An Ontological Model for Digital Preservation

Panos Constantopoulos[1,2] and Vicky Dritsou[2]

[1]Information Systems and Databases Laboratory

Athens University of Economics and Business

[2]Digital Curation Unit, R.C. Athena

DigCCurr2007, 18-20 April 2007, Chapel Hill, NC, USA

Digital preservation -1

- Two kinds of perils face digital content
  - *Physical*: destruction of file systems, corruption of digital media, fire, earthquake
  - *Technological*: obsolete and incompatible systems, software, formats
- Physical perils are more straightforward to address
  - multiple copies of digital content:
    - On different media
    - At different geographic locations
- Technological hazards require a more complex policy
  - preservation strategies

Digital preservation -2

- Digital preservation strategies for technological hazards
  - Information migration
  - Technology emulation
  - Technology preservation
  - Backwards compatibility
  - Reliance on standards
  - Encapsulation
  - Transformation to non-digital form
  - Digital archeology

Metadata

- Preservation strategies usually require some information to be collected and stored: *metadata*
- Metadata kinds
  - Descriptive
  - Structural
  - Administrative
- Several metadata sets for preservation exist

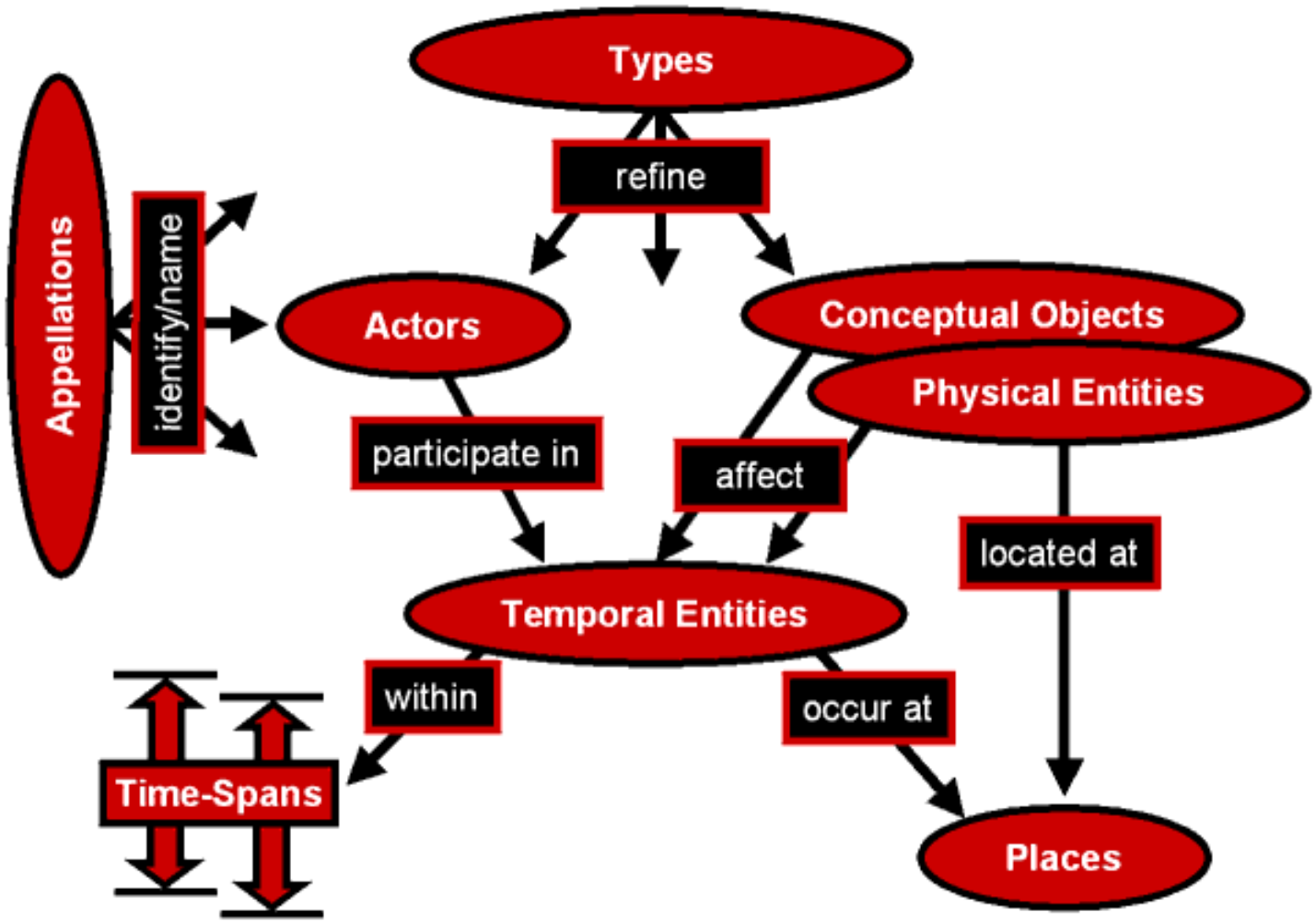
Preservation metadata sets

- We have focused on five widely known ones:
  - Dublin Core
  - Open Archival Information Systems (OAIS)
  - Curl Exemplars Digital Archives (CEDARS)
  - Pittsburgh Project
  - National Library of Australia (NLA)
- Discussion
  - DC: Access-oriented, inadequate
  - OAIS, CEDARS: very detailed, difficult to use
  - PP: detailed, necessary/optional elements, use instructions
  - NLA: Structured elements, object types
  - None contains inter-related concepts (element lists)

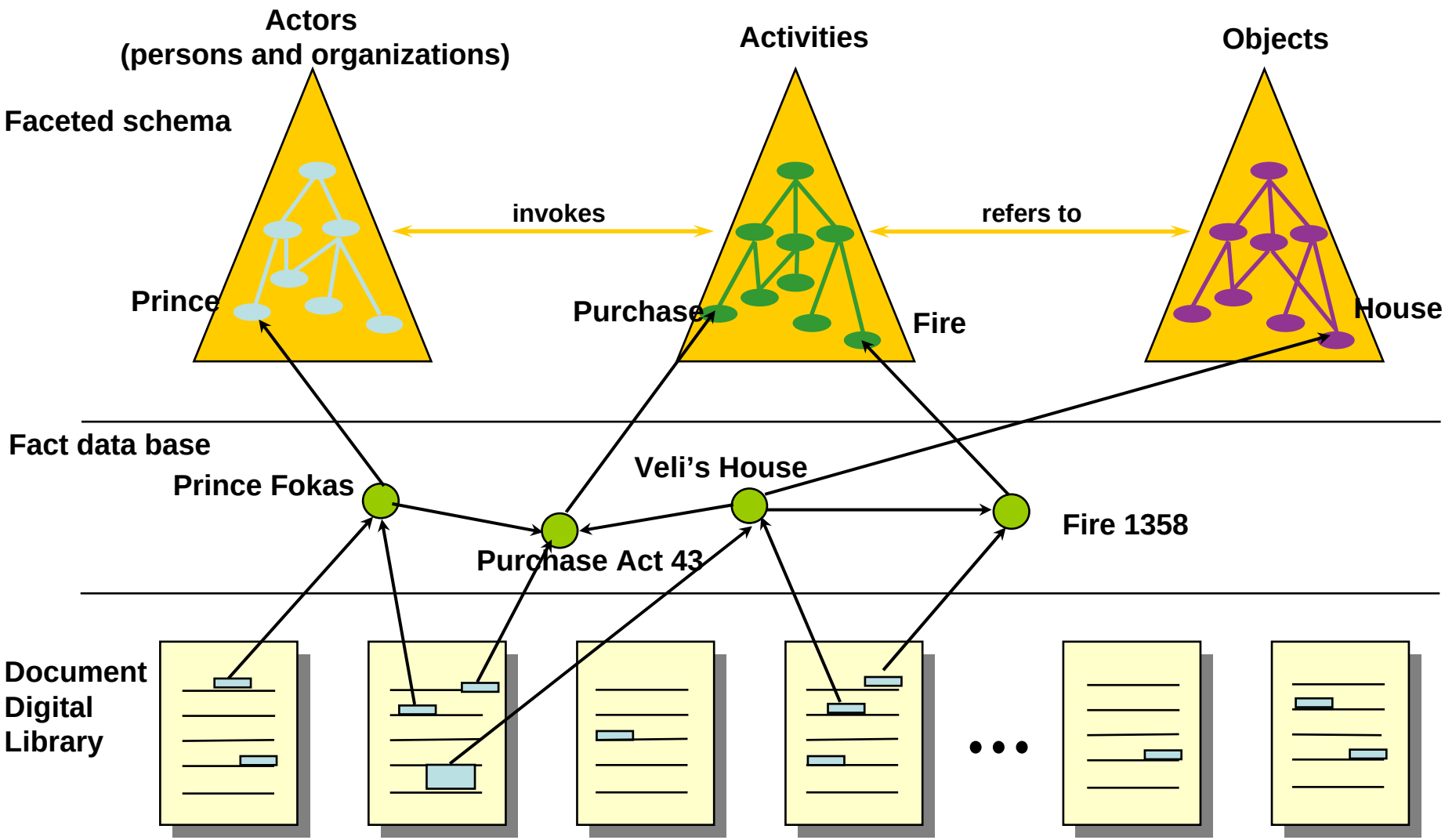
A conceptual model for preservation

- Motivation:
  - Documenting preservation activity at a finer semantic grain than unary property assignment can support useful inferences.
    - In cultural documentation this issue has been addressed.
  - Preserved digital objects can be considered as cultural objects themselves, therefore digital preservation is the counterpart of preserving collections of objects.
- Approach:
  - Define a preservation conceptual model compatible with CIDOC CRM / ISO 21127, the cultural domain ontology.
  - Ensure that the model covers the information requirements emerging from comparing the above proposed metadata sets.

CIDOC CRM: general structure



Reasoning over ontology-based knowledge networks



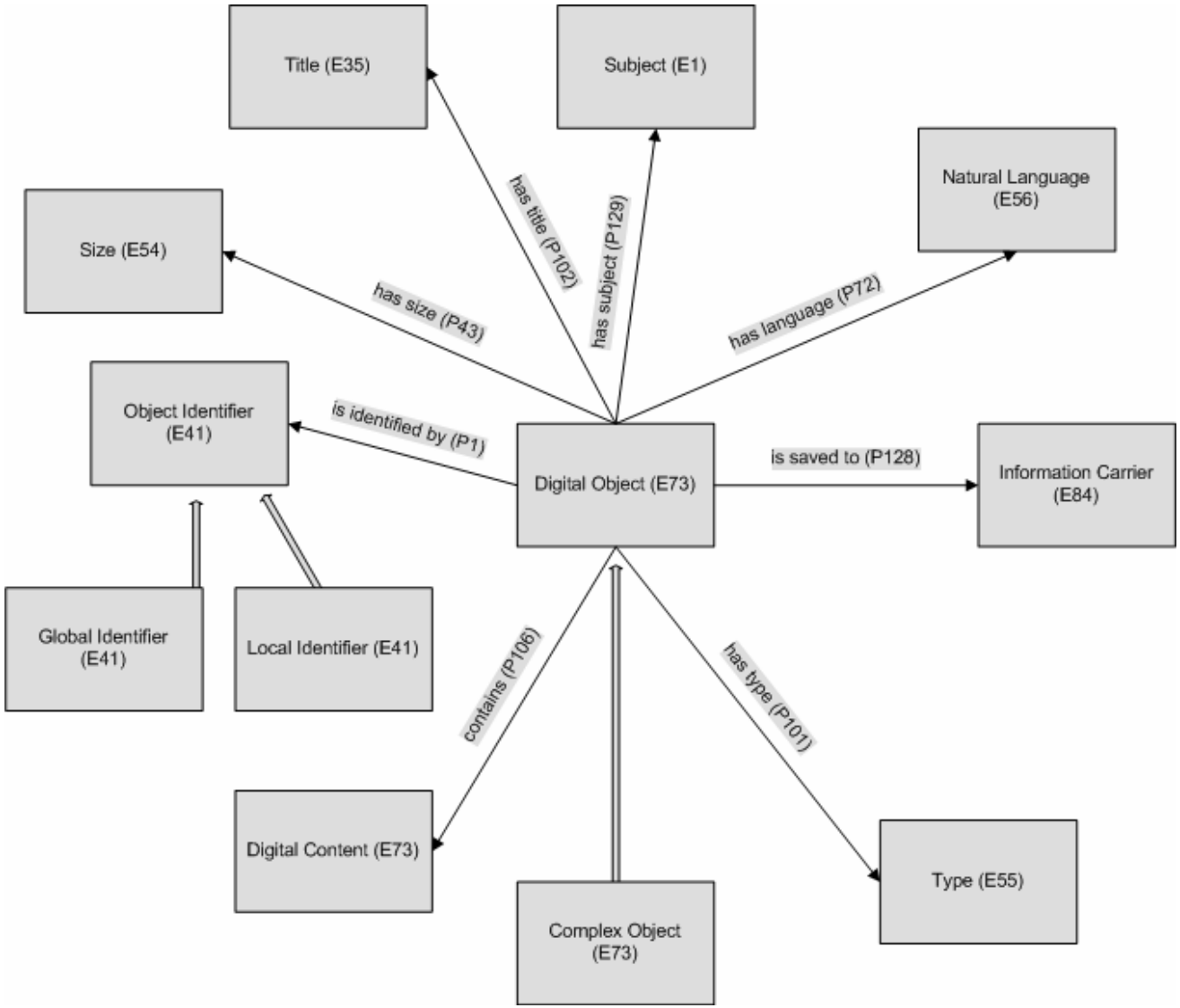
Modelling

- The following metadata elements are represented as entities in the model:
  - Title
  - Identifier
  - Subject
  - Language
  - Type
  - Format
  - Technical Equipment
  - Information Carrier
  - Activity
  - Right
  - Actor
  - Effect
  - History
- Relations between entities
- Model elements declared as subclasses of appropriate CIDOC CRM classes.
- A few elements are not derived from CIDOC CRM.
- An application ontology for digital preservation
  - Independent from preservation strategy

Model concepts -1

- Main concept: `Digital Object`
  - Subclass of `E73 Information Object`
  - Has attributes: `Title, Subject, Type, Size, Identifier, Language, Digital Content`
    - Identifiers may be local or global (unique)
    - Digital Content allows separation of content from descriptive/administrative aspects
  - Stored in an `Information Carrier`
  - Digital Objects can consist of other digital objects (`Complex Object`)
  - `Type`: image, text, sound, multimedia,…

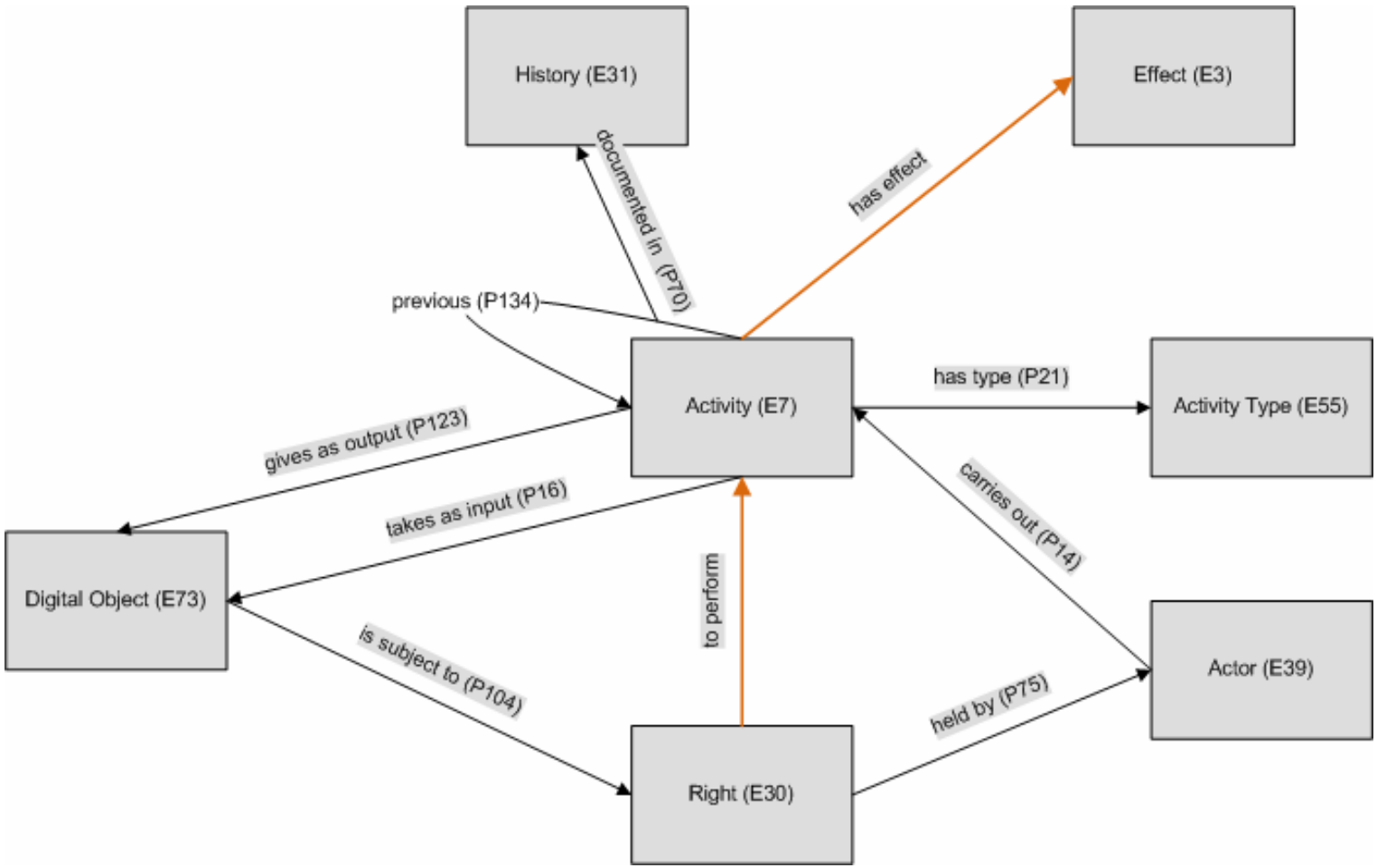
Schema -1


Title (E35)
Subject (E1)
Natural Language (E56)
Size (E54)
has title (P102)
has subject (P129)
has size (P43)
has language (P72)
Object Identifier (E41)
is identified by (P1)
Digital Object (E73)
is saved to (P128)
Information Carrier (E84)
Global Identifier (E41)
Local Identifier (E41)
contains (P106)
has type (P101)
Digital Content (E73)
Complex Object (E73)
Type (E55)

Model concepts -2

- `Activities` have digital objects as `input` and `output`, are carried out by `Actors` and are subject to `Rights`
- `Activity types`:
  - Create
  - Delete
  - Modify
  - Alter
  - Copy
  - Read
    - In all of them, except Read and Deletion, the output is a new object
- The sequence of performed Activities is recorded by `previous` and documented in `History`
- `Effects` can be used as a space-saving device when versions need not be kept

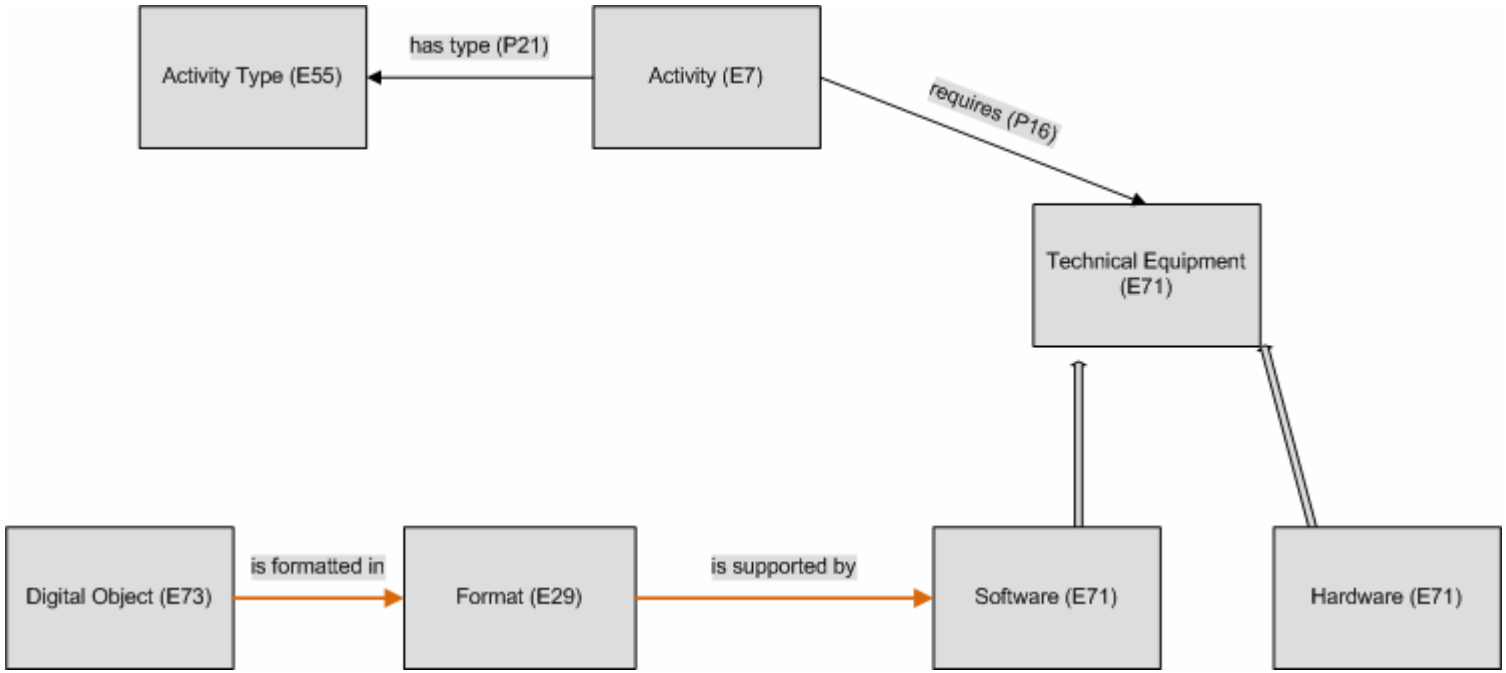
Schema -2


History (E31)
Effect (E3)
documented in (P70)
has effect
previous (P134)
Activity (E7)
has type (P21)
Activity Type (E55)
gives as output (P123)
takes as input (P16)
carries out (P14)
Digital Object (E73)
to perform
Actor (E39)
is subject to (P104)
held by (P75)
Right (E30)

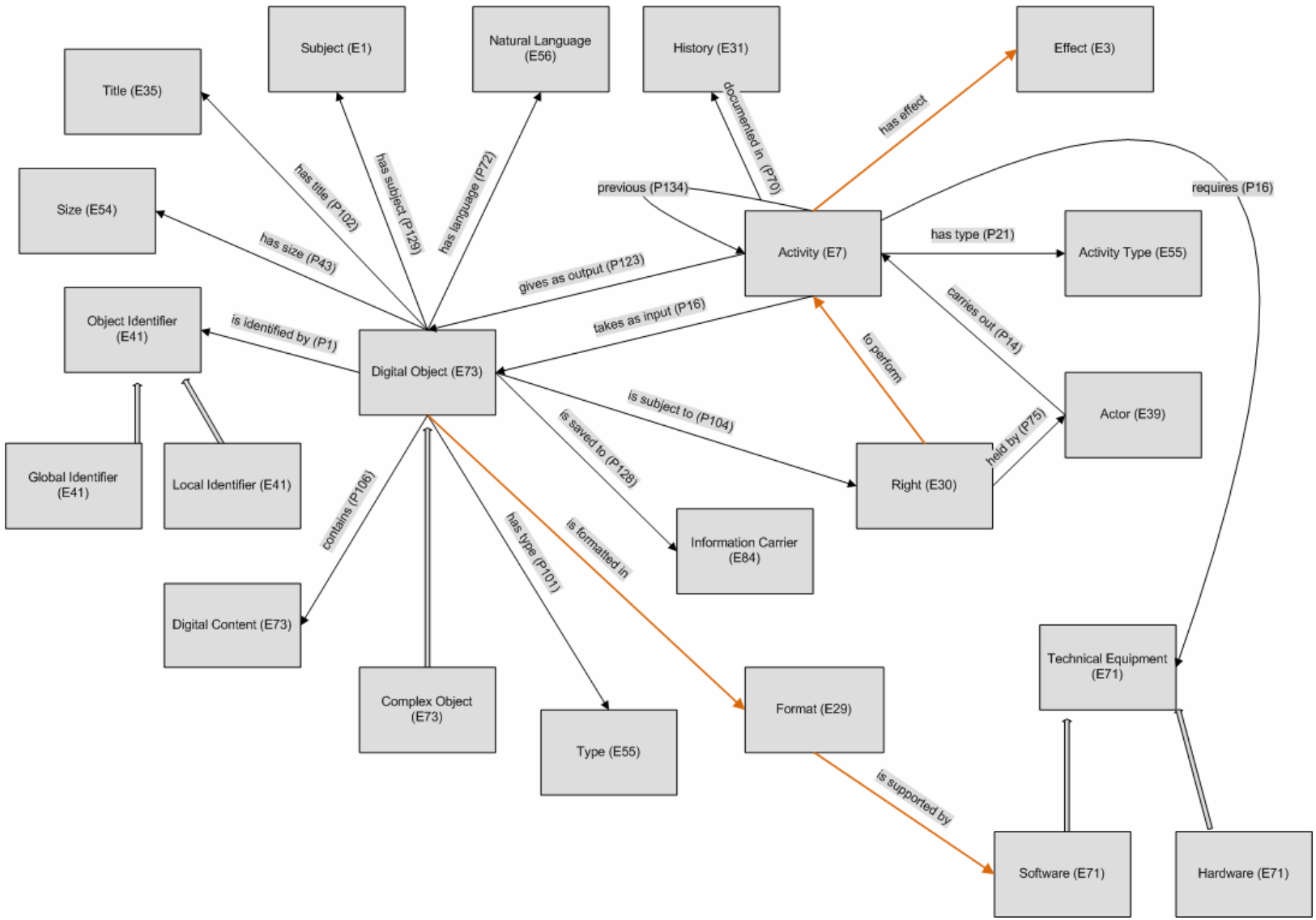
Model concepts -3

- Each object follows a specific `Format`
  - This Format is supported by specific `Software` products
- Activities require the appropriate `Technical Equipment` to be performed (`Software, Hardware`)
  - These are all specializations of `E71 Man-Made Thing`

Schema -3


Activity Type (E55)
has type (P21)
Activity (E7)
requires (P16)
Technical Equipment (E71)
Digital Object (E73)
is formatted in
Format (E29)
is supported by
Software (E71)
Hardware (E71)

The complete model



Model entities and parent CIDOC CRM concepts

Preservation Model Entity	Parent CIDOC CRM Entity
Digital Object	E73 Information Object
- Complex Object	E73 Information Object
Digital Content	E73 Information Object
Object Identifier	E41 Appellation
- Global Identifier	E41 Appellation
- Local Identifier	E41 Appellation
Size	E54 Dimension
Title	E35 Title
Subject	E1 CRM Entity
Natural Language	E56 Language
Type	E55 Type
Format	E29 Design or Procedure
Information Carrier	E84 Information Carrier
Technical Equipment	E71 Man-Made Stuff
- Software	E71 Man-Made Stuff
- Hardware	E71 Man-Made Stuff
Activity	E7 Activity
Activity Type	E55 Type
Actor	E39 Actor
Right	E30 Right
Effect	E3 Condition State
History	E31 Document

Property Name	Domain	Range	Parent CIDOC CRM Property
contains	Digital Object	Digital Content	E73 Information Object. P106 is composed of (forms part of): E73 Information Object
is identified by	Digital Object	Object Identifier	E1 CRM Entity. P1 is identified by (identifies): E41 Appellation
has size	Digital Object	Object Identifier	E70 Stuff. P43 has dimension (is dimension of): E54 Dimension
has title	Digital Object	Title	E71 Man-Made Stuff. P102 has title (is title of): E35 Title
has subject	Digital Object	Subject	E73 Information Object. P129 is about (is subject of): E1 CRM Entity
has language	Digital Object	Natural Language	E33 Linguistic Object. P72 has language (is language of): E56 Language
has type	Digital Object	Type	E70 Stuff. P101 had as general use (was use of): E55 Type
is saved to	Digital Object	Information Carrier	E24 Physical Man-Made Stuff. P128 carries (is carried by): E73 Information Object
is formatted in	Digital Object	Format	---
is supported by	Format	Software	---
carries out	Actor	Activity	E7 Activity. P14 carried out by (performed): E39 Actor
is subject to	Digital Object	Right	E72 Legal Object. P104 is subject to (applies to): E30 Right
held by	Right	Actor	E39 Actor. P75 possesses (is possessed by): E30 Right
to perform	Right	Activity	---
takes as input	Activity	Digital Object	E7 Activity. P16 used specific object (was used for): E70 Stuff
gives as output	Activity	Digital Object	E81 Transformation. P123 resulted in (resulted from): E77 Persistent Item
hat type	Activity	Activity Type	E7 Activity. P21 had general purpose (was purpose of): E55 Type
requires	Activity	Technical Equipment	E7 Activity. P16 used specific object (was used for): E70 Stuff
has effect	Activity	Effect	---
previous	Activity	Activity	E7 Activity. P134 continued (was continued by): E7 Activity
documented in	previous	History	E31 Document. P70 documents (is documented in): E1 CRM Entity

Conclusion

- Metadata elements drawn from existing metadata sets
- Conceptual model for digital preservation
  - Previous works included only lists of metadata elements
  - Extensible as needed
- Compatible with CIDOC CRM

  Digital objects as
  - digital surrogates of non-digital objects
  - cultural objects by themselves

Further work

- Historical processes:
  - interpretation
  - CIDOC CRM domain of application
- Preservation processes:
  - decision and production processes
  - Prescription and monitoring

➢ Explore differences in modelling requirements

The Digital Curation Unit / R.C. Athena

Launched by the **Athena** Research and Innovation Center in Information, Communication and Knowledge Technologies, Athens, Greece in Jan. 2007.

- Mission:

  To conduct research, develop technologies and applications, provide services and training, and act as a national pole in the field of digital curation.

  – Technological and organizational support for archiving and preservation
  – Best practice, standards and methodologies
  – Semantically enhanced information and communication services based on curated collections
- Will involve research laboratories, cultural organisations, archives, libraries, scientific data repositories, government agencies and private institutions
- Interdisciplinary approach
- Spans the entire lifecycle of digital assets:
  – production of high quality, dependable digital assets
  – organisation, archiving and long-term preservation
  – generation of added value from digital assets by means of resource-based knowledge elicitation
- Contact: Prof. Panos Constantopoulos, Director

  Email: panosc@aueb.gr, Tel./fax: +30-210-8203551