Data integration and metadata structures

CLARIN/Language Bank Seminar 15/12-08

Christian-Emil Ore

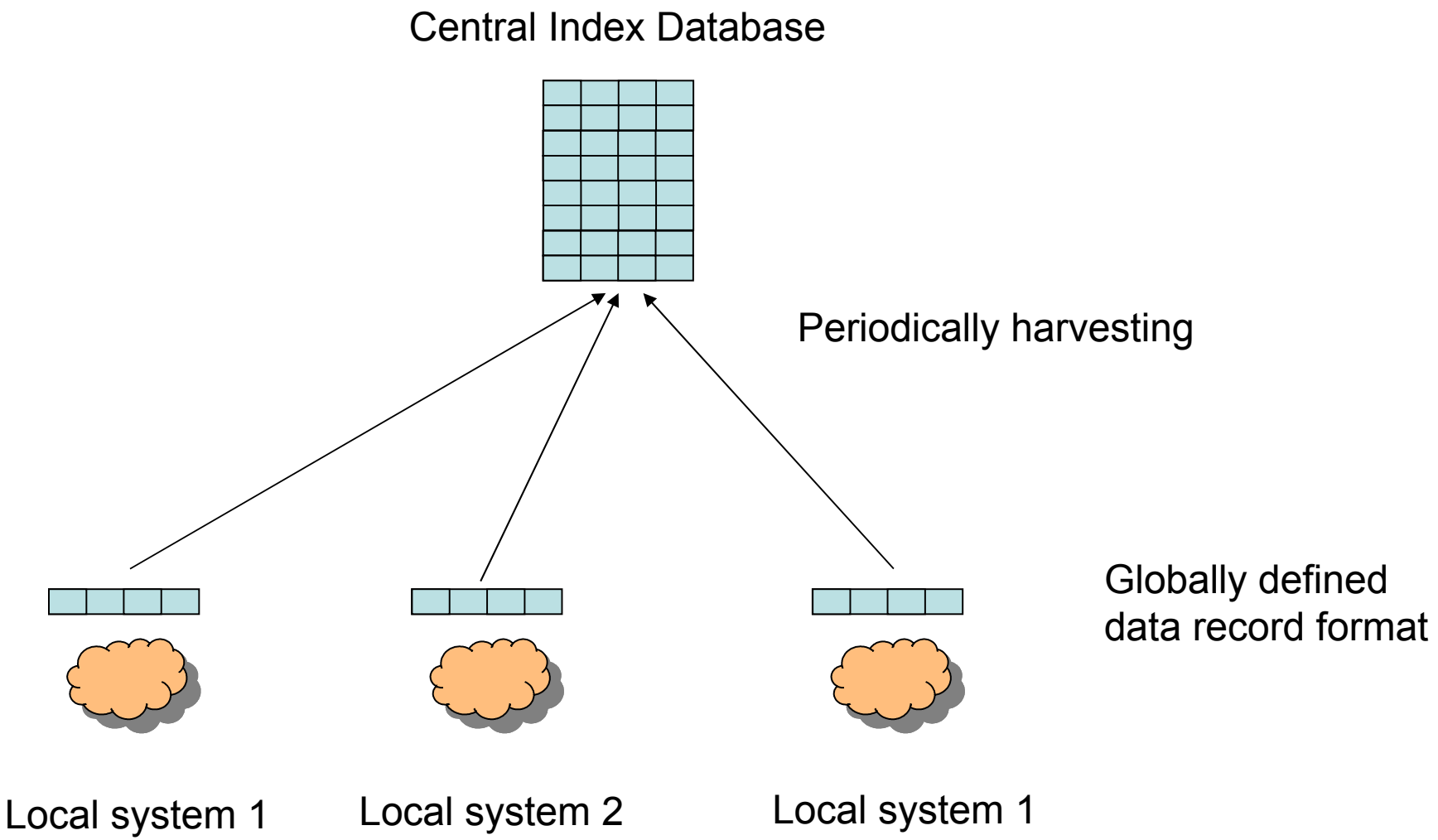
University of Oslo

CLARIN Objectives?

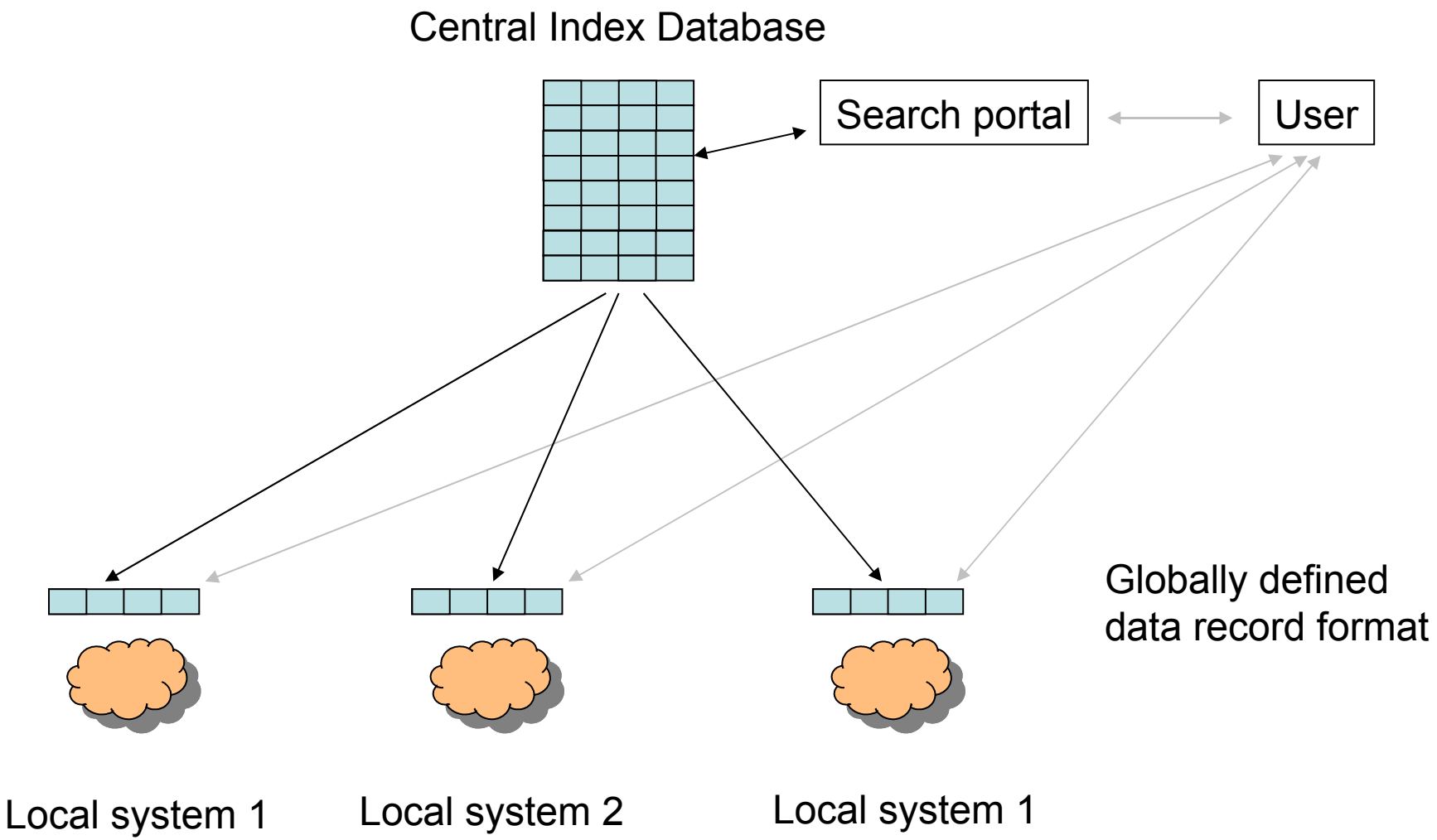
- Collection of separate LRT-Capsules (Language Resources and Technologies)?
  - Texts of all sorts which can be digitized medieval sources, web-sites, newspapers, digitized books etc
  - Multimedia recordings (audio/video) and time series recorded during communication (data glove, eye tracking, etc)
  - Various types of manually or automatically created annotations on texts, media streams etc
  - Tools such as aligners, speech recognizers, tokenizers, part-of-speech taggers, parsers, manual annotators, viewers etc
  - Various types of knowledge sources encapsulating knowledge about resources and languages such as metadata descriptions, GIS, lexica, concept registries, ontologies, etc

- Databases with knowledge extracted from such sources?

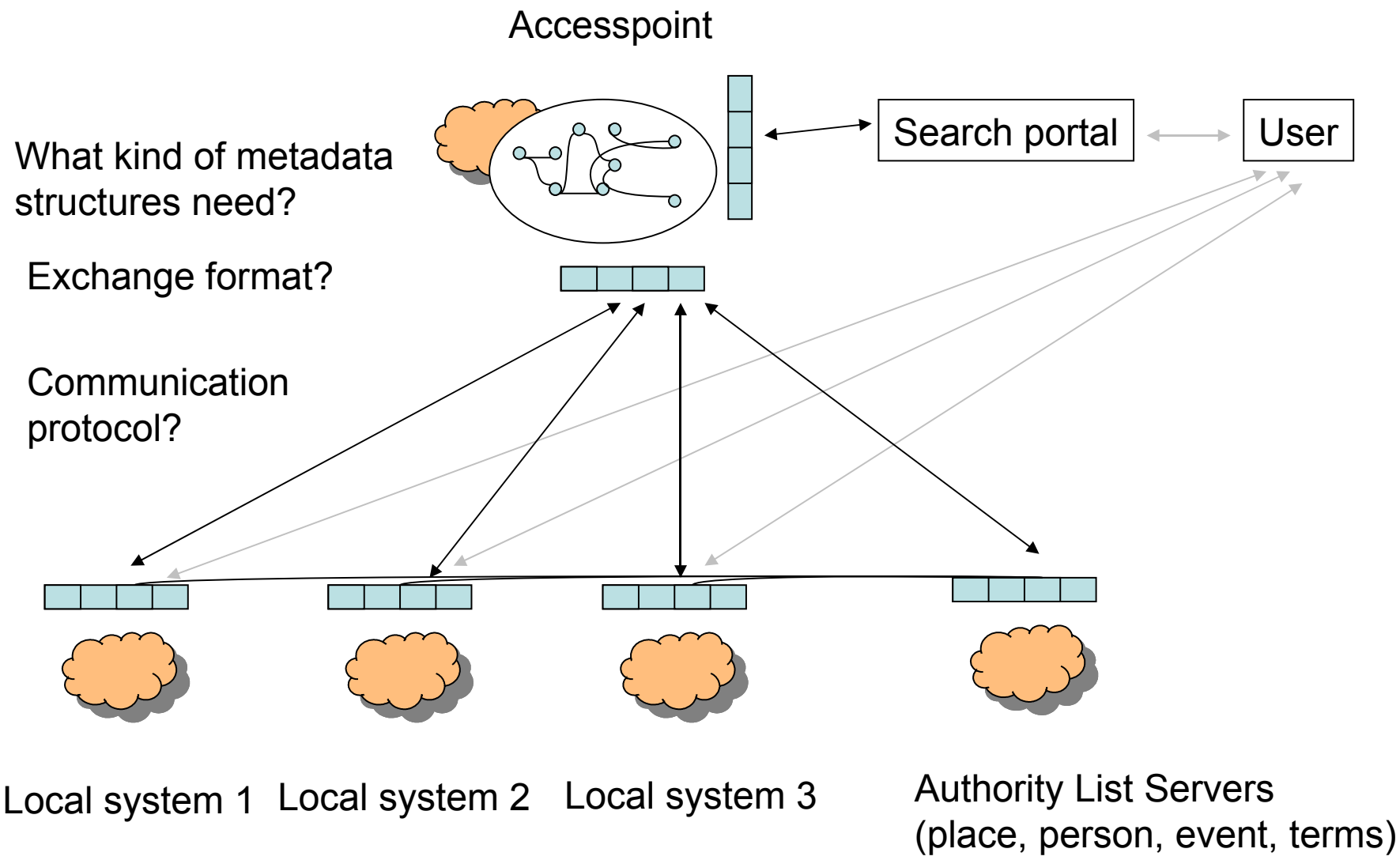
Data integration – simple OAI style architecture



Data search – simple OAI style architecture



Data integration – architecture



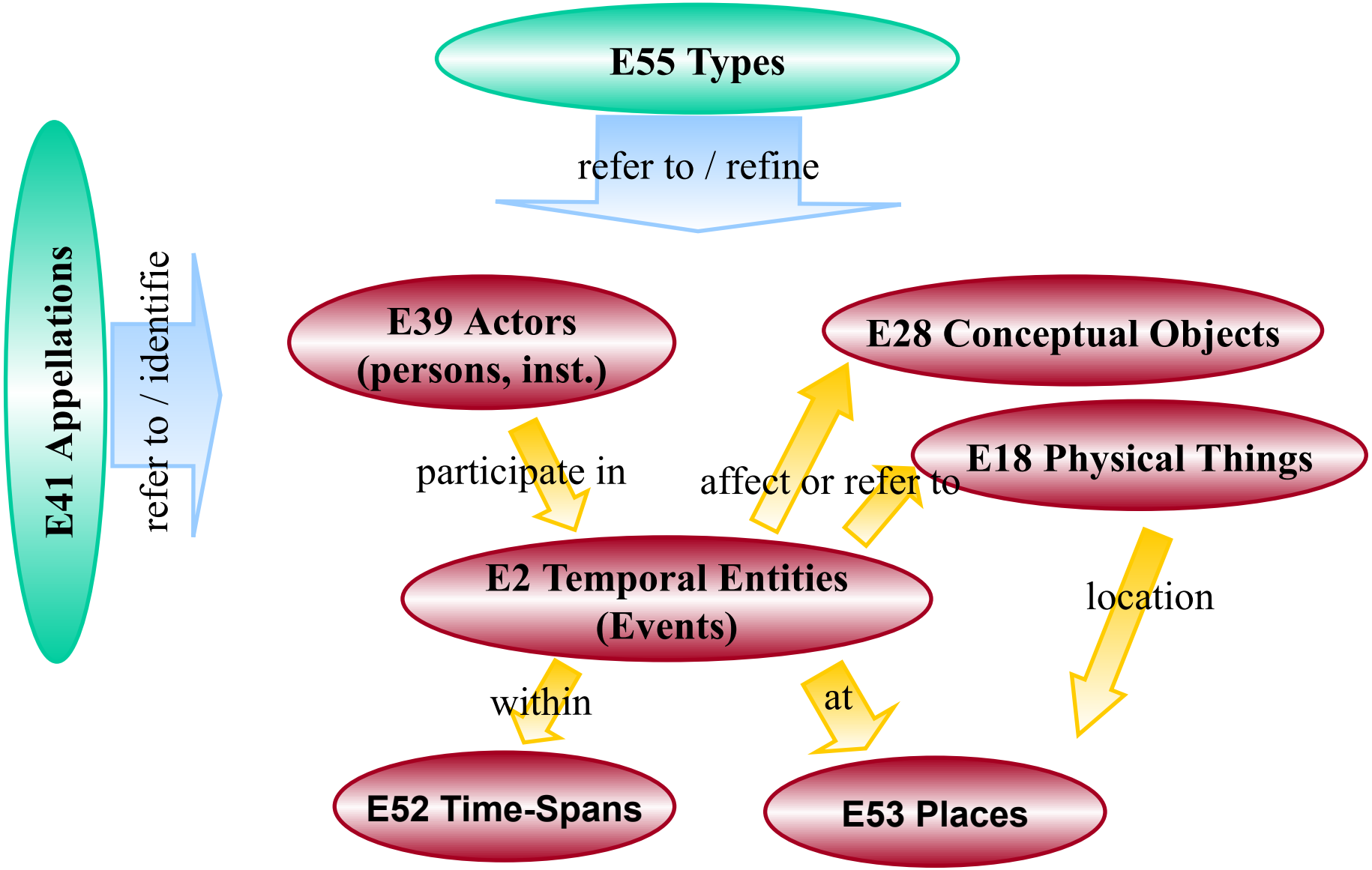
The CIDOC Conceptual Reference Model
(cidoc.ics.forth.gr)

- What is the CIDOC CRM?
  - An object oriented ontology developed by ICOM-CIDOC, 1996-2005
  - Accepted as ISO-21127 in September 2006
  - About 80 classes and 130 properties for cultural and natural history
  - CRM instances can be encoded in many forms: RDBMS, ooDBMS, XML, RDF(S), Topic Maps, DL, OWL.

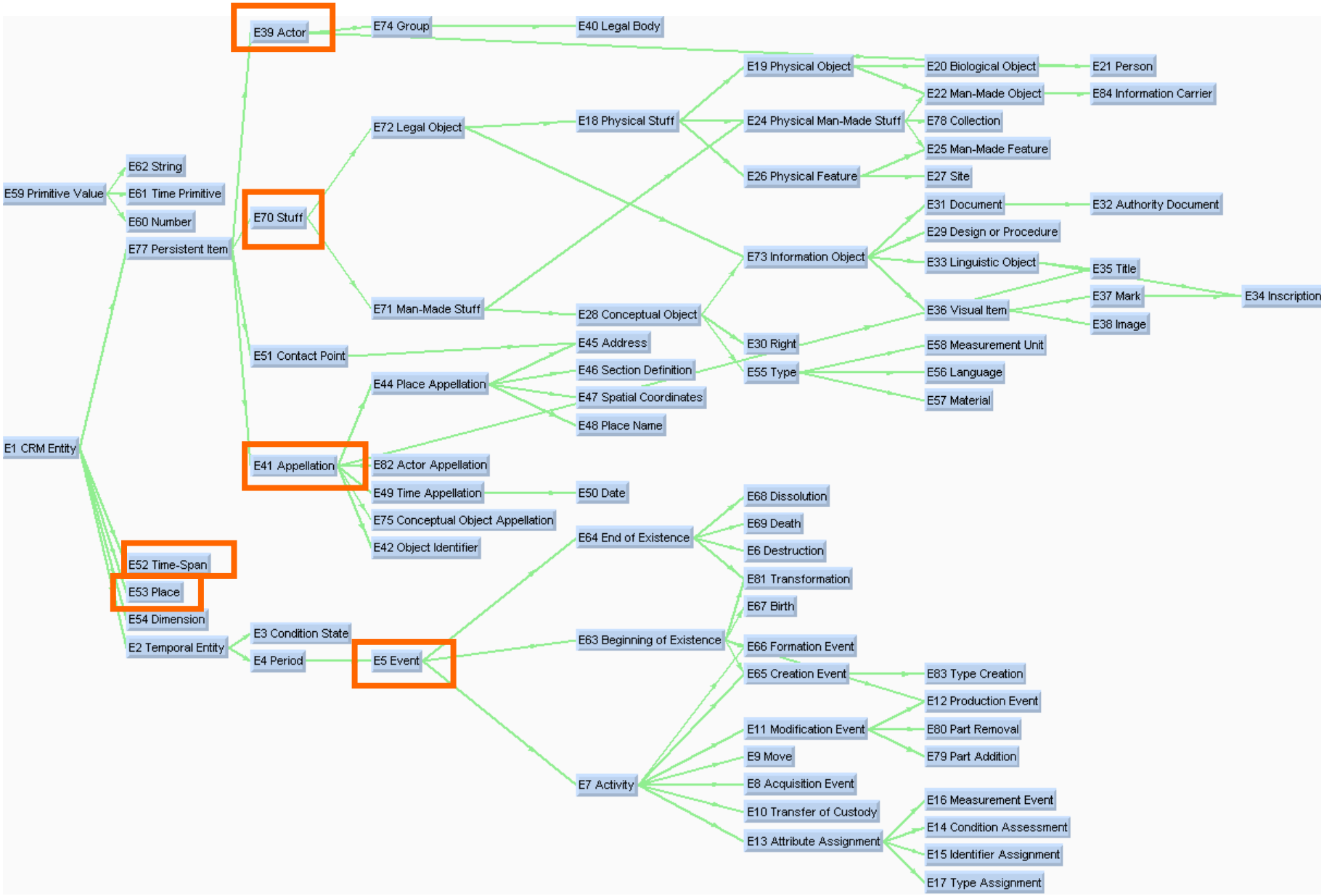
- What is the CIDOC CRM for?
  - A language for analysis of existing sources and models for data integration (mapping)
  - Intellectual guide to create schemata, formats, profiles
  - Best practice guide
  - Transportation format for data integration / migration /Internet

The CIDOC CRM

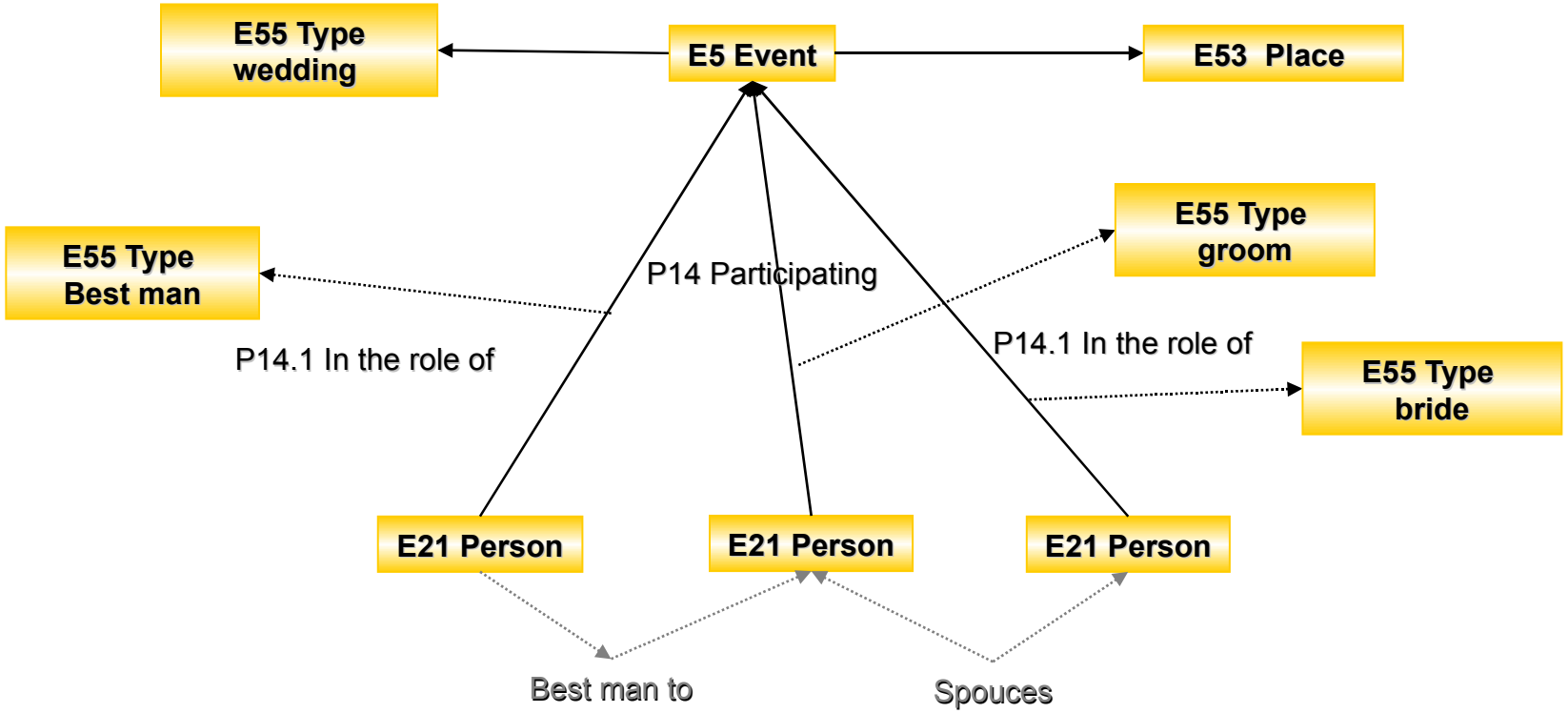
Top-level Classes relevant for Integration



CIDOC CRM: Class hierarchy



Relations between event, place and person



Data extrtaction
Motivation: Grey literature in Museums

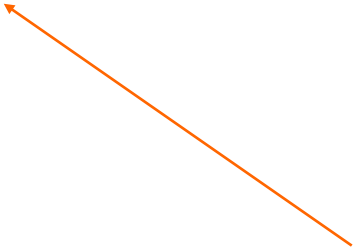
The excavation in Wasteland in 2005 was performed by Dr. Diggey. He had the misfortune of breaking the beautiful sword (C50435) into 30 pieces.

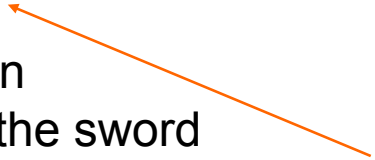
Information extraction

Actor: Dr. Diggey
Relation: performed
Event: E1
Type excavation
Place: Wastland
Time- span 2005

Actor: Dr. Diggey
Relation: performed
Event: E2
Type: Modification
Descr: Breaking the sword into 30 pieces
Relation: part of E1
Relation: in presence of
Object: Sword
Relation: identified by

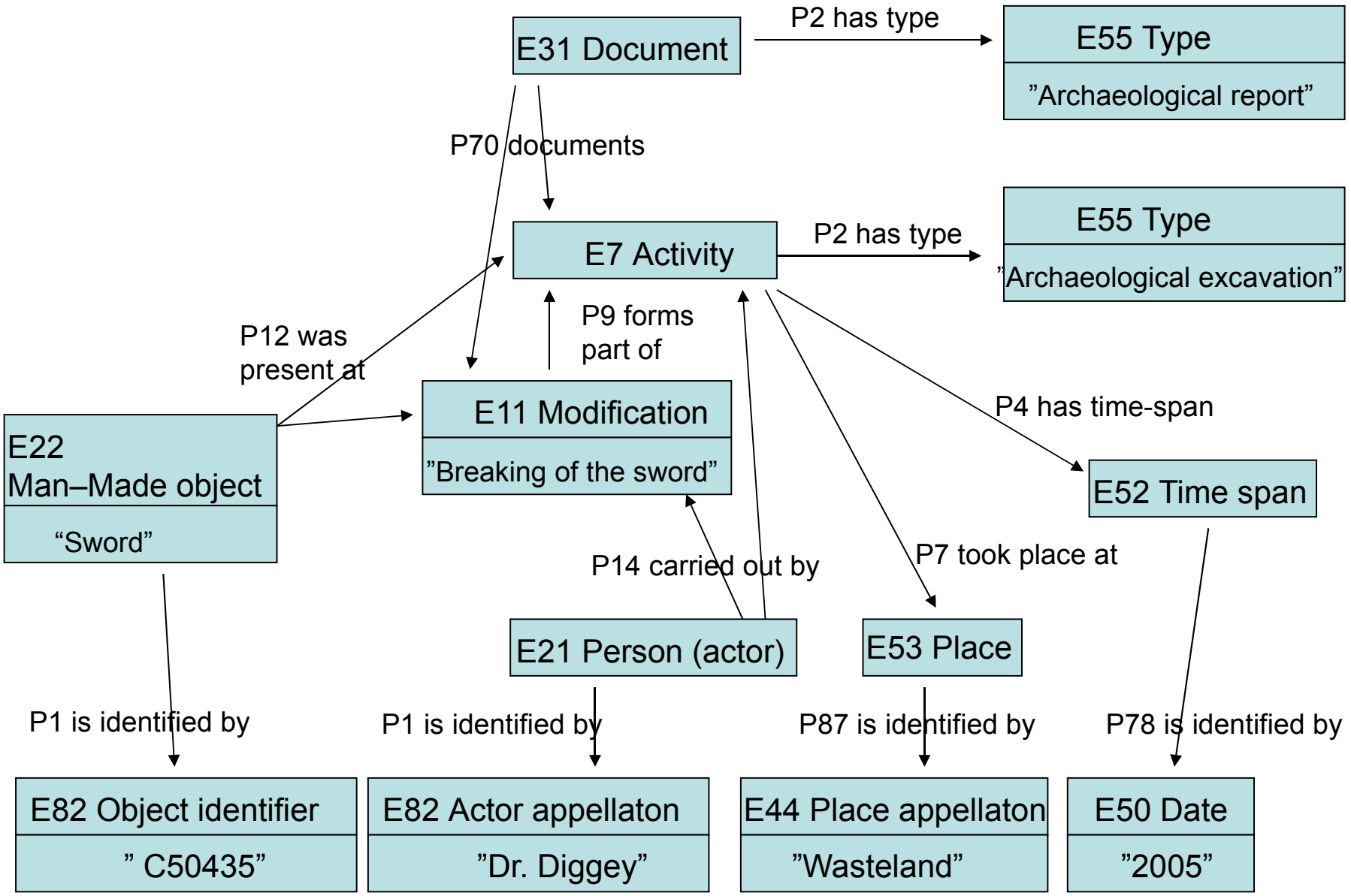
Identifier: C50435





```
<TEI>
<teiHeader>
…
</teiHeader>
<text>…
<p id="p1">
<rs id="e1">The excavation in
<name type="place" id="n1">Wasteland
</name> in <date id="d1">2005</date></rs>
was performed by
<name type="person" id="n2">Dr. Diggey
</name>.
He had the misfortune of  <rs id="e2">
breaking  <rs id="o1">the beautiful sword
<rs id="o_id1">(C50435)</rs></rs> into 30
pieces</rs>.
</p>
…
</text></TEI>
```

The content of the text expressed in the CIDOC-CRM



The CIDOC-CRM: Images – Visual Items

From the collection of art plates at the University Library, Oslo

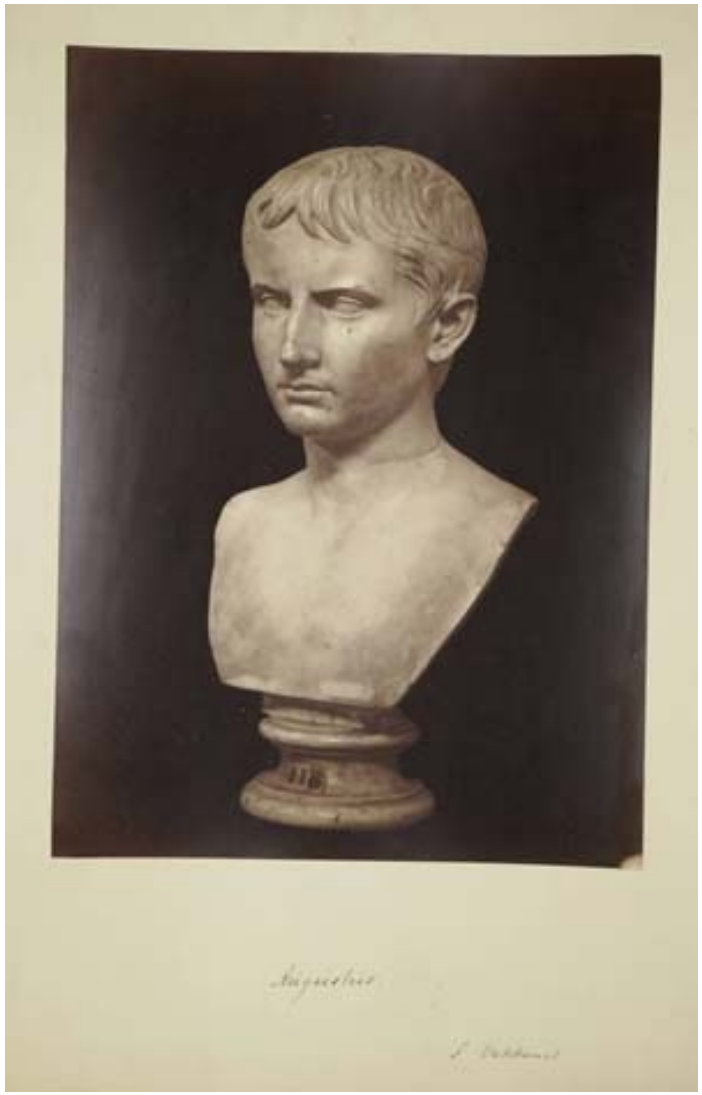
The young Augustus

[Bottom middle with ink:] Augustus // In the Vatican //

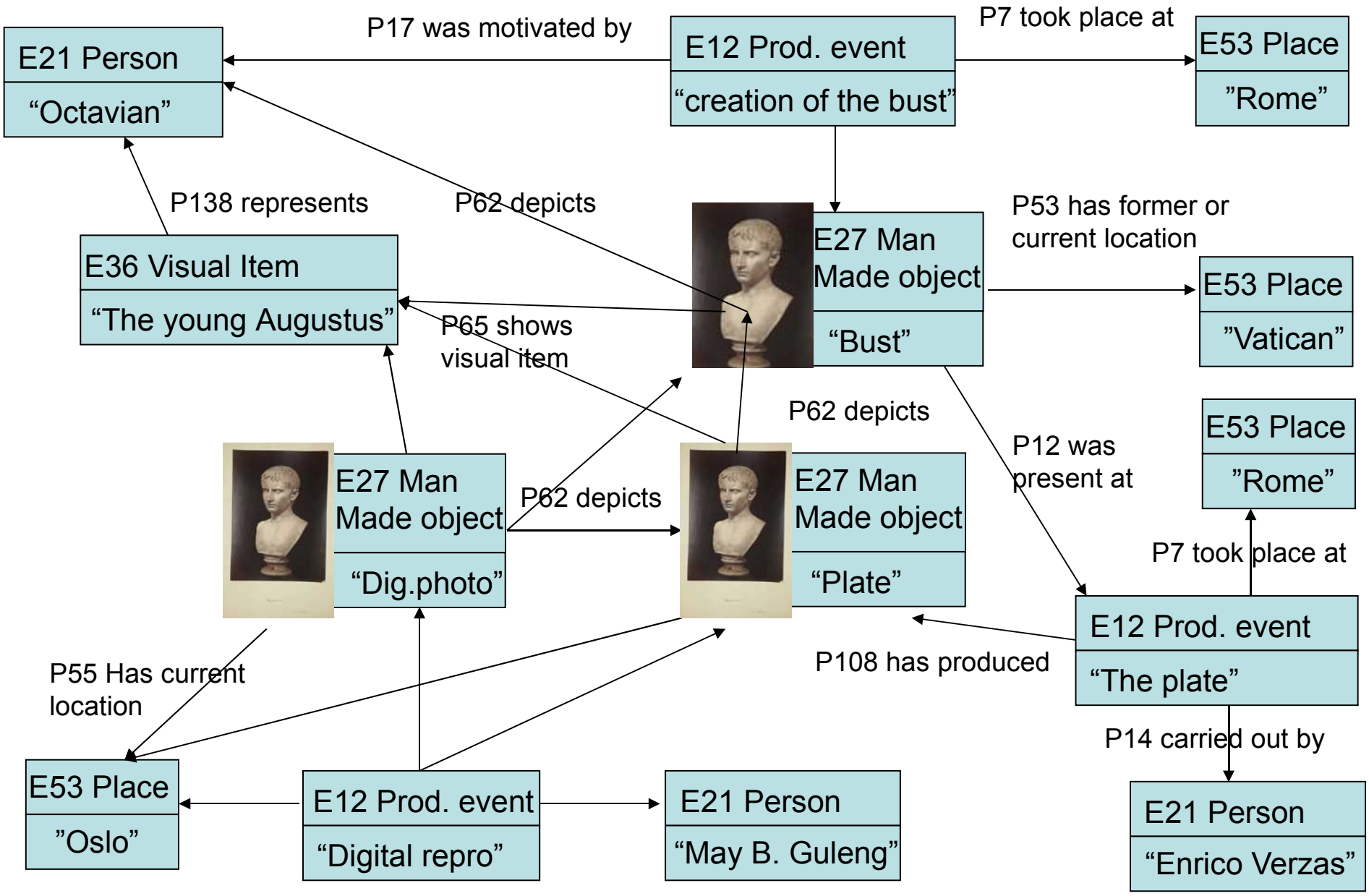
[Relief in the paper: bottom right:] ENRICO VERZASCHI / EDITORE FOTOGRAFO / ROMA / VIA DEL CORSO 133 A 136

Before 1877

Black background, ¾ format



The CIDOC-CRM: Images – Visual Items



The CIDOC CRM
Integration of Historical Archives

Type:	**Text**
Title:	**Protocol of Proceedings of Crimea Conference**
Title.Subtitle:	**II. Declaration of Liberated Europe**
Date:	**February 11, 1945.**
Creator:	**The Premier of the Union of Soviet Socialist Republics** **The Prime Minister of the United Kingdom** **The President of the United States of America**
Publisher:	**State Department (USA)**
Subject:	**Postwar division of Europe and Japan**

Metadata

Documents

"The following declaration has been approved:
The Premier of the Union of Soviet Socialist Republics,
the Prime Minister of the United Kingdom and the President of the United States of America have consulted with each other in the common interests of the people of their countries and those of liberated Europe. They jointly declare their mutual agreement to concert…
….and to ensure that Germany will never again be able to disturb the peace of the world…… "

(acc. M.Doerr & S.Stead)

The CIDOC CRM
Integration of Historical Archives

Type:	**Image**
Title:	**Allied Leaders at Yalta**
Date:	**1945**
Publisher:	**United Press International (UPI)**
Source:	**The Bettmann Archive**
Copyright:	**Corbis**
References:	**Churchill, Roosevelt, Stalin**

Photos, Persons

Metadata

About…

(acc. M.Doerr & S.Stead)

The CIDOC CRM
Integration of Historical Archives

TGN Id: 7012124
Names: Yalta (C,V), Jalta (C,V)
Types: inhabited place(C), city (C)
Position: Lat: 44 30 N,Long: 034 10 E
Hierarchy: Europe (continent) <– Ukrayina (nation) <– Krym (autonomous republic)
Note: …Site of conference between Allied powers in WW II in 1945; ….
Source: TGN, Thesaurus of Geographic Names

Places, Objects

About…

Title: Yalta, Crimean Peninsula
Publisher: Kurgan-Lisnet
Source: Liaison Agency

(acc. M.Doerr & S.Stead)

The CIDOC CRM
Integration of Historical Archives

- Problem 1, Identity:
  - Actors, Roles, proper names:
    - The Premier of the Union of Soviet Socialist Republics
      Allied leader, Allied power, Joseph Stalin, ...
  - Places
    - Jalta, Yalta,
    - Krym, Crimea
  - Events
    - Crimea Conference, “Allied Leaders at Yalta”,
      “… conference between Allied powers” “Postwar division”
  - Objects and Documents:
    - The photo, the agreement text

(acc. M.Doerr & S.Stead)

The CIDOC CRM
Integration of Historical Archives

- Solution to Problem 1, Identity:
  - Local Vocabulary control – local authorities (thesauri, gazetteers)
    - e.g. Conference 1: “Yalta Conference”, “Crimea Conference”…
  - Global Authority Registers
    - e.g. TGN id 7012124
    - Connect all local authorities to global ones
  - Authority Registers must be rich in
    - synonyms
    - distinct attributes for identification (e.g. geo-coordinates)
  - Persistent collection identifiers
    - history of all identifiers

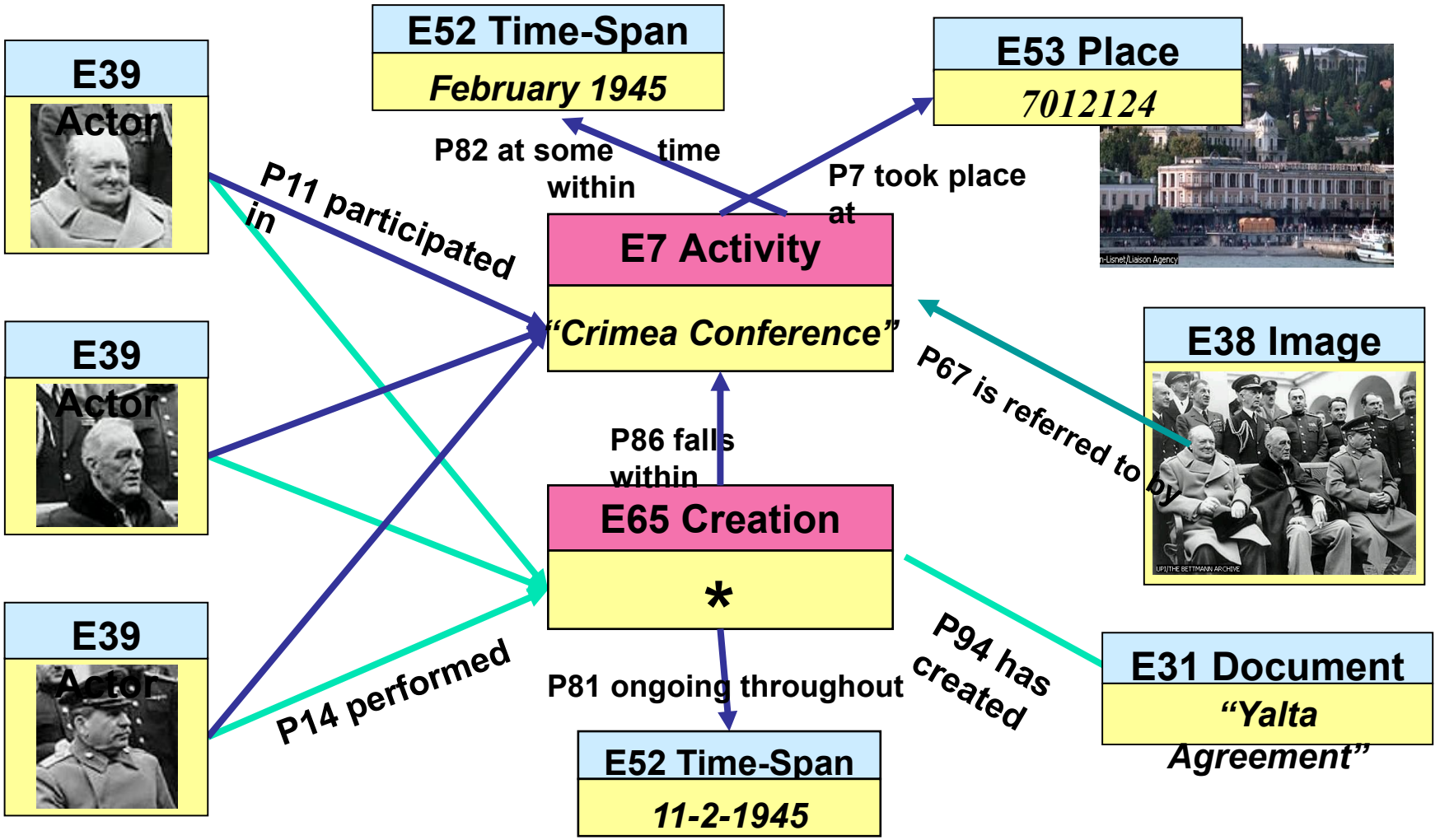
(acc. M.Doerr & S.Stead)

The CIDOC CRM
Integration of Historical Archives

- Problem 2, hidden entities (typically found in “title field”):
  - Actors
    - Allied leader, Allied power
  - Places
    - Yalta, Crimea
  - Events
    - Crimea Conference, “Allied Leaders at Yalta”,“… conference between Allied powers” “Postwar division”
- Solution:
  - Change metadata structures: but what are the relevant elements?

(acc. M.Doerr & S.Stead)

The CIDOC CRM
Explicit Events, Object Identity, Symmetry



(acc. M.Doerr)

The CIDOC CRM – FRBR Harmonization

- The CIDOC Conceptual Reference Model (CRM)
  - developed since 1996 by CIDOC / ISO TC46, ISO 21127 by 2006
  - a core ontology aiming to integrate cultural heritage information
- Innovations
  - centre descriptions not around the things, but around the events that connect people, material and immaterial things in space-time.
  - explicit description of the discourse on relations between identifiers and the identified.
  - typologies modeled both as classification means and as objects of the cultural-historical discourse
- Lacks: a model of intellectual work

The CIDOC CRM – FRBR Harmonization

- The Functional Requirements for Bibliographic Records (FRBR)
  - developed 1992-1997 by IFLA, now being complemented by the Functional Requirements for Authority Data (FRAD)
  - A core ER model to integrate library objects by content relation
  - Might result in a new library practice
- Innovations:
  - Definition of stages/ abstraction levels of intellectual products: Work, Expression, Manifestation, Item.
  - Clusters publications and items around the notion of derivation and common conceptual origin across stages / abstraction levels.
- Lacks: any explicit notion of the processes behind. Partially ambiguous definitions (overgeneralization).

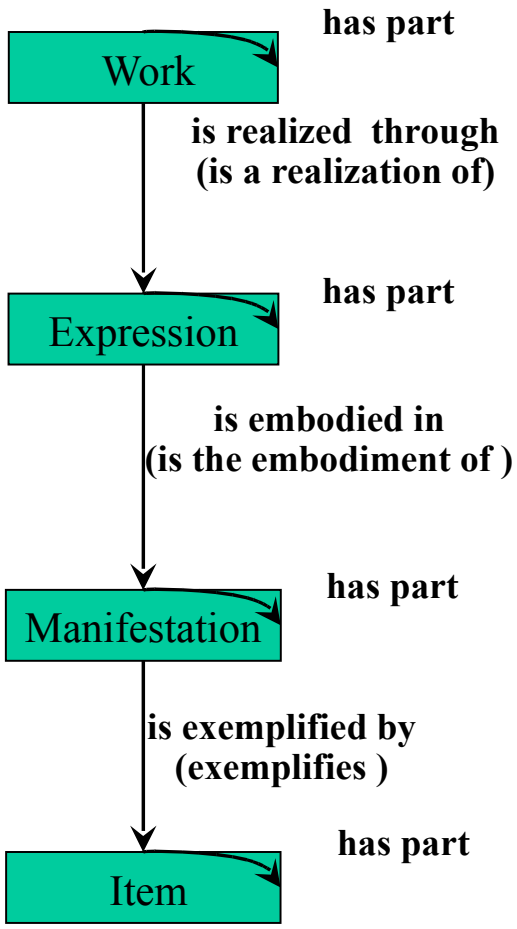
The FRBR - CRM Harmonization
FRBR : Abstraction Levels

"a distinct intellectual or artistic creation… there is no single material object one can point to as the work..."

"the intellectual or artistic realization of a work in the form of alpha-numeric, musical, or choreographic notation, sound, image, object, movement, etc"

"the physical embodiment of an expression of a work…all the physical objects that bear the same characteristics…

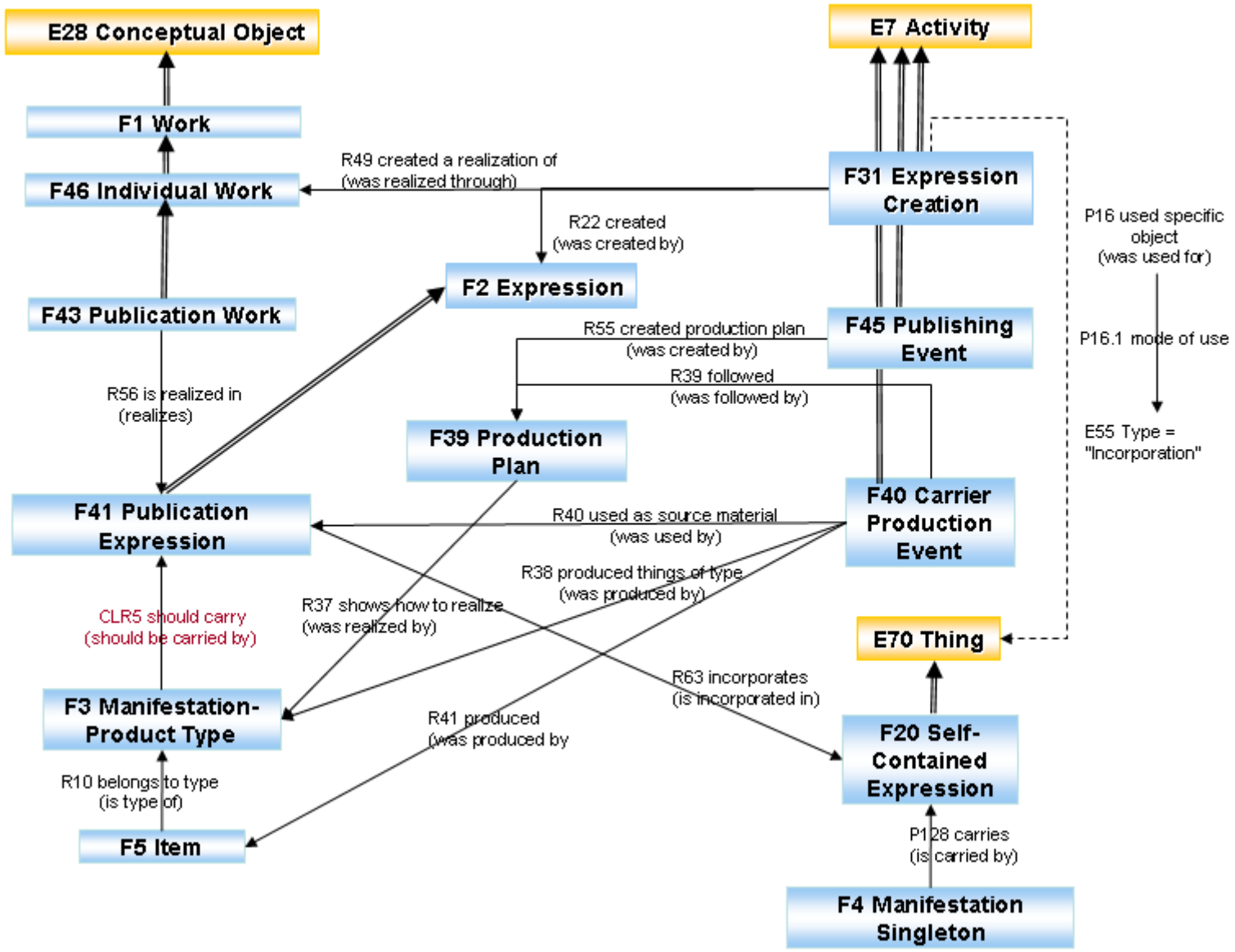
"a single exemplar of a manifestation..."



has a complement
has a successor
has a summary
has a supplement
has a transformation
has adaptation
has an imitation

has a complement
has a successor
has a summary
has a supplement
has a transformation
has adaptation
has an imitation

From Expression to Publication



Data integration – architecture