



SCAPE Preservation Planning and Watch

Kresimir Duretec

Vienna University of Technology

SCAPE information day at ONB

Vienna, 05 May 2014

- 3 workpackages
 - Automated Watch
 - Policy Representation
 - Automated Planning



- **Measure** : the result of measurement of an attribute (Kulovits et al. Open Preservation Data: Controlled vocabularies and ontologies for preservation ecosystems)
- **Attribute**: inherent property or characteristic of an entity that can be distinguished quantitatively or qualitatively by human or automated means (ISO/IEC 15939:2002)

Measures in Digital Preservation

- Format : open/not open
- Collection size : 10 TB
- SSIM : 0.9
- Costs per file per year : 1 €
- Image height equals : yes
- Running hardware costs per year : 2 €
- Expected size of a repository in 10 years : 20 TB
-

What do we do with those measures in PW ?

1. Produce tools that provide measures automatically (scalability !)
2. Monitor attributes and their measures over time
3. Provide a vocabulary so different tools can speak to each other
4. Provide a language for expressing different objectives
5. Make a well documented and trustworthy decision based on some (important) measures

- Different categories of measures that can be covered
 - Characterization of single files
 - Costs
 - Quality Assurance
 - Formats / Tools
- We are interested in collections -> **Collection profiling**
- Due to the size and heterogeneity of today's collections this task raises significant scalability challenges

Clever Crafty Content Profiling of Objects(C3PO)

- Tool for :
 - Analysing large scale collections
 - Discovering file format , file size and any other kind of distributions
 - Drilling down to specific parts of your collection
 - Selecting sub-collections to enable experiments
 - Reducing conflicts between different tools



- Uses different characterization tools such as FITS (not really a characterization tool) and Apache Tika
- Collects extracted data into a MongoDB database
- Provides a detailed analysis of collected data
- <https://github.com/openplanets/c3po>

- Measures change over time
 - Repositories are growing
 - Costs are increasing (or decreasing ?)
 - Actions are being executed with different results
 - New formats introduced
 -
- We need to monitor all of those changes over time so we can react on time when an important risk or opportunity is detected
- How do we monitor things that are scattered over the web ?

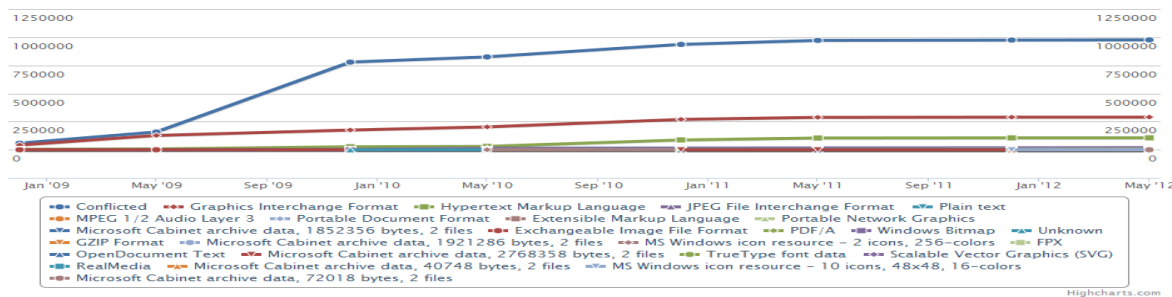
- Knowledge base
 - Entities and their properties
 - Measures of properties over time
 - Triggers define conditions and events
- Flexible and extensible
 - A well-defined, flexible data model
 - Adaptors for different information sources
- Monitoring Capabilities
 - Internal Monitoring
 - External Monitoring
 - Monitor compliance, risks and opportunities

Preservation Monitoring System(Scout)

- Tool to monitor your :
 - Content (collection profiles from C3PO)
 - Repository events
 - Environment
 - Files and formats
 - Other content holders content

Property history

This property has changed in time as represented in the chart below. Click on the chart dots for more information.



Value provenance

The current value for this property was measured on the following dates.

← Previous

1-1 of 1

Next →

Date

Source

Adaptor

Tuesday, 1 May 2012, 1:00:00 am

Internet Memory Foundation

IM Content Profile (c3po-v0.0.7)

← Previous

1-1 of 1

Next →

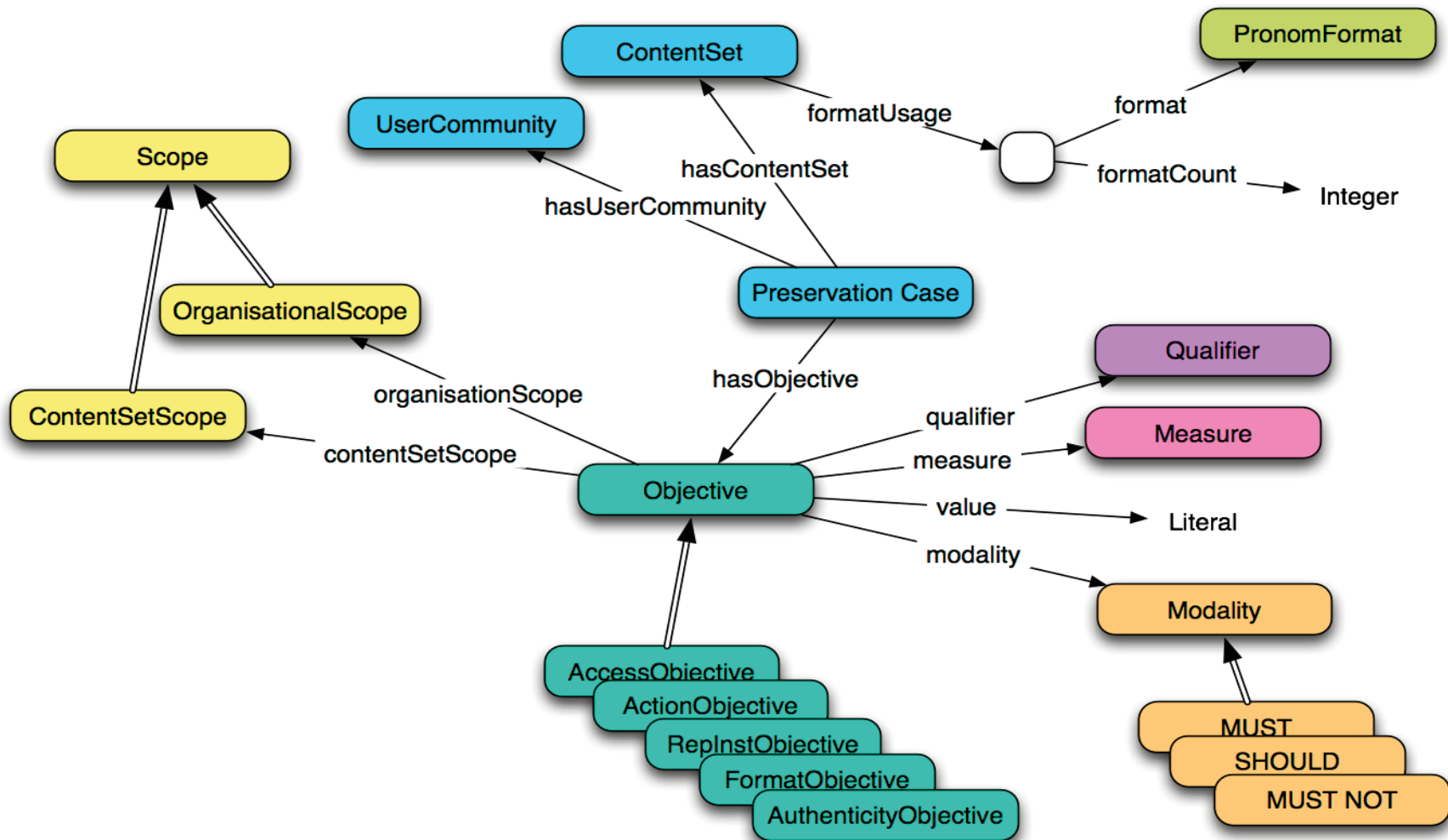
Provide a vocabulary so different tools can speak to each other

- Different
 - Units : €, £, \$
 - Scales : KB, GB, TB
- Sometimes it is not clear what is the measure and its scale and units
- Vocabulary: purl.org/dp/quality/measures
- Already ~ 400 measures registered
- You can require a new measure at :
<https://github.com/openplanets/policies/issues>

- Often we have some objectives (requirements) about certain attributes and their measures
 - Formats need to be open
 - Files should not contain compression
 - Costs per object should be under 1 €
 - Image height must be equal (when migrating images)
 - ...
- We need a language to express these objectives in a manageable and standard way so different components can react on them

- Control policies : practicable elements of governance that relate to clearly identified entities in a specified domain model (Kulovits et al. Open Preservation Data: Controlled vocabularies and ontologies for preservation ecosystems)
- Preservation case
 - Links a content set to a user community with particular objectives
- Links elements defined in the vocabulary
- Ontology

Control policies



(Kulovits et al. Open Preservation Data: Controlled vocabularies and ontologies for preservation ecosystems)

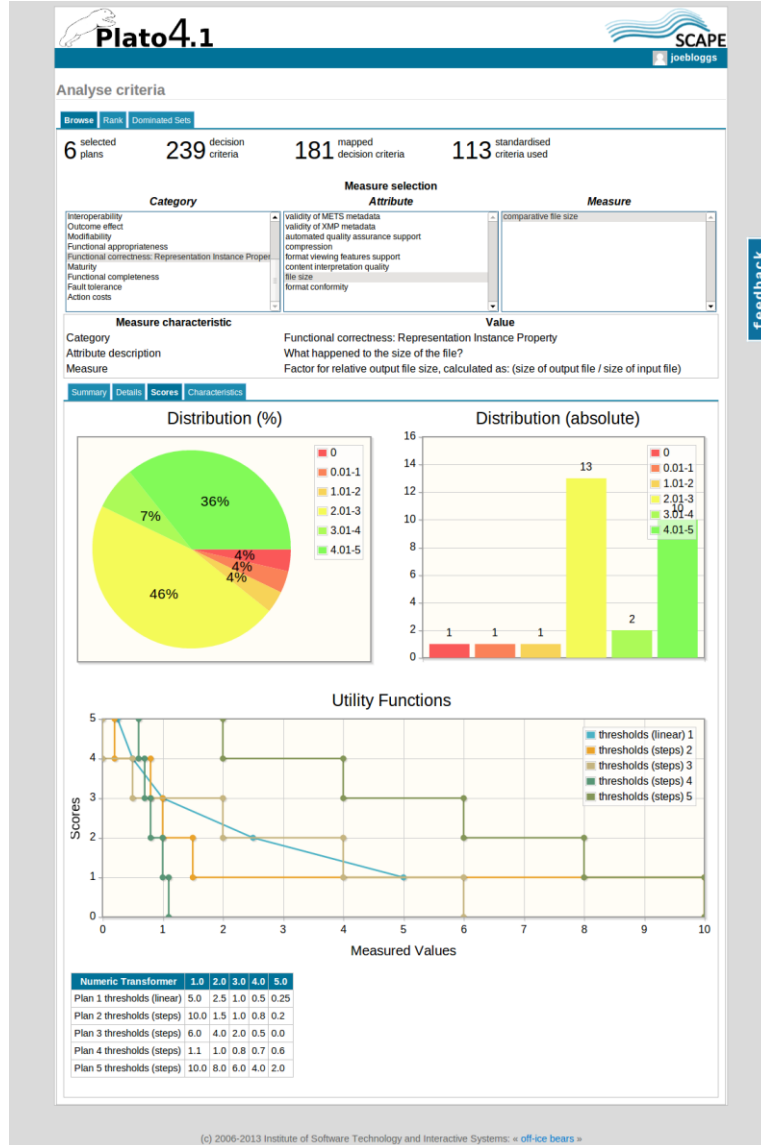
- Business Policies exist to govern; that is, control, guide, and shape the Strategies and Tactics. They define what can be done and what must not be done, and may indicate how, or set limits on how, it should be done (Object Management Group (2010). *The Business Motivation Model v1.1*)
- *Defined on the management level*
 - *Not directly applicable on the technical level*
 - *Need to be translated (not an easy job)*
- *Provide a catalogue of policy elements*
 - *<http://www.scape-project.eu/deliverable/d13-2-catalogue-of-preservation-policy-elements-draft>*

Make a well documented and trustworthy decision based on some (important) measures

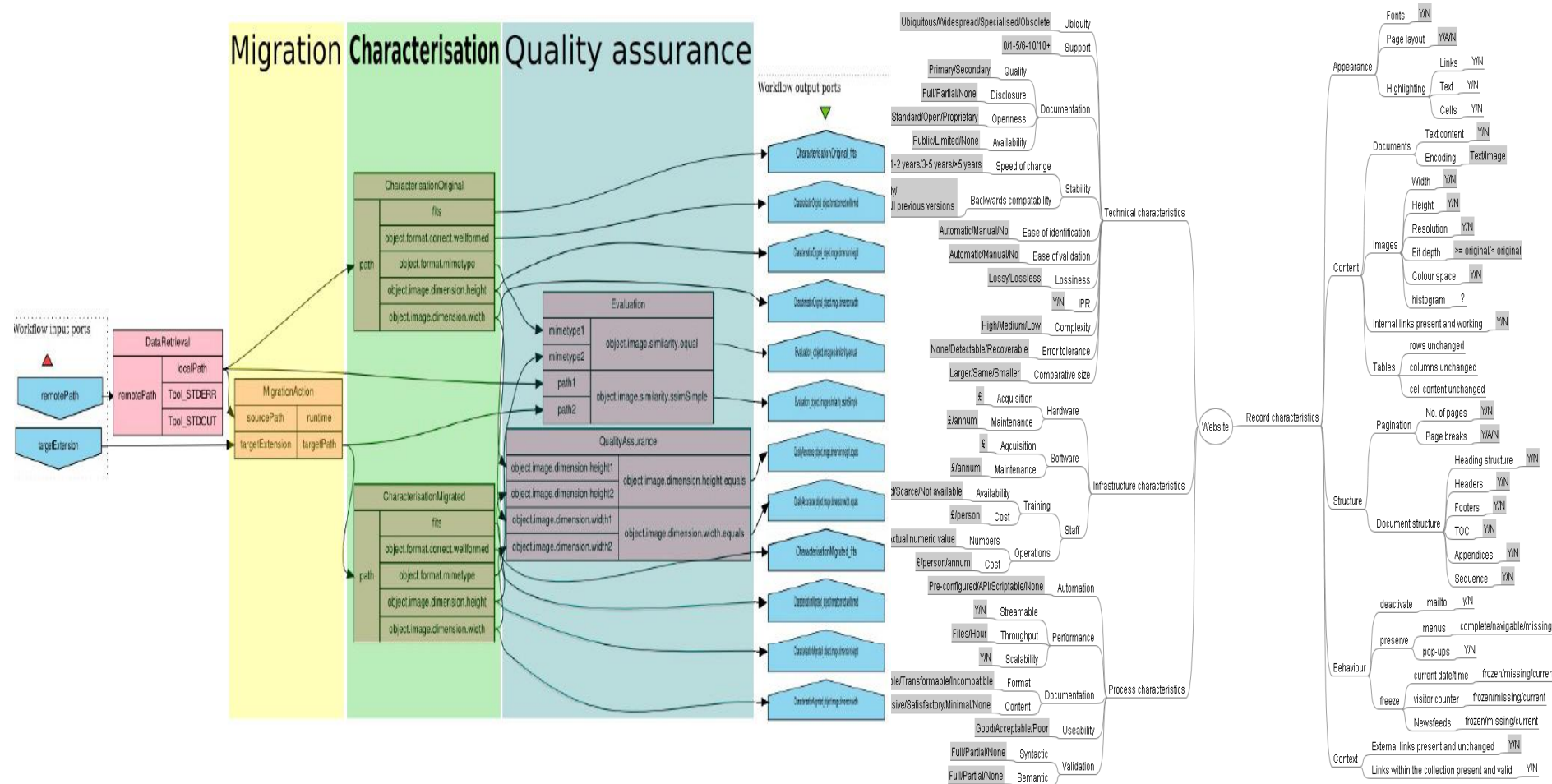
- When preserving digital content you need to make certain decisions on what to do
- This decision will be based on different requirements covering domains such as :
 - costs of the actions
 - technical characteristics (speed, ...)
 - quality of the results produced by the action
- We need a way how to make a decision in a trustworthy and scalable way

Preservation planning (Plato)

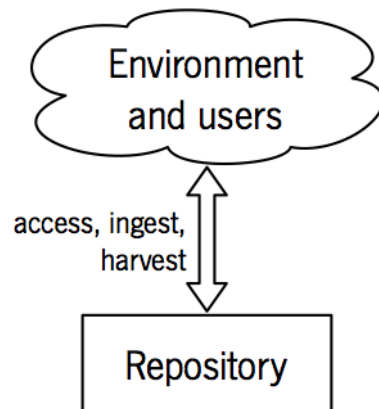
- Tool for :
 - Planning your preservation actions
 - Helps you to decide which action with which settings is most appropriate for your collection under certain conditions
 - Creating an executable preservation plan which can be executed without any intervention



Preservation planning (Plato)

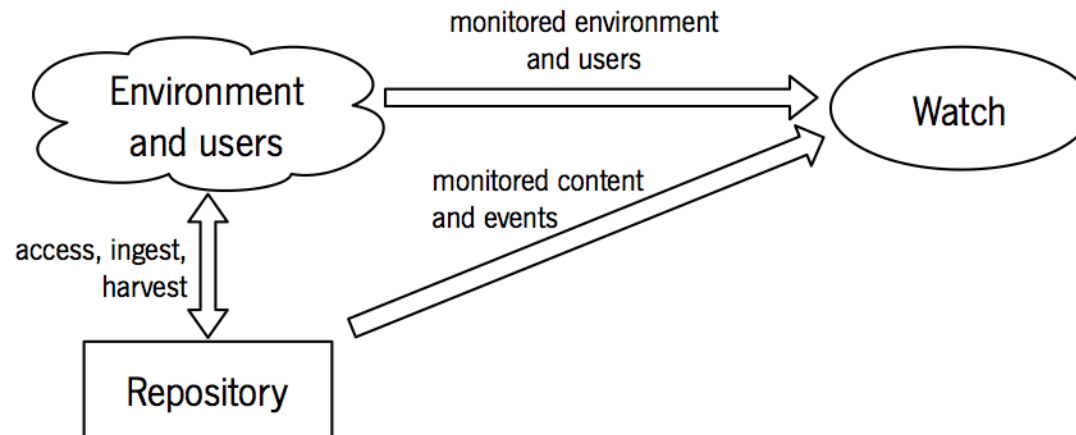


Bringing everything together



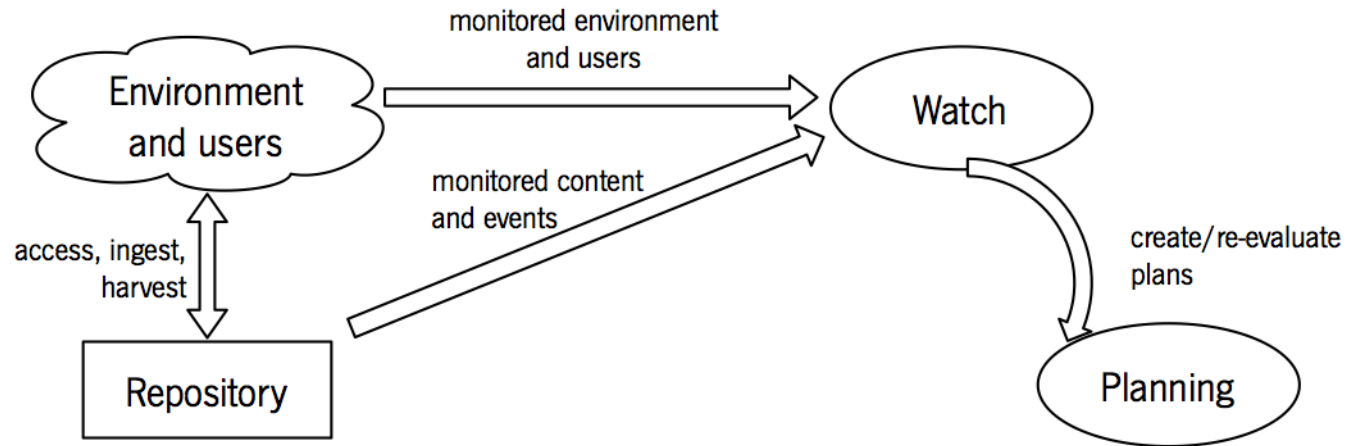
Author of the diagram Luis Faria

Bringing everything together



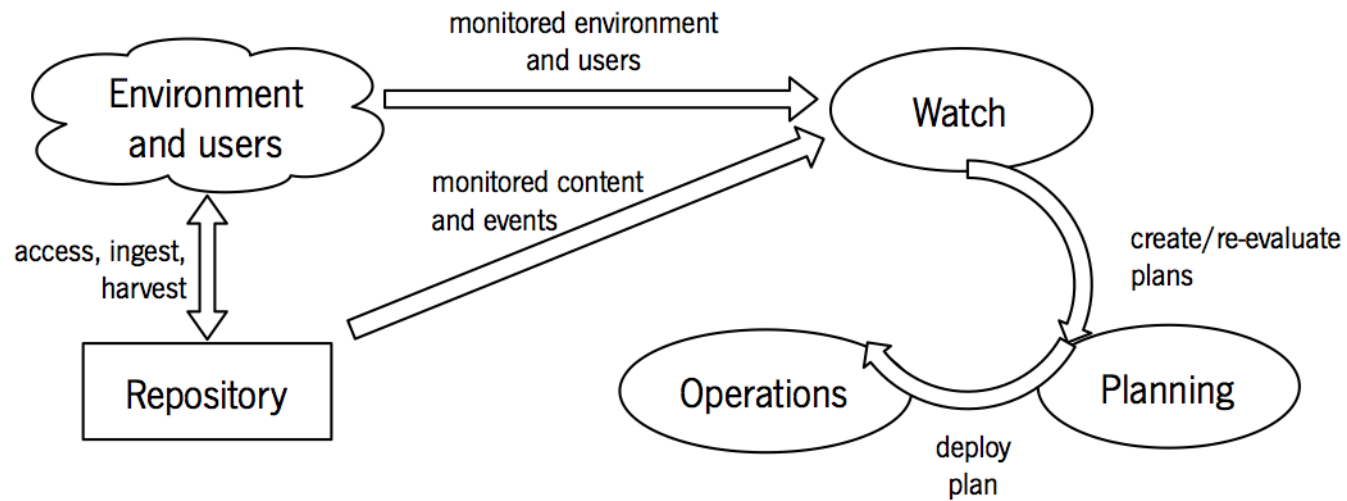
Author of the diagram Luis Faria

Bringing everything together



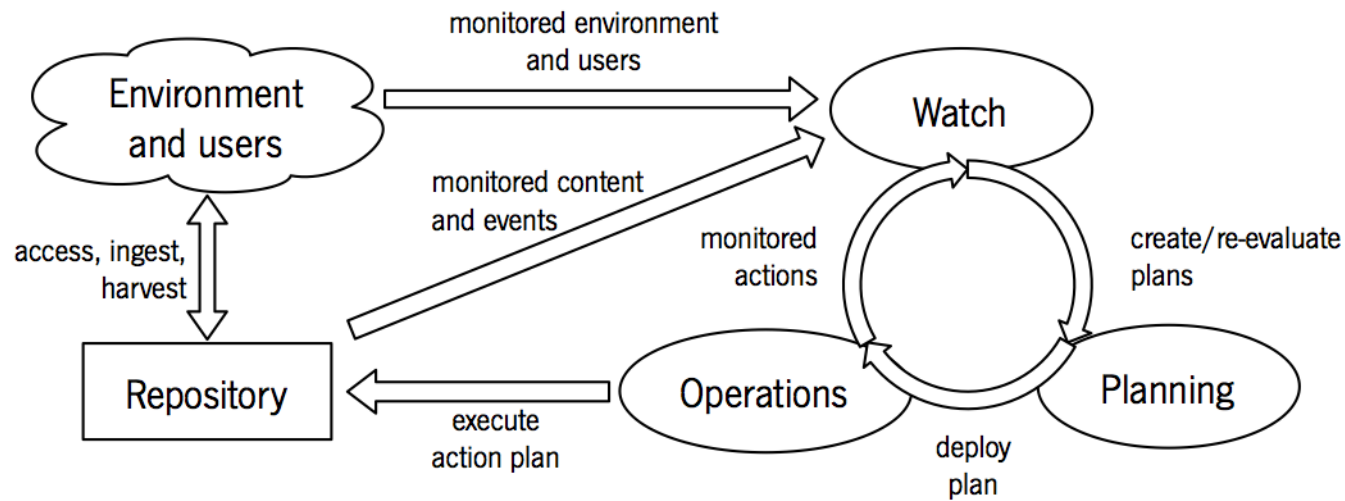
Author of the diagram Luis Faria

Bringing everything together



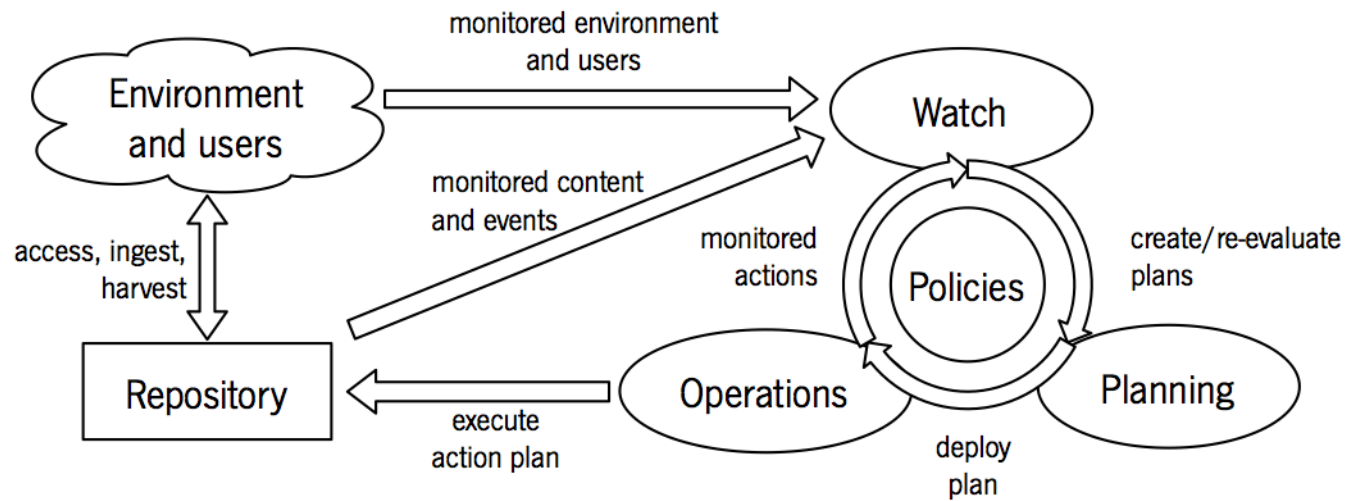
Author of the diagram Luis Faria

Bringing everything together



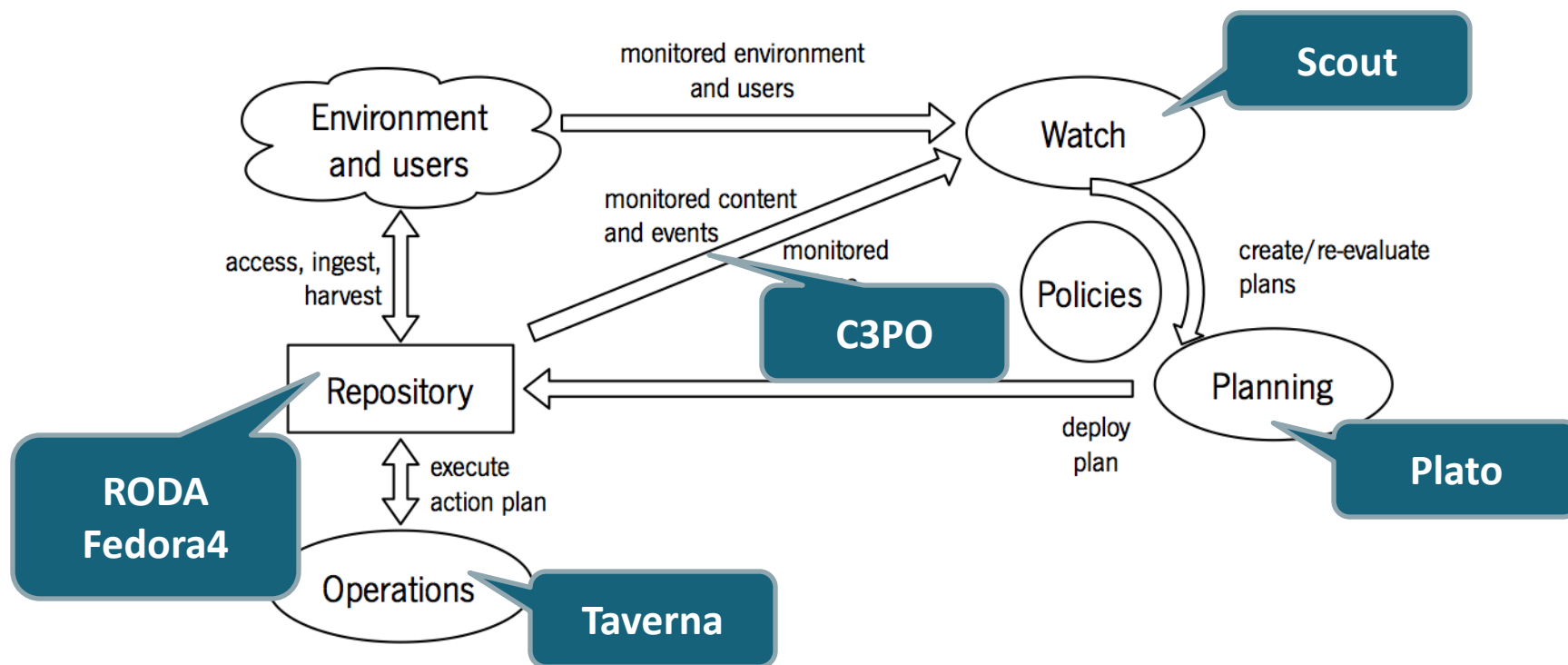
Author of the diagram Luis Faria

Bringing everything together



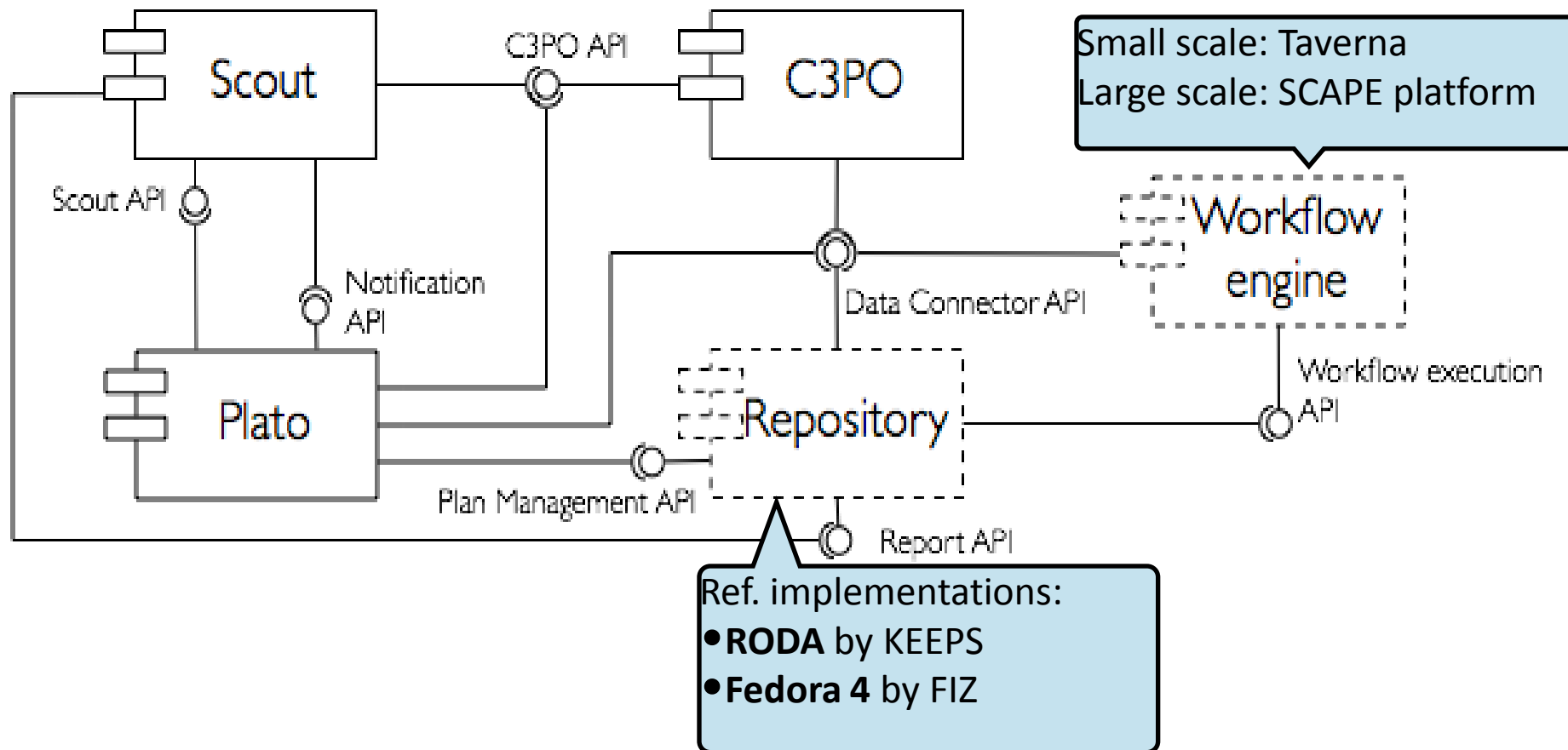
Author of the diagram Luis Faria

Bringing everything together



Author of the diagram Luis Faria

Technical implementation



Author of the diagram Luis Faria

- C3PO: <https://github.com/openplanets/c3po>
- Policies: <https://github.com/openplanets/policies>
- Scout: <https://github.com/openplanets/scout>
- Plato: <https://github.com/openplanets/plato>
- Public instance: <http://www.ifs.tuwien.ac.at/dp/plato>