Digital Preservation Process: Preparation and Requirements

Hans Hofman
Nationaal Archief Netherlands

Training session, 26 March 2009, Barcelona

Overview

- What is the problem? Defining the issues
- Preservation Planning Process: scope and role, preservation plan
- What’s needed? How to identify requirements?
  - The organisational/ business context
  - Usage requirements, and
  - Collection profile

The issue / challenge

- The enormous and rapidly increasing amount of digital information
  - Fragile resources
- The rapid evolution in technology
- The risk of obsolescence and therefore corruption and/or loss of valuable information
- ***(Pro-)active*** and ongoing attention / maintenance required
- Potential solutions still fragmented
  - infrastructure
  - not comprehensive

Stakeholders

- Memory institutions (‘content holders’)
  - archives, libraries
- (Scientific) data centres
- Government organisations (record creators)
- Business companies (record creators, intellectual capital)
- Individuals (e.g. family pictures)

Terminology

- Preservation vs. curation
  - recordkeeping,
  - archiving
- Framework, policy, strategy
- Integrity, fixity
- Plan, action, method
- Metadata, representation information
- Object, deliverable unit, record, collection
- Properties, characteristics
- **Do we actually understand each other when we talk about digital preservation?**

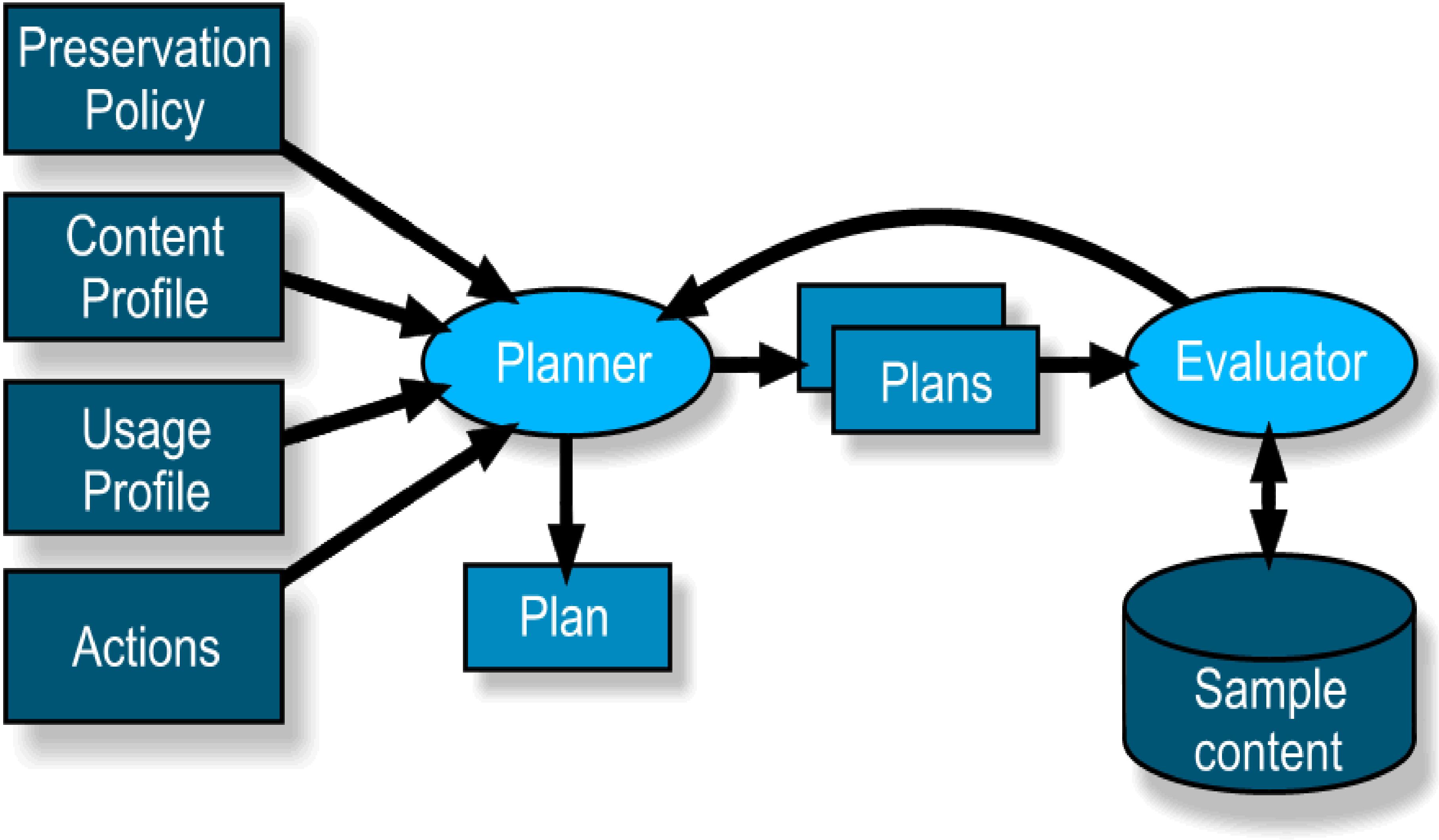
How to prepare preservation actions?

- Understanding the organisational context
  - mandate/ legislation
  - the organisational policy
  - user community
- Understanding the objects
  - (collection of) digital objects: characteristics
    - authenticity requirements
- Understanding the infrastructure
  - technology (past, present, future), infrastructure
  - people, knowledge, skills
- Available options
  - potential methods/ strategies + tools and their quality
- Decision making process: preservation planning

Objectives of Preservation Planning

- Identify and analyse the organisational context
  - including a risk assessment
  - define a framework for preservation / policy
- Support decision-making about digital preservation including
  - Identifying criteria for preservation within that context
  - Defining workflow for evaluating/ defining preservation plans
  - Developing methodologies for assessing the risks of applying different preservation strategies for different types of digital objects
- Enable formulation, evaluation and execution of high-quality and cost-effective preservation plans that suit the organisational (e.g. repository) needs
- Support the on-going evaluation of the results of executing preservation plans and provide a feedback mechanism
- Document the planning process carefully

Preservation planning decision-making process



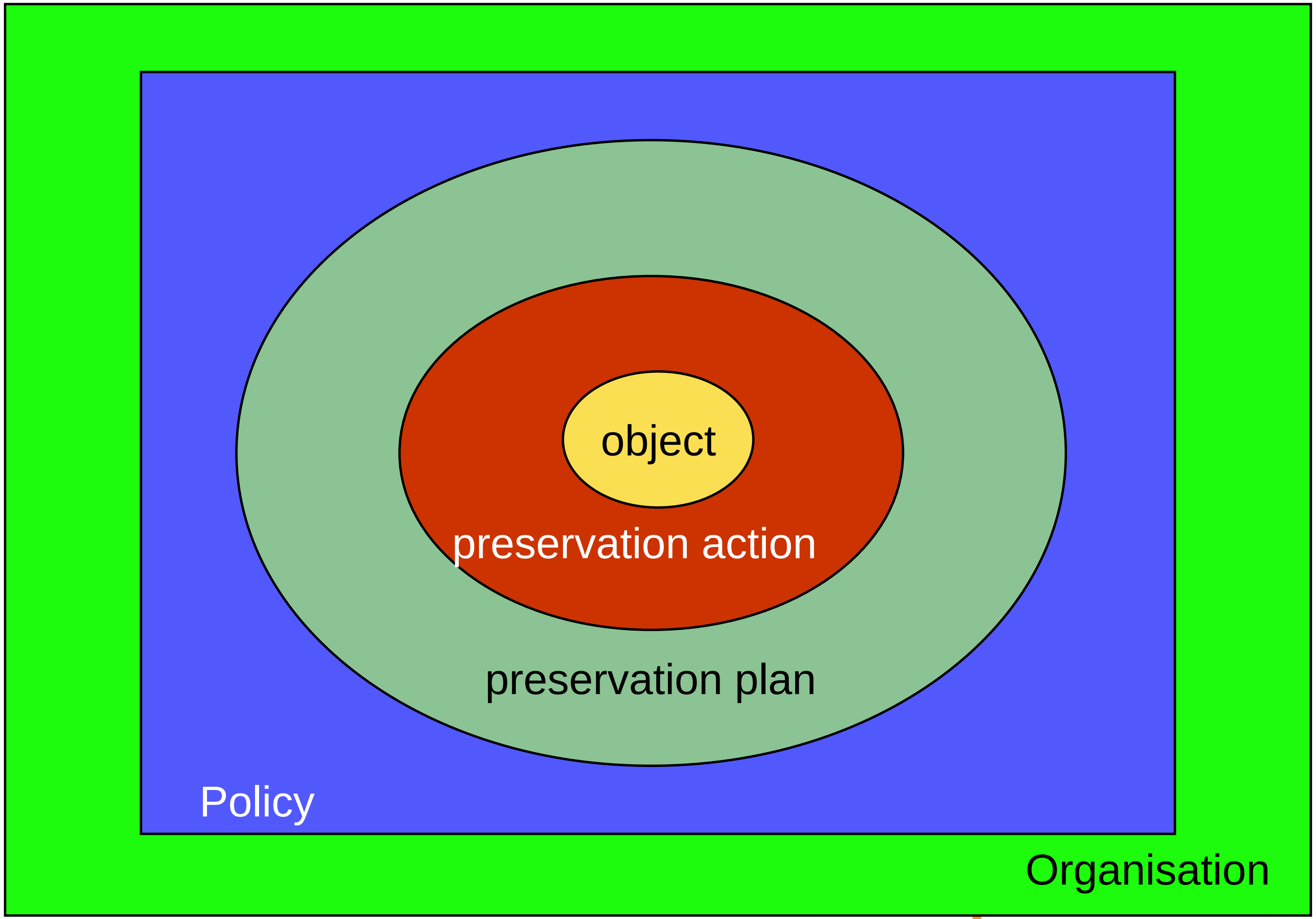
Policy model

Model for organisational requirements

- Approach:
  - Literature study and interviews with decision makers
  - Extraction of requirements and building of conceptual model
  - Create model from first principles
- First version of model
  - describes and positions policy requirements in wider context of Planets data model
  - indicates potential policy requirements on different levels and in relation to various types of preservation objects (e.g. collection, deliverable unit, manifestation, byte stream, etc.)
  - (thus) is able to cover policy requirements that are relevant for all types of preservation actions in different organisational settings
- Translate organisational constraints and requirements into a machine interpretable model
  - an organisation can choose from them according to its needs

Preservation policy: situation now

- Trying to understand what organisations do in this area:
  - Large institutions are accumulating expertise and are building trusted digital repositories
  - Small institutions generally lack expertise and funding to build a digital repository
  - Large institutions have formulated various *requirements* –as can be discovered in different types of documents
- No coherent picture (yet)
- Very high level and abstract



Ownership
Awareness
Responsibility

Preservation policy: examples

- Example policy statements of institutions with a digital preservation programme
  - – State Library of Victoria
  - – National Archives of Australia

State Library of Victoria (Australia)

- < Digital Preservation Policy, State Library of Victoria
- http://www.slv.vic.gov.au/about/information/policies/digitalpreservation.html
- **“Storage.** Born-digital objects published on disk (CD-R or DVD) are considered the archival copy and will be stored appropriately. When needed and authority granted, the physical format data may be copied to another storage carrier in order to preserve its contents. The master TIFF files shall be stored appropriately in a secure location on the Library's LAN, and back-ups made in accordance with TSD policy.”
- What does this choice mean in practice? Three examples:
  - CD-R and DVD should be stored in appropriate places (with appropriate temperatures and relative humidity).
  - It is allowed –if needed– to make copies to ensure long-term preservation.
  - For files stored on the Local Network, back-ups are made.

National Archives of Australia

- < An Approach to the Preservation of Digital Records
- http://www.naa.gov.au/images/an-approach-green-paper_tcm2-888.pdf
- p. 14: “The digital preservation program must be able to preserve any digital record that is brought into National Archives’ custody regardless of the application or system it is from or data format it is stored in.”
- What does this choice mean in practice? One example:
  - all records that are accepted, should be preserved, regardless file format, medium, application, etc.
  - transform to open standard + keep ‘original’ format

Policies: issues

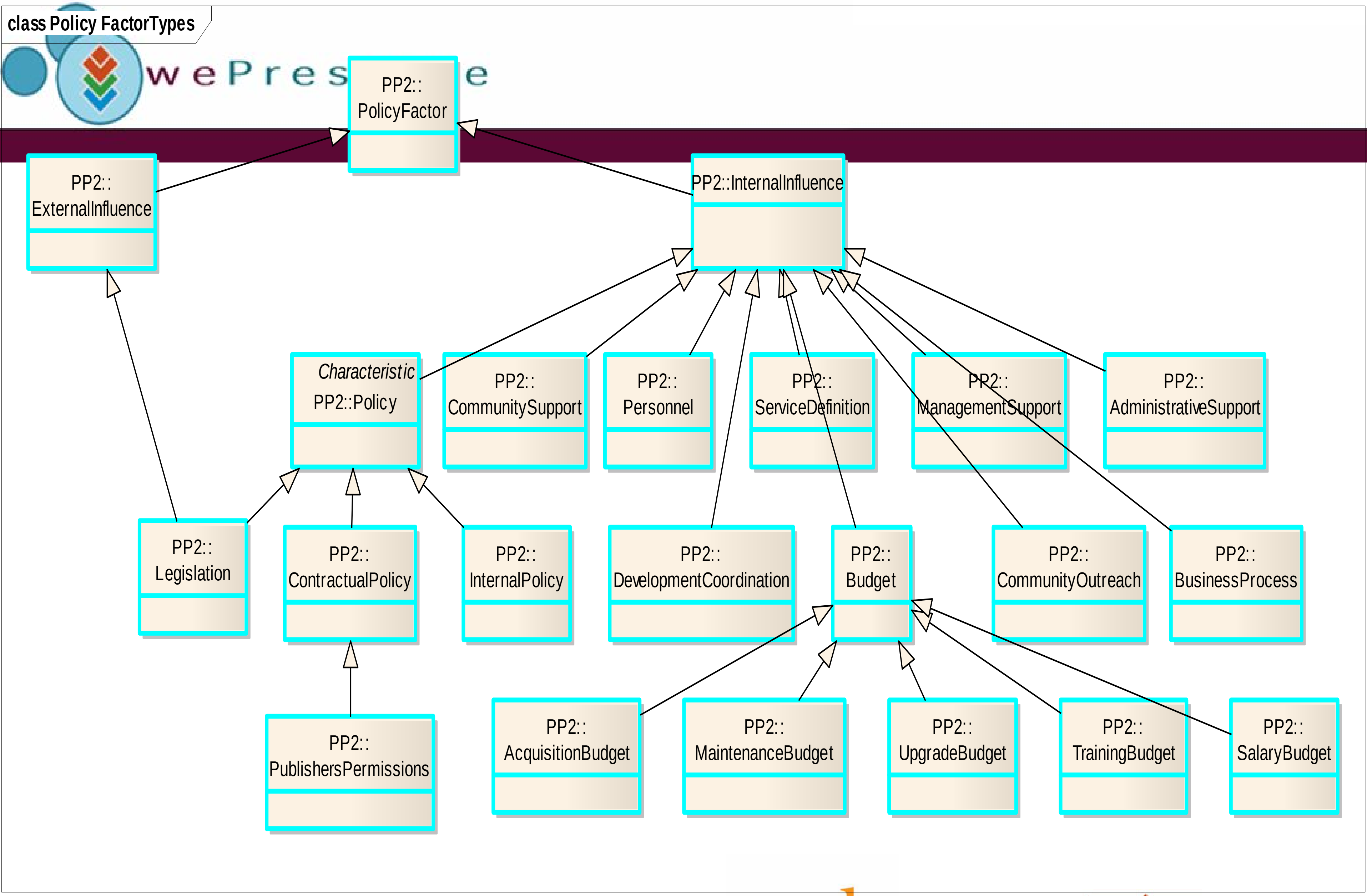
- The framework developed in Planets allows that various (sets of) requirements are identified
- Framework (requirements) are identified at a high level
- However, on this high level, it is possible and feasible that some requirements are already made explicit. Examples:
  - choice for one strategy (e.g. migration to open document format)
  - choice that some types of records/documents can be denied because e.g. an *exotic* file format is used
- Some decisions/choices may prove to be difficult to implement
- At this moment, lack of information about some strategies

  -> possibility to miss interesting new solutions …
- Some strategies are excluded as viable solutions: no investments have to be made

TOOL: Table – Digital Preservation Policy: AREAS OF COVERAGE

AREAS	Authority and responsibility
	Conversion and reformatting
	Appraisal, selection and acquisition
	Storage and maintenance
	Access and dissemination
	Implementation
	Standards
	Procedures
	Quality control, auditing and benchmarking
	Cooperation
	Technical infrastructure

class Policy FactorTypes



PP2::PolicyFactor

PP2::ExternalInfluence

PP2::InternalInfluence

Characteristic
PP2::Policy

PP2::CommunitySupport

PP2::Personnel

PP2::ServiceDefinition

PP2::ManagementSupport

PP2::AdministrativeSupport

PP2::Legislation

PP2::ContractualPolicy

PP2::InternalPolicy

PP2::DevelopmentCoordination

PP2::Budget

PP2::CommunityOutreach

PP2::BusinessProcess

PP2::PublishersPermissions

PP2::AcquisitionBudget

PP2::MaintenanceBudget

PP2::UpgradeBudget

PP2::TrainingBudget

PP2::SalaryBudget

Preliminary conclusion

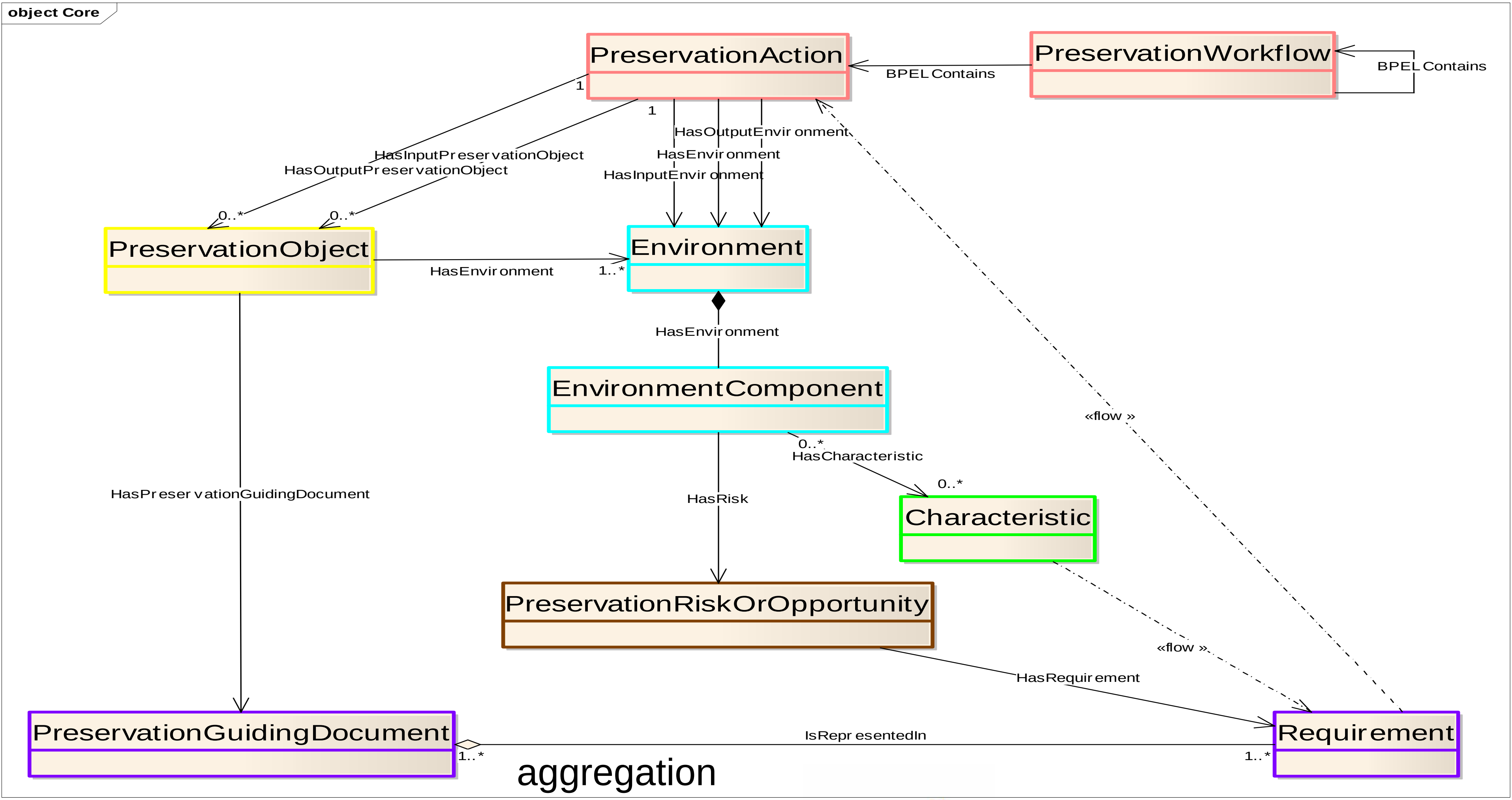
- The variety of requirements in institutional policy documents is very large
- Some of these requirements are very general, others are a bit more specific.

- Questions
  - How does one make these requirements operational for preservation planning purposes?
  - What are 'valuable' requirements in a process of preservation planning?
  - …

Types of requirements

- Risk specifying requirements
  - setting (acceptable) values/parameters for when objects are at **risk**
- Preservation guiding and action defining requirements
  - desirable types of **preservation actions** **given the significant characteristics**
- Preservation process guiding requirements
  - setting parameters for the preservation action **process**
  - Preservation infrastructure requirements
    - setting parameters for the **infrastructure** needed for carrying out actions

A generic model



object Core

PreservationAction

PreservationWorkflow

BPEL Contains

BPEL Contains

1

1

HasOutputEnvir onment

HasEnvir onment

HasInputEnvir onment

HasInputPr eser vationObject

HasOutputPr eser vationObject

0..*

0..*

PreservationObject

Environment

HasEnvir onment

1..*

HasEnvir onment

EnvironmentComponent

0..*

HasCharacteristic

0..*

Characteristic

«flow »

HasPr eser vationGuidingDocument

HasRisk

PreservationRiskOrOpportunity

«flow »

HasRequir ement

PreservationGuidingDocument

IsRepr esentedIn

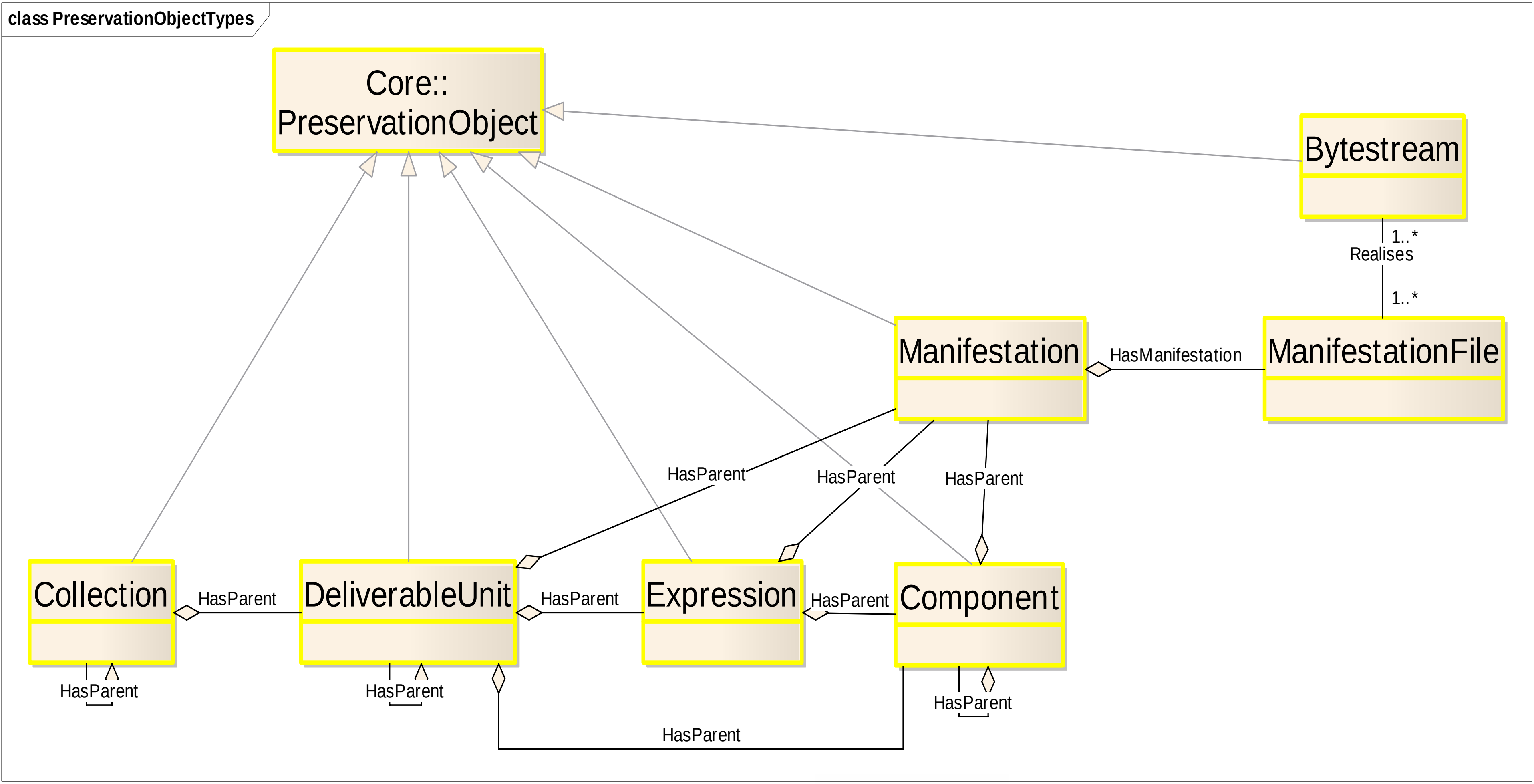
Requir ement

1..*

1..*

aggregation

The digital object



Background

- Describing the objects:
  - – extraction of characteristics (characterisation)
  - – mainly technical (intellectual characteristics will be done manually (after creation) or automatically (at creation?)
  - – enable
    - to compare
    - to validate results of actions
    - to automate preservation action processes

Usage model

- Preliminary model with user requirements has been created on the basis of results of probe approach
- Model shows some variances between different types of users
- Usage requirements
  - Performance, Usability, Presentation, Authenticity
  - Understandability, Rights, Costs
- To what extent relevant for digital preservation?

- These requirements are in first instance relevant for presentation files, but will have an impact on preservation files
- Used as input into PLATO
- Article on methodology published in D-Lib Magazine article, May/June 2008.

User perspective

- Goal of digital preservation is to serve (future) users in providing usable and authentic information

- What are needs/requirements of users?
  - easy access
  - knowledge about origin of documents/ to be able to interpret them
  - to use them for their own convenience

- Example requirements
  - some users prefer that all information is presented in a uniform way
  - some users prefer that they can search full-text in documents
    - consequence: don’t migrate texts to image files
  - …

User perspective (2)

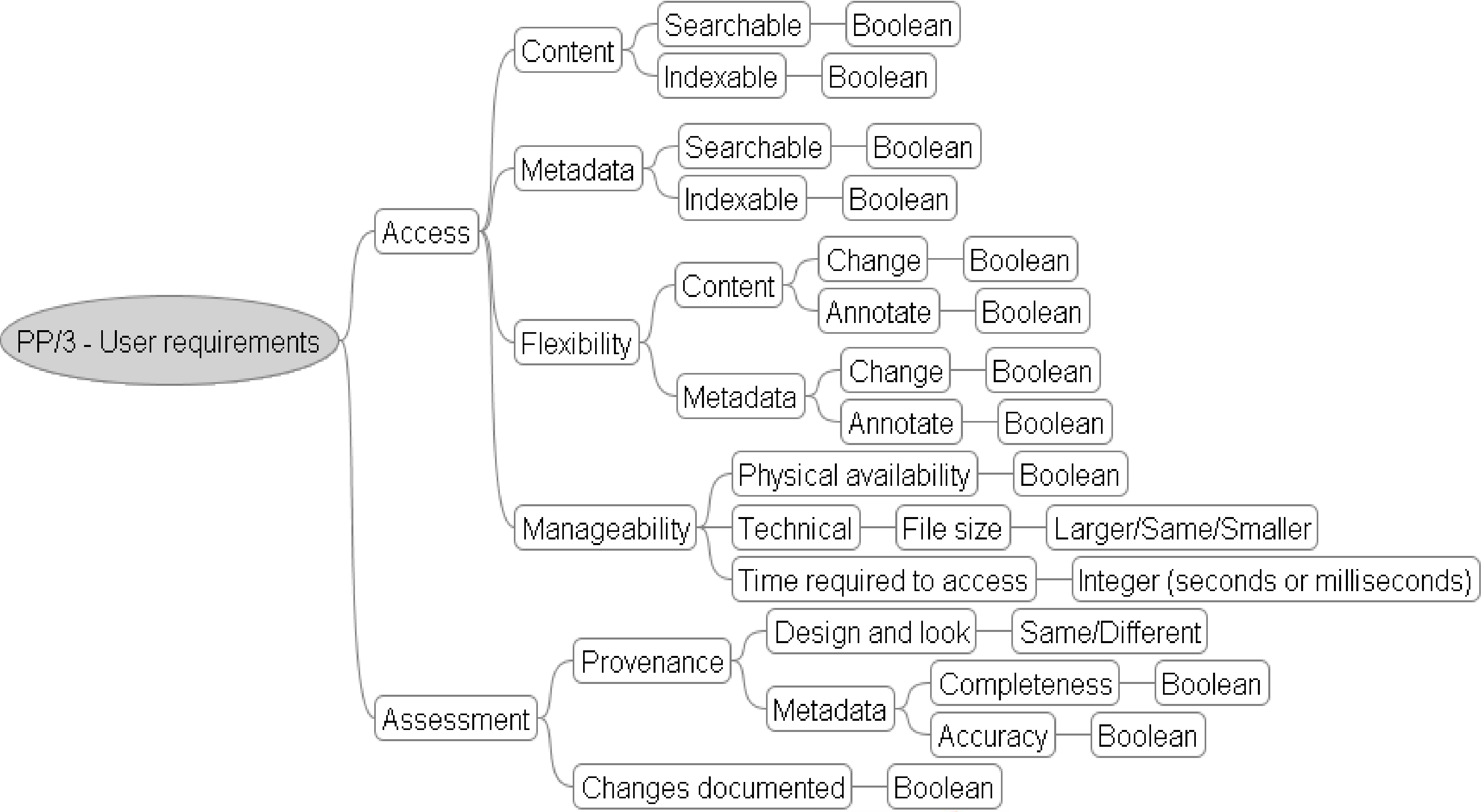
- Some user requirements will affect decisions for preservation actions
  - Different manifestations of ‘deliverable units’?
- However, difference between preservation and presentation copy
  - Not necessarily the same ‘object’/ format...
  - Should it be the same, in order to reduce costs for preservation strategies/ actions?
  - Providing presentation copy ‘on the fly’/ on demand?
- Not all users necessarily have the same requirements
  - requirements may be based on user segmentation

User perspective (3)

- Difficult to assess what users want in the digital era
  - because they are (often) not used to work with digital information/documents
  - because they are possibly not aware of the possibilities of different ways of presentation

- Compilation of requirements is based on user studies in which the participants combine the use of paper and digital documents

- Predict what users in the future want, is impossible….!



- PP/3 - User requirements
  - Access
    - Content
      - Searchable — Boolean
      - Indexable — Boolean
    - Metadata
      - Searchable — Boolean
      - Indexable — Boolean
    - Flexibility
      - Content
        - Change — Boolean
        - Annotate — Boolean
      - Metadata
        - Change — Boolean
        - Annotate — Boolean
    - Manageability
      - Physical availability — Boolean
      - Technical — File size — Larger/Same/Smaller
      - Time required to access — Integer (seconds or milliseconds)
  - Assessment
    - Provenance
      - Design and look — Same/Different
      - Metadata
        - Completeness — Boolean
        - Accuracy — Boolean
    - Changes documented — Boolean

‘Technical’ requirements

- These could be based on requirements as identified by users or the institution

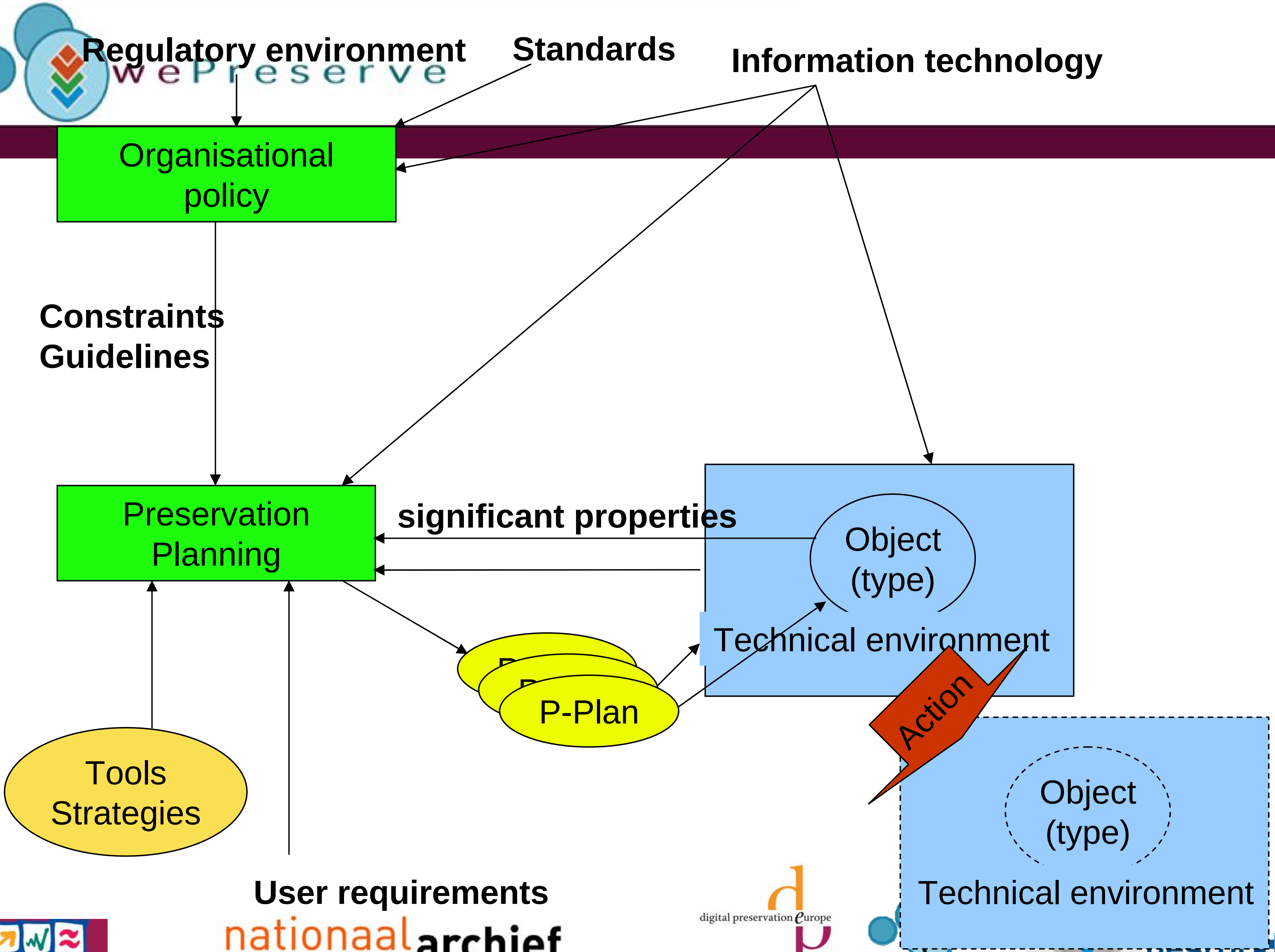
- Is it –from a user’s point of view- necessary that the page layout is as close as possible to the ‘original’?
  - Yes -> page layout, fonts, links, headers & footers, titles & subtitles should be preserved
    - migrate
    - emulate
  - No -> only the textual content is important
    - don’t bother about emulation
    - migrate to an ‘easy’ format
  - Not necessarily a preservation format..!

Collection profile

- What types of objects (both technical and intellectual aspects)?
- Technical: file formats
  - registries (e.g. PRONOM, …)
- Intellectual: for instance documentary form, structure, look and feel, 'behaviour'
  - objective tree 'templates'
  - an (intellectual) object may consist of different computer files
    - what strategy then?
- What needs to be preserved?
- What criteria for determining these essential characteristics?

Requirements for objects

- **Authenticity**
  - to be what it purports to be,
  - to have been created or sent by the person purported to have created or sent it, and
  - to have been created or sent at the time purported
- **Reliability**
  - contents can be trusted as a full and accurate representation of the transactions, activities or facts to which they attest and can be depended upon in the course of subsequent transactions or activities
- **Integrity**
  - being complete and unaltered
- **Usability**
  - can be located, retrieved, presented and interpreted, so retrievable, readable, interpretable
- **Accuracy**
  - the degree to which data, information, documents or records are precise, correct, truthful, free of error or distortion or pertinent to the matter.



Regulatory environment

Standards

Information technology

Organisational policy

Constraints
Guidelines

Preservation Planning

significant properties

Object (type)

Technical environment

P-Plan

Action

Object (type)

Technical environment

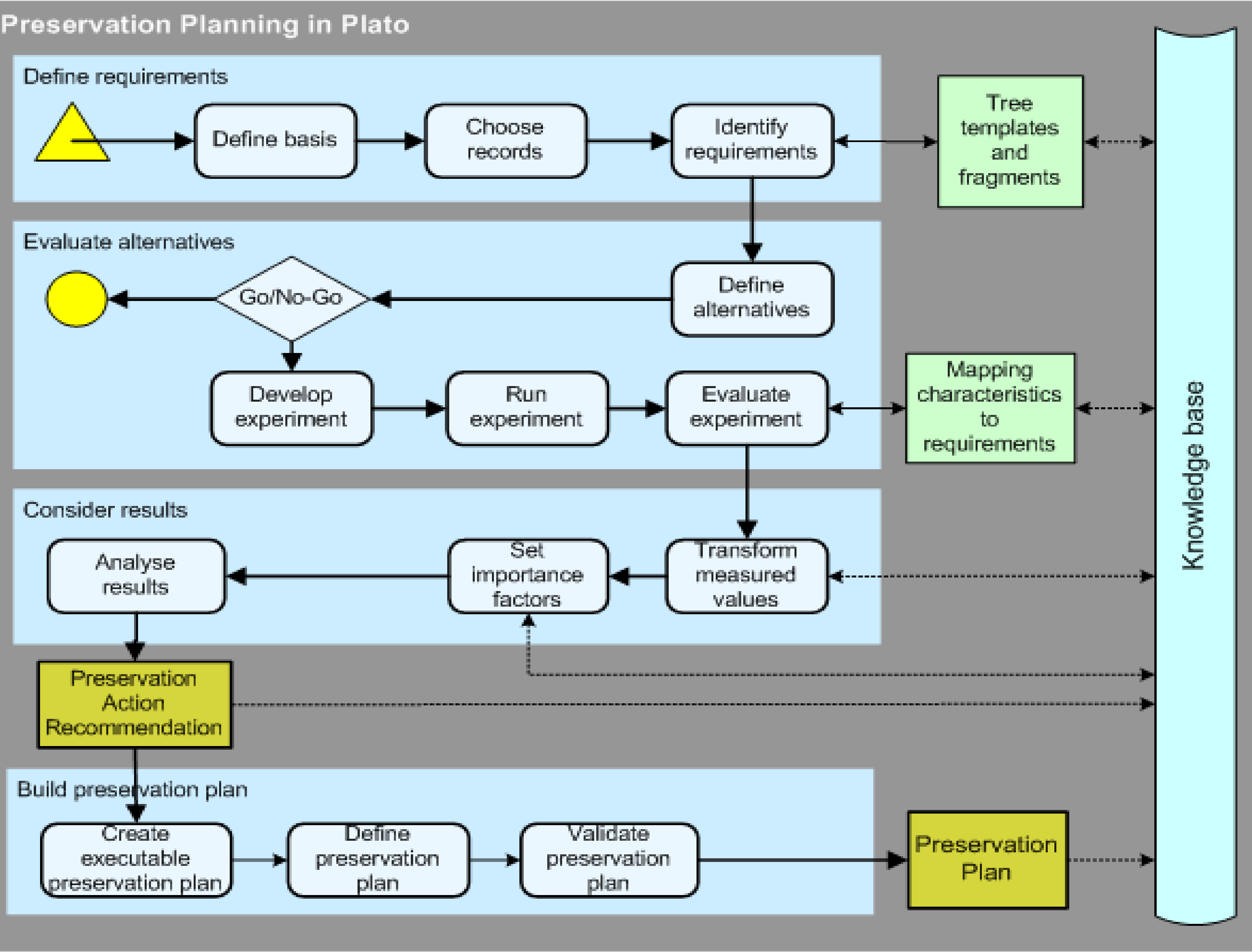
Tools
Strategies

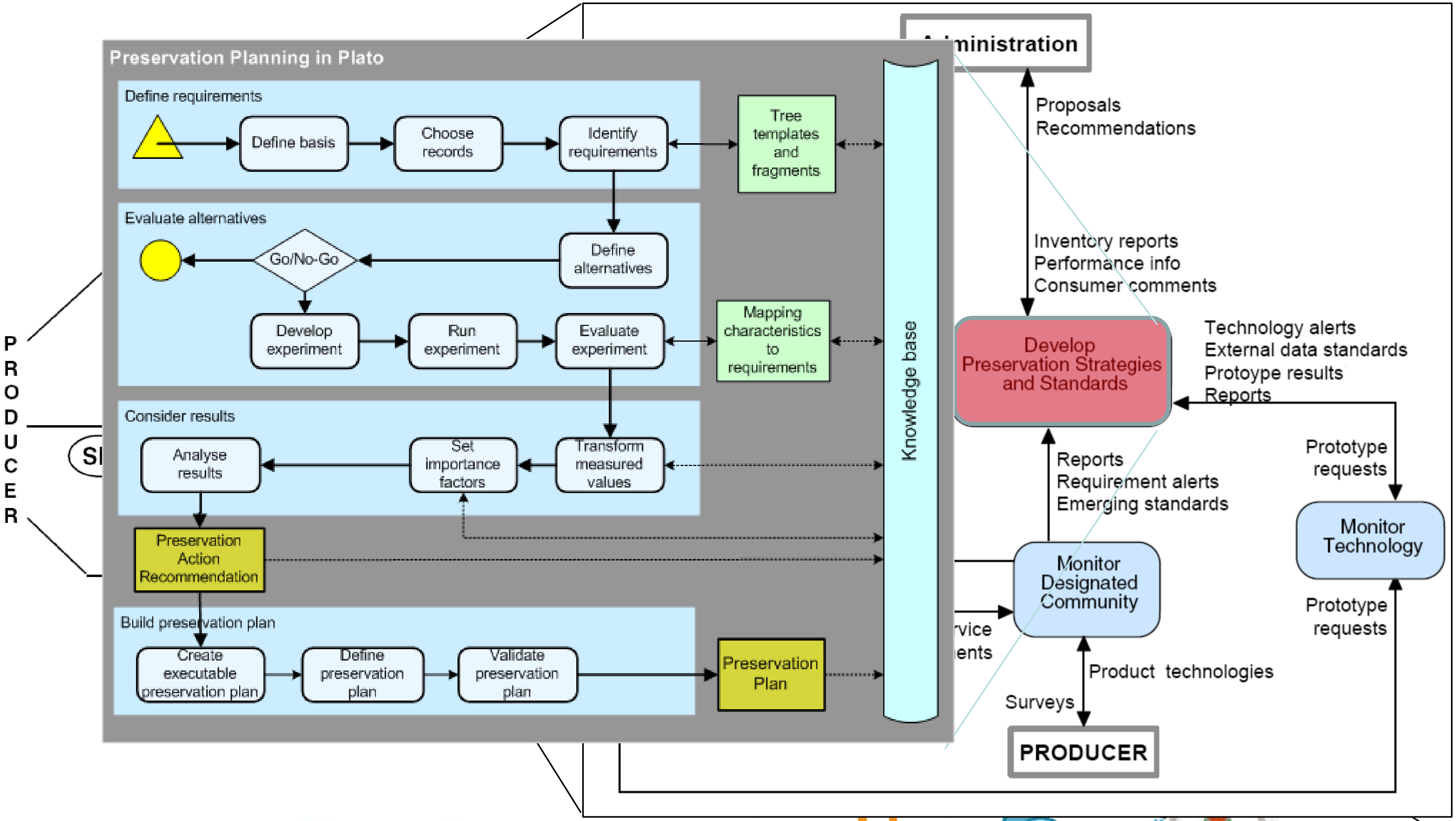
User requirements

Issues

- How to develop a natural and logical flow of questions to be answered?
  - e.g. each question narrowing/excluding and/or including next steps
- How to translate that into a decision, documented in a preservation plan?
  - what is a preservation plan?
- How to enable the automated execution of the plan?
- How to evaluate the result of the execution of the plan?

Preservation Planning in Plato





Preservation Planning in Plato

Define requirements

Define basis

Choose records

Identify requirements

Tree templates and fragments

Evaluate alternatives

Define alternatives

Go/No-Go

Develop experiment

Run experiment

Evaluate experiment

Mapping characteristics to requirements

Consider results

Transform measured values

Set importance factors

Analyse results

Preservation Action Recommendation

Build preservation plan

Create executable preservation plan

Define preservation plan

Validate preservation plan

Preservation Plan

Knowledge base

Administration

Proposals
Recommendations

Inventory reports
Performance info
Consumer comments

Develop Preservation Strategies and Standards

Technology alerts
External data standards
Protoype results
Reports

Prototype requests

Monitor Technology

Prototype requests

Reports
Requirement alerts
Emerging standards

Monitor Designated Community

rvice
ents

Product technologies

Surveys

PRODUCER

P R O D U C E R

SI

Definition of a Preservation Plan

- ‘A ***preservation plan*** defines a series of preservation actions to be taken by a responsible institution to address an identified risk for a given set of digital objects or records (called collection).‘
- The Preservation Plan takes into account the preservation policies, legal obligations, organisational and technical constraints, user requirements and preservation goal. It also describes the preservation context, the evaluated alternative preservation strategies and the resulting decision for one strategy, including the rationale of the decision

Characteristics of a preservation plan

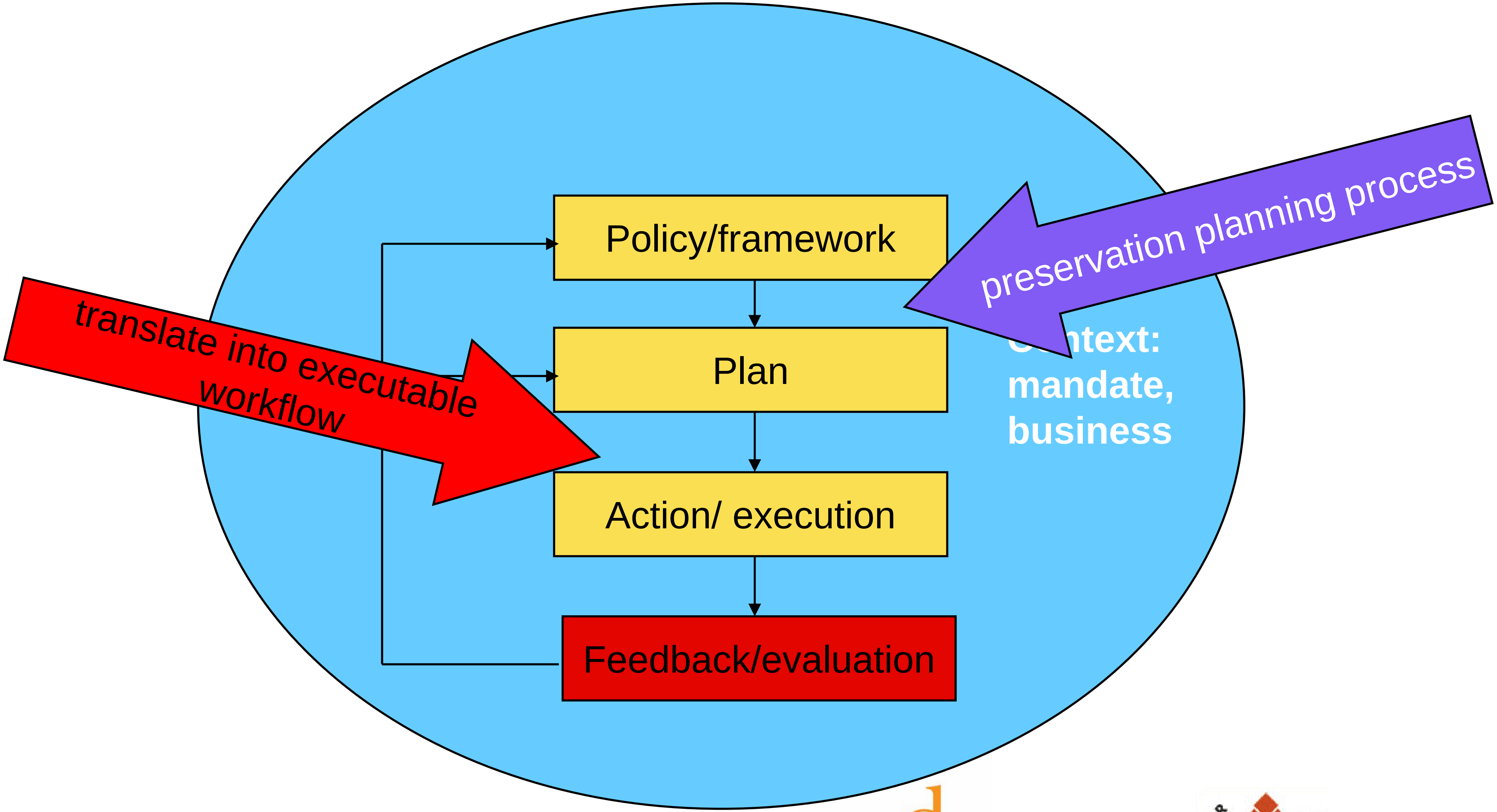
- It is a concrete translation of a preservation policy how to handle/treat a certain type of digital objects in a given institutional setting
- New plans will be needed over time due to
  - ✓ changes in technology
  - ✓ changes in organisational setting
  - ✓ changes in user requirements
  - ✓ changes in available tools
  - ✓ changes in preservation methods
- It also specifies a series of steps or actions along with responsibilities and rules and conditions for execution
  - ✓ This is called **preservation action plan.** It is in the form of an executable workflow definition, detailing the actions and the required technical environment
  - ✓ Relationship with a specific action
  - ✓ The preservation plan provides the context/ background of the preservation action plan

The content of a preservation plan

1. Identification
2. Status
   - ✓ What was the immediate reason for this plan?
   - ✓ Has it been approved and if so, when and by whom
   - ✓ How does it relate to other P-plans related to a specific type of objects?
3. Description of institutional setting
4. Description of the collection (digital objects)
5. Purpose and requirements
6. Evidence of decision for a specific preservation action
   - ✓ what is the foundation of the decision
   - ✓ description of evaluation of possible actions
7. Costs considerations
8. Trigger for re-evaluation
9. Roles and responsibilities
10. Preservation action plan
    - ✓ executable program

draft

From preservation policy to action


Policy/framework
Plan
Action/ execution
Feedback/evaluation
translate into executable workflow
preservation planning process
Context: mandate, business

Models and tools in PP

- Conceptual model of potential preservation requirements (characteristics), both organisational and object-oriented
- Conceptual model of PP-process
- Validation framework (how to measure whether a preservation action has been successful?)
- Definiton of preservation (action) plan

- Tools:
  - Machine interpretable models for usage, collection profile, policy requirements/constraints
  - Plato tool for decision making: v.2 (November 2008)
  - Validation Framework - comparator - metrics
  - Technology watch service
  - Recommender service
  - Collection profiling service
    - technical (related to PRONOM) and intellectual aspects
  - Risk assessment service
    - technical aspects (related to PRONOM registry)

Summary

- Understanding of context
  - analysis of organisational needs, user needs, legal requirements
- Identify criteria for preservation
  - how long, restrictions of formats, standards, …
  - *Risk analysis !*

- Determine what to keep/maintain
  - essential characteristics (objective trees), characterisation of computer files
- Evaluate available strategies (actions) against criteria
  - identify best strategy
  - well-founded and documented decision
  - create/finalise preservation plan
- Execute plan when needed
- Evaluate what happened/performance
- Re-iterate when technology changes or review when policy and/or collection and/or usage changes
- Automated decision process support (?)

Thank you for your attention!

Questions?

www.planets-project.eu

hans.hofman@nationaalarchief.nl