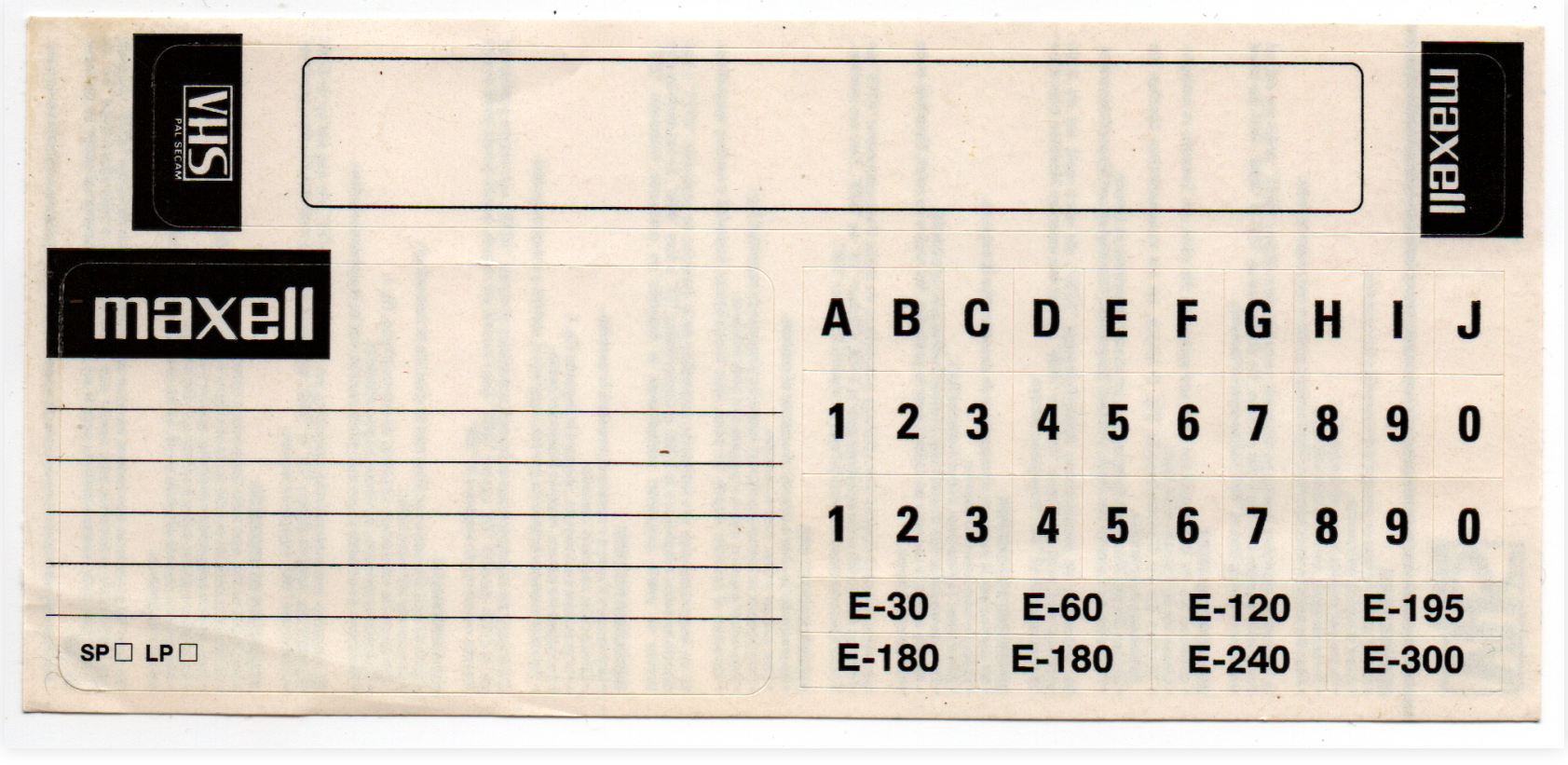
Topic 2 - Metadata

Examples? Ideas?

Retro Classic



Rockstars of

- Artist
- Title
- Album
- Date
- Cover art
- ...

MD Types for Preservation

- Descriptive
- Technical
- Environment
- Agent
- Rights
- Event
- Provenance
- Structure

Speaker notes

Name some examples for each one of these types.

btw: Metadata is sometimes abbreviated as "MD". Examples: techMD, rightsMD, sourceMD, digiprovMD etc

Examples:

- Descriptive: title, album
- Technical: resolution, bitrate
- Environment: tech required
- Agent: person, organization, tool
- Rights: copyright, ownership
- Event: an action that has a date
- Provenance: who owned it before
- Structure: file is part of a sequence

Important things

- Controlled vocabulary!
- Date/time: ISO 8601
  (2019-05-30T10:34:47+00:00)
- Standards: Interoperability

Speaker notes

Controlled vocabulary: Be very greedy with "free text" fields, as they lead to chaos, disorder and mayhem! On the other hand: It's always good to have "some" field where to put the "doesn't go anywhere else" stuff. But use with caution... ;)

Date/time: Different datetime formats: No fun, Dr. Jones! Therefore, ISO8601 is a great improvement. Please use it.

Standards & Interoperability: Having (meta)data stored in the same standard format, but written by > 1 system, doesn't mean they're automatically "plug-compatible".

For example: "color information" (tech MD).

- "YUV, 4:2:2, 10bpc"
- yuv422p10le "Same-same, but different".

CoVoc Examples

- 35mm = 35 mm = 35 millimètre
- dup pos = duplicate positive
- de = deu = german = German = alemán
- YUV,4:2:2,10bpc = yuv422p10le
- Director = Directed by

Speaker notes

Controlled vocabulary: Be very greedy with "free text" fields, as they lead to chaos, disorder and mayhem!

On the other hand: It's always good to have "some" field where to put the "doesn't go anywhere else" stuff. And it may also be useful/necessary to be able to store/preserve the original terms "as-is", since mapping different sources to a vocabulary may not be as straight forward or exact as one would hope.

Oh and other mapping fun: * Typos do happen! "35mn" anyone? * 1967-06-07-12 (Date with a "sort index" hack added in Freetext field)

Normalize date/time

- 4. May 2019
- May 4th, 2019
- 4.5.2019
- 5/4/2019
- ...
- ISO 8601
  (2019-05-30T10:34:47+00:00)

Speaker notes

Date/time: Different datetime formats: No fun, Dr. Jones! Therefore, ISO8601 is a great standard that greatly improves the situation. Please use it :)

...and we haven't even started with different year offsets, like the Islamic Calendar yet.

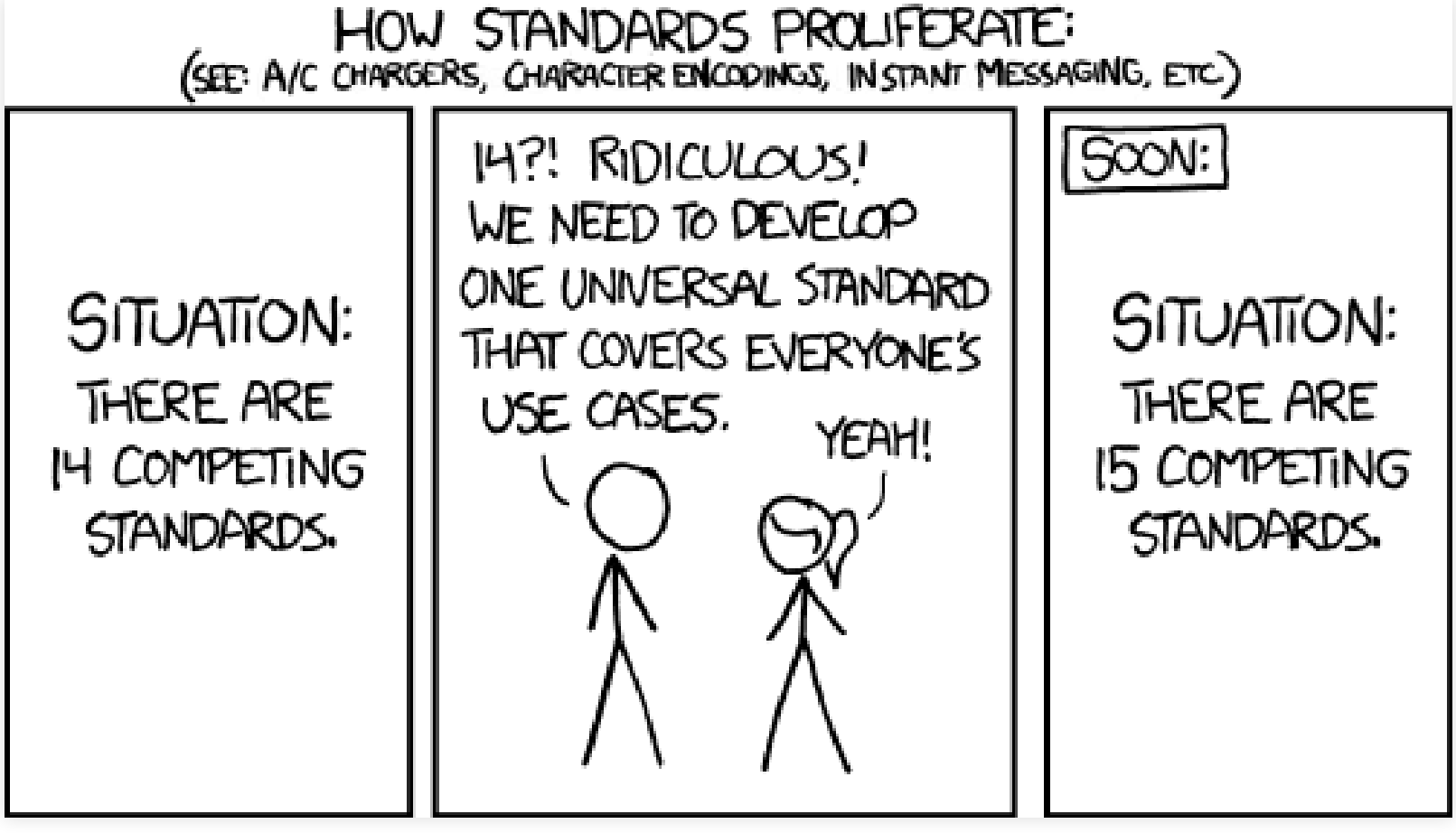
Controlled Vocabulary

Your suggestions for:

- Descriptive?
- Technical?
- Other?

Standards

One To Rule Them All?



Exercise

Imagine some regular (preservation) workflows and identify:

- **Events**
  Which action is done?
- **Agents**
  Who is (involved in) doing it?
- **Objects**
  What does it apply to?
- **Rights**
  Who holds which rights to what?
- **Relationships**
  How do these entities relate to each other?

Exercise

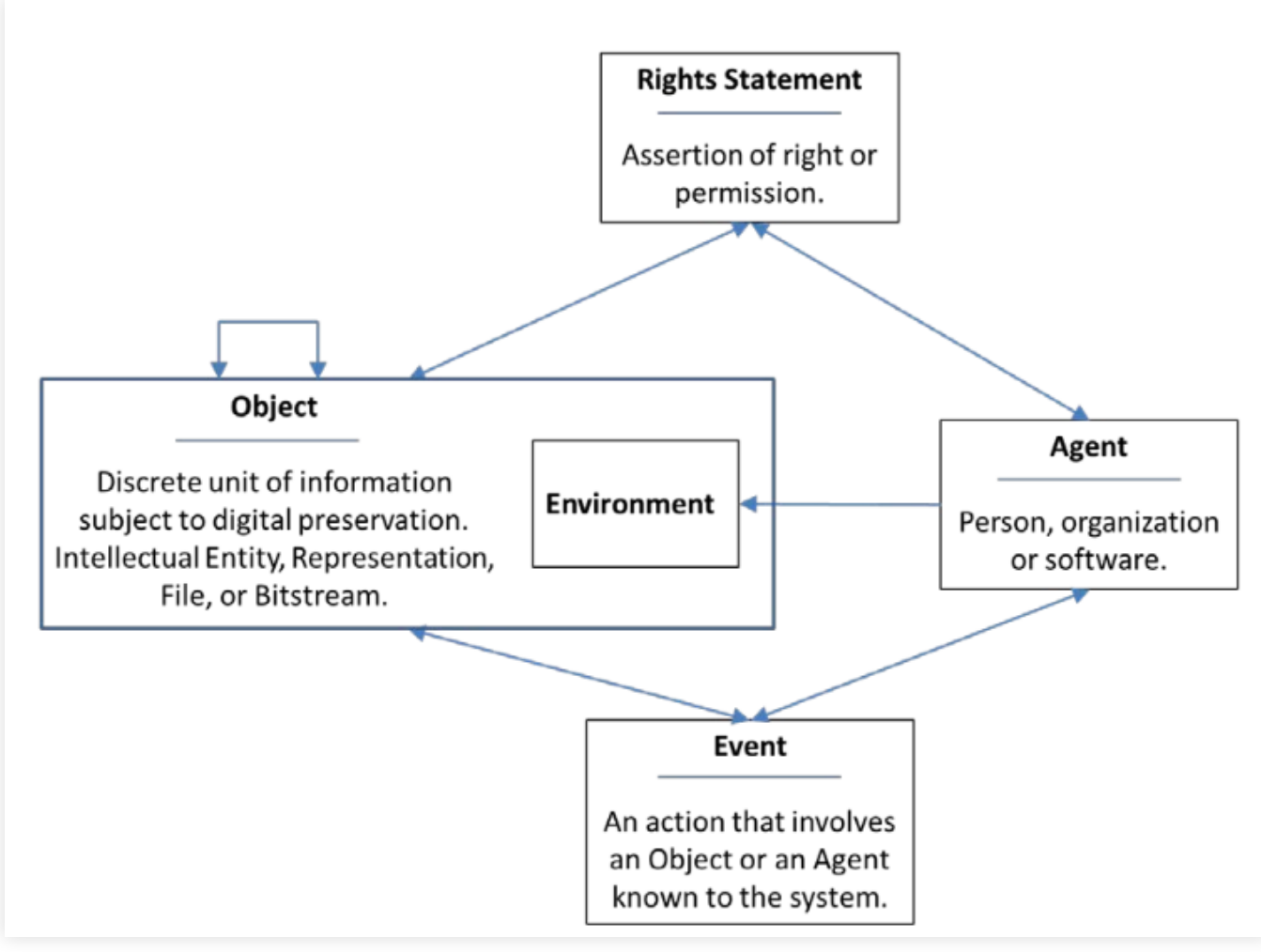
- **Event** eg: "transcode"
- **Agent** eg: "tool:ffmpeg"
- **Object** eg: "preservation file"
- **Rights** eg: "item owner"
- **Relationship** eg: "converts", "is copy of"

Draw (and name) relationships between entities (one way/arrow). Write down what metadata you think should be captured in that entity.

Speaker notes

- **Object** (or Digital Object): a discrete unit of information subject to digital preservation.
- **Environment:** Technology (software or hardware) supporting a Digital Object in some way (e.g. rendering or execution). Environments can be described as Intellectual Entities and captured and preserved in the preservation repository as Representations, Files and/or Bitstreams.
- **Event:** an action that involves or affects at least one Object or Agent associated with or known by the preservation repository.
- **Agent:** person, organization, or software program/system associated with Events in the life of an Object, or with Rights attached to an Object. It can also be related to an environment Object that acts as an Agent.
- **Rights Statement:** assertion of one or more Rights or permissions pertaining to an Object and/or Agent.

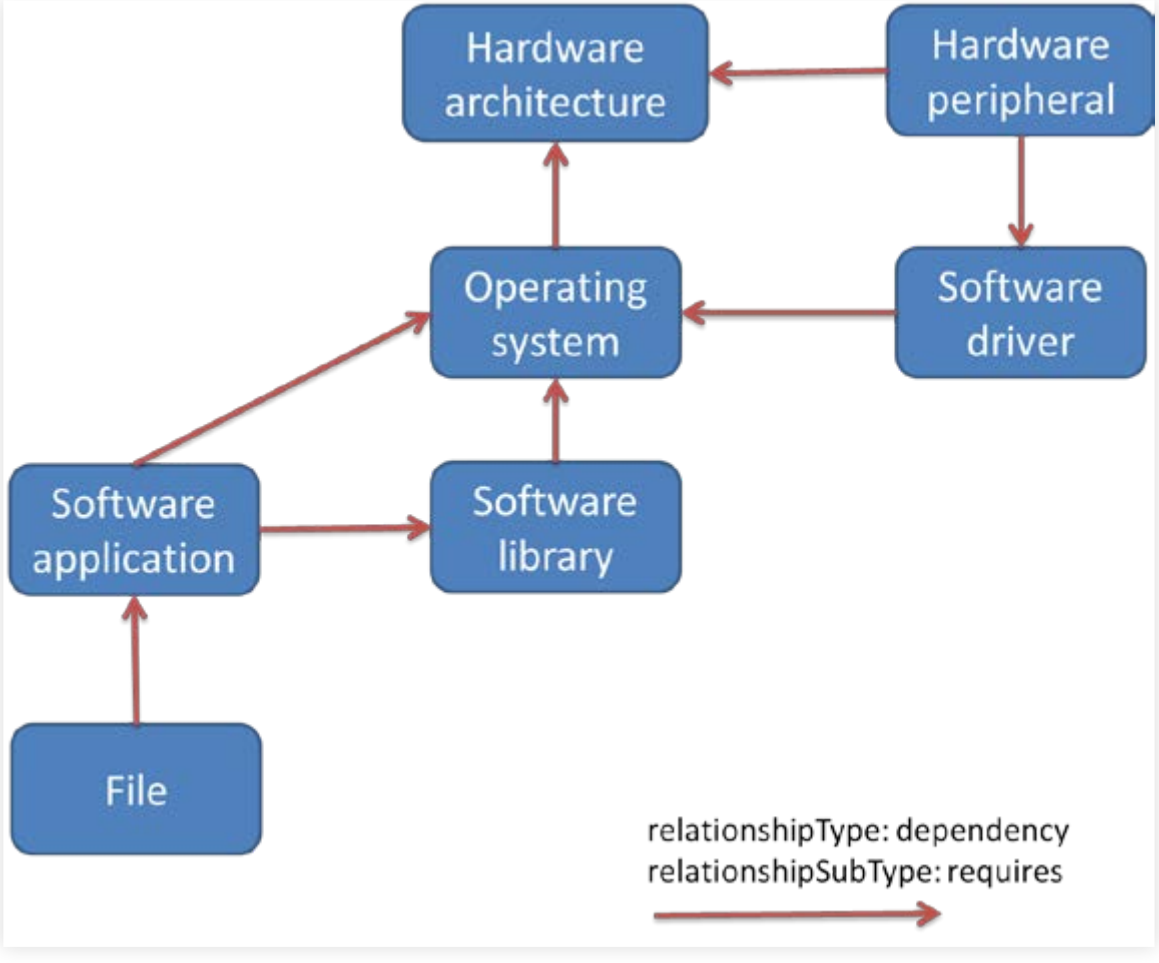
PREMIS


Rights Statement
Assertion of right or permission.
Object
Discrete unit of information subject to digital preservation. Intellectual Entity, Representation, File, or Bitstream.
Environment
Agent
Person, organization or software.
Event
An action that involves an Object or an Agent known to the system.

Speaker notes

Looks familiar? Congratulations! You've just derived the PREMIS data model on your own.

PREMIS - Environment Stack



Speaker notes

This is an example of an environment stack necessary to access a file stored on a physical carrier (hardware). This matches the example layers/components I've mentioned in the introduction.

Good to Know

- Dublin Core: Core fields (& their names)
- METS: (MD container) descriptive, administrative, structural
- PREMIS: A metadata framework
- EBUCore: descriptive & technical (broadcast use case focus)
- CEN EN 15907: Comprehensive description of cinematographic works
- FRBR ("furbur"): Comprehensive description of bibliographic works
- Mediainfo XML: Technical metadata (AV)

Speaker notes

So you see: Metadata itself needs to be stored in "a format" too. Same rules and properties as for the other data formats apply (open, accessible, etc)

PREMIS: PREservation Metadata: Implementation Strategies METS: Metadata Encoding and Transmission Standard EBUCore: European Broadcast Union (EBU) - based on Dublin Core EN 15907: Metadata standard for Cinematographic Works FRBR: Functional Requirements for Bibliographic Records

Dublin Core: Elements

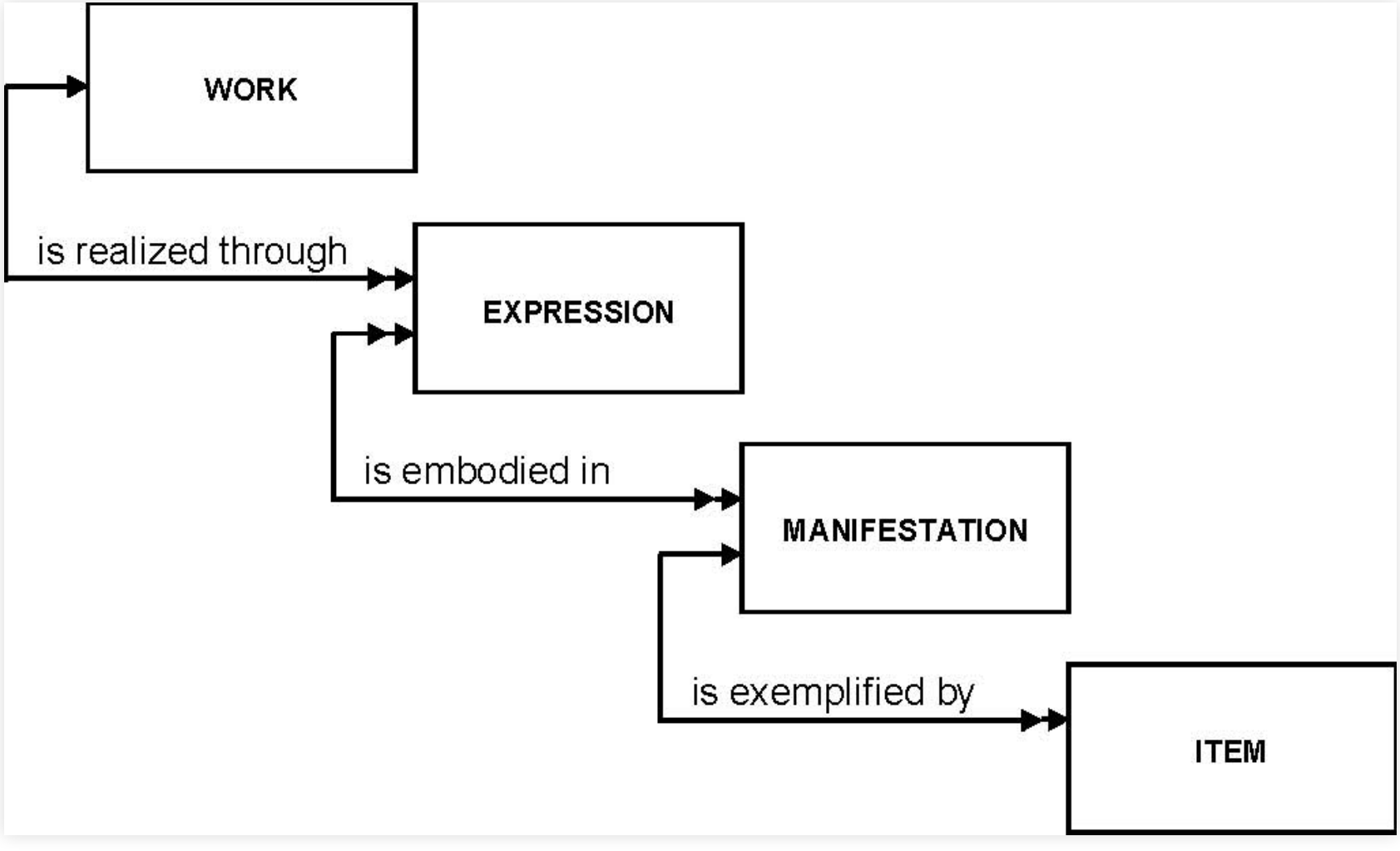
1. Contributor
2. Coverage
3. **Creator**
4. **Date**
5. Description
6. Format
7. Identifier
8. Language
9. Publisher
10. Relation
11. Rights
12. Source
13. Subject
14. **Title**
15. Type

See: DC Reference Description

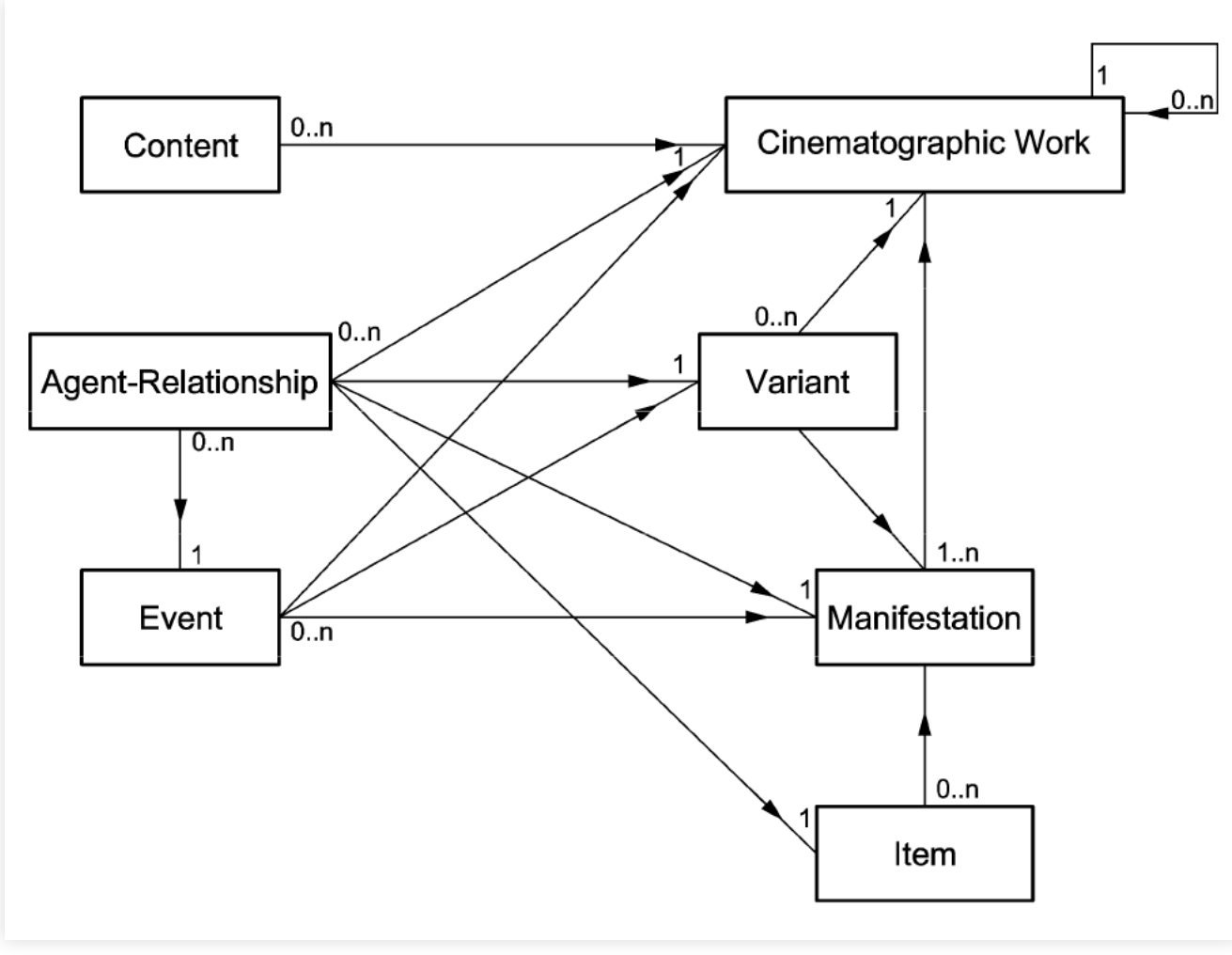
Speaker notes

1. Contributor - “An entity responsible for making contributions to the resource.”
2. Coverage - “The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.”
3. Creator - “An entity primarily responsible for making the resource.”
4. Date - “A point or period of time associated with an event in the lifecycle of the resource.”
5. Description - “An account of the resource.”
6. Format - “The file format, physical medium, or dimensions of the resource.”
7. Identifier - “An unambiguous reference to the resource within a given context.”
8. Language - “A language of the resource.”
9. Publisher - “An entity responsible for making the resource available.”
10. Relation - “A related resource.”
11. Rights - “Information about rights held in and over the resource.”
12. Source - “A related resource from which the described resource is derived.”
13. Subject - “The topic of the resource.”
14. Title - “A name given to the resource.”
15. Type - “The nature or genre of the resource.”

FRBR


WORK
is realized through
EXPRESSION
is embodied in
MANIFESTATION
is exemplified by
ITEM

EN 15907


Content
0..n
1
Cinematographic Work
1
0..n
1
0..n
Agent-Relationship
0..n
1
Variant
0..n
1
1
1..n
Event
0..n
1
Manifestation
0..n
1
Item

Interoperability

- **Use a standard.**
  (at least as basis)
- **Use English field terms.**
  (at least in the background)

Metadata creation and capture

- When is which metadata created?
  - Ingest
  - Modifications (e.g. transcoding, rewrapping)
  - Restoration
  - Meta-Metadata:
    Who edited/updated which MD-field?
  - ...

Metadata creation and capture

- How can metadata be captured?
  - Automatic
  - Manual

Speaker notes

Some metadata can be added or extracted later on, but some metadata is only available in a certain time frame. e.g.: Which capture workstation with which settings of the equipment - which ADConverter settings, etc.

Metadata: Usage & Storage

- Whatfor is metadata kept?
- How is metadata stored?

Plain Text

```
Sherlock
========================
Released in 2010 in United Kingdom
Actors: Benedict Cumberbatch, Martin Freeman
```

Speaker notes

Plain text files are great! If you know the encoding, it's almost certain they can be opened and read by any computer/operating system/machine - even in the future.

But unstructured text makes it hard for a machine to index/find/structure things.

CSV

Comma Separated Value

```
"Title", "Release Date", "Country", "Actor 1", "Actor 2"
"Sherlock", 2010-07-25, "UK", "Benedict Cumberbatch", "Martin Fre
```

The CSV file

Speaker notes

Technically, CSV is also plain text. But there's some structure defined on a meta level.

XML

```
<?xml version="1.0" encoding="UTF-8"?>
<work version="0.1">
    <title> Sherlock </title>
    <date type="release"> 2010-07-25 </date>
    <country code="iso-3166" type="production"> gb </country>
    <actor>
        <first_name> Benedict </first_name>
        <last_name> Cumberbatch </last_name>
    </actor>
    <actor>
        <first_name> Martin </first_name>
        <last_name> Freeman </last_name>
    </actor>
</work>
```

The XML file

DOC / PDF?

- Pros?
- Cons?

Speaker notes

btw: Office document formats (.doc, .docx, .pdf, .odt, etc) are way harder to read for automation. Also they need complex applications to read/open them.

MS-Office formats (.doc, .xls, etc) are also not publicy documented. If you cannot make sense of a file in a plain text editor, it's very likely that it cannot easily be understood/used without a special application. Future proof is always: The simpler, the better.

Fun fact: DOCX and ODT are additionally Zipped. If you rename them to ".zip" you can extract their contents to a bunch of XMLs and other data that is inside :)

If you only need to store and use text-based information, stay as close to plain-text formats as you can. If you need any kind of formatting, consider HTML rather than DOC or PDF.

Embedded vs Sidecar

- Embedded =
  inside the media file

- Sidecar =
  additional files **next** to media file

Speaker notes

Both have their advantages and disadvantages.

Embedded is more comfortable for files that are often moved around (less likely to get lost), but harder to change/augment/update/correct, because it requires applications to support read/write of these metadatas. This is not always the case.

For media objects, this often requires the whole file to be rewritten, regardless how trivial the metadata change was.

Furthermore: Any change to the media file changes their fixity information (hashcode, timestamp, etc), which must be dealt with accordingly. Consider runtimes of: read/write and hashcode generation!

Sidecar must be taken care of separately, but can more easily be changed/augmented/updated/corrected (because it's often text-based: CSV, XML, TXT, etc). May follow separate format specifications than the media object. This allows separate format migrations of the sidecar metadata.

Also: The sidecar data is usually significantly smaller than the media object (especially for AV): So if fixity information needs to be updated the large files may stay in place/untouched.

Again: Keep it as simple as possible and as complicated as necessary.

Good practice

- Use controlled vocabulary
- ...from public/common sources
- Prefer structured text (CSV or XML)
- ...over DOC(X) and PDF
- Embed at least basic info to ID the file
- ...and select which data may go better as sidecar
- Machine *and* human readable
- Field names ("backstage"): English, please.

Questions?

Comments?

Example XMLs

- Mediainfo
- METS
- PREMIS
- EBUCore

Links

- ISO 8601(Wikipedia)
- Date and time format - ISO 8601(ISO 2019)