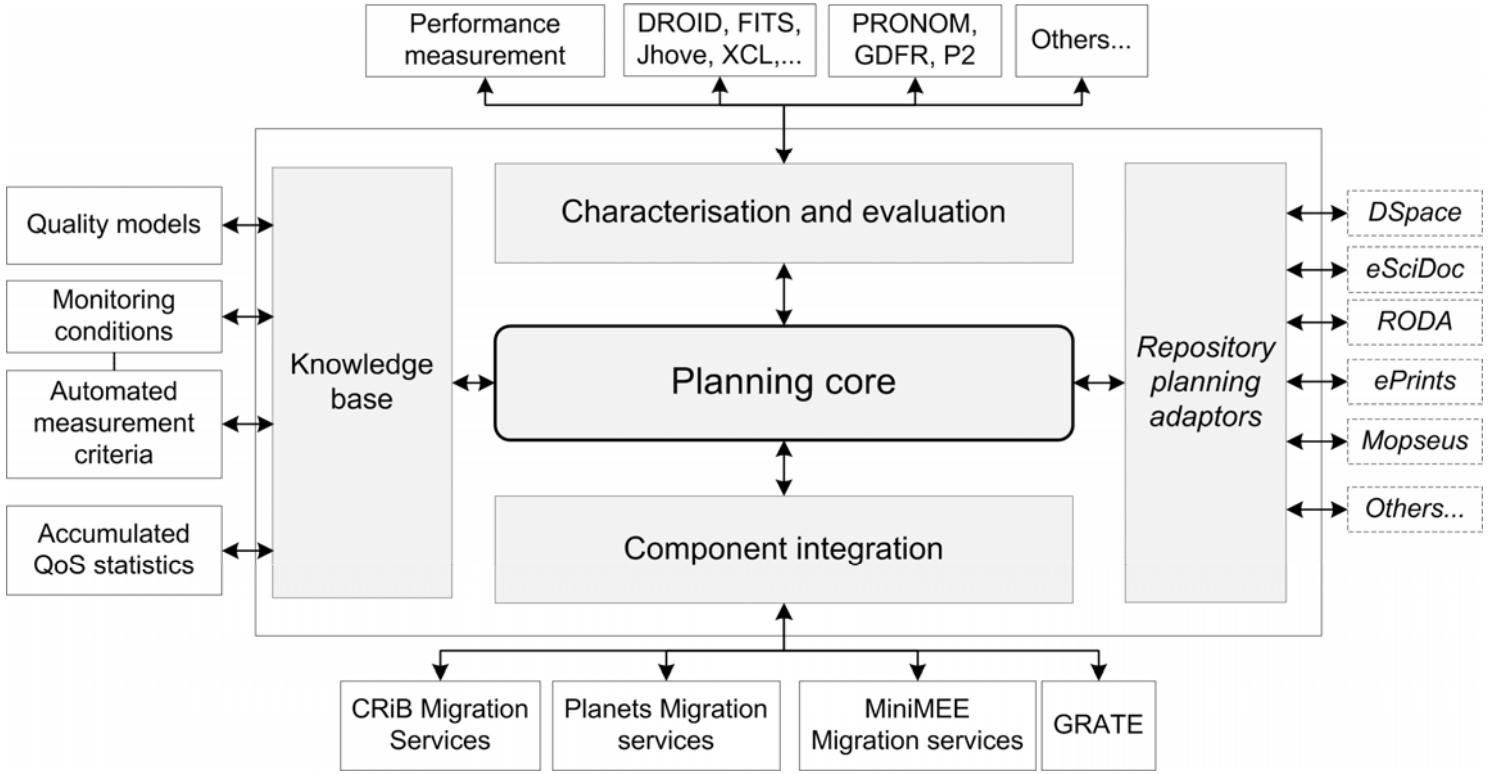
# Decision criteria

- Each criterion concerns either the action or its outcome
- **Outcome**
  - **Object** (authenticity, editability, …)
  - **Format** (licensing, standardisation, complexity…)
  - Effect (Costs…)
- **Action**
  - **Runtime** properties (performance, stability, logging…)
  - Static (price, license…)
  - Judgement (configuration interface usability…)



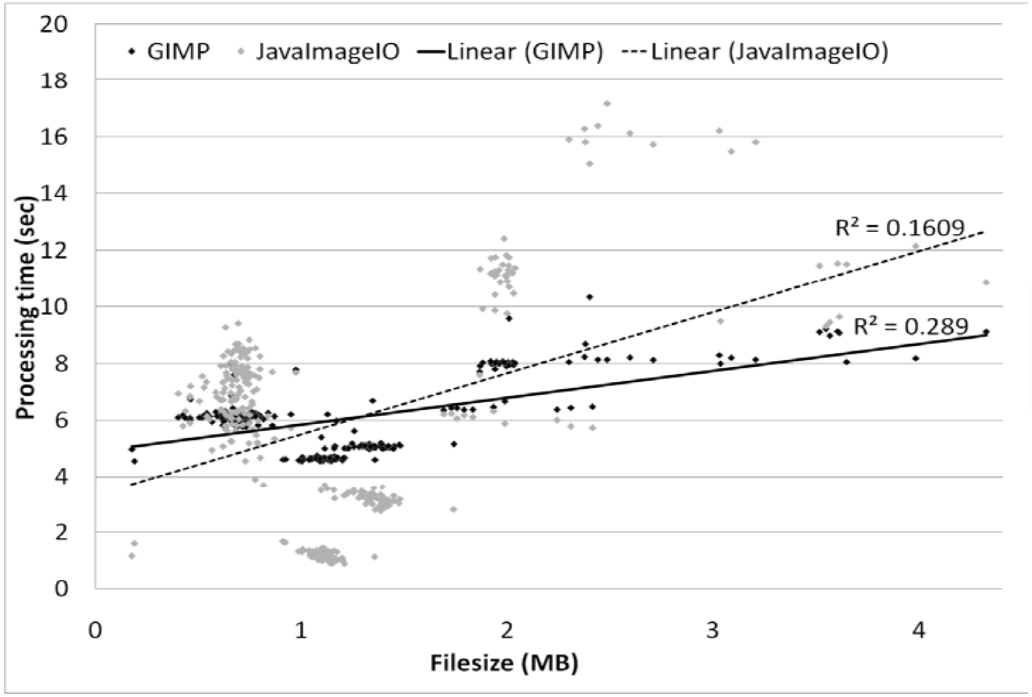
# Controlled experimentation in DP

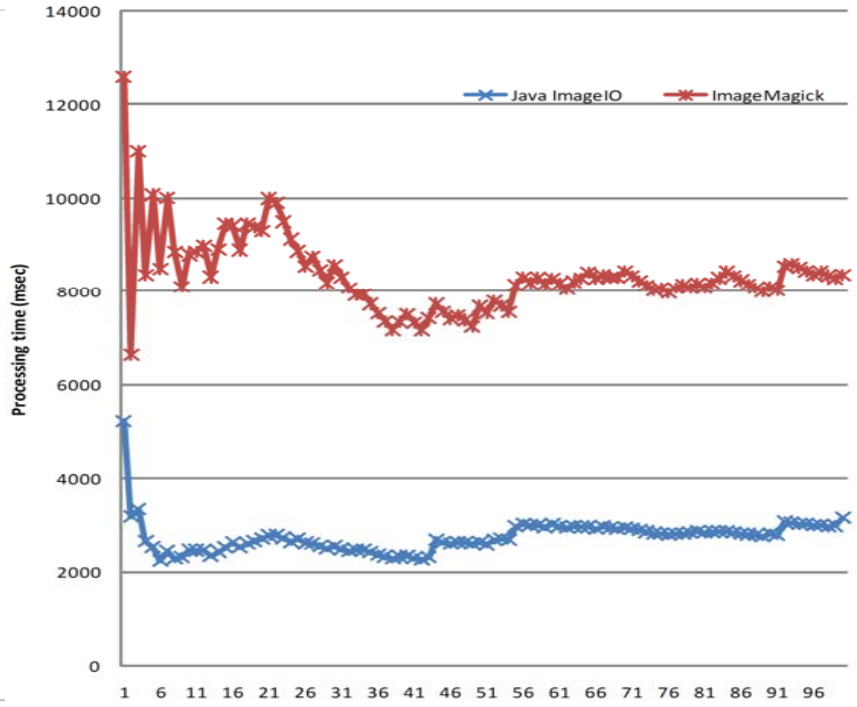
- SOAs for digital preservation
  - Dynamic discovery and invocation of preservation actions
  - Web services for loose coupling and flexible integration
  - Little or no information about process on provider side
  - Intermediaries, Sniffing, Probing, Instrumentation



# MiniMEE: A quality-aware migration engine

- Non-invasive provider-side service instrumentation
- Migration engine monitors components at runtime
- Transparent invocation in controlled environment
- Performance information delivered to the requester
- Experience accumulation, benchmarking

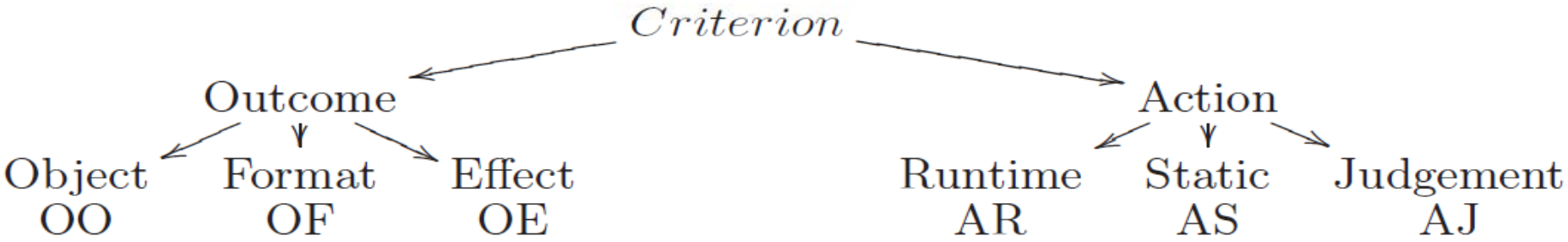


# Extensible measurement framework

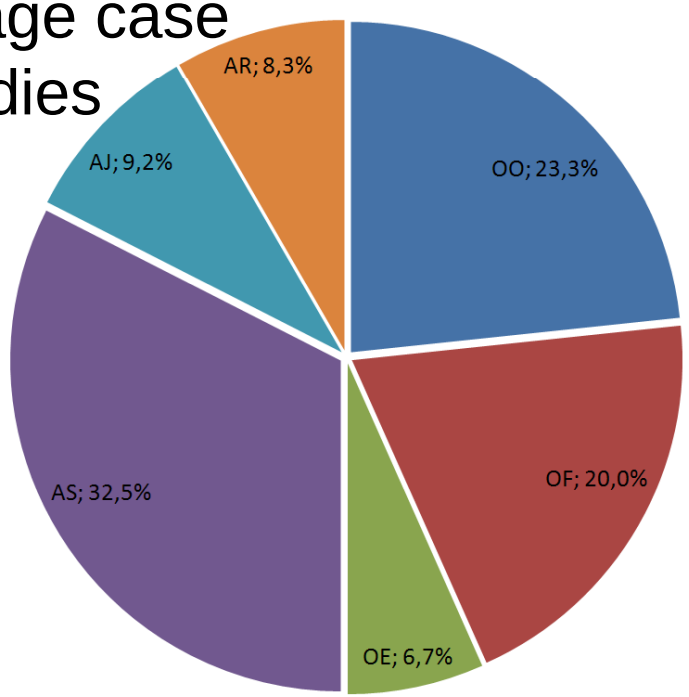
- Evaluators declare the properties they can measure
  - Decision criteria linked to measurements
  - Integration of characterisation tools
    - *Object format conformance validation*
  - Comparison of objects
    - *Image similarity algorithms*
    - *Embedded metadata consistency validation*
  - Runtime profiling of actions in controlled environments
    - *Peak memory usage, used cpu cycles...*
  - Access to external information sources
    - *How many tools are able to render this object?*
- Substantial improvement in repeatability, scalability,and trustworthiness of decisions
  - *Full coverage of “expensive” criteria for images*

# Case studies revisited



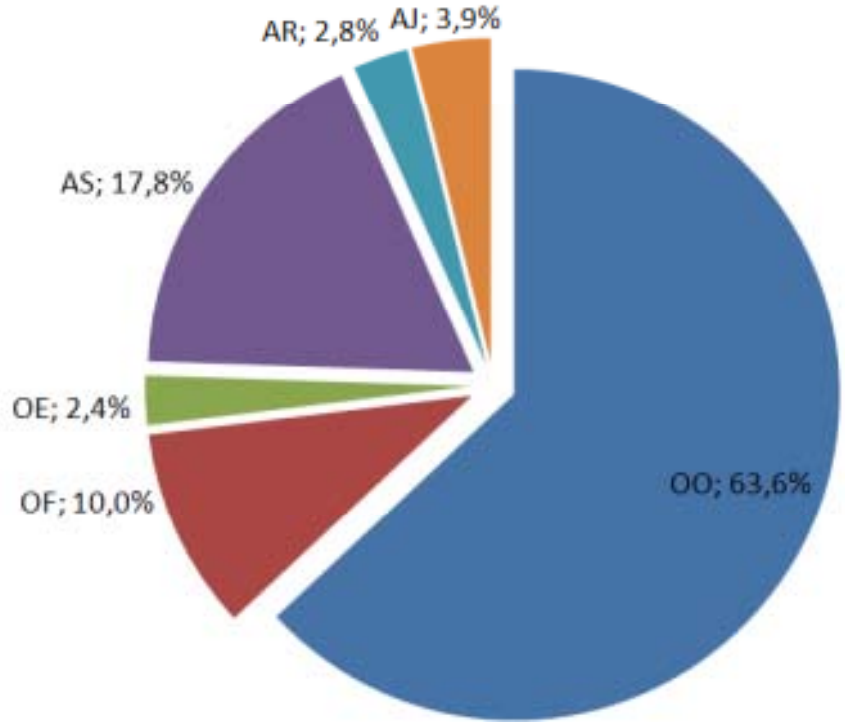

- Distribution of criteria across taxonomy
  - 13 case studies, 617 criteria

- Image case studies →



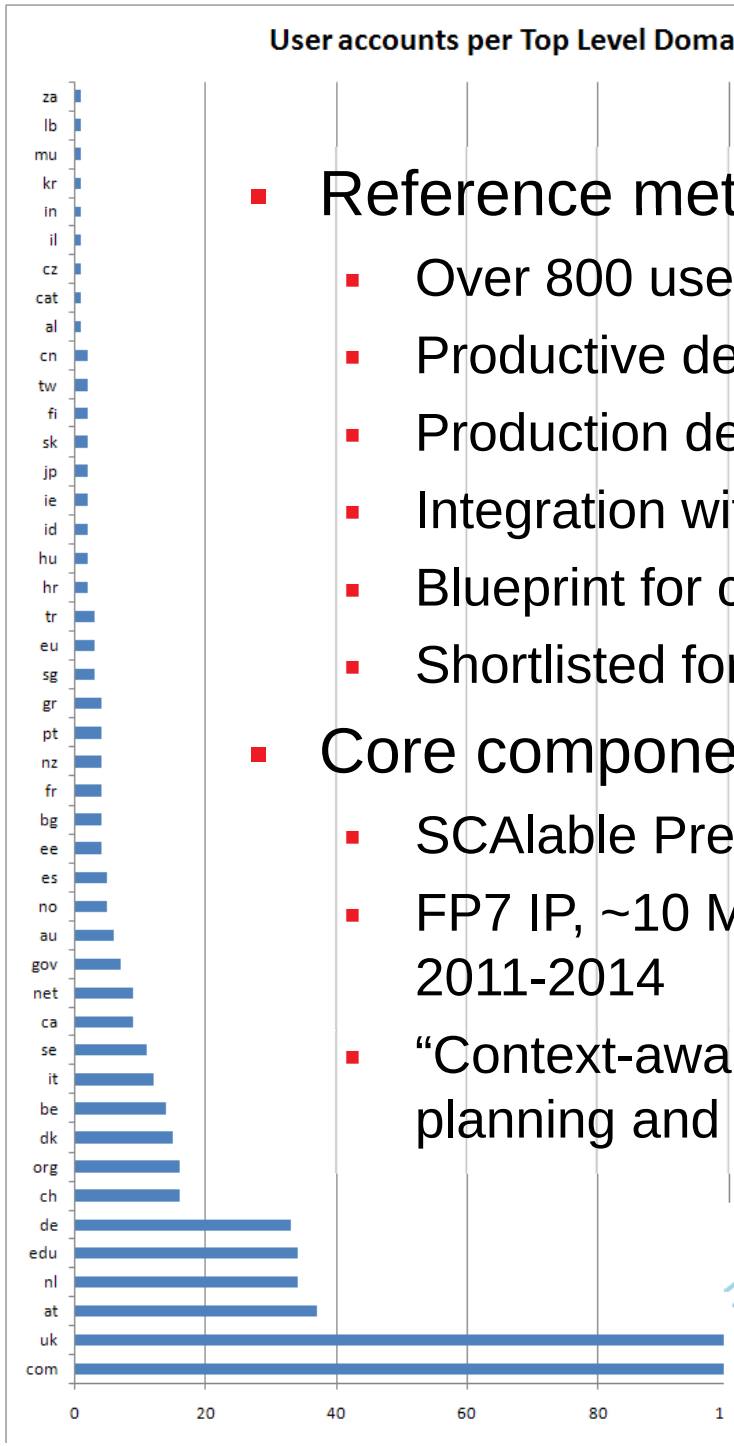

…

All cases:



# Takeup

- Reference method and tool for preservation planning
  - Over 800 users from 46 TLDs signed up for Plato
  - Productive decisions in national institutions
  - Production deployment at ÖSTA
  - Integration with ePrints, eSciDoc, RODA…
  - Blueprint for commercial implementation
  - Shortlisted for DP Award 2010
- Core component of new FP7 IP SCAPE
  - SCAlable Preservation Environments
  - FP7 IP, ~10 Mio. €, 15 partners, 2011-2014
  - “Context-aware planning and watch”




# Next: Scalability, QA measures

- Scale down: Planning-as-a-Service
- Scale up: Scalable preservation planning
  - Heterogeneous, large-scale, complex content
  - Planning lifecycle and round-trip monitoring
  - From ad-hoc decisions to a continuous management activity
- Integration of DP with IT Governance, Governance/Risk/Compliance
- Given a set of objects, find *n* representative samples
- Improve measurement coverage and techniques
  - Challenge: Similarity between two (performances of) objects
  - Static analysis vs. perceptual-level analysis
  - Quality-aware emulation
  - **Ground truth and benchmarking**
    (The black box problem)

# Danke für die Aufmerksamkeit!

## Fragen?

www.ifs.tuwien.ac.at/~becker
www.ifs.tuwien.ac.at/dp/plato

**"We felt that this project will be of relevance across the entire digital preservation community and will provide immediate, practical assistance to organisations with varying levels of experience of digital preservation."**

**"The PLATO tool puts sophisticated preservation capability into the hands of a broader swathe of information specialists. There would seem to be great potential here for further integration and development with other tools and frameworks."**

**"The value of Plato 3 lies in its applicability to a wide range of repositories and formats and the fact that it gives preservation and repository managers the confidence and tools to tackle what previously appeared as the very daunting task of preservation."**

**"Staff at the Parliamentary Archives felt that this project demonstrated the clearest practical benefit to practitioners."**


- Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A. & Hofman, H. (2009). Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans, *International Journal on Digital Libraries.* 10(4), 133-157.
- Becker, C. and Rauber, A. (2010) Improving component selection and monitoring with controlled experimentation and automated measurements. *Information and Software Technology, 52(6):641–655, June 2010.*
- Becker, C. and Rauber, A. (2011) Decision criteria in digital preservation: What to measure and how. *Journal of the American Society for Information Science and Technology* 62 (6): 1009-1028, June 2011.