SHAMAN

Sustaining Heritage Access through Multivalent ArchiviNg

Pericles

FP7 Digital Preservation

Digital preservation: research and practice

Elena Maceviciute

Co-funded by the European Union

SEVENTH FRAMEWORK PROGRAMME

Securing communication with the future

The aim of this lecture is to:

Sum up some main issues that you have heard today

Explain the research into the digital preservation issues

Present some ideas from one recent and one on-going research projects: SHAMAN and PERICLES

Outline

- The research into digital preservation: why and what
- Practical applications: where and how
- SHAMAN project – emphasis on re-use within a distributed storage
- PERICLES project – emphasis on significant environment information and semantics

Digital preservation research (EU)

Project name / duration	Project name / duration
ERPANET – 2001-2004	DigitalPreservationEurope – 2005-2009
CASPAR – 2005-2009	Parse.Insight – 2006-2010
PLANETS 2006-2010	LiWA – 2007-2011
Protage – 2006-2010	PrestoPrime – 2008-2012
SHAMAN – 2007-2011	KEEP – 2007-2011
BlogForever – 2009-2013	ARCOMEM – 2009-2013
APARSEN – 2010-2014	ENSURE – 2010-2014
TIMBUS – 2010-2014	SCAPE – 2010-2014
PRESTO4U – 2010-2014	4C – 2011-2015
PERICLES – 2012-2017	DIACHRON – 2011-2016

The preservation context

- Whether *born digital* or *digitized* the digital record is associated with:
  - the software employed
  - the hardware employed to produce
  - the record format or formats (e.g., pdf files with embedded jpg files)
  - the hardware for reading the format(s)
  - the software for reading the format(s)
  - the institutional policies on documentation
  - the standards observed in all of these

Digital preservation systems at work

- Deutches NationalBibliothek – KOPAL and KoLibri
- LOCKSS – Lots of copies keep stuff safe (user community consists of networks of libraries and publishers)
- Office of the Chief Archivist of Lithuania

SHAMAN: Sustaining Heritage Access through Multivalent Archiving

- International RTD project (2008-2011)
  - Participants: 16 partners from 7 EU countries + USA:
    - INMARK Estudios y Estrategias; InConTec GmbH; Xerox Research Centre Europe; Philips Innovation Lab; Industrious Media, Globale Informationstechnik GmbH
    - Deutsche Nationalbibliothek, Georg-August-Universität Göttingen Stiftung Öffentlichen Rechts; Instituto de Engenharia de Sistemas e Computadores

      Investigacao e Desenvolvimento em Lisboa
    - University of Liverpool ; SSLIS; FernUniversität in Hagen; University of Strathclyde; Otto-von-Guericke Universität Magdeburg, University of Glasgow; University of Illinois
  - Instrument: Collaborative Project
  - Funding agency: EU FP7

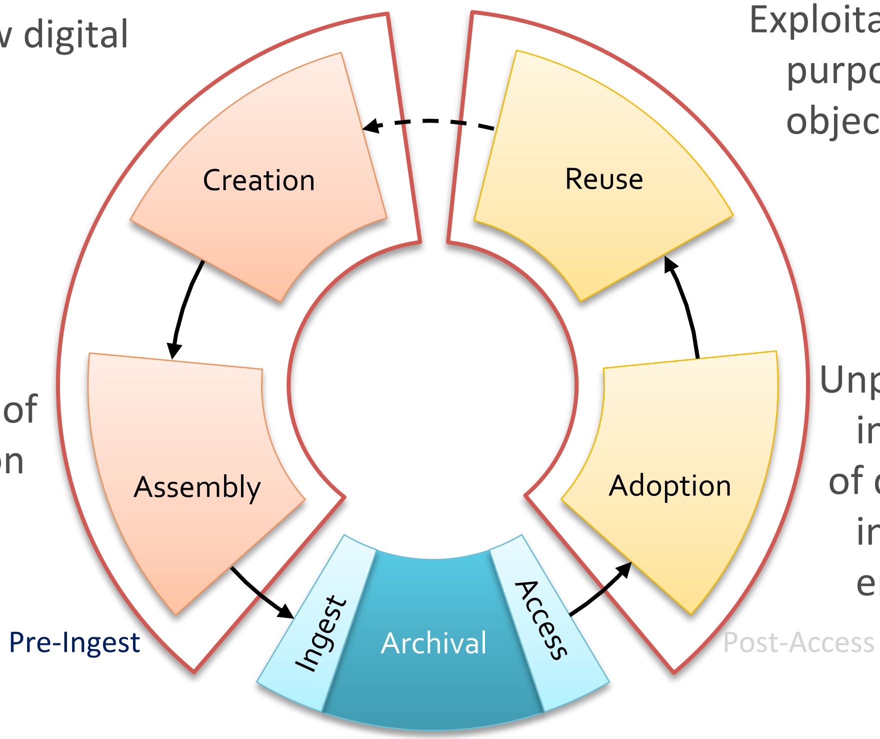
SHAMAN: aims and objectives

- *Communicate with the future* by means of securing the integrity of digital objects
- *Create a technology environment* to manage the storage, access, presentation, and manipulation of potentially any digital object over time
- *Build prototype* by a combination of grid technology, persistent archives and digital libraries for three user communities:
  - Memory institutions,
  - Industrial engineering companies
  - E-science

The preservation life-cycle

Production of new digital objects

Exploitation and re-purposing of digital objects



Appraisal and packaging of objects for preservation

Unpacking and integration of digital objects into working environment

Management of content for use over the long-term.

Preservation policies

- Effective preservation strategies require organizational policies for:
  - Selection of what needs to be preserved.
  - Identification of the intended user groups.
  - Selection of appropriate standards.
  - Access modes and rights.
  - Appropriate staff recruitment.
  - Collaborative agreements.
  - Migration and emulation.

Migration and emulation in SHAMAN

- Automatic migration of numerous files
  - File and metadata migration
  - Checking the quality of migrated materials

- Access to preserved bit stream using Multivalent technology
  - Generation of metadata on content

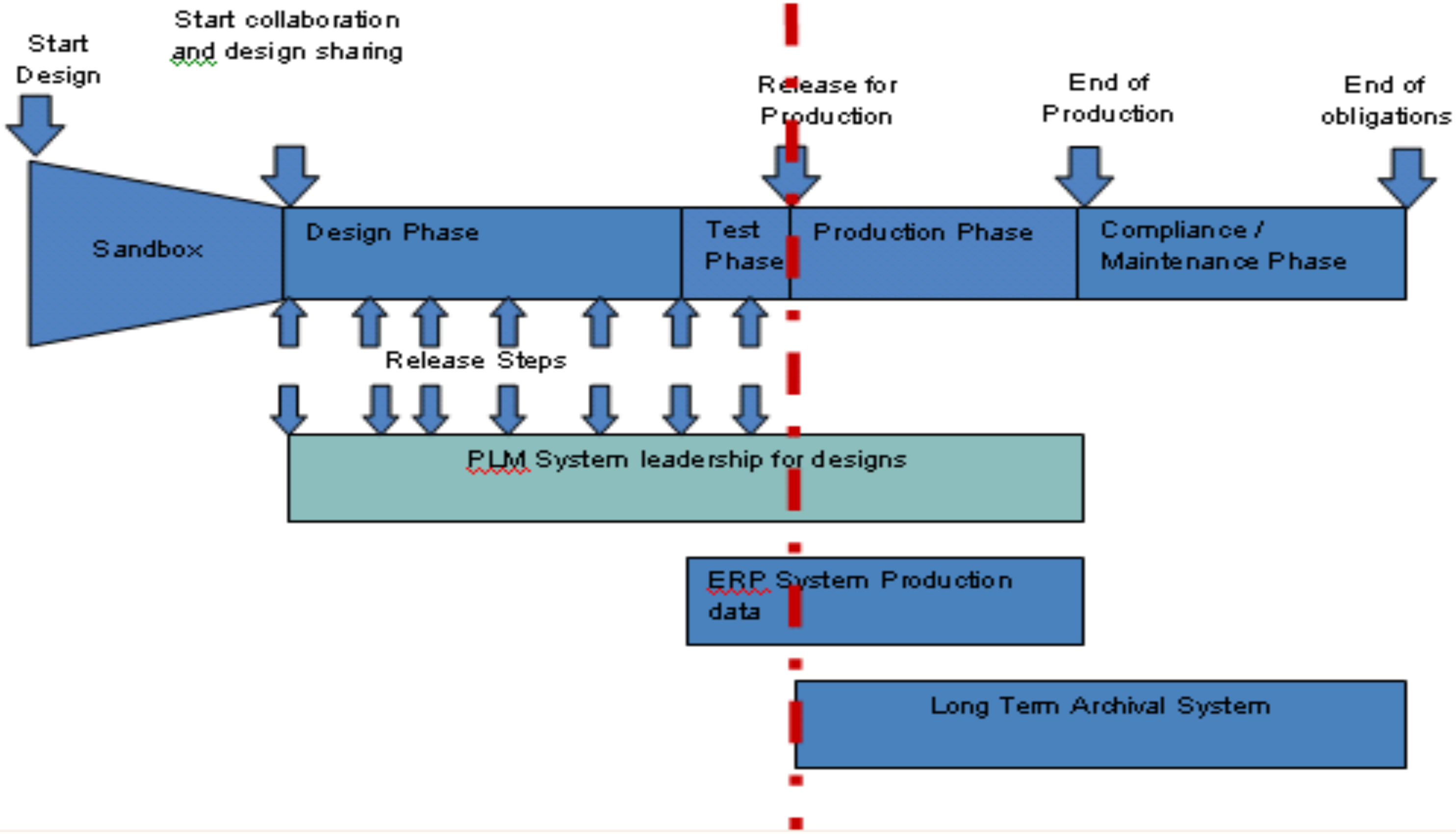
Access and use

- SHAMAN is using specific technology from Multivalent to represent preserved objects, by reading the bit-stream and the preserved context.
- The Multivalent browser allows the user of the archive to access and read the preserved records, whether the formats of those records continue to exist, or not.

The digital storage issue

- In SHAMAN *Grid technology* will enable interoperability among storage depositories located on local data grid networks.
- On the Grid we have:
    - data distributed across multiple sites and storage systems;
    - data managed independently of the storage system;
    - consistent management of file properties;
    - persistent identifiers and access controls; and
    - a scalable storage environment.

Engineering process and data management


Start Design
Start collaboration and design sharing
Release for Production
End of Production
End of obligations
Sandbox
Design Phase
Test Phase
Production Phase
Compliance / Maintenance Phase
Release Steps
PLM System leadership for designs
ERP System Production data
Long Term Archival System

PERICLES: Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics

- International RTD project (2013-2017)
  - Participants: 11 partners from 6 EU countries:
- Kings College London; University of Borås, Georg-August-Universität Göttingen Stiftung Öffentlichen Rechts; University of Liverpool; University of Edinburgh
- Xerox Research Centre Europe; Multimedia Knowledge and Social Media Analytics Laboratory; Dotsoft: Technology, Projects, Solutions
- TATE, Belgian User Support and Operation Centre; Space Applications Services NV
  - Instrument: Integrated Project
  - Funding agency: EU FP7 (ICT 9 call)

PERICLES: aims and objectives

To enable trusted access to digital content that is complex, heterogeneous, highly-interconnected, and subject to change and to facilitate continued understanding and reuse of those objects across all phases of the lifecycle.

- Developing a model based on a linked data paradigm for describing the resources in preservation environments.
- Addressing the evolution of digital ecosystems and their dependencies, and developing an associated framework and tools.
- Developing methods for identifying and capturing preservation-related information from digital content and its environment.

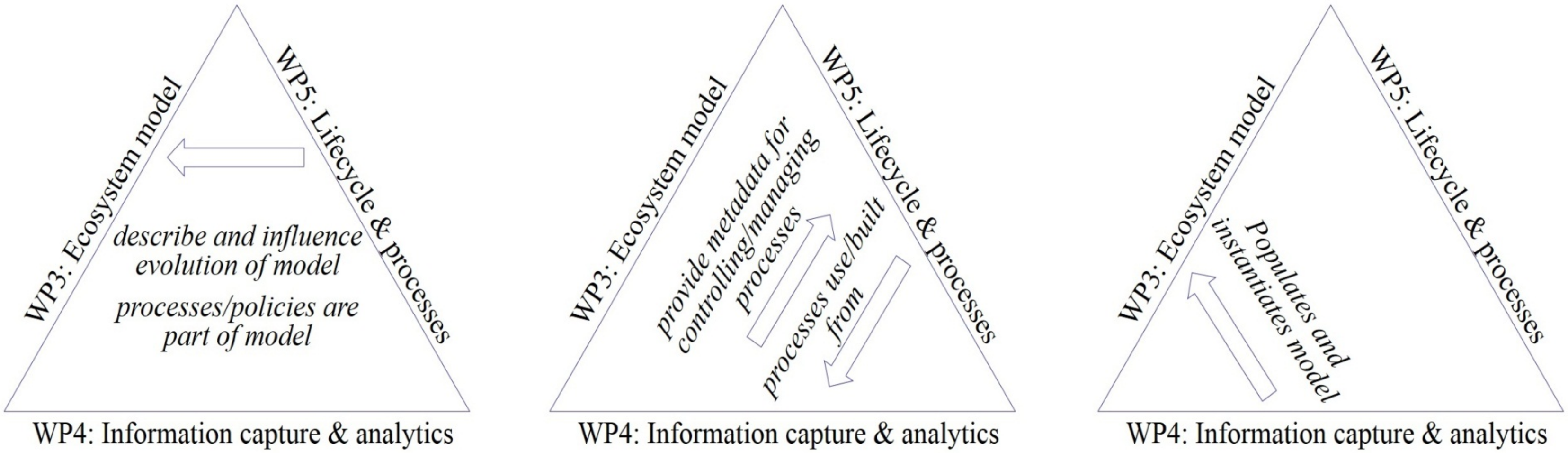
Domains of research: digital art and media

- TATE digital art and time-based media
  - interactive software-based installations
  - digital videos
  - digital multimedia
  - Websites as art objects

Domains of research: experimental space science

- BUSOC and SpaceApps: *Loss of earth science data has already occured: 19th century tropospheric ozone is only very partly preserved and might not be correctly represented in climate models.*
  - Solar data from sensors
  - Calibrated science data (After you have obtained your level 1(L1) data, the next step in data analysis is to use your calibrator and target observations to produce calibrated visibilities)
  - Integration of the data with a variety of documents

Main research activities



WP3: Ecosystem model

WP5: Lifecycle & processes

describe and influence evolution of model

processes/policies are part of model

WP4: Information capture & analytics

WP3: Ecosystem model

WP5: Lifecycle & processes

provide metadata for controlling/managing processes

processes use/built from

WP4: Information capture & analytics

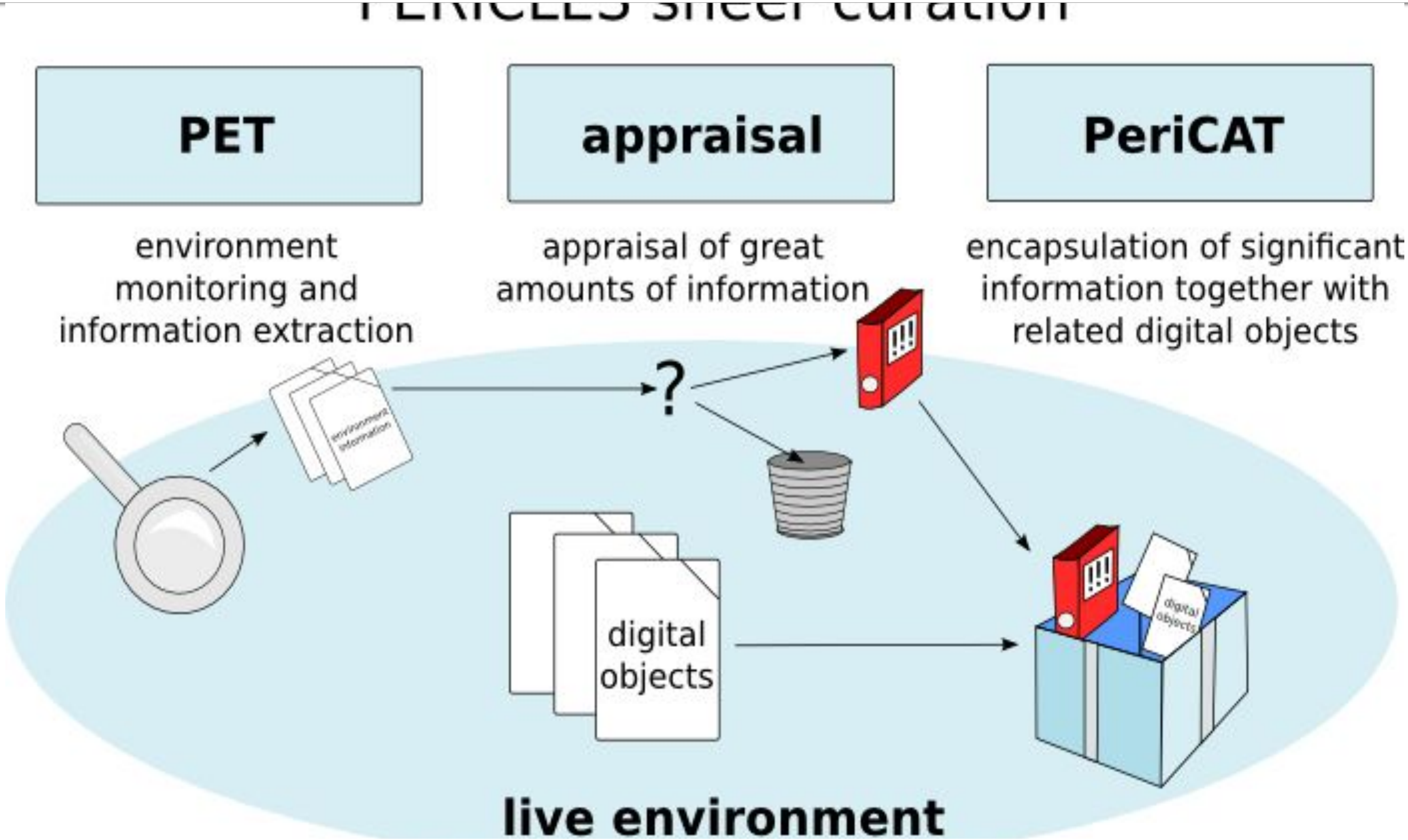
WP3: Ecosystem model

WP5: Lifecycle & processes

Populates and instantiates model

WP4: Information capture & analytics

PERICLES in sheer curation



PET

environment monitoring and information extraction

appraisal

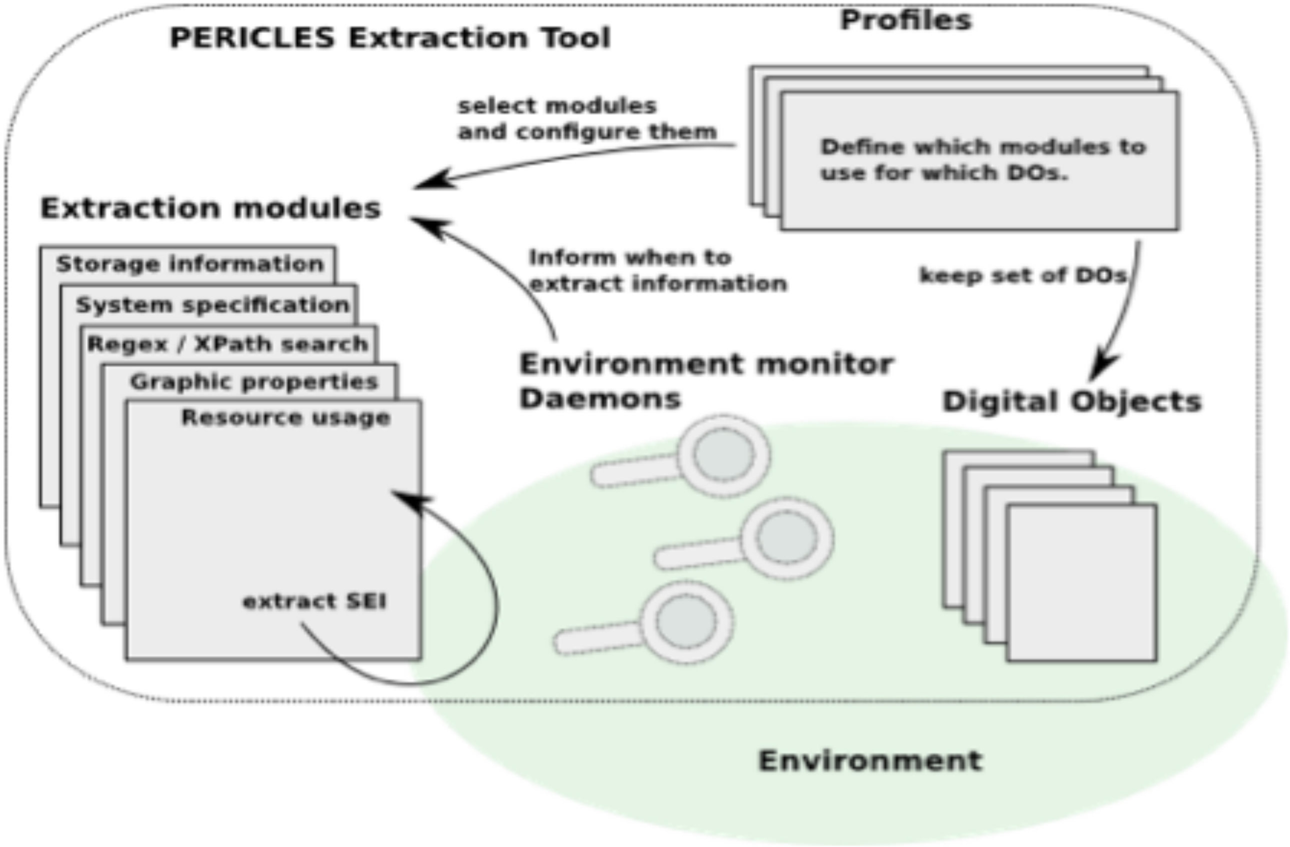
appraisal of great amounts of information

PeriCAT

encapsulation of significant information together with related digital objects

live environment

PET – extraction tool



Semantic drift

- Two main partners work on understanding how semantic drift occurs in the context of long-term digital preservation and how it may influence the technologies for access and re-use of the digital objects.
- Application: retrieval of preserved objects, management of archives over time, making sense of retrieved objects…

Conclusion

- In most cases there is a continuity between the digital preservation research projects (SHAMAN capturing metadata in engineering environment – PET development)
- Most of research projects search for high level solutions, and do not develop actual systems. The tools produced in the process in most cases are open source and available for further development

Big preservation questions:

- Do we need to preserve everything?
- Can we preserve everything?
- Can we ensure power supply in the long run?
- What institutions can be truste to do long-term preservation?
- What should be preserved?

https://www.youtube.com/watch?v=Z7SGxB88mEk