GRANT AGREEMENT: 601138 | SCHEME FP7 ICT 2011.4.3
Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics [Digital Preservation]

Pericles
FP7 Digital Preservation

PERICLES – Management of change to enable long term reuse

Simon Waddington (King's College London)

Our Digital Future, Cambridge, 14th–15th March, 2016

SEVENTH FRAMEWORK PROGRAMME

PERICLES Project

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

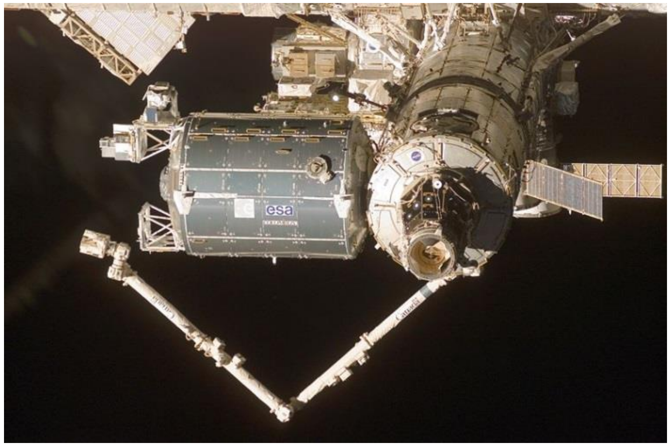
- EC FP7 Integrated Project, Digital Preservation (Feb. 2013– Jan. 2017). 11 partners.

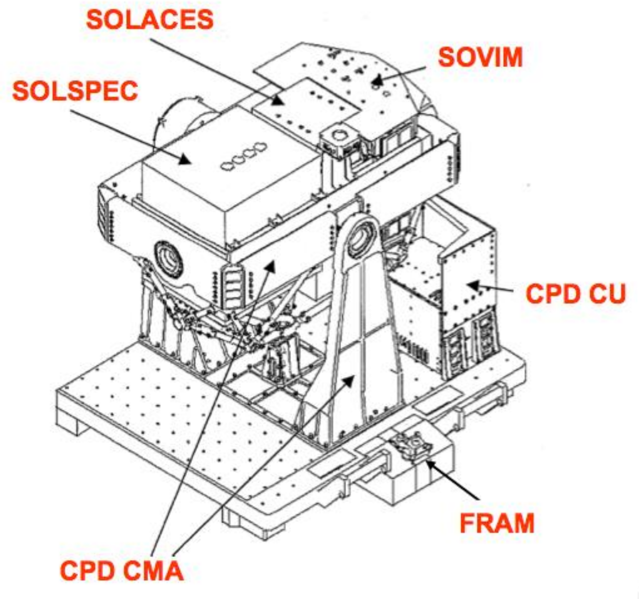
Objectives

- Facilitate continued understanding, access to and reuse of digital objects that are:
  - Heterogeneous, volatile and complex (highly interconnected)
- Enforce policies that govern management and evolution of content
- Integrated test beds
  - Addressing primarily space science and digital art domains
- Aim to develop reusable components to support ongoing reusability
  - Not a repository system

Science case study

- Science data originating from International Space Station
- SOLAR
  - Experiment that monitor the sun's spectral variability
  - Raw data and telemetry are captured by instrument
  - Data are calibrated by solar scientists
  - Dataset is made available to
    - Scientists in other fields (e.g. climate)
    - Users of other instruments
- Complex dependencies
- Long timeframes





Media case study

- Software-based artworks
  - Self-contained or networked systems
  - Comprise hardware and software elements
    - Proprietary/open source/custom software
  - Typically involve cutting edge technology
    - Unique and challenging to maintain
  - Unlike physical artworks, often necessary to replace elements
    - Works can exist in multiple versions
  - Synergies with the space science experiments
    - Complex dependencies

Sow Farm by John Gerrard

Brutalism, by Jose Carlos Martinat

Approach

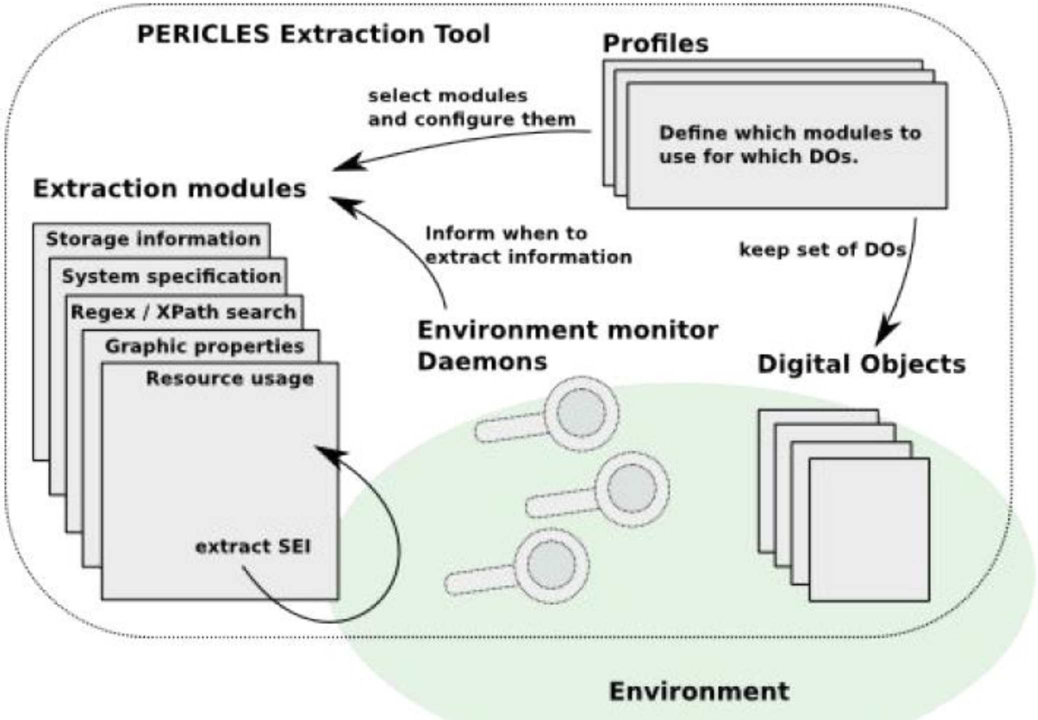
- Capture and representation of the environment
  - Understand the wider context around digital objects that impacts their long-term reuse

- Digital ecosystems
  - Analogy with biological systems
  - Evolving systems of interdependent entities

- Model-driven approach
  - Abstraction of complex systems as models that can be manipulated independently
  - Models are computational – not merely descriptive

- Continuum approach
  - Merging of active-life and archival phases
  - Non-custodial

Capture of the environment



PET – PERICLES Extraction Tool
https://github.com/pericles-project/pet
Apache Licence 2.0

- Available and used system resources
- File format identification and checksums
- Currently running processes
- Event information (file and network) from processes
- Graphic configuration information
- MS Office and PDF font dependencies
- Native commands

Why Digital Ecosystem?

- "Digital Ecosystem" represents the surrounding environment of a digital object that impacts reuse
  - Now or at a later point in time

- Digital ecosystem can include data objects, software, user communities, processes, technical services
  - Includes dependencies between entities

- Scope
  - The scope of the digital ecosystem depends on the particular use case

Types of change

- Archiving versus preservation

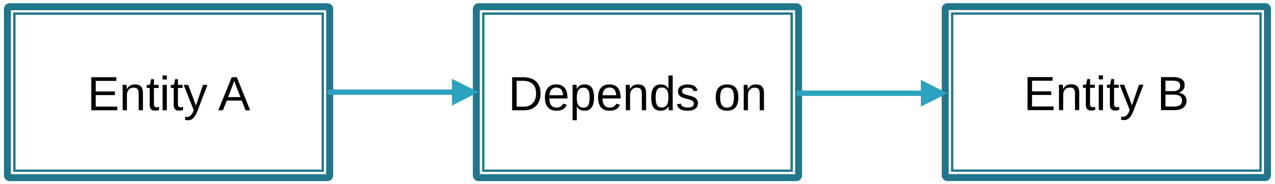
- Behavioural change
  - E.g. technological change, policy change, which have an impact on other entities through dependencies

- Semantic change
  - E.g. User community knowledge and practices

- If significant change occurs, it may impair or obstruct data reuse, access or interpretation

Dependency and change

- *Given objects A and B. A is dependent on B if changes to B have a significant impact on the state of A, or if changes to B can impact the ability to perform function X on A.”*

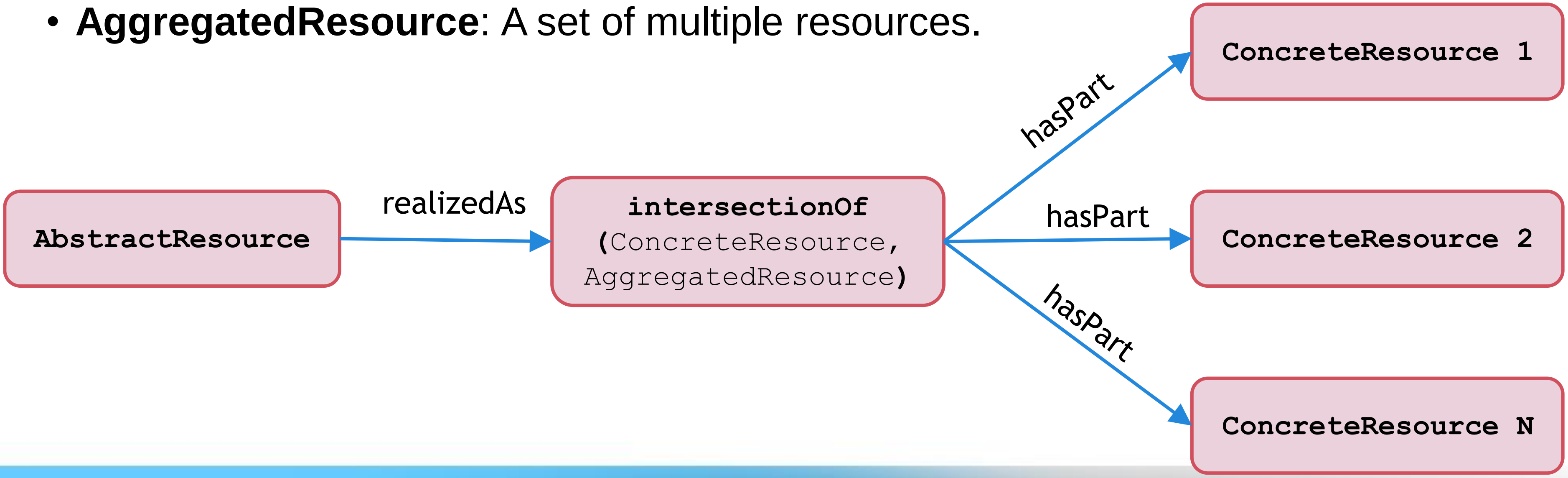


- PERICLES modelling language
  - Linked Resource Model (LRM) –Upper OWL ontology for modelling linked resources
  - DEM – formalism for digital ecosystems
  - Domain ontologies

LRM Resource

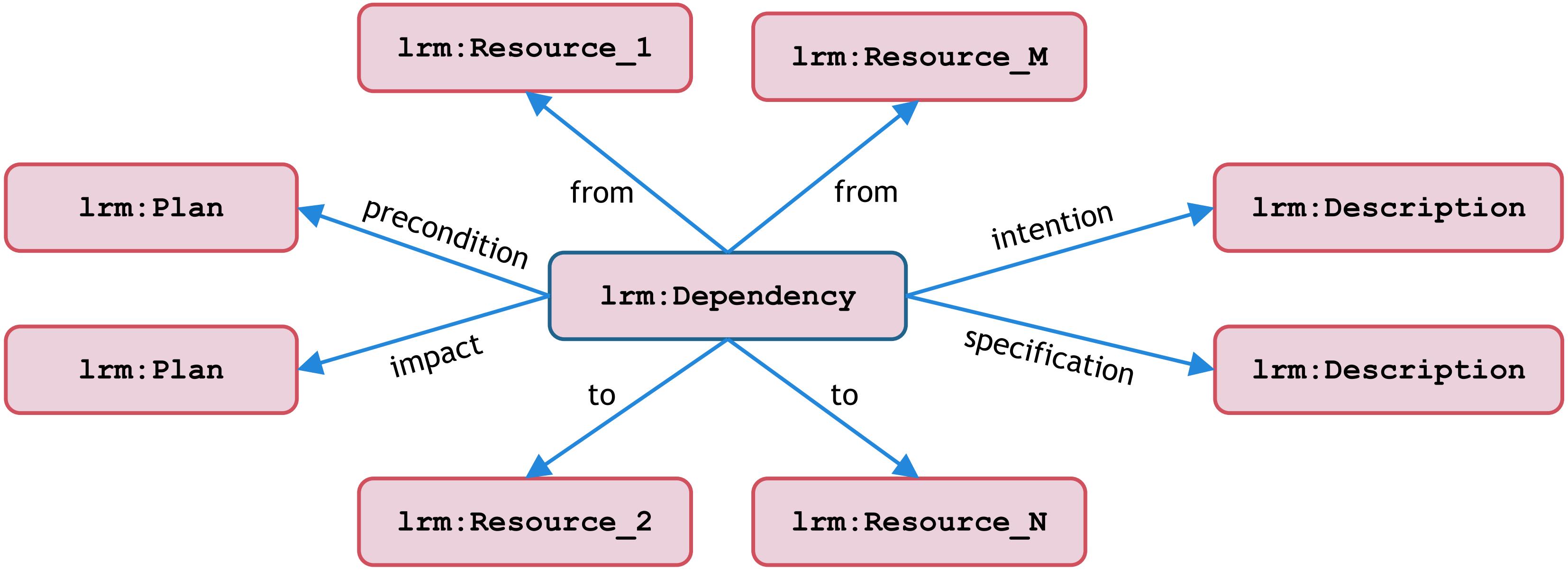
Any physical, digital, conceptual, or other kind of entity and in general comprises all things in the universe of discourse of the LRM Model.

- **AbstractResource**: Conceptual representation of an entity.
- **ConcreteResource**: Concrete realization of an abstract resource (with a physical extension).
- **AggregatedResource**: A set of multiple resources.

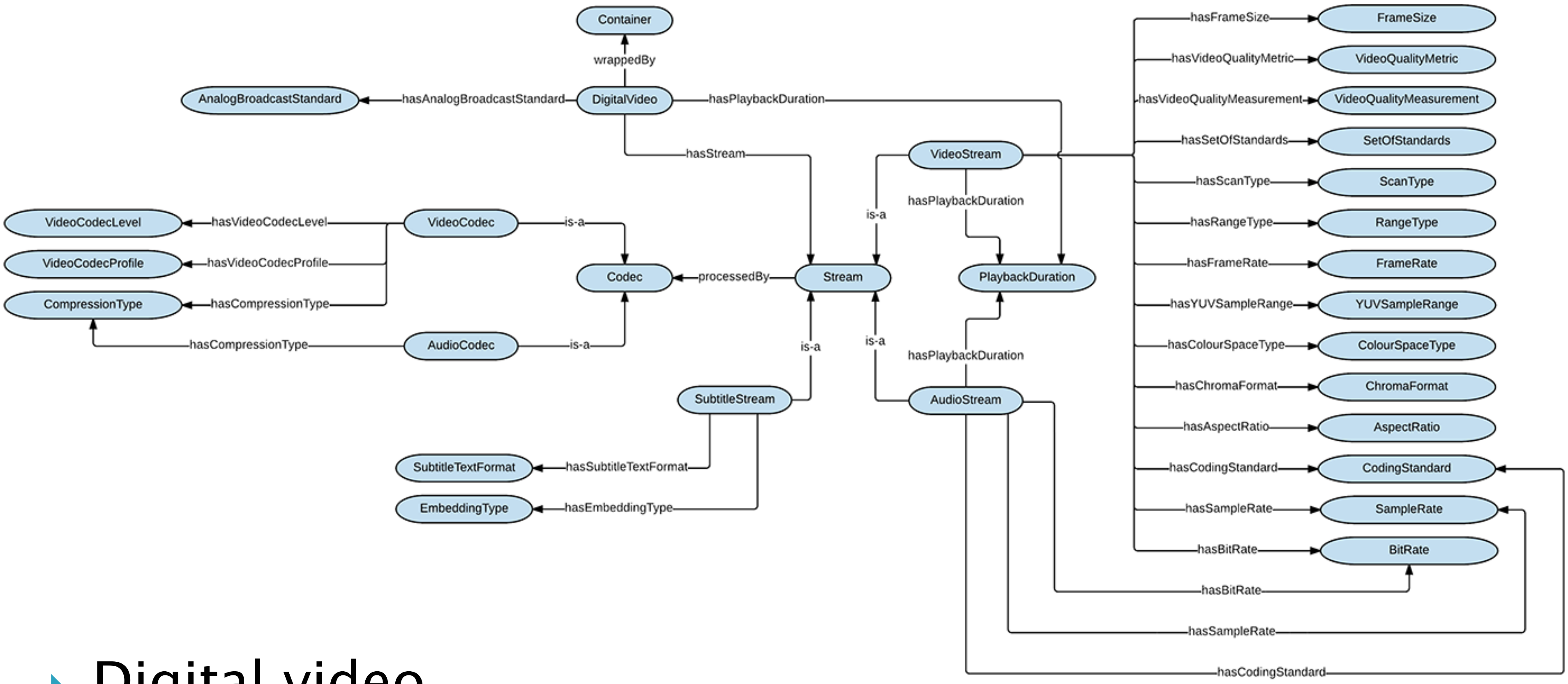


LRM Dependency

- Context under which change in one or more entities has an impact on other entities of the ecosystem



Ontology design patterns



Container

wrappedBy

AnalogBroadcastStandard

hasAnalogBroadcastStandard

DigitalVideo

hasPlaybackDuration

hasStream

VideoStream

hasPlaybackDuration

is-a

VideoCodecLevel

hasVideoCodecLevel

VideoCodec

is-a

VideoCodecProfile

hasVideoCodecProfile

CompressionType

hasCompressionType

Codec

processedBy

Stream

PlaybackDuration

hasCompressionType

AudioCodec

is-a

is-a

is-a

hasPlaybackDuration

SubtitleStream

AudioStream

SubtitleTextFormat

hasSubtitleTextFormat

EmbeddingType

hasEmbeddingType

hasFrameSize → FrameSize

hasVideoQualityMetric → VideoQualityMetric

hasVideoQualityMeasurement → VideoQualityMeasurement

hasSetOfStandards → SetOfStandards

hasScanType → ScanType

hasRangeType → RangeType

hasFrameRate → FrameRate

hasYUVSampleRange → YUVSampleRange

hasColourSpaceType → ColourSpaceType

hasChromaFormat → ChromaFormat

hasAspectRatio → AspectRatio

hasCodingStandard → CodingStandard

hasSampleRate → SampleRate

hasBitRate → BitRate

hasBitRate

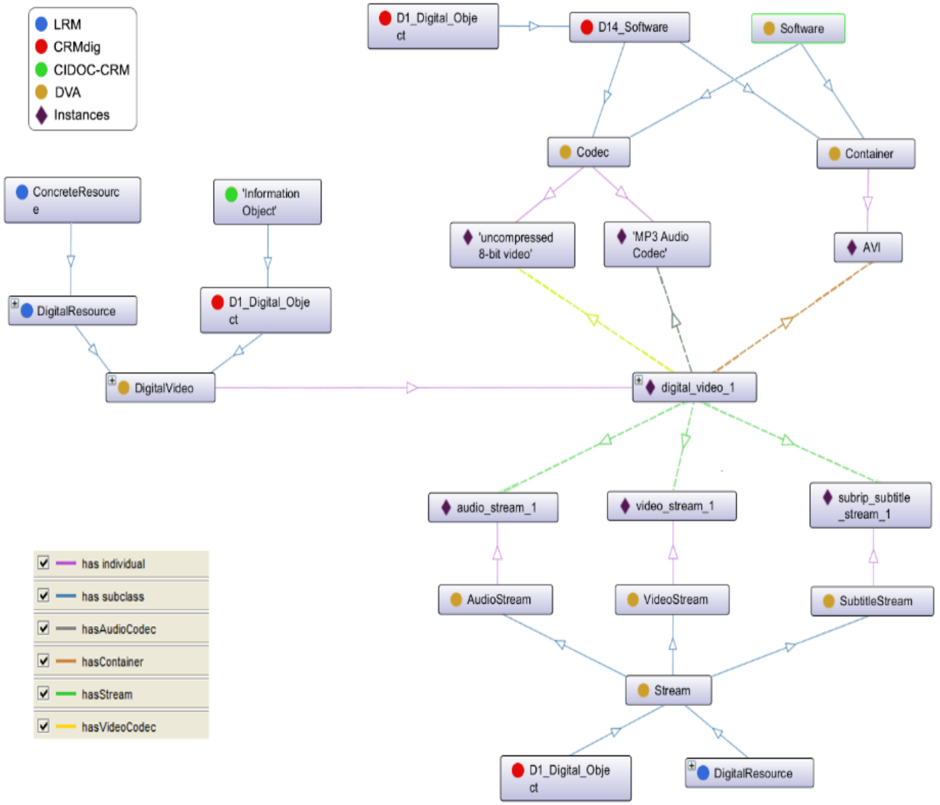
hasSampleRate

hasCodingStandard

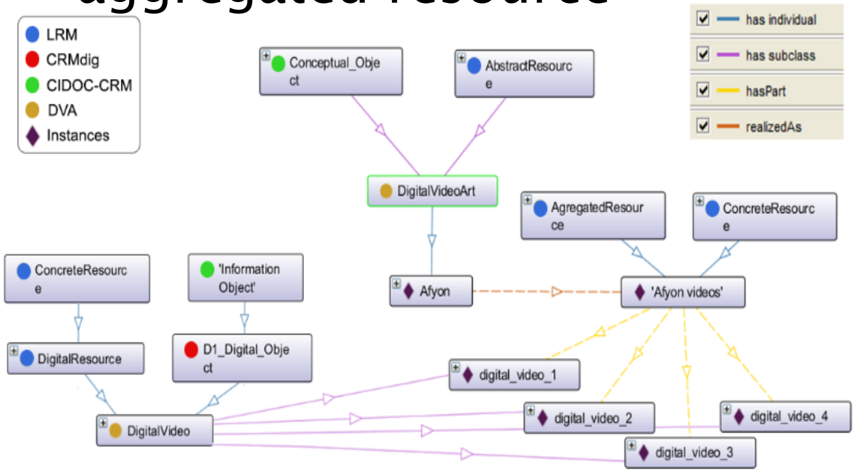
- Digital video
  - http://ontologydesignpatterns.org/wiki/Submissions:DigitalVideo

Example – Consistent video playback

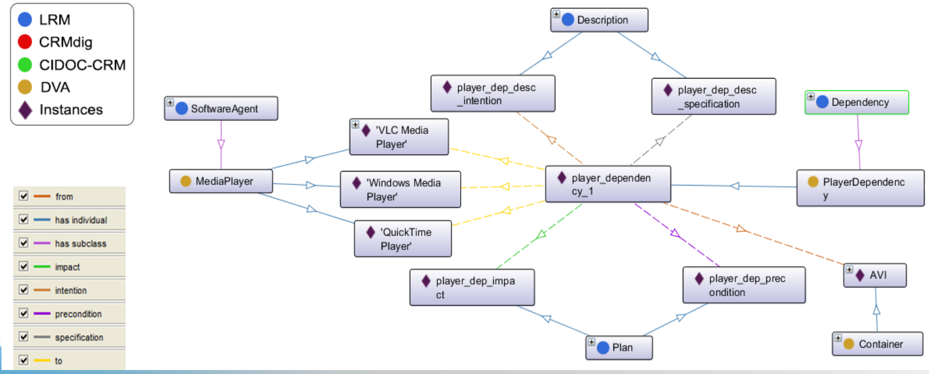
- Digital video playback
  - Representation of a digital video resource



- Video artwork as an aggregated resource



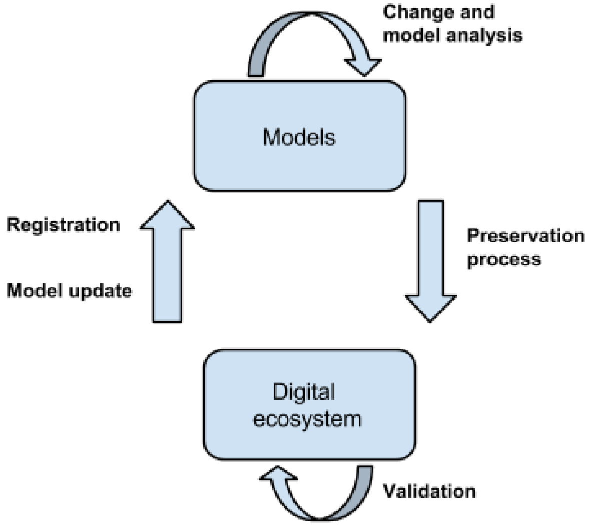
- Player dependency

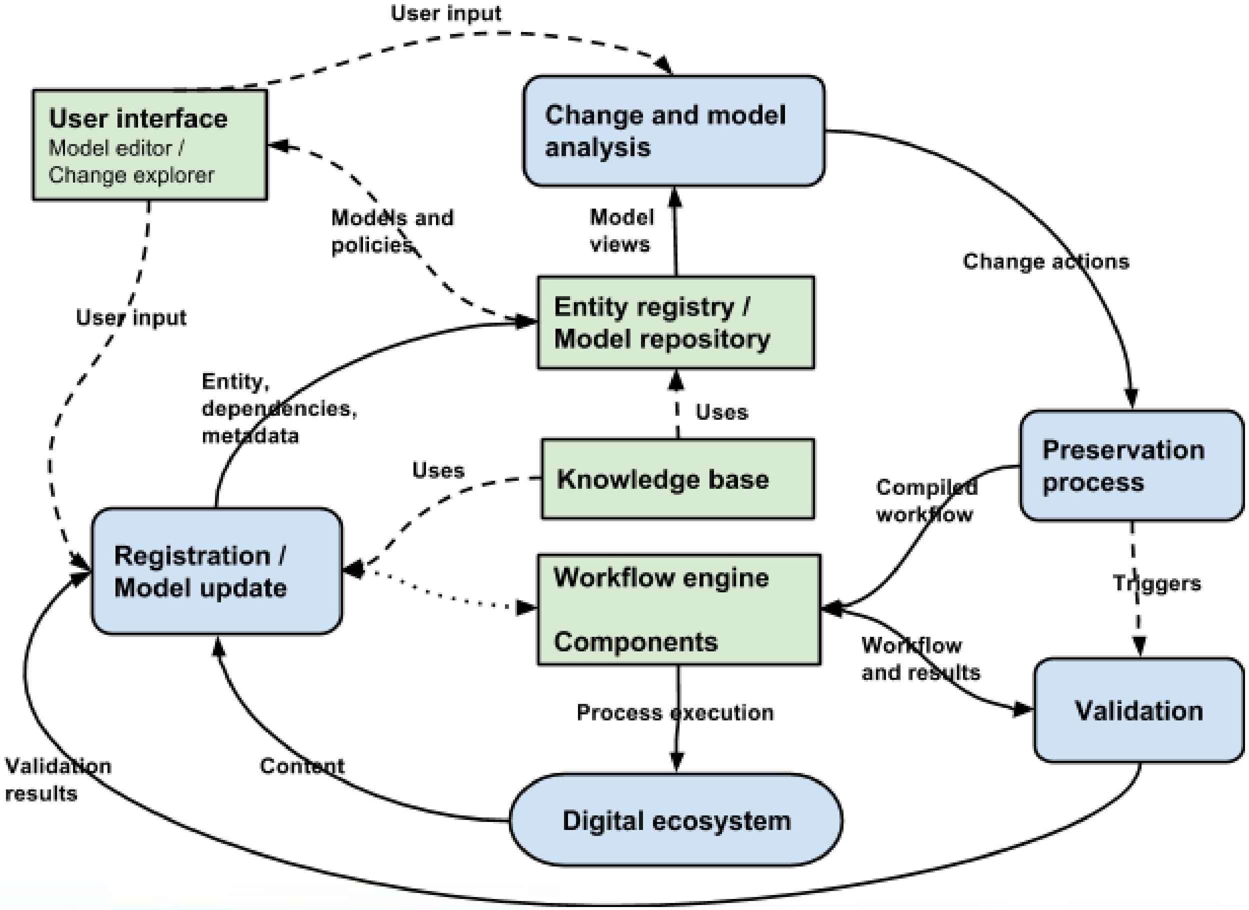


Populating the models

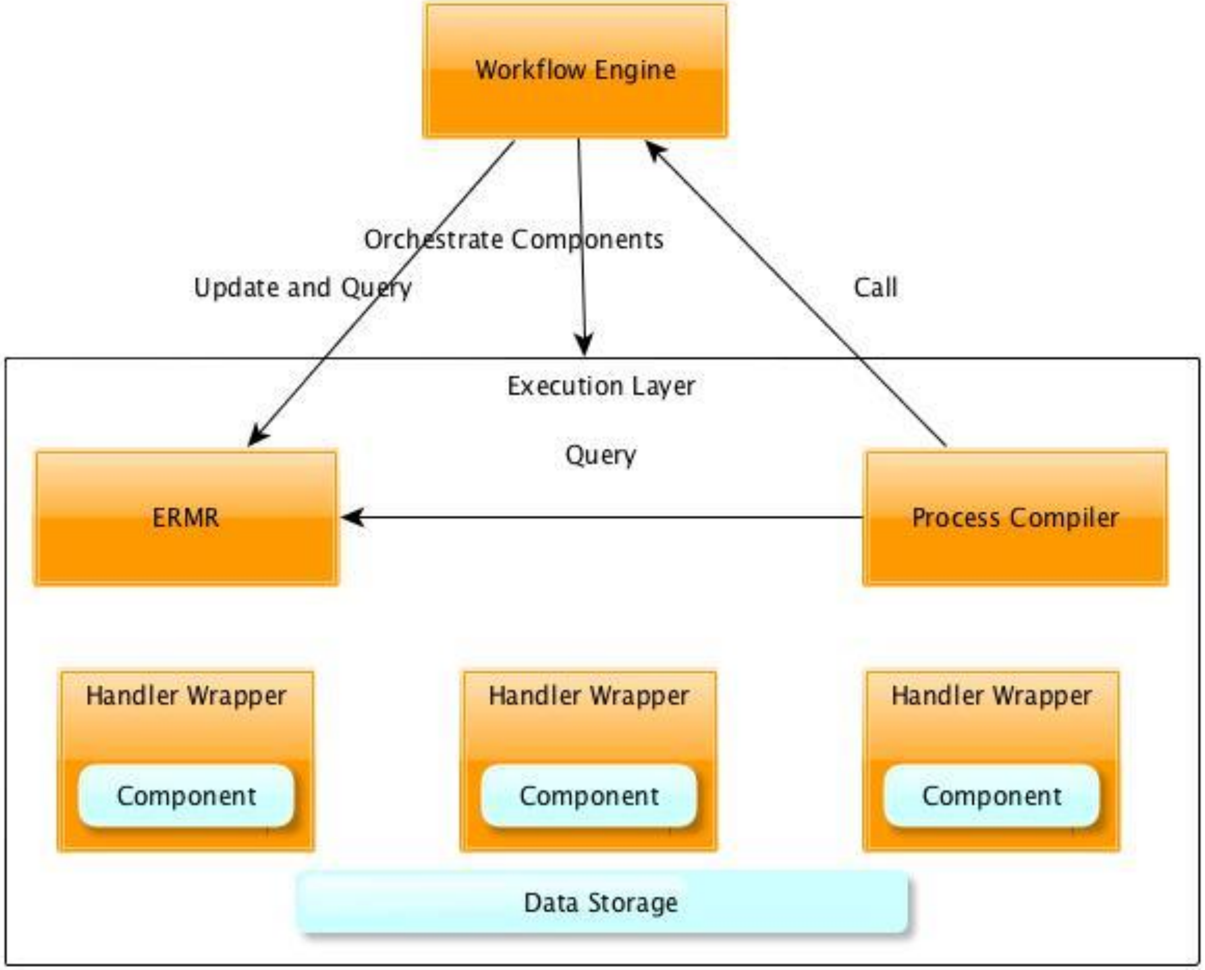
- Ontology design patterns
  - Reusable components that can be used across models
- Model editor
  - Manual editing through a GUI
- PET tool
  - Sheer curation tool running in background
  - PET2LRM
- Semantic extraction from text
  - Populating the ontologies with instances
- VERGE
  - Scalable feature extraction and feature processing from images and video

Model-driven preservation





Implementation – Test bed



- Technologies
  - Jenkins
  - Docker
  - Python-based Web Service Wrapper
  - jBPM

Jenkins + docker

Application of models

Predictive versus reactive strategies

- Reactive
  - Implement changes when there is a known failure or obsolescence (technology watch)
  - Disadvantages
    - Don’t enable forward planning or value assessment
    - Could result in loss of availability if major actions are required

- Predictive
  - “What if scenarios”
  - Manipulate independently of digital objects
  - Reduce “brute force” processing

Technical appraisal

- Can we preserve?
  - Risk due to hardware failure, software obsolescence, format obsolescence, semantic change etc.
- Three main dimensions
  - Risk – probability of an entity being unusable
  - Impact – potential loss of functionality and cost of mitigating actions
  - Proximity – time frame in which we consider risk/impact
- Models enable estimation of secondary risks
- MICE (Model Change Impact Explorer) tool
  - Visualisation of digital ecosystem and change impact

Lessons learned

- Model-driven approach
  - Upfront cost of building models versus benefits
  - Mitigated by reusability across different use cases
  - Use of design patterns
  - Ability to make predictions as well as react
  - Trade-off between high and low resolution models
  - Reflexive models – model the underlying preservation system

- Automation
  - For heterogeneous, volatile, complex objects, automation is essential
  - Decision-support not decision-making

Further information

- Contact: Simon Waddington
  - simon.waddington@kcl.ac.uk
- Website
  - http://pericles-project.eu/
- Public wiki
  - https://projects.gwdg.de/projects/pericles-public/wiki
- Twitter
  - https://twitter.com/PericlesFP7
- PERICLES Community of Practices