Softwaretechnologie für Fortgeschrittene
Wohce 1

Systementwicklung

(with contributions from Christian-Emil Ore, Jon Holmen, and other colleagues at the Unit for Digital Documentation, University of Oslo
and from Martin Dörr and Stephen Stead, CIDOC-CRM SIG)

The media server

What are the requirements for a system for media objects?

- Upload
- Storage
- Metadata
- Presentation
- Long term preservation

Will focus on images but equally relevant for other media types

Upload operations

- Connect to the storage
- Find a logical place to put the data
- Submit metadata
- Establish a stream connection
- Upload bits
- Check result
- Get a receipt and an identifier back
- *The client may be:*
  - *a human*
  - *a computer programme*

Storage operations

- Receive a request with metadata
- Return an ID for the stream
- Receive the stream
- Receive further metadata
- Enter metadata into database
- Store file based on stream on disk
- Establish link from database to disk file

Metadata

- Technical metadata such as
  - file format
  - shoot date and time
  - size
  - location and direction
- Content metadata such as
  - motive
  - classification
  - date and time
  - source
  - location

Presentation

- User interfaces for
  - searching
  - listing metadata and/or thumbnails
  - delivering images in different formats
  - protecting images
  - ordering images
  - payment
- Different platforms
- Different user groups
- Different contexts

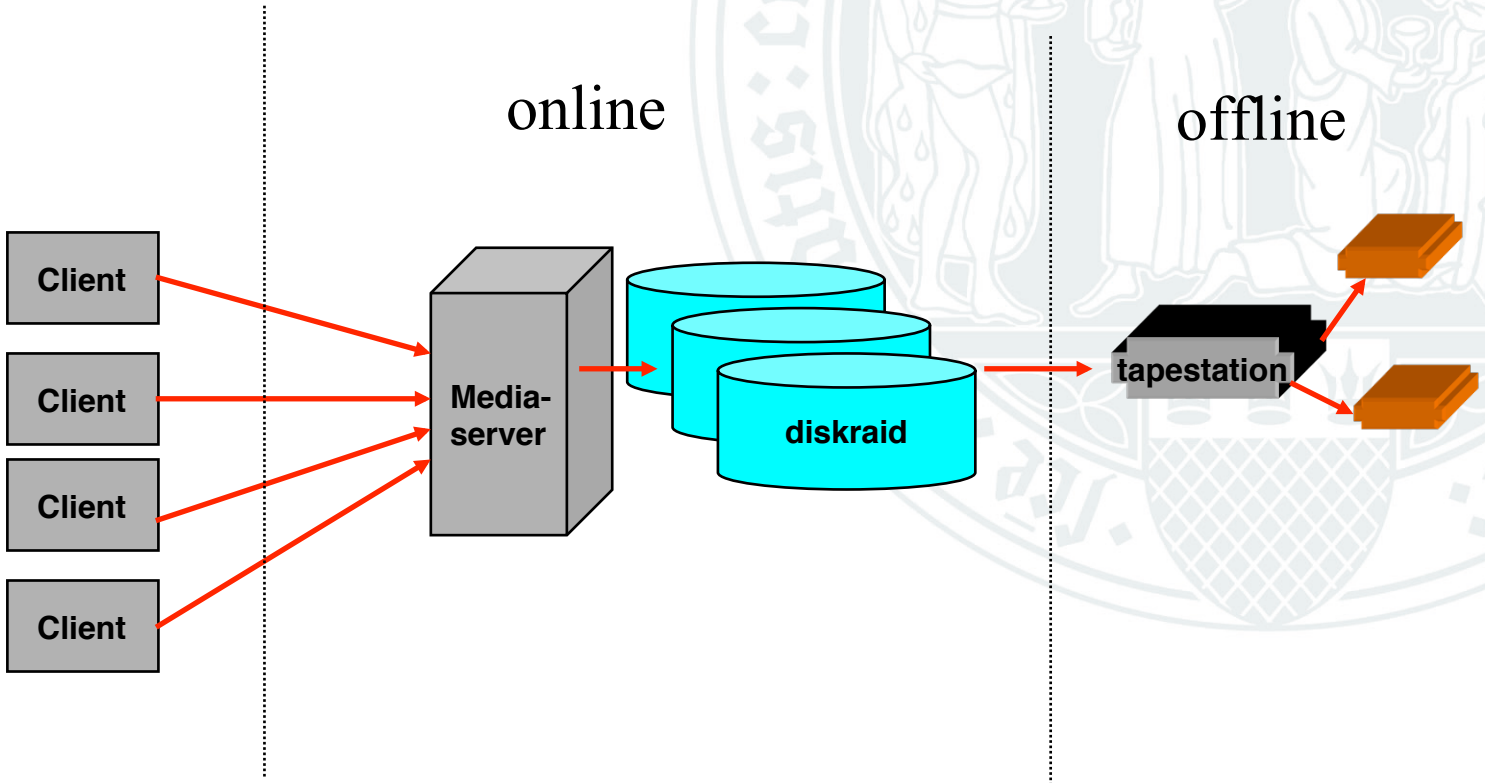
Long term preservation

- Make sure the data survives for the future
- *Long* term – not just 10 or 30 years
- Preservation
  - bitstreams
  - meaning
  - context
  - usability
- Technology
- Administration
- Politics

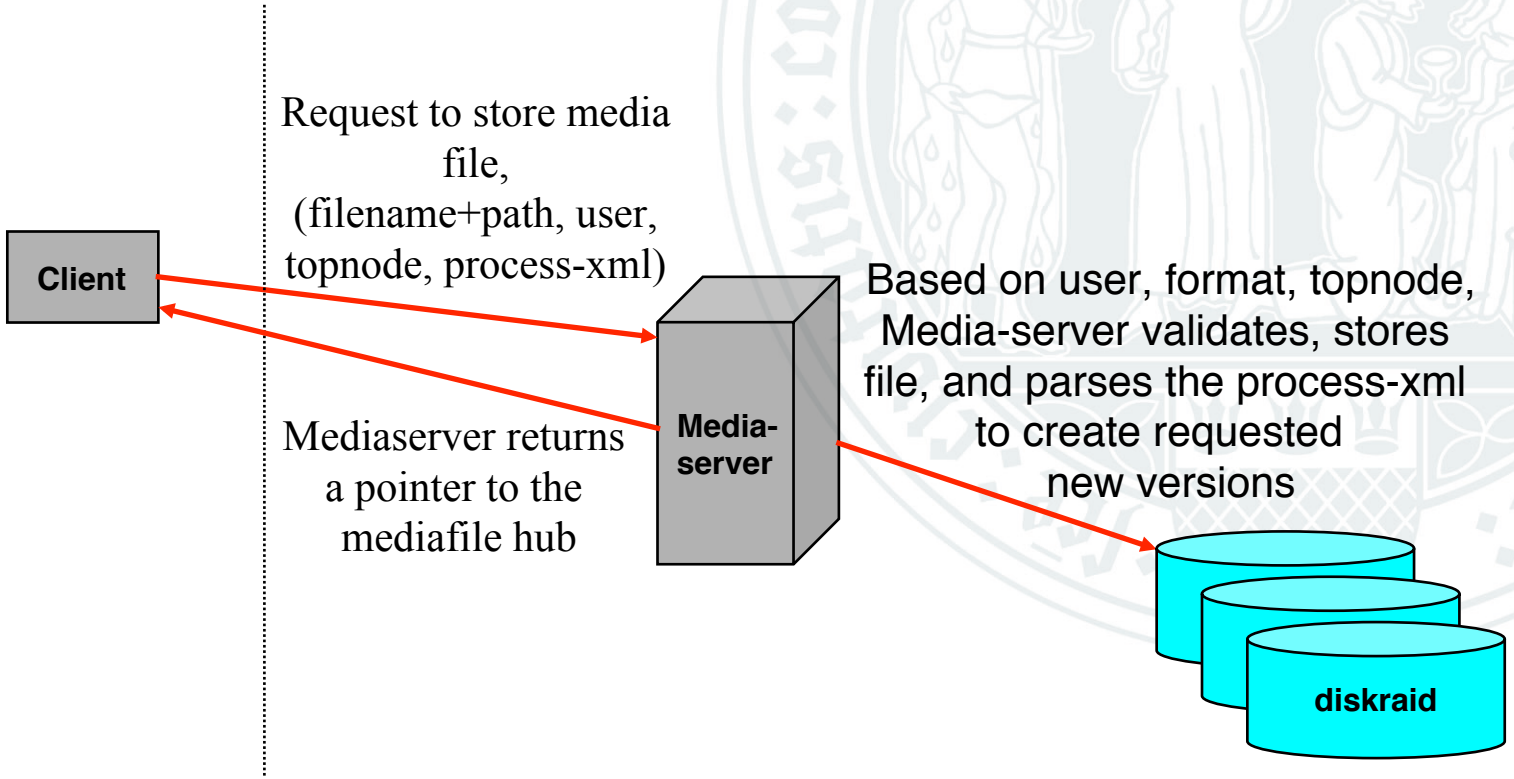
The system presented here

- Image collections at Norwegian universities
  - University history
  - Art history
  - Cultural history
  - Archaeology
  - Natural history
  - …
- Document archives
  - archaeology
  - dialectology
  - …
- Sum:
  - 1-2 000 000 traditional digitised photos
  - 3 000 000 document facsimiles
  - *(figures a few years old)*

Overall architecture



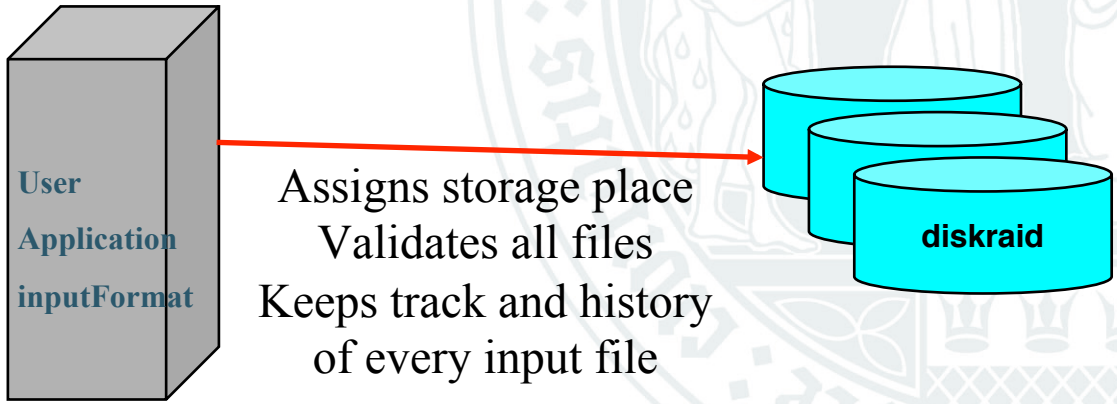
Data flow


Request to store media file, (filename+path, user, topnode, process-xml)
Client
Media-server
Mediaserver returns a pointer to the mediafile hub
Based on user, format, topnode, Media-server validates, stores file, and parses the process-xml to create requested new versions
diskraid

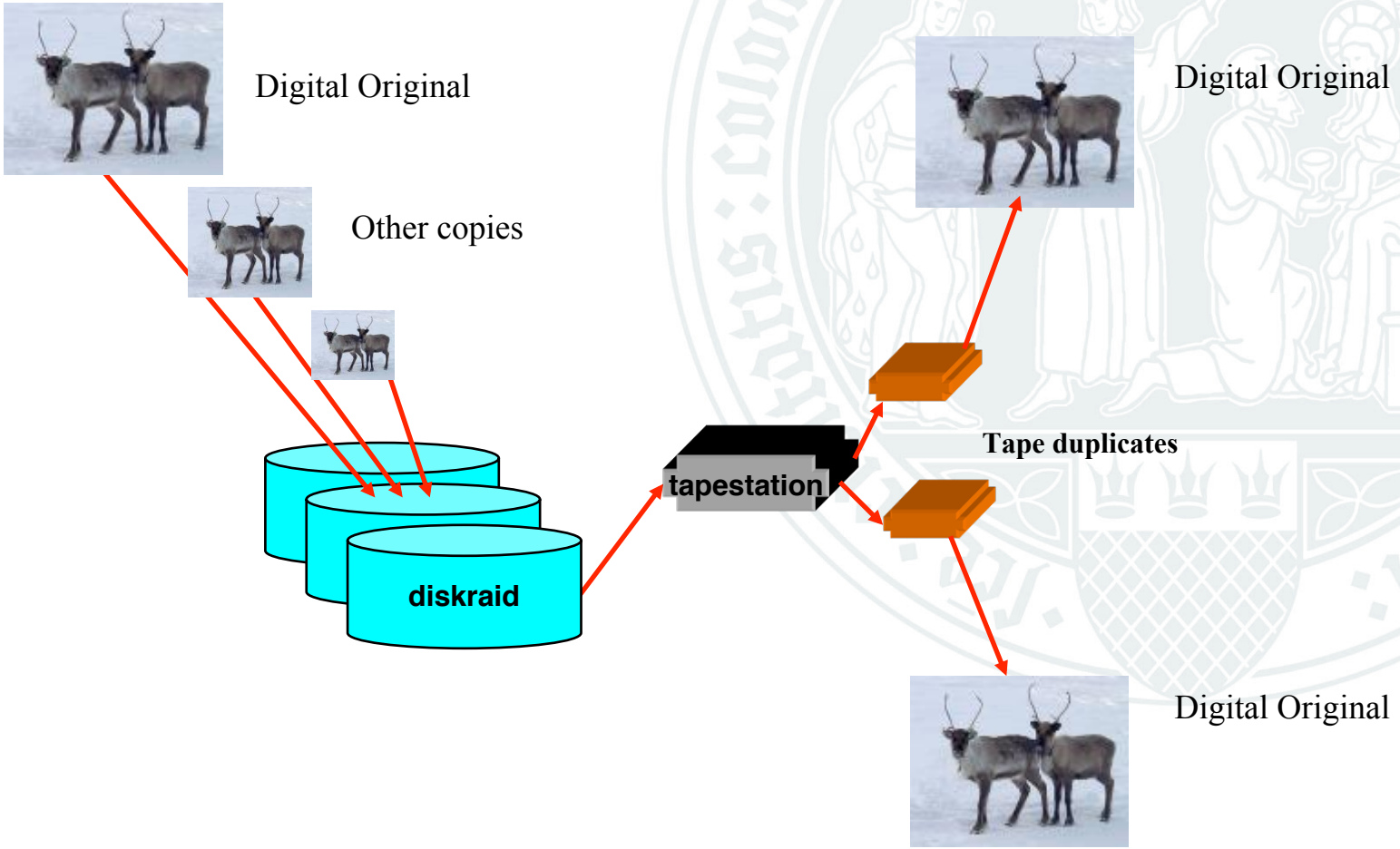
User applications

- A GUI user application is a frontend for:
  - cataloguing pictures (metadata)
  - importing pictures
  - changes and updates
- A command line application is a frontend for:
  - running import scripts
  - file list as parameter
  - meant for expert users
  - meant for large volumes
  - metadata as XML files
  - can link to pre-existing metadata
- Always connected to one discipline schema

Storage keeper



Long term preservation



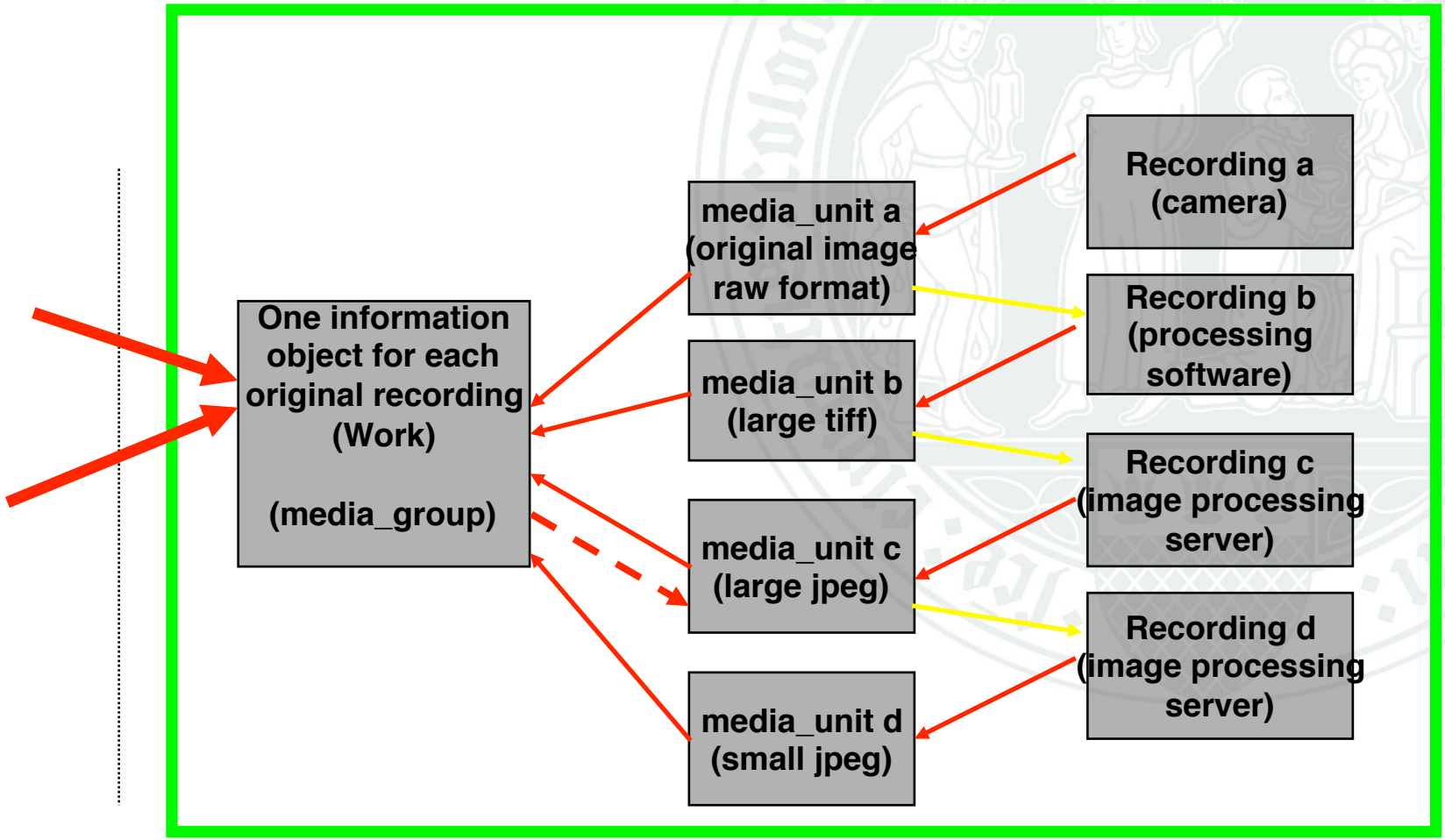
Database

Discipline schema 1
• image_card – media_group_id – concent_metadata • *subject specific tables*

Common schema
• media_group – id – process_xml – tech_metadata • media_unit – tech_metadata – default {0, 1} • schema_setup – process_spec – delete_limit • process – process_spec – status {0, 1, 2, 9}

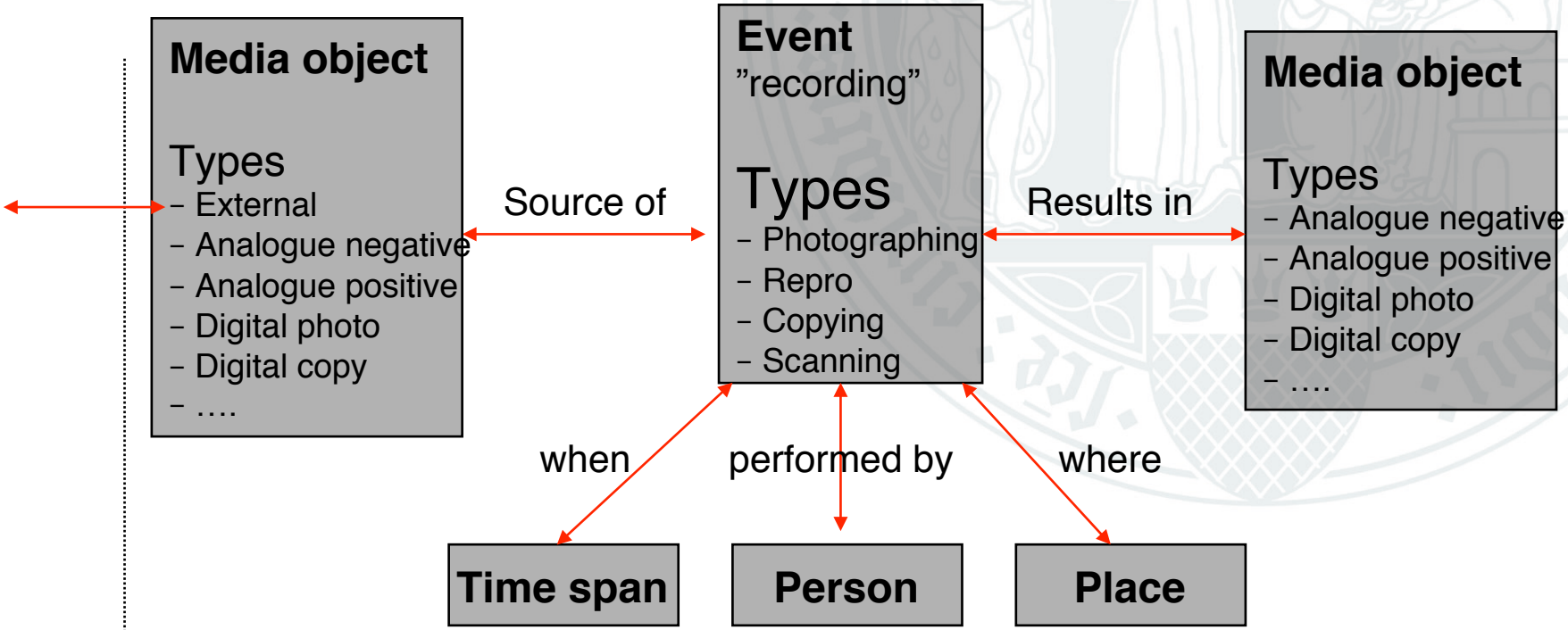
Discipline schema 2
• image_card – media_group_id – concent_metadata • *subject specific tables*

Example work flow (digital image)



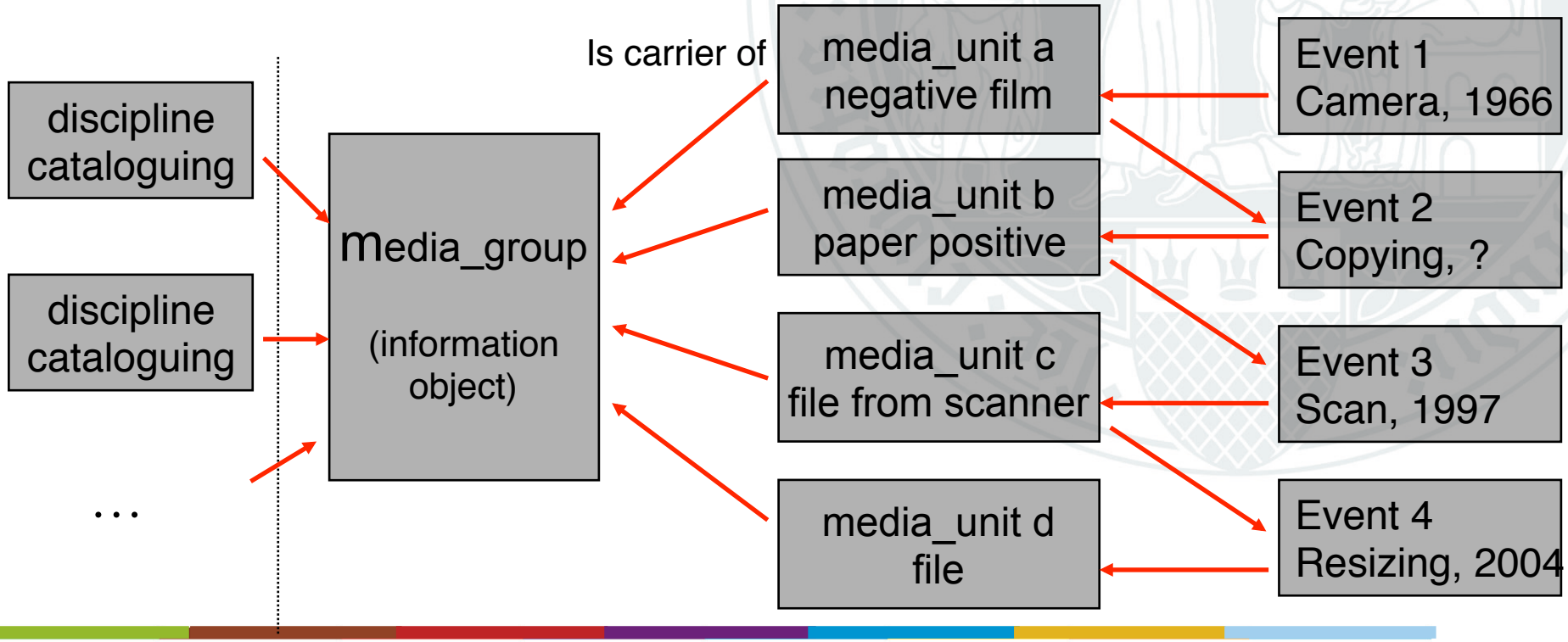
Event centric data model

source → recording → result



Data model example: digitised image

Separation between “information object” and “information carriers”



File processing

- Parse processing XML
- Set up production line
  - any conversion path with possible conversions can be added
- Matrix of in and out formats and default scripts
  - can be overridden.
- Scripts run in background
  - queue handling
  - load balancing

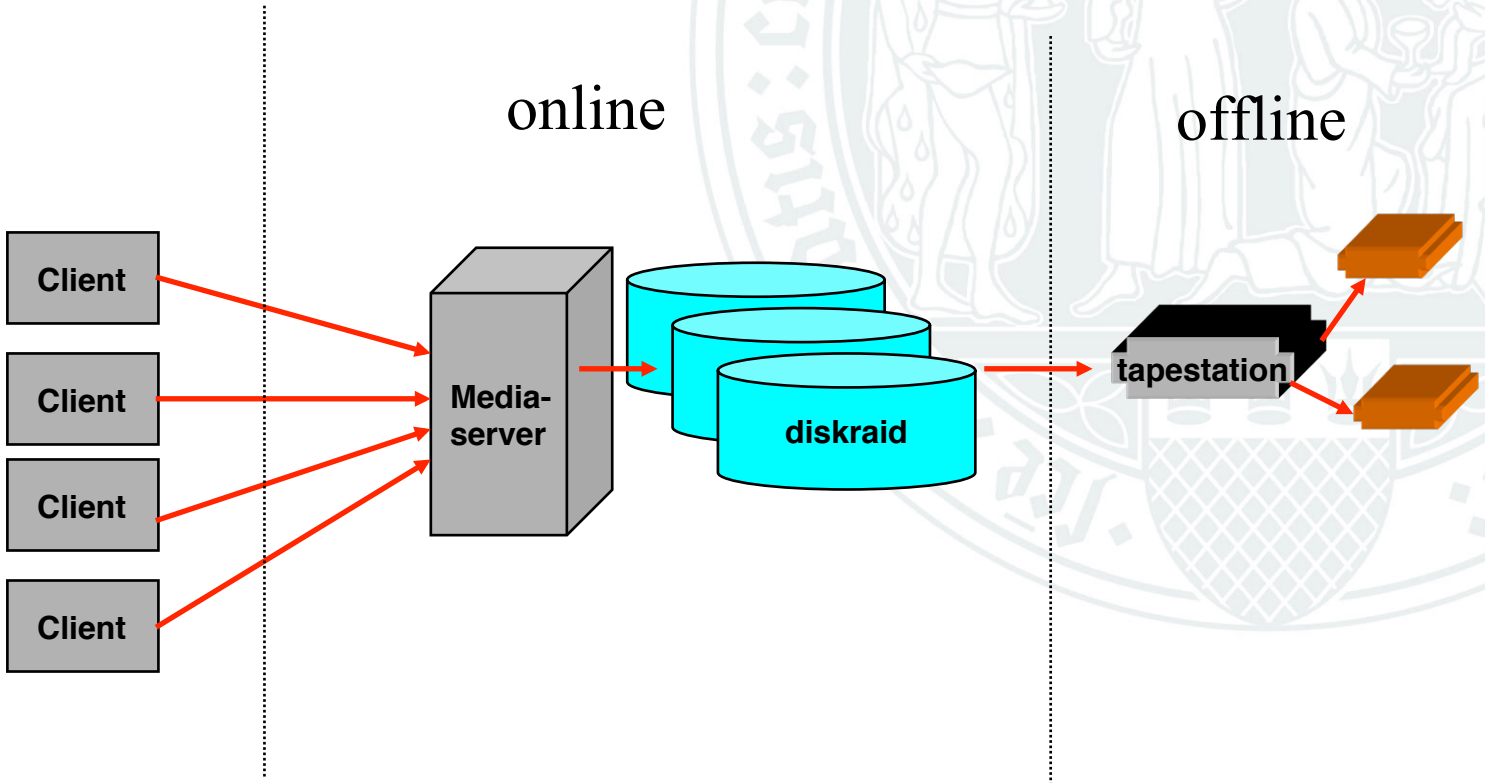
Monitoring

- Zombie jobs
- Server load
- Memory consumption
  - leakage
  - fine on a PC, but with uptime > 100 days…
- Database
  - error messages
  - instability
  - abnormal behaviour
- Must be monitored by humans (using tools)
  - email messages with control data

Extending metadata

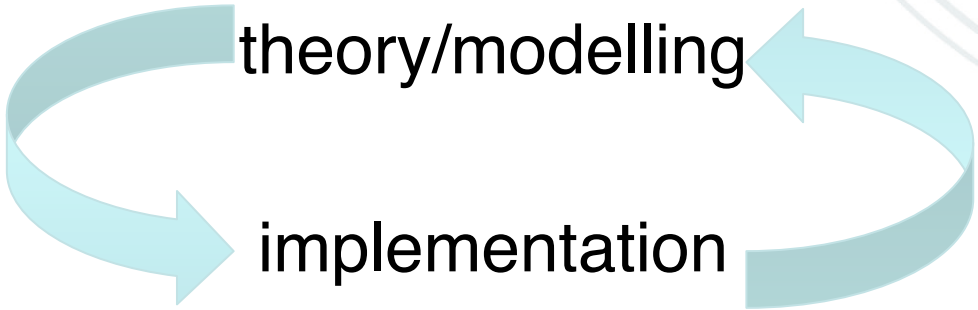
- Some metadata must be kept
  - e.g., process records
  - format details
- Some metadata can be changed
  - classification
  - motive description
- But keep old versions
  - institution history
  - legal liability
    - historical institutional/governmental racism
    - land rights

Overall architecture



Towards abstract modelling

- How can this method be generalised?
  - Some preliminary notes
- Learning strategies
- The role of theory



What is an image?

- An image can be found
  - in a data file
  - on a 35mm film
  - at a paper positive
  - at a glass plate
  - …
- We do not care, they are all images
  - modelled as media_units
  - connected to media_groups
- If something is
  - another media_unit to an existing media_group, or
  - a derived media_group
- is a scholarly (content based) choice

The transformation event

- Each transformation event happened in time
  - may or may not know when
  - actor(s) may be know or unknown
- Transfer events from
  - analogue to analogue
  - analogue to digital
  - digital to digital
  - (digital to analogue)
- are recorded in the same way

Chains of events

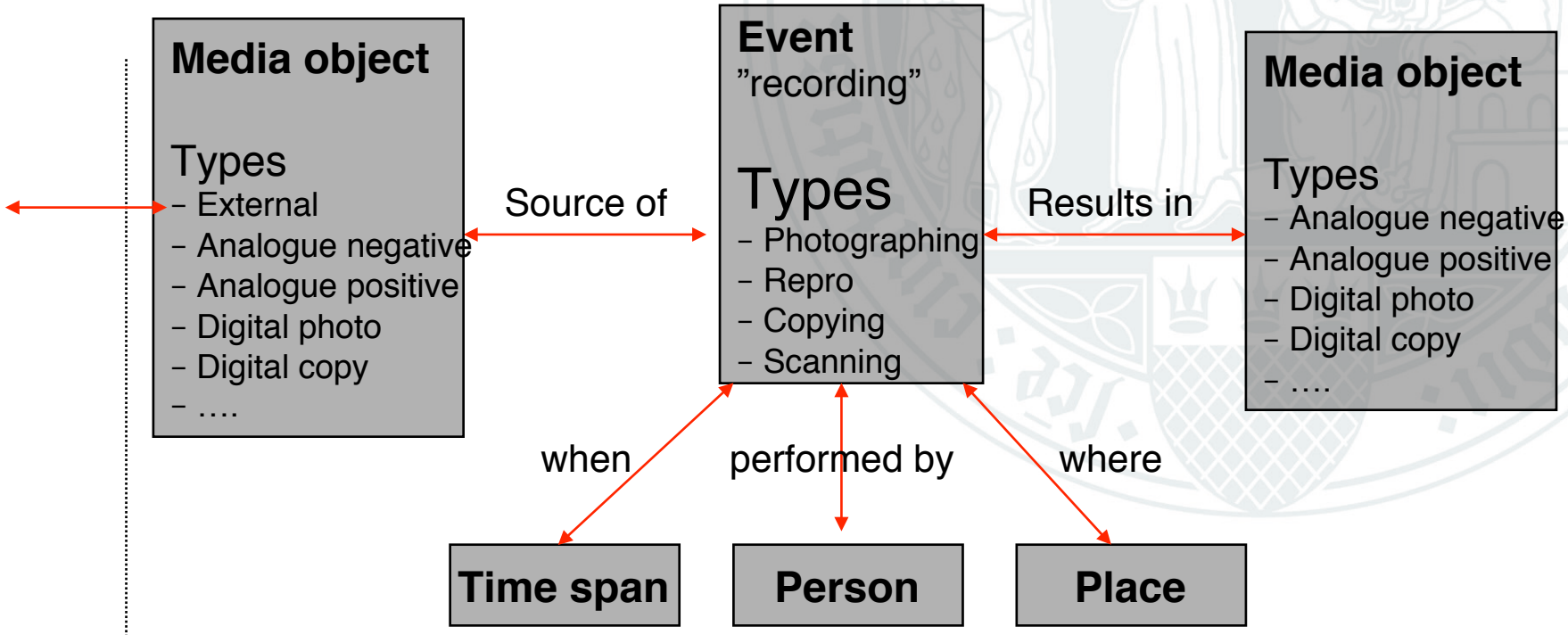
- Connected to a image there is a chain of events
  - like the passport of a person with stamps
- Can see where the image comes from
- Can step in at any point to re-do processing
- Some images can be seen as caches
  - but the distinction between cached and not is not central
  - rather: some processes can be re-done
  - but be aware of detail differences
    - program versions
    - libraries

The memory of events

- Storing events: writing history
- This is obviously important for old stuff
  - museums try to track provenience
- But all new will become old
- History is made by our scripts
  - we can record it or let it go

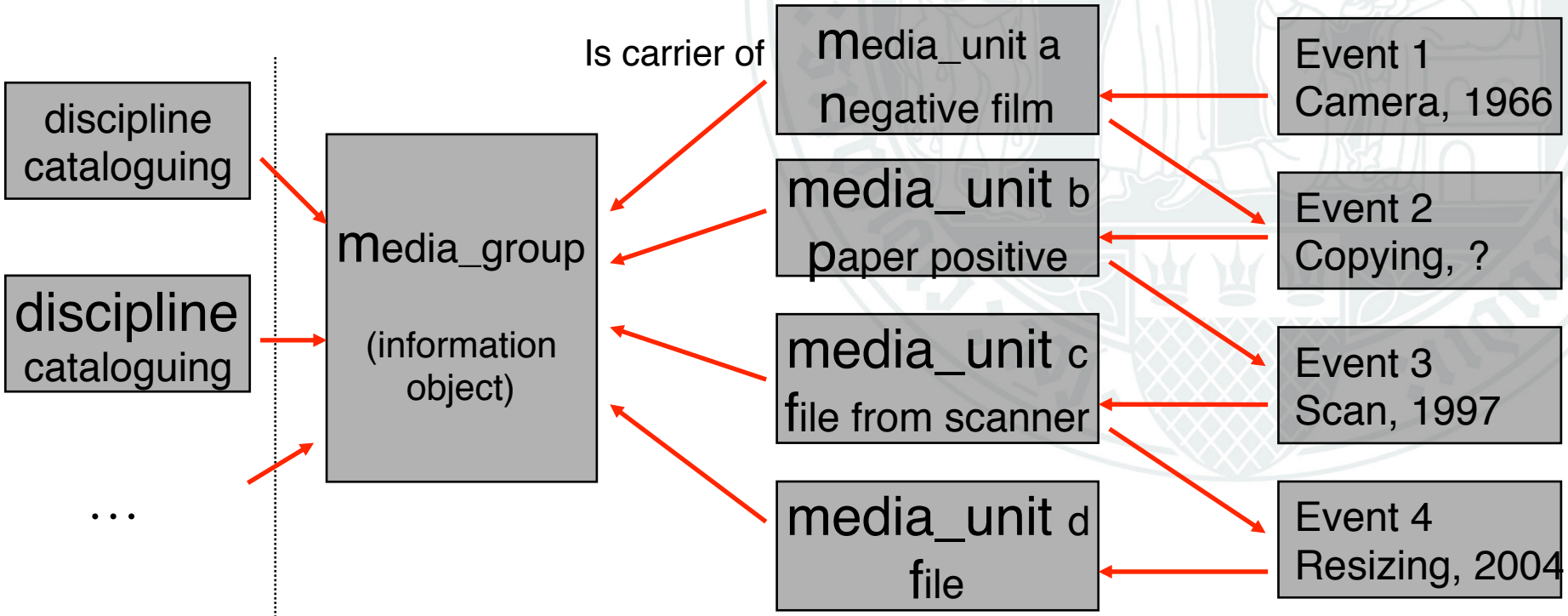
Reminder: Event centric data model

source → recording → result



Reminder: Data model (example)

Separation between “information object” and “information carriers”



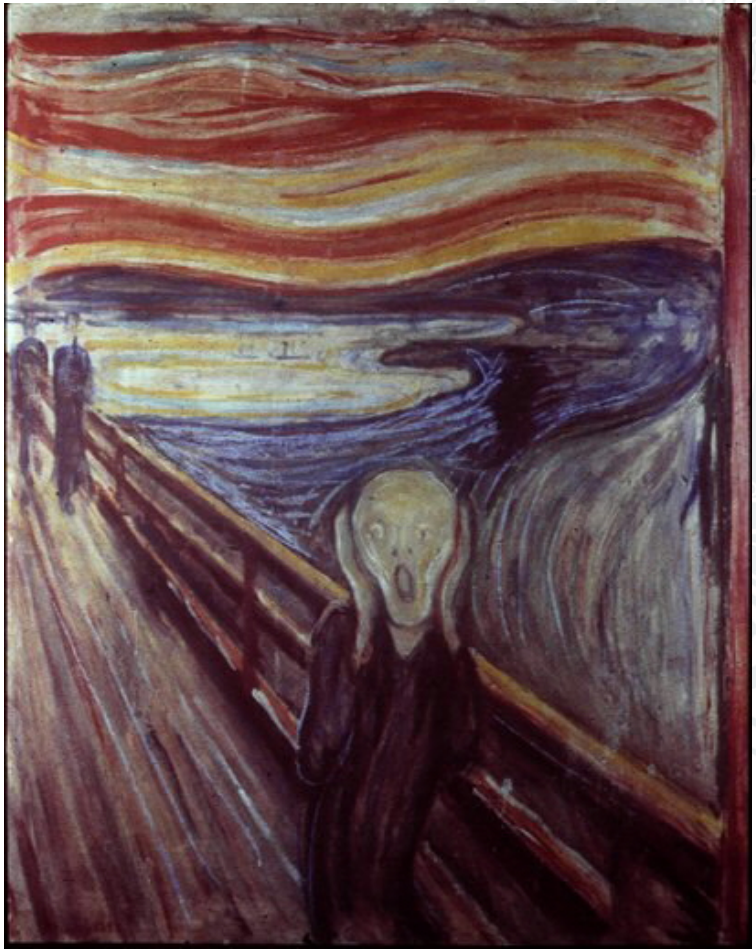
Remember: Database

Discipline schema 1
• image_card – media_group_id – concent_metadata • *subject specific tables*

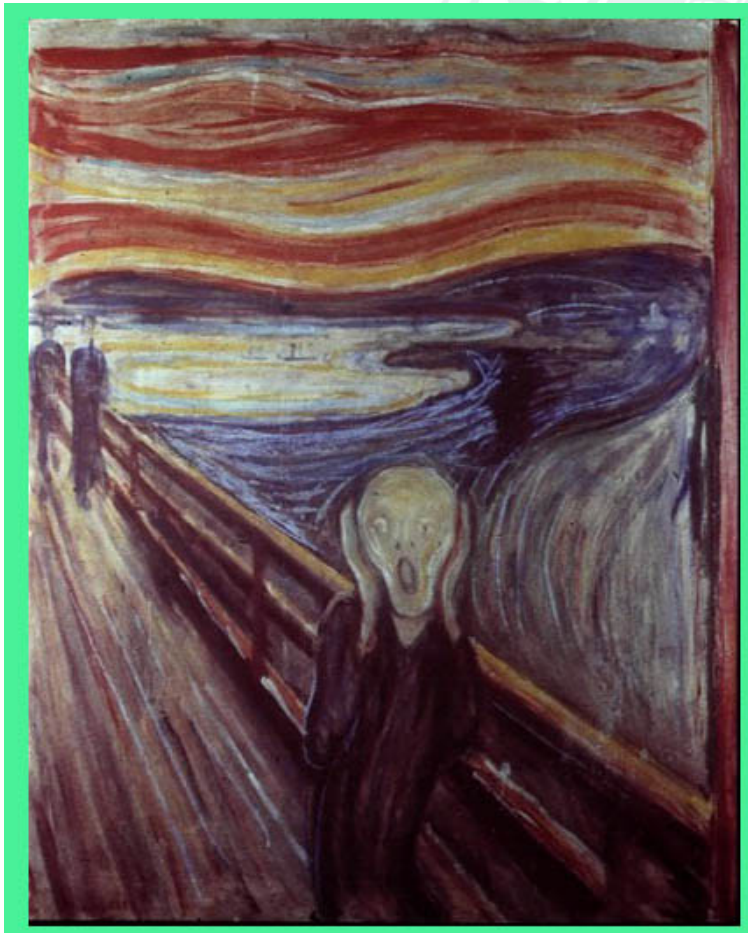
Common schema
• media_group – id – process_xml – tech_metadata • media_unit – tech_metadata – default {0, 1} • schema_setup – process_spec – delete_limit • process – process_spec – status {0, 1, 2, 9}

Discipline schema 2
• image_card – media_group_id – concent_metadata • *subject specific tables*

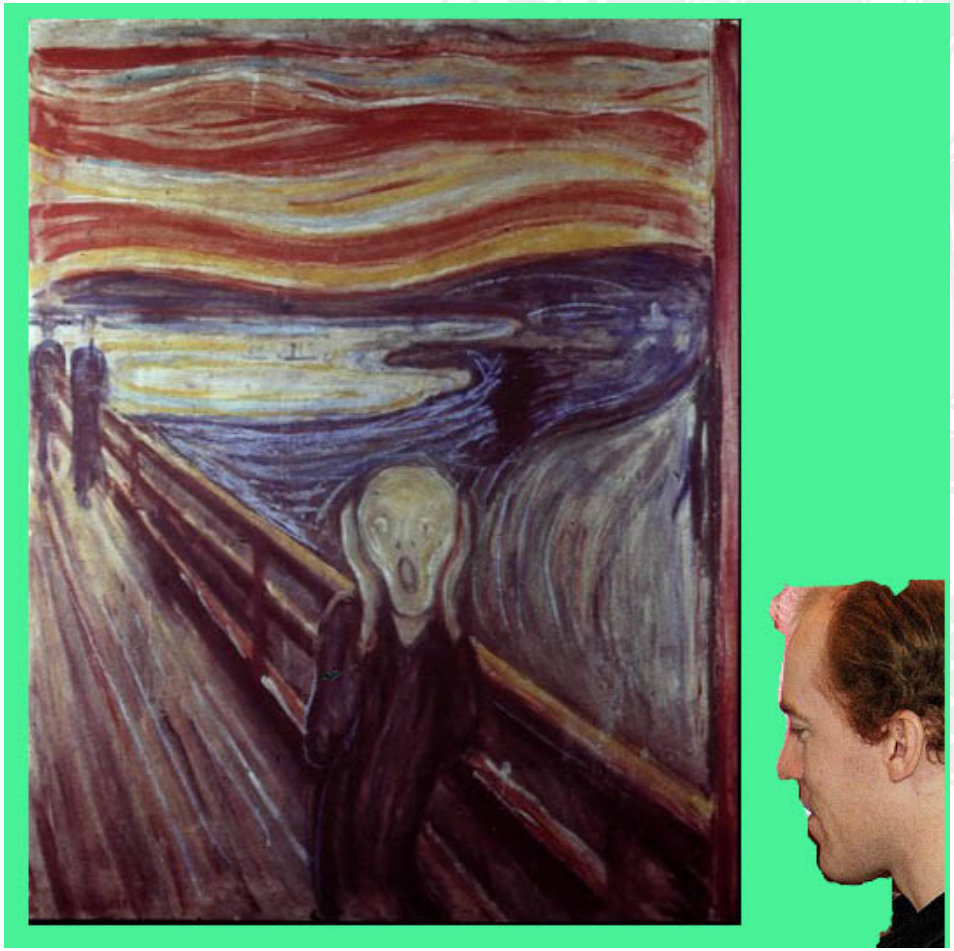
Which discipline?



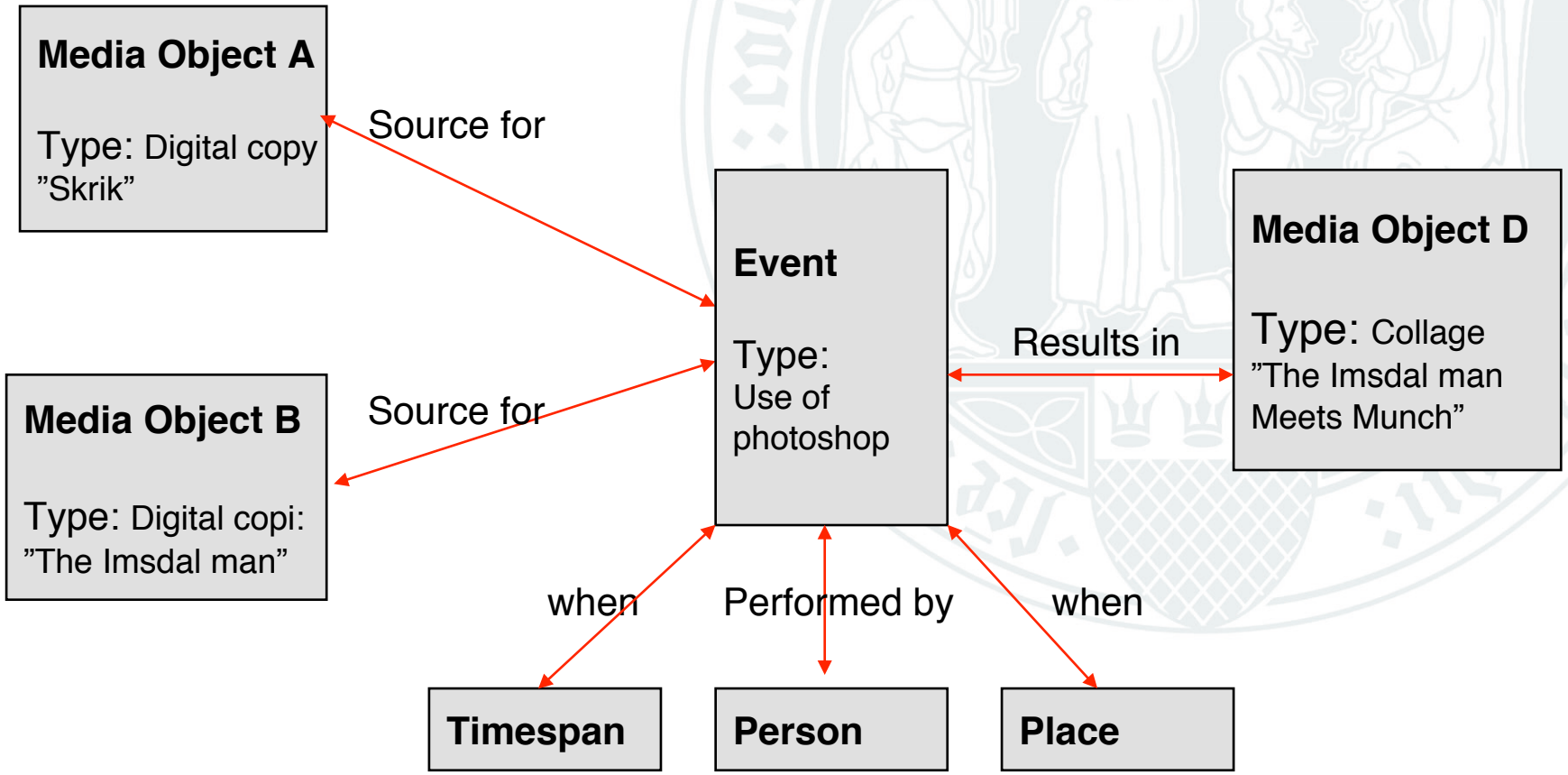
Which discipline?



Which discipline?



Event centric data model



Date model, concepts and things