VISUAL HISTORY OF THE HOLOCAUST
Rethinking Curation in the Digital Age

Deliverable D4.2

Metadata Integration Concept

Lead-beneficiary	*OFM*
Work Package No. and Title	*WP4 Curation and Enrichment of Metadata*
Work Package Leader	*JLU*
Relevant Task	*Task 4.2 Integration of existing metadata*
Task Leader	*OFM*
Main Author(s)	*OFM (Peter Bubestinger-Steindl, Michael Loebenstein)*
Contributor(s)	*LBI (Ingo Zechner)*
Reviewer(s)	*DIF (Kerstin Herlt, David Kleingers)* *LBI (Ingo Zechner)*
Dissemination Level	*Public*
Due Date	*M12 (2019-12)*
Version (No., Date)	*V1.4, 2019-12-31*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822670. www.vhh-project.eu

VISUAL HISTORY OF THE HOLOCAUST
Rethinking Curation in the Digital Age

Metadata Integration Concept

Table of Contents

1. Introduction

This document presents the outcomes of our evaluation of the most suitable metadata model to capture, manage and enrich cataloguing information relating to a broad range of filmic, and non-filmic heritage materials used in the project.

It presents a proposed metadata schema – the **VHH-EFG Metadata Schema** – that will effectively:

- serve the purpose of capturing existing metadata on archival materials, in particular film heritage materials;
- meet the requirements of new database technologies;
- remain compatible with existing relevant standards and schemas, in particular the ones employed by VHH Consortium members and widely adopted in the European flagship program European Film Gateway (EFG).

This document is intended to inform the acquisition, management and enrichment of metadata pertaining to film materials and related non-filmic archival materials in film heritage institutions (FHIs), other collecting institutions, or academic institutions. Indeed, as the FIAF Cataloguing Manual states "accurate, well-organized descriptions of both filmographic and technical information" about the heritage materials acquired, ingested and used "constitute the key to accessing collections by external users such as scholars, researchers and the general public – both now and for future generations."[1]

This document furthermore also serves as a guideline for the technical partners involved in VHH to implement the most suitable database engine within the VHH Media Management and Search Infrastructure (VHH-MMSI).

The purpose of this document is to

- provide a schema to manage all film- and non-film related heritage material metadata;
- inform the database implementation in the VHH-MMSI;
- serve as a best practice guide to institutions evaluating models and schemas to manage rich time-based metadata.

This document is a living document to be used together with the following deliverables:

- D4.1 Data Management Plan
- D2.1 Advanced Digitization Tool Kit
- D4.1 Controlled Vocabularies Specification
- D5.3 System Design v1 Report

[1] The FIAF Moving Image Cataloguing Manual. Maria Assunta Pimpinelli and Thelma Ross, edited by Linda Tadic and Natasha Fairbairn. Indiana: IUP, 2016, p.6

2. The VHH-EFG Schema: Methodological Background

Metadata integration is central to providing effective technical and curatorial solutions to engage users in the curation and enrichment of digital heritage materials. Existing technical and filmographic metadata pertaining to the filmic heritage materials – originating either in the archival source records (f.e. catalogue records, databases) or created during advanced digitization (such as scanner logs, checksums, etc.) – therefore needs to be aggregated, analyzed, normalized, mapped onto the VHH metadata schema and imported into the VHH-MMSI for further enrichment. A specific metadata set – the *Cinematography of the Holocaust* plays a key role and will be updated and expanded by project partner DIF to serve as the filmographic core of the VHH-MMSI.

A decision was made early on to base the VHH metadata schema on the following criteria:

- It must be aligned with modern cataloguing and metadata standards recommended by the relevant peak bodies in the field.
- It must allow for capturing and documenting the specificities of moving image (film) works.
- It should be well documented, with implementations available to evaluate its suitability for use in an archive/library/museum context.
- It should be interoperable with business systems (databases or media asset management systems) used in related projects or by VHH Consortium members.
- It must allow for the implementation of the project-specific controlled vocabularies and entities.

Following from this approach concepts such as FRBR (Functional Requirements for Bibliographic Records)[2] and its entities-relationship model, and published standards such as EN 15907, PBCore and RDA (Resource Description and Access) were evaluated for their suitability for implementation.

Based on the FRBR model, EN 15907 is the de-facto standard for moving image material description[3]. In a further step the team analyzed and evaluated existing implementations ("flavors") of the standard to see if any existing schema was fit for use in the project or could be meaningfully expanded to accommodate for project-specific extensions.

Another area of address were three aspects of EN 15907 which are currently underdeveloped:

- While highly suitable for describing moving image works it is less developed for the description of non-film materials (f.e. paper documents, books, still photographs etc.).

[2] https://www.ifla.org/publications/functional-requirements-for-bibliographic-records (31.12.2019)
[3] http://filmstandards.org/fsc/index.php/EN_15907 (18.12.2019)

- Its hierarchy of entities foresees no instance smaller than “Item” which means that instances for the time-based annotation (“Frame”, “Shot”, “Segment”, “Transition”) need to be created as extensions.
- Its model of relationships between entities is somehow rudimentary.

3. Evaluation of existing Metadata Models and Schemas

In a next step we reviewed all relevant metadata schemas and standards that are either EN 15907-compliant, related or potentially suitable for augmenting it (e.g. describing film-related material). In the course of this evaluation a "Metadata Workshop" was hosted by DIF in Frankfurt on October 21-22, 2019 to further analyze and discuss the shortlisted standards and schemas, and to consider workflow issues in regard to metadata integration.

The outcome of these considerations was that the **EFG Metadata Schema**[4] was identified as the most suitable and fit-for-reuse metadata schema. It will serve as the core upon which our implementation – the **VHH-EFG Metadata Schema** – is based.

3.1. Considered Formats/Standards

- EFG (European Film Gateway)
- EDM (Europeana Data Model)[5]
- RDA (Resource Description and Access)[6]
- RDF (Resource Description Framework)[7]
- DIF-XML (filmportal.de Schema)[8]
- PBCore (Public Broadcasting Metadata Dictionary)[9]
- DC (Dublin Core)[10]
- EAC-CPF (Encoded Archival Context for Corporate Bodies, Persons, and Families)[11]
- IIIF (International Image Interoperability Framework)[12] + W3C Open Annotation Data Model[13]

3.2. Summary of Findings

- The EFG schema is very well structured, and addresses some requirements not covered by EN 15907 such as treating non-film materials ("NonAVCreation")

[4] https://www.efgproject.eu/guidelines_and_standards.php (18.12.2019)
[5] https://pro.europeana.eu/page/edm-documentation (18.12.2019)
[6] https://www.loc.gov/aba/rda/pdf/core_elements.pdf (18.12.2019)
[7] https://www.w3.org/RDF/ (18.12.2019)
[8] http://www.filmstandards.org/schemas/de-dif/zf-fw-view-1.5/ (31.12.2019)
[9] https://pbcore.org/what-is-pbcore (18.12.2019)
[10] https://www.dublincore.org (31.12.2019)
[11] http://www.loc.gov/ead/ (18.12.2019)
[12] https://iiif.io/api/presentation/3.0 (18.12.2019)
[13] https://www.w3.org/TR/annotation-model/ (18.12.2019)

analog to film works ("AVCreation"). It is multi-lingual by default and provides extensive compatible vocabularies.

- RDA is a bit fuzzy to grasp, since it seems to be intended as a schema and also a cataloging ruleset for a librarian context. It is closely related to the idea of RDF, but lacks consistent semantic structuring/grouping. Its "RDF registry" is filled with terms, which may provide as a useful source for LoD and/or controlled vocabulary terms. Yet, its data structure and design is "flat" (instead of hierarchical/typed). The term collection seems more like rows in a table than structured text. This raises doubts if it would actually be good to use RDA as a source/guideline.
- The DIF-XML (as used by DIF on filmportal.de as export format) is clean, but there seem to be no benefits over EFG, yet its German XML elements/attributes – even including umlauts – would rather suggest not to use it.
- It makes sense to think in RDF triplets and keep machine-readability in mind, yet it is better not to store data directly in this format, but rather provide interfaces that can generate RDF-compatible output on demand.
- PBCore is a very good and practical format to describe audiovisual resources. PBCore is well documented and also well structured, but except for a subset (e.g. for technical metadata), it doesn't seem to have advantages over EFG for the cataloging of filmographic works.
- Dublin Core field terms can and should be used in all parts where there is no other definition against using it.
- EAC-CPF looks like a proper schema to use for CPF, yet for use in VHH, we prefer restricting it to a suitable subset of fields instead of supporting the whole standard. It also contains flaws that hinder proper machine-readability when not using "vocabularySource" in certain elements (see translation issue in the paragraph about EAC-CPF).
- A decision was made furthermore to use IIIF and its definition of W3C's "Annotation Data Model" as an extension to cover time/based annotation in the "Content" entity, thus addressing the fundamental gap mentioned previously in the EN 15907 standard and its various implementations.

A more detailed analysis is provided in Appendix B of this Metadata Integration Concept.

4. The VHH-EFG Schema: Description

4.1. Deviations from plain EFG

As already mentioned, the main use case of EFG, and what is was tailored for, is not cataloging, but aggregation of metadata for a common metadata hub and then rendered for display on a website.

We therefore decided to take the liberty to revert some EFG-specific definitions back to their CEN (European Committee for Standardization) origins in EN 15907[14].

Wherever EFG's definitions gravitate predominantly towards a data-exchange/web-hub use case, CEN's definition will be preferred over EFG. This also includes re-adding some fields or entities from CEN which were dropped in EFG.

For field naming, EFG's wording will be preferred over CEN wherever possible.

Overview of Adaptations

- Identifier & Scheme: allow to use Identifier and Scheme for any kind of archival object identifier
- Agent: generic CEN Agents instead of three hardcoded types
- Event: added *PreservationEvent and LegalChange (new)*
- Item: use Item from CEN with field-wording from EFG (where possible)
- AVCreation: skipped fields, added DescriptionLevel and use YearOfReference instead of ProductionYear
- Thumbnail images: optional, not mandatory
- Content: not used but handled by the content annotation schema

Identifier & Scheme

In EFG, the unique identifier for an entity is defined as being automatically generated by the EFG database. This is only suitable for the EFG web use case.

For regular cataloging, the generic CEN definition is preferred, as it allows to use institution-specific identifiers.

The Identifier- and Identifier.Scheme field is defined for all entities in EFG. This decision therefore applies to all entities (AVCreation, AVManifestation, Agent, etc).

[14] Cf. Film identification – Enhancing interoperability of metadata – Element sets and structures (ÖNORM EN 15907 Edition 2010-10-01); for easy access cf. http://filmstandards.org/fsc/index.php/EN_15907 (31.12.2019).

Quote from the tabular definition of EFG:

Identifier

"Identifier of the [Entity] that will be automatically generated by the EFG database (GUID or ID chosen from an external naming schema)."

Identifier.Scheme

"The scheme by which the EFG identifier will be generated. This can be an internationally known identification scheme, an XML namespace identifier, or a URN namespace (e.g. URN: UUID)."

In CEN, *Identifier* is defined as follows:

"**An unambiguous reference to the resource within a given context**, where possible the International Standard Audio-visual Number (ISAN), otherwise a specific number issued by a government department or other official body in an individual country, **or an archive's inventory number**."

The *Identifier.Scheme* is defined as follows:

"**An unambiguous reference of the scope within which the identifier is unique.** References may take the form of an internationally known identification scheme, an XML namespace identifier […], or a URN namespace (e.g. URN:UUID or URN:ISAN).

This change is CEN compatible, while providing the option to store an EFG UID as a parallel Identifier, if an entity is present in the EFG database and therefore had an UID generated.

When importing entity data from external sources, that originated from another institution or database, CEN as well as EFG define a separate field, called "RecordSource.SourceID", which will not be used in the VHH-EFG Schema. Source information will be handled by Changelog and DataSource instead.

Agent

In CEN, there is the entity *Agent*with the field *Agent.Type* to distinguish between individuals, corporations, etc. In EFG on the other hand, there are three hardcoded types of agent: *Agent.Person*, *Agent.CorporateBody*, *Agent.Group*

These three Agent types are hardcoded and all their common fields implemented in parallel. A comparison has shown that they have 17 fields identically in common and only these four differ:

- Place
- Place.Type
- Sex
- ViewBiography

These can be further reduced to two, since "Place" (and its type) is only defined for Person and CorporateBody, but also possibly makes sense for any other Agent type which leaves only "Sex" (Gender) and "ViewBiography" to be Person-specific.

It therefore seems better to generalize this back to the classic *CEN Agent* definition, as it leads to a simpler and less rigid practical implementation.

The four fields that mostly only apply to *Agent.Perso*n can still exist for each *Agent.Type*. Simple cataloging rules will be used to declare which of these four fields may not be used for which *Agent.Type*.

Both versions, EFG's 3-type and CEN's generic Agent definition can be converted to one another without ambiguity or data loss.

Event

PreservationEvent

Since EFG was defined only for displaying filmographic information on a website, it has omitted the Preservation.Event type.

This event type will be added, according to its definition in CEN.

LegalChange

Rights information is only rudimentarily covered in EFG and except for Intellectual Property Rights (IPR) registration, there is no place defined in CEN as well to capture relevant information.

An additional event type was therefore added which is intended to record changes that affect the legal status of heritage materials referenced in the VHH database. This event type is called "LegalChange".

The currently considered (but not limited to these) use cases are:

- Change of ownership
- Change of content license
- Change of IPR

The event has the following elements:

- **Type:**
  The type of legal change that was done. This shall be a controlled vocabulary.
  Examples: ownership, content license, IPR
  Occurence: 1
- **Detail:**
  Any information describing the process in greater detail. Freetext.
  Occurence: 0/1

Item

Use Item from CEN

The Item in EFG is purely designed for dealing with aggregation of files. CEN's Item is designed to describe physical film material as well as digital files.

We therefore prefer the CEN definition over EFGs, in order to be able to represent analog audiovisual items too.

EFG has a completely different Item than CEN, with no common fields (except for an identifier, but under different names). In order to keep the VHH-EFG Schema consistent, CEN-Item fields will be renamed to EFG terms where it makes sense.

CEN Item.sourceID = EFG.Identifier

For consistency inside the schema, EFG's Identifier definition will be used instead of CEN's "Item.sourceID". Identifier can depict the same information if necessary, while staying EFG/CEN compatible.

Use CEN InstantiationType instead of EFG Type

In EFG, Item.Type would be used to declare the media type of an item (text, image, video, sound), whereas in CEN Item.InstantiationType can be used for this, while being able to be used for digital as well as analog items.

Adding EFG Note

All other EFG entities have "Note" defined – except for Item. For consistency, this field (and its attributes) will be added to Item.

Excluded fields in Item

The following elements are defined in EFG but are surplus to requirements in a cataloging context and in VHH specifically. We therefore decided to exclude them in VHH-EFG.

- Attribute: HighQuality
- isShownBy
- isShownAt
- Provider
- Aggregator
- URI
- Country
- DigitalFormat

Note The information to be stored in *Item.DigitalFormat* are valuable, but also described in *AVManifestation.Format.DigitalFormat* and therefore superfluous here.

AVCreation

DescriptionLevel

The attribute "Work.DescriptionLevel" was added to AVCreation.

YearOfReference instead of ProductionYear

The more generic definition "Year of Reference" from CEN (6.6) is preferred over being able to only store the year of production.

Excluded fields in AVCreation

The following elements are defined in EFG but are surplus to requirements in a cataloging context and in VHH specifically. We therefore decided to exclude them in VHH-EFG.

- ViewFilmography

Thumbnail images

The "thumbnail" field in AV- and NonAVManifestation are defined mandatory in EFG. This makes sense for having web presence as a main use case, but for regular cataloging, this field shall be optional.

Content

Content related elements defined in EFG shall not be used but handled by the content annotation schema.

The structure and layout in which these fields will be implemented in IIIF shall conform to the definition in EFG in order to allow easy exchange with conventional catalogs that are CEN/EFG compatible.

This affects the following metadata elements, currently defined:

- (Non)AVCreation:
  - 4.5 Keywords
  - 4.6 Description
  - 4.7 User Tag
- (Non)AVManifestation:
  - Coverage

Keywords are Tags

For cataloging, keywords and tags can be considered the same thing. The EFG specification of "4.5 Keywords" and "4.7 User Tag" elements make it clear that:

- **Keyword:**
  Is for describing the content of (Non)AVCreation (usually by a cataloger)
- **UserTag:**
  "tags created by the [web-]user through the EFG interface"

Since content description values will be stored in a different schema and location in the database, there is no need to distinguish between "keywords" and "tags". They will therefore be stored only as "keywords" in IIIF.

Content annotation in VHH has its creator recorded and stored by default, therefore this information will be used to distinguish between the EFG use cases of "Keywords vs UserTags".

It therefore stays compatible with EFG-only systems.

4.2. Changelog and DataSource

The following features are required for recording the provenance of information entered into fields, either on import or entered manually by annotators/curators. It helps to ensure proper scholarly research standards are met, in particular if information deviates between two sources, and in cases where annotators manually edit and alter information in a field (corrections etc.):

- write changelogs on edit
- option to add a comment when changing a record (similar to a "commit message" in version control)
- option to enter/view "DataSource" per field

They are however independent of the data model used and therefore considered "beyond" the metadata schema: they are considered a feature-request for the cataloging engine.

4.3. Content Annotations

As noted in our evaluation of metadata models, EFG/CEN schemas are predominantly designed and used for capturing filmographic data. The "Content" entity by default is designed for recording static (non-time based) keywords (or subject terms) and cataloging descriptions or commentaries. More complex cataloging needs are addressed by individual implementations.

For VHH a different data model and mechanism is required for the time-based annotation of content.

For VHH, classical text-only time-based annotation (technically similar to closed captions) is insufficient: A structured text schema must be defined which allows to create content description for different scholarly demands yet is machine readable and interoperable.

It is desirable to use existing standards, especially ones that are designed for- and used in the preservation domain and which are technically well-designed.

Although IIIF API definitions for moving image is relatively new and officially still in BETA state, its goals align perfectly with the needs of VHH.

The following components are part of IIIF and will be used in VHH:

- IIIF for segmentation and structural information15
- W3C Web Annotation Data Model16 (aka "Open Annotation17") for content annotation.
- JSON-LD (Linked Data)18 and IIIF Notes on JSON-LD19

Linking Annotations to EFG Entities

In order to assign content annotations to its related EFG entries, the same "Relationship" model as used throughout CEN/EFG is used.

To quote the CEN standard specification:

"A relationship associates an instance of an entity with another instance of an entity. For each relationship defined in this standard, a cardinality of zero or more is assumed for both ends, resulting in a many-to-many association. This may be restricted to a one-to-many association in suitable cases (e.g. for HasVariant)."

This newly defined Relationship for content annotation is called:

- HasContentAnnotation

A Relationship can have its own metadata attributes. For linking content annotation as relationship, the following attributes are defined:

- Motivation: [Controlled vocabulary] The OpenAnnotation "Motivation" value as used in the IIIF Manifest (see below). This allows easier differentiation between different annotations in the catalog database engine, without requiring to parse IIIF/OpenAnnotation data structures on this level.

---

15 https://iiif.io/technical-details/ (18.12.2019)
16 https://www.w3.org/TR/annotation-model/ (18.12.2019)
17 http://openannotation.org/ (18.12.2019)
18 https://json-ld.org/spec/latest/json-ld/ (18.12.2019)
19 https://iiif.io/api/annex/notes/jsonld/ (18.12.2019)

- Type: [Controlled vocabulary] A more concrete information about the annotation type. Examples include transition, shot, audio, wikidata, etc.

Annotations: Motivation and Purpose

IIIF uses so called "Motivation" and "Purpose" in its Manifests to describe different kinds of annotations, as defined in W3C's Annotation Data Model's "Motivation and Purpose"20.X

Here is a list of Motivations/Purposes defined by W3C:

- **assessing**: The motivation for when the user intends to assess the target resource in some way, rather than simply make a comment about it. For example: to write a review or assessment of a book, assess the quality of a dataset, or provide an assessment of a student's work.
- **bookmarking**: The motivation for when the user intends to create a bookmark to the Target or part thereof. For example: an annotation that records the point in a text where the reader stopped reading.
- **classifying**: The motivation for when the user intends to classify the Target as something. For example: to classify an image as a portrait.
- **commenting**: The motivation for when the user intends to comment about the Target. For example: to provide a commentary about a particular PDF document.
- **describing**: The motivation for when the user intends to describe the Target, as opposed to commenting on it. For example: describing the above PDF's contents, rather than commenting on their accuracy.
- **editing**: The motivation for when the user intends to request a change or edit to the Target resource. For example: an annotation that requests a typo to be corrected.
- **highlighting**: The motivation for when the user intends to highlight the Target resource or segment of it. For example: to draw attention to the selected text that the annotator disagrees with.
- **identifying**: The motivation for when the user intends to assign an identity to the Target. For example: to associate the IRI that identifies a city with a mention of the city in a web page.
- **linking**: The motivation for when the user intends to link to a resource related to the Target.
- **moderating**: The motivation for when the user intends to assign some value or quality to the Target. For example: annotating an Annotation to moderate it up in a trust network or threaded discussion.

[20] https://www.w3.org/TR/annotation-model/#motivation-and-purpose (18.12.2019)

- **questioning**: The motivation for when the user intends to ask a question about the Target. For example: to ask for assistance with a particular section of text or to question its veracity.
- **replying**: The motivation for when the user intends to reply to a previous statement, either an Annotation or another resource. For example: providing the assistance requested in the above.
- **tagging**: The motivation for when the user intends to associate a tag with the Target.

Selectors

In VHH annotation goes beyond creating static (non-time based) metadata. Annotators and curators need to be able to define segments (shots, sequences) in audiovisual material in order to describe "content components" – for example persons, objects, events etc. – represented within these segments. The same principle applies to non-time-based digital heritage materials such as books, documents and photographs: annotations refer to certain pages or paragraphs from longer documents, or to regions in an image.

The W3C Annotation Data Model defines so called "Selectors" to select a part of a resource: a so called "Segment of Interest".

The quote from W3C's Selectors definition[21] describes and matches this requirement well:

"Many Annotations refer to part of a resource, rather than all of it, as the Target. We call that part of the resource a Segment (of Interest). A Selector is used to describe how to determine the Segment from within the Source resource. The nature of the Selector will be dependent on the type of resource, as the methods to describe Segments from various media-types will differ. Multiple Selectors can be given to describe the same Segment in different ways in order to maximize the chances that it will be discoverable later, and that the consuming user agent will be able to use at least one of the Selectors."

The concept of Selectors is generic and requires different Selector-Types to handle different media types.

FragmentSelector

The FragmentSelector is a type of Selector that currently supports defining "positions" of parts within the following media types.

The content of this table is taken from W3C's FragmentSelector definition[22] and shows the fragment specification links, as well an example for how to define the fragment.

[21] https://www.w3.org/TR/annotation-model/#selectors (18.12.2019)
[22] https://www.w3.org/TR/annotation-model/#fragment-selector (18.12.2019)

Name	Fragment Specification	Example
Plain Text	http://tools.ietf.org/rfc/rfc5147	char=0,10
PDF	http://tools.ietf.org/rfc/rfc3778	page=10&viewrect=50,50,640,480
Media	http://www.w3.org/TR/media-frags/	t=ss.xx#xywh=50,50,640,480

Name	Fragment Specification	Example
HTML	http://tools.ietf.org/rfc/rfc3236	namedSection
XML	http://tools.ietf.org/rfc/rfc3023	xpointer(/a/b/c)
RDF/XML	http://tools.ietf.org/rfc/rfc3870	namedResource
CSV	http://tools.ietf.org/rfc/rfc7111	row=5-7
SVG	http://www.w3.org/TR/SVG/	svgView(viewBox(50,50,640,480))
EPUB3	http://www.idpf.org/epub/linking/cfi/epub-cfi.html	epubcfi(/6/4[chap01ref]!/4[body01]/10[para05]/3:10)

For time-based media, such as audiovisual material, the position syntax used is: t=time_in[,time_out] ("time_out" being optional and used only if defining a range). The time value is defined as XML dateTime literal[23], with the syntax being [hh:mm:]ss.xx. For example: http://example.org/video1#t=30,60

It can be combined with xywh parameter to describe a visual area at a certain point in time. Please see the Media Fragments Syntax Section[24] of the W3C "Media Fragments URI specification" for more details on how to combine and phrase more complex fragments.

Other Selectors useful for VHH

There are more Selectors than just "FragmentSelector" defined in W3C's Annotation Data Model, but the following ones may be particularly interesting and useful for VHH:

- Text Quote Selector25
- Text Position Selector26

[23] https://www.w3.org/TR/xmlschema11-2/#dateTime (18.12.2019)
[24] https://www.w3.org/TR/media-frags/#media-fragment-syntax (18.12.2019)
[25] https://www.w3.org/TR/annotation-model/#text-quote-selector (18.12.2019)
[26] https://www.w3.org/TR/annotation-model/#text-position-selector (18.12.2019)

- Range Selector[27]

VHH Annotation Types

- Content Summary (keywords, abstract, synopsis, table of contents, index, etc.)
- Content Components (object, situation, action, event, organization, nationality, audio, written words, etc.)
- Frame
- Shot
- Segment
- Transition
- Page

Links and References

- IIIF: JSON-LD Implementation Notes[28]
- Data Catalog Vocabulary (DCAT, W3C)[29]
- JSON-LD (Linked Data)[30]
- JSON-LD Framing[31]
- IIIF Design Patterns[32]
- IIIF Annotation Library (NCSU)[33]
- IIIF Text Granularity Technical Specification Group Charter[34]
- IIIF Annotation Studio (2D)[35]
- Issue #764: Granularity Parameter in Search API[36]
- Issue #32: Tagging of content. e.g. link to streetview of the scenes[37]

[27] https://www.w3.org/TR/annotation-model/#range-selector (18.12.2019)
[28] https://iiif.io/api/annex/notes/jsonld/ (18.12.2019)
[29] https://www.w3.org/TR/vocab-dcat/ (18.12.2019)
[30] https://www.w3.org/TR/json-ld/ (18.12.2019)
[31] https://json-ld.org/spec/latest/json-ld-framing/ (18.12.2019)
[32] https://iiif.io/api/annex/notes/design_patterns/ (18.12.2019)
[33] https://ncsu-libraries.github.io/iiif-annotation/ (18.12.2019)
[34] https://iiif.io/community/groups/text-granularity/charter/ (18.12.2019)
[35] https://github.com/atomotic/iiif-annotation-studio (18.12.2019)
[36] https://github.com/IIIF/api/issues/764 (18.12.2019)
[37] https://github.com/IIIF/iiif-av/issues/32 (18.12.2019)

5. Metadata Integration: Concept and Workflow Principles

All metadata for use in the VHH project is ingested into the VHH-MMSI which consists of the project database and a number of software services as described in D5.3 "System design v1".

We differentiate between:

- pre-existing metadata;
- metadata created by users of the VHH-MMSI who interact with project-specific data and/or metadata.

Pre-existing metadata originates from a variety of contributors, in various schemas and file formats. It includes information provided by content providers as well as data created in external systems during processes of automated analysis.

For ingest into the VHH-MMSI this metadata is:

- aggregated and corrected;
- mapped onto the VHH-EFG schema;
- bulk-uploaded by the responsible aggregating partners (DIF, LBI, OFM and TUW) via XML exports;
- or alternatively harvested via an automated data exchange interface (such as OAI-PMH) by the VHH-MMSI.

Created metadata includes confirmation, correction and enrichment of non-time-based and time-based metadata performed manually via the VHH-MMSI functionality ("annotation") by a number of contributors (annotators) including test users during the trial period. An annotation manual will be prepared by LBI as an internal working paper to ensure data quality and consistency.

Roles of the Beneficiaries in Metadata Creation

Aggregation, normalization and mapping of external pre-existing metadata is performed by several consortium members according to the most effective, efficient and economical use of project resources.

- **DIF:** provides metadata sets of the Cinematography of the Holocaust (CDH). The metadata is aggregated, quality-checked and normalized to DIF-ZDB standard in DIF's business systems. It is then mapped onto VHH-EFG Schema and exported to the VHH-MMSI or alternatively harvested via API (to be determined).
- **OFM:** aggregates metadata sets from content providers of digital film heritage materials, including technical metadata from advanced digitization processes in OFMs business system, the Digital Film Master Repository (DFMR). It is then mapped onto VHH-EFG Schema and uploaded to the VHH-MMSI.
- **LBI:** aggregates metadata sets from its own and from Consortium members' literature research, and from content providers of digital non-film heritage

materials, in LBIs business systems, Citavi and MS Excel. This metadata is then mapped onto VHH-EFG Schema and uploaded to the VHH-MMSI.

- **TUW:** creates time-based metadata in the course of the automated analysis of digital heritage materials, in TUW's business systems. It is then mapped onto the VHH-EFG Schema and exported to the VHH-MMSI or alternatively harvested via API (to be determined).

A test mapping of common source metadata schemas and formats was performed as part of the development of the VHH-EFG Schema.

Information on common data file formats collected, processed and stored in the course of the project can be found in the most current version of D1.4 Data Management Plan.

Appendix A. The VHH-EFG Schema: Table Overview

Tabular overview of the VHH-EFG metadata model

This spreadsheet file contains several sheets, listing all entities and elements which compose the EFG/CEN based data model used in the „Visual History of the Holocaust“ (VHH) Project.

Each sheet represents an entity and lists all elements that belong to it. In all cases a short description as well as the occurrence status of the element is provided.

The syntax for occurrence status is:

0/1 = zero or one
0+ = zero or more
1+ = one or more

Data types: When controlled vocabulary from EFG is listed it only serves as a suggestion.

Exists in CEN (EN15907) and EFG. Field names may, but function is identical.

Exists in CEN, and was added to EFG for regular film cataloging

Field is an Attribute (of an Element)

Newly added field that did not originate from CEN

Not used in VHH Schema

Mandatory element

AVCreation

EFG 3.1, CEN 4.1

This entity refers to the concept of Cinematographic Work as defined in EN 15907 (clause 4.1).
An audiovisual creation should only have properties that remain constant throughout all of its manifestations.

Element			Description	Status	Datatype
Identifier			An unambiguous reference to the resource within a given context, where possible the International Standard Audio-visual Number (ISAN), otherwise a specific number issued by a government department or other official body in an individual country, or an archive's inventory number.	1	
	Scheme		An unambiguous reference of the scope within which the identifier is unique. References may take the form of an internationally known identification scheme, an XML namespace identifier [...], or a URN namespace (e.g. URN:UUID or URN:ISAN).	1	[CVOC] Efg:IDScheme
RecordSource			A reference to the EFG content provider and the local IDs. This is a wrapper element for all elements and attributes listed hereunder.	1+	
	SourceID		The local identifier of the audiovisual creation. If this does not exist in the content provider´s database the value is "undefined".	0/1	
	Provider		The name of the archive supplying the record.	1	
		Identifier	An unambiguous reference to the archive supplying the record. This is preferably the institution's identifier, which is registered by the ISIL agency. If the institution is not registered, then the official acronym of the institution will be sufficient.	1	
		IdentificationScheme	Name of the registration scheme encoding the institution name ("ISIL code" or "EFG instit	1	
Note			This element includes all data derived from content providers, that cannot fit into other elements.		
	Source		This attribute specifies the source that produced the note element.		
DescriptionLevel			A one-character symbol indicating the role of the cinematographic work within part-whole relationships: Analytic (component part), Monographic, Serial and Collection.	1	[CVOC] {a,m,s,c}
Title			Titles can occur several levels of description and can change during the lifecycle of a creation. Titles can be a word, phrase or character, naming the work or a group of works, a particular manifestation or an individual item. The title on the Creation level is a derivate of the best known manifestation title. Film titles in EFG will be managed through a common authority registry. The title element is modelled after the EN 15907 (clause 6.3). Title is a wrapper element for all elements and attributes listed hereunder.	1+	
	Language		The language of the title. (If no value is available in the source archive, the value must be set to "unspecified".)	1	
	TitleText		The textual expression of the title.	1	
	PartDesignation		A combination of the name of a structuring unit and the count value that identifies the current creation as an individual part of a complex work. For complex works with multi-level numberings (e.g. "Season 2, Episode 12"), an instance of this element should be created for each level. Non-numeric values such as letter symbols (e.g. "Part B" or calendar dates - e.g. "August 1956 issue") are permitted.	0/1	
		Unit	The name of the unit that is represented within a series, serial, or multi-part work. (E.g. "Part", "Episode", "Issue", etc)		
		Value	An ordinal number, numeration symbol, or date value that identifies the individual creation within a complex work.		

	TitleRelationship		The type of relationship between the title and the entity to which it is assigned. (e.g. "Original title", "Distribution title", "Translated title" etc)	0/1	
	TemporalScope		For titles that were changed, or introduced after the date of issue, the time span in which the title was used should be indicated.	0/1	
	GeographicScope		Indication to which region(s) the title applies if a work was distributed under different titles in distinct geographic regions.	0/1	
		Code	Region code.	0+	
		RegionName	Name of a region other than a country.	0+	
IdentifyingTitle			A short phrase for identifying the audiovisual creation, to be used e.g. in human-readable result lists from database queries. The identifying title should not be constructed by a cataloguer according to cataloguing rules but algorithmically during the metadata conversion process for EFG.	1	
	Origin		Acronym or other identifier indicating the origin of the element content. For identifying titles created automatically, this should be the name and version of the software agent.		
CountryOfReference			The geographic origin of a audiovisual creation. This should be the country or countries where the production facilities are located. If production information is missing, this element can refer to countries where the audiovisual creation was filmed or distributed, or where copies are known to exist in archives.	1+	
	Reference		The relationship between a geographic area and the audiovisual creation. Defaults to "production", use of this attribute is only required if the point of reference is different from the country of production.		[CVOC] efg:CountryReference
YearOfReference			A year associated with an event in the life cycle of the cinematographic work, typically associated with its creation, availability or registration (for example for copyright purposes). A typical use of this element is chronological ordering of lists of cinematographic works. The year of reference is expressed as a four digit value, optionally followed by a dash (Unicode value 002Dhex) and another year to denote a span of years.	1+	4 digit value (optional: year range)
	Reference		A name for the event the Year refers to. If omitted, Year will be interpreted as the production year.		[CVOC]
	DerivedFrom		If this attribute is empty or omitted, then it is assumed that the “Production Year” has been validated as a real production year. Otherwise, it may contain a textual string which specifies how it has been derived (e.g. inferred from censorship).		Freetext

~~ViewFilmography~~			~~An unambiguous URL reference to the full filmographic entry of a film on the content provider web site.~~		
~~UserTag~~			~~This is a tag created by the user through the EFG interface.~~		
	~~Language~~		~~This indicates the language of the tag.~~		
	~~Creator~~		~~This indicates the ID of the user that has created the tag.~~		
	~~Date~~		~~This indicates the date on which the tag has been created (“YYYY-MM-DD”).~~		

Moved to Content Annotation (IIIF), but kept here as reference for implementing an EFG-matching field layout:

Keywords			A term or set of vocabulary terms describing the content of an audiovisual or non-audiovisual creation. This can be keywords or other vocabularies to describe subjects. Controlled and uncontrolled terms can be used together, but not within a single set of terms. Likewise, if more than one controlled vocabulary is used, then terms from each of these must be contained in a separate instance of this element. A separate instance is also required for each language if terms in more than one language are taken from a multilingual vocabulary. This is a wrapper element for all elements and attributes listed hereunder.	0+	
	Language		The language of the content of each subject. Can be omitted if the language is evident from the scheme.		
	Type		Type of information described by the keywords (e.g. "city", "person", "lake" etc.) or the subject (e.g. "Genre", "Category").		[CVOC] efg:KeywordType
	Scheme		A unique identifier denoting the controlled vocabulary (preferably URI). If the subject terms are not from a controlled vocabulary, the value of this element should be set to “uncontrolled”.		[CVOC] efg:IDScheme
	Term		An element containing a single term. This can be the textual value of the term. For non-textual terms the classification codes, preferably a combination of the code and the verbal class description should be indicated. (If no value is available in the source archive, the value must be set to "not available in the source archive".)	1+	
		ID	A non-text identifier that can be combined with the scheme ID from a unique resource identifier for the term within a controlled vocabulary.		
Description			Textual descriptions include synopses, plot summaries, reviews, transcripts or shot lists. They can occur in more than one language and they can have statements of authorship or references to external resources.		
	Language		The language of the description text.		
	Type		A keyword denoting the type of description (e.g. "Synopsis", "Shot list", "Review" etc.).		
	Source		Either the name of the institution or an URI identifying the source directly or via a reference system such as an on-line catalogue.		

AVManifestation

EFG 3.2, CEN 4.3

This entity groups all properties that can change during the lifecycle of an audiovisual creation without affecting the identity of a film work as such. A manifestation is the physical embodiment of an audiovisual creation. Examples are archival copies (analogue or digital) or distribution files. This entity was modelled by using the EN 15907 (clause 4.3).

Element			Description	Status	Datatype
Identifier			An unambiguous reference to the resource within a given context, where possible the International Standard Audio-visual Number (ISAN), otherwise a specific number issued by a government department or other official body in an individual country, or an archive's inventory number.	1	
	Scheme		An unambiguous reference of the scope within which the identifier is unique. References may take the form of an internationally known identification scheme, an XML namespace identifier [...], or a URN namespace (e.g. URN:UUID or URN:ISAN).	1	[CVOC] Efg:IDScheme
RecordSource			A reference to the EFG content provider and the local IDs. This is a wrapper element for all elements and attributes listed hereunder.	0+	
	SourceID		The local identifier of the audiovisual creation. If this does not exist in the content provider´s database the value is "undefined".	0/1	
	Provider		The name of the archive supplying the record.	1	
		Identifier	An unambiguous reference to the archive supplying the record. This is preferably the institution's identifier, which is registered by the ISIL agency. If the institution is not registered, then the official acronym of the institution will be sufficient.	1	
		IdentificationScheme	Name of the registration scheme encoding the institution name ("ISIL code" or "EFG insti	1	
Note			This element includes all data derived from content providers, that cannot fit into other elements.	0+	
	Source		This attribute specifies the source that produced the note element.		
Title			Titles can occur several levels of description and can change during the lifecycle of a creation. Titles can be a word, phrase or character, naming the work, a group of works, a particular manifestation or an individual item. The title on the Creation level is a derivate of the best known manifestation title. Film titles in EFG will be managed through a common authority registry. The title element is modelled after the EN 15907 (clause 6.3). Title is a wrapper element for all elements and attributes listed hereunder.	0+	
	Language		Language of the manifestation's title. (If no value is available in the source archive, the value must be set to "unspecified".)	1	
	TitleText		The textual expression of the manifestation's title.	1	
	PartDesignation		A combination of the name of a structuring unit and the count value that identifies the current manifestation as an individual part of a complex work. For complex works with multi-level numberings (e.g. "Season 2, Episode 12"), an instance of this element should be created for each level. Non-numeric values such as letter symbols (e.g. "Part B" or calendar dates - e.g. "August 1956 issue") are permitted.		
		Unit	The name of the unit that is represented within a series, serial, or multi-part manifestation. (E.g. "Part", "Episode", "Issue", etc)		
		Value	An ordinal number, numeration symbol, or date value that identifies the individual creation within a complex manifestation.		
	TitleRelationship		The type of relationship between the title and the entity to which it is assigned (e.g. "Original title", "Working title", "Distribution title" etc.).		

	TemporalScope		For titles that were changed, or introduced after the date of issue, the time span in which the title was used should be indicated.		
	GeographicScope		Indication to which region(s) the title applies if a work was distributed under different titles in distinct geographic regions.		
		Code	Region code.		
		RegionName	Name of a region other than a country.		
Language			The language of the spoken, sung or written content.	0+	[CVOC] Efg:iso639-1/2LanguageCS
	LanguageUsage		This indicates the relationship between the language and the manifestation (e.g. "original spoken dialogue", "dubbing", "subtitles", "voice-over commentary" etc.). Defaults to “unspecified”.		[CVOC] Efg:LanguagePurposeCodeCS
Dimension			The total physical dimension of the manifestation represented as numeric value, with decimal places if required. For example, if the carrier type is a DVD, it represents the DVD size. If the carrier type is a file, it represents the size in bytes, etc.	0+	
	Unit		Unit of the physical dimension of the manifestation (e.g. “bytes”, “meters”, “inches”).		[CVOC] Efg:DimensionUnitCS
	Reference		If the manifestation is composed of more than one physical carrier for which individual lengths are known, then this attribute should be used to refer the length information to a particular carrier (e.g. “reel 3”).		Freetext
Duration			The running time of the audiovisual manifestation measured in minutes and seconds.	0/1	
	Framerate		Optional value for the projection speed, given in frames per second, to which the given duration refers.		Freetext
Format			The description of the physical artefact or digital file on which an audiovisual manifestation is fixed. This is a wrapper element for all elements and attributes listed hereunder.	0+	
	DigitalFormat		The format of the digital file.	0/1	
		Container	A container or wrapper format is a file format, or often a stream format whose specifications regard only the way data are stored but *not* coded within the file (e.g. “AVI”, “MP4”, “3GP”, “RealMedia”).	0/1	
		Coding	Specifies the codification of the data (e.g. “WMA”,”WMV”, “MPEG-4”, “RealVideo”).	0/1	
		OriginalStatus	Specifies if the object is originally digital or derived from a physical object (e.g. “born digital” or “digitised”).	0/1	
	CarrierType		The manifestation's physical carrier. Digital manifestations may be bound to a physical carrier (e.g. "DVD") or exist without a defined carrier (e.g. "online streaming media", "podcasts").	0/1	
	Gauge		The width of the film stock or other carrier (such as magnetic tape) used for the manifestation. Should include value and unit (e.g. "35 mm", "1/2 inch" etc).	0/1	
	AspectRatio		The ratio between width and height of the image (e.g. "full frame", "cinemascope", "1:1,33").	0/1	
	Sound		The name of the system by which sound is recorded either on the carrier, on a separate medium, or as part of the digital encoding (e.g. Western Electric (Westrex), Movietone, RCA Photophone, Tobis, Dolby Digital, etc.).	0/1	
		HasSound	A boolean value indicating if the manifestation includes recorded sound ("true" or "false").		
		RecordingSystem	A boolean value indicating if the system and/or method was used for the primary recording ("true" or "false").		

	Colour		The name of the colour system or process. For digital formats the name of the colour space. For analogue videos the name of the video colour system.	0/1	
		HasColour	A boolean value indicating if the manifestation includes recorded sound ("true" or "false").		
RightsHolder			Name of the copyright holder.		Freetext
	URL		If available, URL to the homepage of the copyright holder.		Text (URL)
RightsStatus			Specifies the copyright status of the digital item. Rights statement is controlled by the vocabulary "Copyright protected" or "Not copyright protected".		[CVOC] Efg:Rightsstatus
Provenance			Organisation which owns or has custody of the item (analog or born digital).		
Thumbnail			Link to the reduced-size image of the manifestation (link or “unknown”).	0/1	Text (URL)

Moved to Content Annotation (IIIF), but kept here as reference for implementing an EFG-matching field layout:

Coverage			The spatial or temporal topic of the audiovisual manifestation.	0+	
	Spatial		This may be a named place, a location, a spatial coordinate or a named administrative entity.		Freetext
	Temporal		This may be a period, date or range date.		Freetext OR Controlled syntax (ISO8601)

NonAVCreation

EFG 3.2, CEN -

Non-audiovisual creations in EFG can be pictures, photos, correspondence, books or periodicals. The descriptions of NonAVCreations and their associated manifestations follow an early version of the Europeana Semantic Elements specification. For the data exchange with the Europeana system mappings to the current ESE version will be constructed. Digital representations of non-film works available for viewing are accommodated in an associated instance of the item entity.

Element			Description	Status	Datatype
Identifier			An unambiguous reference to the resource within a given context, where possible the International Standard Audio-visual Number (ISAN), otherwise a specific number issued by a government department or other official body in an individual country, or an archive's inventory number.	1	
	Scheme		An unambiguous reference of the scope within which the identifier is unique. References may take the form of an internationally known identification scheme, an XML namespace identifier [...], or a URN namespace (e.g. URN:UUID or URN:ISAN).	1	[CVOC] Efg:IDScheme
RecordSource			A reference to the EFG content provider and the local IDs. This is a wrapper element for all elements and attributes listed hereunder.	0+	
	SourceID		The local identifier of the audiovisual creation. If this does not exist in the content provider´s database the value is "undefined".	0/1	
	Provider		The name of the archive supplying the record.	1	
		Identifier	An unambiguous reference to the archive supplying the record. This is preferably the institution's identifier, which is registered by the ISIL agency. If the institution is not registered, then the official acronym of the institution will be sufficient.	1	
		IdentificationScheme	Name of the registration scheme encoding the institution name ("ISIL code" or "EFG insti	1	
Note			This element includes all data derived from content providers, that cannot fit into other elements.	0+	
	Source		This attribute specifies the source that produced the note element.		
Title			Titles can occur several levels of description and can change during the lifecycle of a creation. Titles can be a word, phrase or character, naming the work, a group of works, a particular manifestation or an individual item. The title on the Creation level is a derivate of the best known manifestation title. Film titles in EFG will be managed through a common authority registry. The title element is modelled after the EN 15907 (clause 6.3). Title is a wrapper element for all elements and attributes listed hereunder.	1+	
	Language		The language of the title. (If no value is available in the source archive, the value must be set to "unspecified".)	1	
	TitleText		The textual expression of the title.	1	
	PartDesignation		A combination of the name of a structuring unit and the count value that identifies the current creation as an individual part of a complex work. For complex works with multi-level numberings (e.g. "Season 2, Episode 12"), an instance of this element should be created for each level. Non-numeric values such as letter symbols (e.g. "Part B" or calendar dates - e.g. "August 1956 issue") are permitted.		
		Unit	The name of the unit that is represented within a series, serial, or multi-part work. (E.g. "Part", "Episode", "Issue", etc)		
		Value	An ordinal number, numeration symbol, or date value that identifies the individual creation within a complex work.		
	TitleRelationship		The type of relationship between the title and the entity to which it is assigned ("Main title", "Alternative title".).		

	TemporalScope		For titles that were changed, or introduced after the date of issue, the time span in which the title was used should be indicated.		
	GeographicScope		Indication to which region(s) the title applies if a work was distributed under different titles in distinct geographic regions.		
		Code	Region code.		
		RegionName	Name of a region other than a country.		
DateCreated			The point or period of time associated with the creation of the non-audiovisual creation (“YYYY-MM-DD” or “YYYY”).	0+	ISO8601 (YYYY-MM-DD, YYYY, etc)
Language			The language of the major verbal content of the document.	0/1	[CVOC] Efg:iso639-1/2LanguageCS

~~UserTag~~			~~This is a tag created by the user through the EFG interface~~	~~0+~~	
	~~Language~~		~~This indicates the language of the tag.~~		
	~~Creator~~		~~This indicates the user name that has created the tag.~~		
	~~Date~~		~~This indicates the date on which the tag has been created.~~		

Moved to Content Annotation (IIIF), but kept here as reference for implementing an EFG-matching field layout:

Keywords			A term or set of vocabulary terms describing the content of an audiovisual or non-audiovisual creation. This can be keywords or other vocabularies to describe subjects. Controlled and uncontrolled terms can be used together, but not within a single set of terms. Likewise, if more than one controlled vocabulary is used, then terms from each of these must be contained in a separate instance of this element. A separate instance is also required for each language if terms in more than one language are taken from a multilingual vocabulary. This is a wrapper element for all elements and attributes listed hereunder.	0+	
	Type		Type of information described by the keywords (e.g. "city", "person", "lake" etc.) or the subject (e.g. "Genre", "Category").		
	Language		The language of the content of each subject. Can be omitted if the language is evident from the scheme.		
	Scheme		A unique identifier denoting the controlled vocabulary (preferably URI). If the subject terms are not from a controlled vocabulary, the value of this element should be set to “uncontrolled”.		
	Term		An element containing a single term. This can be the textual value of the term. For non-textual terms the classification codes, preferably a combination of the code and the verbal class description should be indicated. (If no value is available in the source archive, the value must be set to "not available in the source archive".)	1+	Freetext or [CVOC]
		ID	A non-text identifier that can be combined with the scheme ID from a unique resource identifier for the term within a controlled vocabulary.		
Description			Textual descriptions include synopses, plot summaries, reviews, transcripts or shot lists. They can occur in more than one language and they can have statements of authorship or references to external resources.	0+	
	Language		The language of the description text.		
	Type		A keyword denoting the type of description (e.g. “Synopsis”, “Shotlist”, “Review” etc.).		
	Source		Either the name of the institution or an URI identifying the source directly or via a reference system such as an on-line catalogue.		

NonAVManifestation

EFG 3.4, CEN -

This entity has the function to handle all film-related materials, such as documents, books, photos, posters, etc.

Element			Description	Status	Datatype
Identifier			An unambiguous reference to the resource within a given context, where possible the International Standard Audio-visual Number (ISAN), otherwise a specific number issued by a government department or other official body in an individual country, or an archive's inventory number.	1	
	Scheme		An unambiguous reference of the scope within which the identifier is unique. References may take the form of an internationally known identification scheme, an XML namespace identifier [...], or a URN namespace (e.g. URN:UUID or URN:ISAN).	1	[CVOC] Efg:IDScheme
RecordSource			A reference to the EFG content provider and the local IDs. This is a wrapper element for all elements and attributes listed hereunder.	0+	
	SourceID		The local identifier of the audiovisual creation. If this does not exist in the content provider´s database the value is "undefined".	0/1	
	Provider		The name of the archive supplying the record.	1	
		Identifier	An unambiguous reference to the archive supplying the record. This is preferably the institution's identifier, which is registered by the ISIL agency. If the institution is not registered, then the official acronym of the institution will be sufficient.	1	
		IdentificationScheme	Name of the registration scheme encoding the institution name ("ISIL code" or "EFG institution code").	1	
Note			This element includes all data derived from content providers, that cannot fit into other elements.		
	Source		This attribute specifies the source that produced the note element.		
Title			Titles can occur several levels of description and can change during the lifecycle of a creation. Titles can be a word, phrase or character, naming the work, a group of works, a particular manifestation or an individual item. The title on the Creation level is a derivate of the best known manifestation title. Film titles in EFG will be managed through a common authority registry. The title element is modelled after the EN 15907 (clause 6.3). Title is a wrapper element for all elements and attributes listed hereunder.	0+	
	Language		Language of the title. (If no value is available in the source archive, the value must be set to "unspecified".)		
	TitleText		The textual expression of the title.		
	PartDesignation		A combination of the name of a structuring unit and the count value that identifies the current manifestation as an individual part of a complex work. For complex works with multi-level numberings (e.g. "Season 2, Episode 12"), an instance of this element should be created for each level. Non-numeric values such as letter symbols (e.g. "Part B" or calendar dates - e.g. "August 1956 issue") are permitted.		
		Unit	The name of the unit that is represented within a series, serial, or multi-part work. (E.g. "Part", "Episode", "Issue", etc)		
		Value	An ordinal number, numeration symbol, or date value that identifies the individual creation within a complex work.		

	TitleRelationship		The type of relationship between the title and the entity to which it is assigned (“Original title”, “Alternative title”.).		
	TemporalScope		For titles that were changed, or introduced after the date of issue, the time span in which the title was used should be indicated.		
	GeographicScope		Indication to which region(s) the title applies if a work was distributed under different titles in distinct geographic regions.		
		Code	Region code.		
		RegionName	Name of a region other than a country.		
Type			The general type of the non-audiovisual manifestation (“image”, “text”, “audio”).	1	
SpecificType			This element further specifies the type of the non-audiovisual manifestation. Here the document type should be indicated (e.g. "photograph", "poster", "letter").	1	
Language			The language of the spoken, sung or written content.	0+	
	LanguageUsage		This indicates the relationship between the language and the manifestation (e.g. original spoken dialogue, dubbing, subtitles, voice-over commentary, etc.). Defaults to “unspecified”.		
Date			A point or period of time associated with the publication of the non-audiovisual manifestation.		ISO8601 (YYYY-MM-DD, YYYY, etc)
	Type		This attribute specifies the type of temporal properties (e.g. issued, digitised).		[CVOC] Efg:ObjectDateType
DigitalFormat			The digital format of the manifestation (RFC 2049 MIME types, e.g. “image/jpg”, ”text/plain” etc.).	0/1	
	Status		Specifies if the object is originally digital or derived from a physical object ("born digital" or "digitized").		
	Size		The dimensions of the digital object expressed in bytes.		
	Resolution		The degree of sharpness of the digital object expressed in lines or pixel.		
PhysicalFormat			The general format of the physical object (e.g. “DIN A4”, “DIN A3”).	0/1	
	Size		The dimensions of the physical object expressed in the standard format width x height, in millimetres.		
Colour			This element can be used to indicate the colour of a non-audiovisual object (e.g. "black and white", "colour", "mixed").		[CVOC] efg:colour
GeographicScope			Indicates the region where the non-audiovisual object was used (e.g. “ISO 3166-1” country codes, except for historical geographic entities).	0+	[CVOC] Efg:iso3166-1CountryCS, efg:XPZ44-002HistoricCountryCS, efg:Regions_LocallyDefined
RightsHolder			Name of the copyright holder.	0+	Freetext
	URL		If available, URL to the homepage of the copyright holder.	0+	Text (URL)
RightsStatus			Specifies the copyright status of the digital item. Rights statement is controlled by the vocabulary "Copyright protected" or "Not copyright protected".	0+	[CVOC] Efg:Rightsstatus
Provenance			Organisation which owns or has custody of the item (analogue or born digital).	0/1	
Thumbnail			Link to the reduced-size image of the manifestation (link or "unknown").	0/1	Text (URL)

Moved to Content Annotation (IIIF), but kept here as reference for implementing an EFG-matching field layout:

Coverage			The spatial or temporal topic of the non-audiovisual manifestation.	0+	
	Spatial		This may be a named place, a location, a spatial coordinate or a named administrative entity.		Freetext
	Temporal		This may be a period, date or date range.		Freetext OR Controlled syntax (ISO8601)

Item

EFG 3.5, CEN 4.4

Item in EFG is purely designed for dealing with aggregation of files, CEN's Item definition can accomodate physical(film) material as well as files. It is therefore preferred to use the CEN definition over EFGs, in order to be able to depict analog audiovisual items too. EFG has a completely different Item than CEN, with no common fields (except for an identifier, but under different names). In order to keep the VHH-EFG schema consistent, CEN-Item fields will be renamed to EFG terms where it makes sense.

Element			Description	Status	Datatype
Identifier			(CEN:SourceID) An identifier for the item-level record, if this exists in the database from which the filmographic record was Produced.	1+	
	Scheme		An unambiguous reference of the scope within which the identifier is unique. References may take the form of an internationally known identification scheme, an XML namespace identifier [...], or a URN namespace (e.g. URN:UUID or URN:ISAN).	1	[CVOC] Efg:IDScheme
RecordSource			A reference to the EFG content provider and the local IDs. This is a wrapper element for all elements and attributes listed hereunder.	0+	
	SourceID		The local identifier of the audiovisual creation. If this does not exist in the content provider´s database the value is "undefined".	0/1	
	Provider		The name of the archive supplying the record.	1	
		Identifier	An unambiguous reference to the archive supplying the record. This is preferably the institution's identifier, which is registered by the ISIL agency. If the institution is not registered, then the official acronym of the institution will be sufficient.	1	
		IdentificationScheme	Name of the registration scheme encoding the institution name ("ISIL code" or "EFG institution code").	1	
Note			This element includes all data derived from content providers, that cannot fit into other elements.	0+	
	Source		This attribute specifies the source that produced the note element.		
Title			Titles should be given on the item level (a) if it is not known if the title found on or in the item has been used for other items of the same manifestation, (b) if no title is known or directly applicable (such as for trims, outtakes and rushes from the production of published works) and the archive has created its own descriptive 10EN 15907:2010 (E) title, or (c) if the item contains a unique combination of more than one work (e.g. a projection reel with added trailers and advertisements).	0+	
Holding Institution			The name of the archive or other institution possessing the copy or authorised to make it available.	1+	Agent
InstantiationType			A term or phrase describing the item type relative to the duplication process (e.g. "original negative", "dup negative", "positive", "original positive (reversal film)", "dupe positive", "Lavender", "image negative", "sound negative", "non-film analogue carrier", "non-film digital carrier").	0/1	
ItemSpecifics			This element should be used for recording any major differences between the individual copy and the properties stated in the associated manifestation record. It is not intended for statements about minor degradations such as scratches. Some examples: "b/w copy of manifestation (in color)", "copy without soundtrack from sound film", "fragment, 1 500 m of 2 400 m (manifestation)", "4 of 5 reels (last reel missing)", Etc.	0/1	
AccessConditions			Any information on how and to whom the item can be made available.	0+	
CatalogueReference			A reference to a more detailed catalogue record describing the item.	0+	

Original EFG fields:					
~~HighQuality~~			~~This element indicates whether an item is appropriate for being highlighted through the EFG web portal. The Boolean value “Yes” means that the metadata related to an item fulfil a minimum set of requirements. Hence, the respective item will be added to a data pool of high-quality items in the EFG database. If no value is given the default is “No”.~~		
~~Provider~~			~~Name of the organisation that sends the data to EFG.~~	~~1~~	
~~Aggregator~~			~~Name of the online portal or IT system that delivers the data to EFG.~~	~~1~~	
~~URI~~			~~An unambiguous URI to the resource within the EFG context. This is a record identifier for the object in the EFG system. It is created based on unique identifiers provided in the source metadata.~~	~~0+~~	
~~Country~~			~~The name of the country in which the content provider is based (or “Europe” in case of Europe-wide projects).~~	~~0/1~~	
~~Type~~			~~Here again the object type should be indicated ("Text", "Image", "Video", "Sound")~~	~~0/1~~	
~~IsShownBy~~			~~An unambiguous URL reference to the digital object on the content provider´s web site in its full best available resolution/quality.~~	~~1~~	
~~IsShownAt~~			~~An unambiguous URL reference to the digital object on the content provider´s web site in its full information context.~~	~~1~~	
~~DigitalFormat~~			~~The file format of the viewing item. (Internet Media Types (MIME); in particular cases, format parameters relative to the MIME type should be included: e.g. resolution, unusual codecs required for viewing, etc.)~~	~~1~~	

Agent

EFG 3.6, CEN 5.1

This entity refers to the concept of Agent as defined in EN 15907 (clause 5.1).
An Agent is defined as an entity that is involved in the creation, realization, curation or exploitation of a Cinematographic Work, Variant, Manifestation or Item.
Typical distinctions between agent types are Person, Corporate Body, Family and Person Group.

Element			Description	Status	Datatype
Identifier			An unambiguous reference to the resource within a given context, where possible the International Standard Audio-visual Number (ISAN), otherwise a specific number issued by a government department or other official body in an individual country, or an archive's inventory number.	1	
	Scheme		An unambiguous reference of the scope within which the identifier is unique. References may take the form of an internationally known identification scheme, an XML namespace identifier [...], or a URN namespace (e.g. URN:UUID or URN:ISAN).	1	[CVOC] Efg:IDScheme
RecordSource			A reference to the EFG content provider and the local IDs. This is a wrapper element for all elements and attributes listed hereunder.	1+	
	SourceID		The local identifier of the audiovisual creation. If this does not exist in the content provider´s database the value is "undefined".	0/1	
	Provider		The name of the archive supplying the record.	1	
		Identifier	An unambiguous reference to the archive supplying the record. This is preferably the institution's identifier, which is registered by the ISIL agency. If the institution is not registered, then the official acronym of the institution will be sufficient.	1	
		IdentificationScheme	Name of the registration scheme encoding the institution name ("ISIL code" or "EFG instit	1	
Name			A name by which the person is (or was) known.	1+	
	Type		This attribute specifies the type of the person's name, e.g. preferred name, bureaucratic name, pseudonym, spelling variant, other name.		[CVOC] Efg:PNameType, CBNameType, GnameType
	Part		Person: A distinct part of a person's name: name prefixes, forename, family name, name suffixes. Corporate Body: A distinct part of a corporate body's name.		[CVOC] Efg:PNamePart
	GeographicScope		Indicates the region in which a particular name was used by the person (e.g. ISO 3166-1 country codes, except for historical geographic entities).		Freetext
	TemporalScope		Time when the name was used by the person.		ISO8601 Range
Date			Temporal properties of the person relating to its existence and activity ("YYYY-MM-DD" or "YYYY").	0+	ISO8601 (YYYY-MM-DD, YYYY, etc)
	Type		This attribute specifies the type of temporal properties ("Date of birth", "Date of death", "Year of activity").		[CVOC] efg:PDateType, CBDateType, GDateType)

RegionOfActivity			A country from the geographic area in which the person is or was active.	0/1	[CVOC] Efg:iso3166-1CountryCS, efg:XPZ44-002HistoricCountryCS, efg:Regions_LocallyDefined
Place			The places related to the person.	0+	Freetext
	Type		This attribute specifies the type of spatial properties (e.g."Place of birth", "Place of death").		[CVOC] Efg:PPlaceType
TypeOfActivity			A film-related activity of the Agent taken from relationship records or from secondary sources.	0+	Freetext

These fields were only for Agent.Person. All other fields are for all Agent types.

Sex			The person´s sex (male, female or unknown).	1	[CVOC] Person=EAC (efg:PsexType)
ViewBiography			An unambiguous URL reference to the full filmographic entry of the person on one or more content providers' websites.	0+	Text (URL)

Event

EFG 3.7, CEN 5.2

Events are defined as primary entities that can occur within the lifecycle of an audiovisual or non-audiovisual creation. This entity was modelled by using EN 15907 (clause 5.2). The EFG metadata model distinguishes between the following event types: Publication event, Decision event, IPR registration, Award and Production event.

Element			Description	Status	Datatype
Common for all Events					
Identifier			An unambiguous reference to the resource within a given context, where possible the International Standard Audio-visual Number (ISAN), otherwise a specific number issued by a government department or other official body in an individual country, or an archive's inventory number.	1	
	Scheme		An unambiguous reference of the scope within which the identifier is unique. References may take the form of an internationally known identification scheme, an XML namespace identifier [...], or a URN namespace (e.g. URN:UUID or URN:ISAN).		
RecordSource			A reference to the EFG content provider and the local IDs. This is a wrapper element for all elements and attributes listed hereunder.	1+	
	SourceID		The local identifier of the audiovisual creation. If this does not exist in the content provider´s database the value is "undefined".	0/1	
	Provider		The name of the archive supplying the record.	1	
		Identifier	An unambiguous reference to the archive supplying the record. This is preferably the institution's identifier, which is registered by the ISIL agency. If the institution is not registered, then the official acronym of the institution will be sufficient.	1	
		IdentificationScheme	Name of the registration scheme encoding the institution name ("ISIL code" or "EFG institution code").	1	
Note			This element includes all data derived from content providers, that cannot fit into other elements.	0+	
	Source		This attribute specifies the source that produced the note element.		
Date			The date on which the event happened, or the action took place that this event describes, as defined by each event type in EFG/CEN.	1	
PublicationEvent					
Type			The type of an exhibition event (e.g. "Premiere", "Cinema release") or type of publication ("e.g. Paperback book with DVD") if not evident with the type of associated audiovisual manifestation.	1	[CVOC] Efg:EventTypes
RegionalScope			The country or large-scale geographic entity where the event took place (e.g. exhibition) or where published items have been made available (e.g. distribution area).	0/1	[CVOC] efg:Iso3166-CountryCS efg:XPZ44-002HistoricCountryCS, efg:MARC21_GeoAreasCS
Place			The name of the city of smaller geographic entity (e.g. building) where the event took place.	0/1	Freetext
EventName			The name of an event that the exhibition was a part of (e.g. name of a film festival, distribution channel of a broadcaster, etc.).	0/1	Freetext

ExhibitionOrganiser			The name of an Agent that was responsible for the exhibition (e.g. name of a cinema company, film club, broadcaster, etc.).	0+	Freetext
Publisher			If not evident from an association of Agent with Manifestation, this element can be used to state the name of the publisher.	0+	Freetext
AccessConditions			Specific restrictions for accessing the content of the manifestation (e.g. “for rental only”).	0+	[CVOC]

DecisionEvent					
Type			Types include e.g. censorship decisions, rating decisions etc.	1	[CVOC] Efg:EventTypes
RegionalScope			The country or large-scale geographic entity for which the decision is (was) valid.	0+	[CVOC] efg:Iso3166-CountryCS efg:XPZ44-002HistoricCountryCS, efg:MARC21_GeoAreasCS
Agency			The name of the agency performing the rating or censorship.	0+	Freetext
CertificateNumber			An identifier issued by the agency uniquely identifying the act of rating or censorship and associated documents such as censorship visa or rating certificates.	0/1	Freetext
Verdict			The outcome of the act of rating or censorship.	0/1	Freetext

IPRRegistration					
RegistrationAgency			Name of agency issuing the registration certificate (e.g. “UK Intellectual Property Office”). If no value is available in the source archive, the value must be set to “"not available in the source archive".	1	Freetext
RegionalScope			The geographic region for which copyright is (was) claimed (e.g. “ISO 3166-2”, “AFNOR XP Z44” for historical countries, “MARC” for extraterrestrial regions).	0+	[CVOC] efg:Iso3166-CountryCS efg:XPZ44-002HistoricCountryCS, efg:MARC21_GeoAreasCS
NameOfApplicant			Name of Agent claiming copyright in the audiovisual or non-audiovisual creation. This is the right holder's name.	0+	Freetext

Award					
NominationOnly			If the audiovisual creation was nominated but not among the winners, then this element should be set to the logical value “true” (“true” or “false”).	0/1	[CVOC] {true, false}
AwardName			The name of the award or trophy, possibly including a numeric designation (e.g. 2nd Price). If no value is available in the source archive, the value must be set to “not available in the source archive".	1	Freetext
Achievement			A phrase describing a specific achievement for which the award was given, if not for the audiovisual creation in total.	0/1	Freetext
Sponsor			Name of the Agent(s) that have sponsored the award or name of the event in the scope of which the prize was awarded.	0+	Freetext

ProductionEvent					
Type			The type of event this element instance refers to. Examples are "outdoor shooting", "indoor shooting", "post-production", etc.	0/1	[CVOC] Efg:EventTypes

Location			Any geographic name or address of the location where the event took place.	0+	Freetext
RegionalScope			The country or other large-scale geographic entity where the event took place (e.g. “ISO 3166-2”, “AFNOR XP Z44” for historical countries, “MARC” for extraterrestrial regions).	0+	[CVOC] efg:Iso3166-CountryCS efg:XPZ44-002HistoricCountryCS, efg:MARC21_GeoAreasCS

PreservationEvent					
Type			The general type of the preservation activity performed (e.g. "restoration", "reconstruction", "transfer", etc.).	1+	[CVOC]
Detail			Any information describing the process in greater detail.	0+	Freetext

LegalChange			A change of any legal status. Typical use cases would be: change of ownership, content license, IPR, etc.		
Type			The type of legal change that was done.	1	[CVOC]
Detail			Any information describing the process in greater detail.	0+	Freetext

Collection

EFG 3.8, CEN -

In the EFG context, a collection is defined by a compilation of creations (audiovisual or non-audiovisual). This entity refers to the notion of collection professionally curated in an archive and functions moreover like a “basket” whenever it makes sense to define any form of grouping.

Element			Description	Status	Datatype
Identifier			An unambiguous reference to the resource within a given context, where possible the International Standard Audio-visual Number (ISAN), otherwise a specific number issued by a government department or other official body in an individual country, or an archive’s inventory number.	1	
	Scheme		An unambiguous reference of the scope within which the identifier is unique. References may take the form of an internationally known identification scheme, an XML namespace identifier [...], or a URN namespace (e.g. URN:UUID or URN:ISAN).		
RecordSource			A reference to the EFG content provider and the local IDs. This is a wrapper element for all elements and attributes listed hereunder.	0+	
	SourceID		The local identifier of the audiovisual creation. If this does not exist in the content provider´s database the value is “undefined”.	0/1	
	Provider		The name of the archive supplying the record.	1	
		Identifier	An unambiguous reference to the archive supplying the record. This is preferably the institution's identifier, which is registered by the ISIL agency. If the institution is not registered, then the official acronym of the institution will be sufficient.	1	
		IdentificationScheme	Name of the registration scheme encoding the institution name ("ISIL code" or "EFG instit	1	
Note			This element includes all data derived from content providers, that cannot fit into other elements.	0+	
	Source		This attribute specifies the source that produced the note element.		
Type			Define the type of this collection. Examples may be: archival curated, user defined, private collection, donation, etc. In EFG, this is only used to define if the collection was defined by the content provider (archive) or on EFG level.	1	[CVOC] Efg:CollectionType
Title			A textual title of the archival collection or the EFG generated collection.	1	
Description			Textual descriptions of the collection. Textual descriptions include synopses, plot summaries, reviews, transcripts or shot lists. They can occur in more than one language and they can have statements of authorship or references to external resources.	0+	
	Language		The language of the description text.		[CVOC] efg:Iso639-1LanguageCS, efg:Iso639-2LanguageCS
	Source		Either the name of the institution or a URI identifying the source directly or via a reference system such as an on-line catalogue.		Freetext or URI
	Type		A keyword denoting the type of description		

Appendix B. Evaluation of existing Metadata Models and Schemas: Detailed Analysis

A summary of this analysis is to be found in chapter 3 of this Metadata Integration Concept.

EFG (European Film Gateway)

The EFG schema is very well structured and fills in some gaps not covered by EN 15907.

EFG introduces new terms for existing concepts – e.g. "AVCreation" instead of "Work" – which causes less confusion due to language-overlap. This is a plus when working with contributors from different fields.

Advantages

- FRBR/EN 15907 structured
- Includes schemas for film-related ("NonAV") materials. It follows an early version of the Europeana Semantic Elements specification (ESE) [EFG D2.2, p. 13]
- Includes some practical extensions to EN 15907 (thumbnail, etc.)
- Includes meta-concept of "Collections" of film- and film-related materials
- Includes concept of adding "keywords"
- Designed for multi-language support (but so is EN 15907)
- Includes concept of "Segmentation" which allows referring to parts of continuous AV content.

Possible Disadvantages

Unable to find referred sources

Sources for the classification schemas referred to in [EFG D2.2] appear to be underdocumented. See "Missing" in the "Resources" paragraph at the end of this chapter.

References to "EFG Database"

The schema definition sometimes refers "Identifier" values to point to "identifier in EFG database" [example EFG D2.2, p.22]. Wherever this is the case, a proper, yet EFG-independent data source might be used instead.

Agent:Type of Activity = Free text field

The schema definition mentions "Agent:Type of Activity" as "Free text" [EFG D2.2, p22], whereas it would be better to use a controlled vocabulary here.

Of course, the same applies to all subtypes of Agent (Corporate, Group, etc.)

Decision Event: Agency = Free text field

It is desirable to refer to an Agency-entity rather than just entering its name as free text [EFG D2.2, p.28]. Same applies to "Event Award: Sponsor" (and possibly other fields, too).

Content description

The EFG schema, like the underlying EN 15907 model is designed and used for capturing filmographic data. Defining the model with which content is described in the "Content" entity is left to organizations and projects, based on their needs and the structure of the data they seek to store and display. This data is usually based on static (non-time based) keywords (or subject terms) and cataloging descriptions or commentaries. For VHH a different data model and mechanism is required for the time-based annotation of content.

"NonAV*"

With each institution or project having a focus on certain type(s) of material, it seems clear to label related material of different types with a "Non" prefix. In the EFG schema, with focus on filmographic material, this is called "NonAV". In library catalogues it may be "NonBook", in other film archives systems it's called "NonFilm", etc.

If different data sources are to be shaped in order to be aggregated in a machine readable interoperable way, it might be better to include each "NonX" material labeled with its proper type (with the ability to use individual format schemata for each). The number of different types is very likely to be rather small/manageable in real life scenarios.

NonAVManifestation = copy only?

"This entity has the function to keep track of copies made of non-audiovisual objects (e.g. a letter gets scanned and then copied, etc.)" – EFG D2.2 (p. 15)[38]

EFG defines NonAVManifestations as always being a copy (due to its web-presence use-case). This limitation is not required when using this entity for regular collection management: Any form (original, copy, etc) of a non-AV object can be depicted here.Typos

- D2.2 (p. 11): "efg:DimenionUnitCS" (missing "s" in "Dimension")

Resources

- EFG D2.2 – Common interoperability schema for archival resources and filmographic descriptions[39]
  NOTE: This document refers to "Milestone 3.2 'Semantic & Syntactic

[38] https://efgproject.eu/downloads/D22_Common_Interoperability_Schema_V3_2_4.pdf (18.12.2019)
[39] https://efgproject.eu/downloads/D22_Common_Interoperability_Schema_V3_2_4.pdf (18.12.2019)

Interoperability Rules'", but the provided link[40] is dead, and web search does not return any results for this document.

Missing/not found yet are definitions for:

- efg:FileFormatCS, efg:DimensionUnitCS (Could be related to MPEG-7 Classification Schemes?)
- Europeana:type
- efg:PNameType, efg:PNamePart, etc (Agent)

[40] http://www.europeanfilmgateway.eu/members/members-workpackage_3.php (18.12.2019)

EDM (Europeana Data Model)

EDM was developed for and during the Europeana Project[41] in order to exchange data of different archival objects from different memory institutions to be aggregated on the Europeana website as a shared search-hub.

Apart from the classes defined in EDM, it also makes use of the following external, pre-existing concepts/schemata:

- CC (CreativeCommons)[42]
- DCAT (Data Catalog Vocabulary)[43]
- ORE (Object Reuse and Exchange)[44]
- RDFS (RDF Vocabulary Description Language)[45]
- SKOS (Simple Knowledge Organization System)[46]

The schema is well-defined, yet it also seems very tailored for the data-aggregation use for the Europeana Project.

It doesn't provide benefits over the EFG schema for the purpose of use in a filmographic database.

Advantages

There seems to be no advantage in using EDM instead of using EFG directly, unless when publishing to Europeana. However, it might be advantageous to define a common denominator for mapping EFG/EN 15907 to EDM for interoperability purposes.

Possible Disadvantages

Too "website-display-specific"

EDM seems very much tailored to the Europeana use cases. This means there are certain fields that are very specific to the intention of displaying an object on a website. This is not a bad thing in itself, but EDM contains certain relationship-types or fields that would seem misplaced or too-website-specific, compared to the intention of using it for general cataloging of works.

---

[41] https://www.europeana.eu (22.12.2019)
[42] http://creativecommons.org/ (18.12.2019)
[43] http://www.w3.org/TR/vocab-dcat/#class-dataset (18.12.2019)
[44] http://www.openarchives.org/ore/1.0/toc (18.12.2019)
[45] https://www.w3.org/2001/sw/wiki/RDFS (18.12.2019)
[46] https://www.w3.org/2001/sw/wiki/SKOS (18.12.2019)

Resources

- EDM Documentation (Europeana Pro)[47]
- EDM Definitions[48]
- EDM Factsheet[49]
- EDM Primer[50]
- EDM Mapping Guidelines[51]
- EDM XML Schema Definition (XSD)[52]

[47] https://pro.europeana.eu/page/edm-documentation (18.12.2019)
[48] https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf (18.12.2019)
[49] https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Factsheet.pdf (18.12.2019)
[50] https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf (18.12.2019)
[51] https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.4_102017.pdf (18.12.2019)
[52] http://www.europeana.eu/schemas/edm/EDM.xsd (18.12.2019)

RDA (Resource Description and Access)

"RDA is a package of data elements, guidelines, and instructions for creating library and cultural heritage resource metadata that are well-formed according to international models for user-focused linked data applications." RDA Steering Committee[53]

The question is whether or not this ruleset, schema and vocabularies can be used to describe film related material (documents, photographs, posters, etc).

The list of "Content Type" values (23 entries) currently listed in the *"RDA Registry"* (see below), gives an impression for which types it is currently being used:

- rdaco:1001 "cartographic dataset"
- rdaco:1002 "cartographic image"
- rdaco:1003 "cartographic moving image"
- rdaco:1004 "cartographic tactile image"
- rdaco:1005 "cartographic tactile three-dimensional form"
- rdaco:1006 "cartographic three-dimensional form"
- rdaco:1007 "computer dataset"
- rdaco:1008 "computer program"
- rdaco:1009 "notated movement"
- rdaco:1010 "notated music"
- rdaco:1011 "performed music"
- rdaco:1012 "sounds"
- rdaco:1013 "spoken word"
- rdaco:1014 "still image"
- rdaco:1015 "tactile image"
- rdaco:1016 "tactile notated music"
- rdaco:1017 "tactile notated movement"
- rdaco:1018 "tactile text"
- rdaco:1019 "tactile three-dimensional form"
- rdaco:1020 "text"
- rdaco:1021 "three-dimensional form"
- rdaco:1022 "three-dimensional moving image"
- rdaco:1023 "two-dimensional moving image"

[53] http://www.rda-rsc.org/ (18.12.2019)

Advantages

There is a registry for RDA vocabulary, called the "RDA Registry".[54] It contains an abundance of listings that can be used to refer to a common vocabulary set, for description of different content types.

In order to get an idea of which properties/fields are used for different types, the RDA Examples (Authority)[55] listings are a useful resource.

While this potentially makes the RDA-referred data compatible for semantic-web applications, the massive set of terms introduces a rather large (and unnecessary) overhead. We believe that data could be described as well, and leaner, in a hierarchical/structured way.

Possible Disadvantages

Flat vs hierarchical/typed structure

The existing vocabulary lists are also used to describe the available properties for each element type. Some lists are quite extensive, which also shows that its "flat" description approach can make it hard to find the proper value/property as these lists seem to be populated "in order of necessity", which makes it somewhat hard to find specific entries in an abundance of similar terms unless a search functionality it introduced to filter entries.

As an example, there are several different date types in the list of 441 Manifestation properties[56], where each one has its own ID:

- P30007 "has copyright date"
- P30008 "has date of distribution"
- P30056 "has date of publication"
- P30099 "has first chronological designation of alternative sequence"
- ...

A machine could not identify these fields as dates. When searching for the string "date", the last entry (P30099) would not be listed.

In comparison, a more hierarchical approach could be:

Property definition:

- P30001 "has date"

---

[54] http://www.rdaregistry.info/termList/AspectRatio/ (18.12.2019)
[55] https://www.rdatoolkit.org/sites/default/files/rsc_rda_complete_examples_authority_april_2016.pdf (18.12.2019)
[56] http://www.rdaregistry.info/Elements/m/ (18.12.2019)

Date types:

- P10000 "copyright"
- P10001 "distribution"
- P10002 "publication"
- P10003 "first chronological designation of alternative sequence"

Each "type" for a date would refer to a separate date-related list. This would allow re-use of similar properties among different elements or entities, as well as make the data more "readable" to machines in a semantic-web way.

This is also described in the semantic web (RDF/SKOS) article on xml.com[57], using the following as an example "The category canals is used instead of [a list of individual terms]".

Another example that depicts potential problems in implementing this approach:

- P70019 "is copyright date of"
- P61108 "is copyright date of"
- P60069 "has copyright date" (=inverse of P61108)

There is a "Domain" entry though that lists P70019 as "timespan"[58], which would give this field a category. Yet, when listing "timespan", it doesn't seem to show date fields that correspond to it.

Audiovisual Use

Terms important for audiovisual material however appear to be underrepresented or underdeveloped in regard to their use in an archival film annotation scenario.

For example, "Aspect Ratio Designation"[59], currently only lists the following three values:

- rdaar:1001 "full screen" @en "An aspect ratio designation for a moving image resource of less than 1.5:1." @en
- rdaar:1002 "wide screen" @en "An aspect ratio designation for a moving image resource of 1.5:1 or greater." @en
- rdaar:1003 "mixed aspect ratio" @en "An aspect ratio designation for a moving image resource that includes multiple aspect ratios within the same resource." @en

While this is understandable due to the origin of RDA in a library context, and this shortcoming might be address in the future, it is an obstacle to using RDA in an AV data management context.

---

[57] https://www.xml.com/pub/a/2005/06/22/skos.html (18.12.2019)
[58] http://rdaregistry.info/Elements/c/C10010 (18.12.2019)
[59] http://www.rdaregistry.info/termList/AspectRatio/ (18.12.2019)

Resources

- LoC RDA Core Elements table[60] (also shows relationship to FRBR structure)
- RDA Registry[61]
- RDA Toolkit (requires subscription)[62]
- RDA Examples (Authority)[63]

[60] https://www.loc.gov/aba/rda/pdf/core_elements.pdf (18.12.2019)
[61] http://www.rdaregistry.info/termList/AspectRatio/ (18.12.2019)
[62] https://access.rdatoolkit.org/ (18.12.2019)
[63] https://www.rdatoolkit.org/sites/default/files/rsc_rda_complete_examples_authority_april_2016.pdf (18.12.2019)

RDF (Resource Description Framework)

RDF is a standard model for data interchange on the internet. RDF has features that facilitate data merging even if the underlying schemas differ. The triples approach specifically supports the evolution of schemas over time allowing webservices to exchange data without a need to change any of their underlying code.

RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triplet"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications.[64]

It makes sense to think in these triplets and keep machine-readability in mind when structuring data and choosing formats. Yet a more current approach is to not store data directly in this format, but rather to provide interfaces that can generate RDF-compatible output on demand.

RDF, and several other data structuring concepts such as "SKOS" (Simple Knowledge Organization System)[65] or "ORE" (Object Reuse and Exchange)[66] are closely related to each other and often cross-referred and re-used wherever RDF-like thinking is implemented or desired.

Example for EN 15907

A theoretical example of mapping EN 15907 fields into RDF XML structure:

```
<RDF>
  <Description
about="http://filmstandards.org/fsc/index.php/EN_15907_Cinematographic_Work">
    <cen:title>Metropolis</cen:title>
    <cen:language usage="spoken">en</cen:language>
  </Description>
</RDF>
```

Another option with "Linked open Data" (LoD) references:

```
<RDF>
  <Description
about="http://filmstandards.org/fsc/index.php/EN_15907_Cinematographic_Work">
    <Description
about="http://filmstandards.org/fsc/index.php/EN_15907_Title">
      <Description about="https://www.imdb.com/title/tt0017136/" />
        <dc:title>Metropolis</dc:title>
      </Description>
```

[64] https://www.w3.org/RDF/ (18.12.2019)
[65] https://www.w3.org/2004/02/skos/ (18.12.2019)
[66] http://www.openarchives.org/ore/ (18.12.2019)

```
    </Description>

    <Description
about="http://filmstandards.org/fsc/index.php/EN_15907_Language">
      <cen:language usage="spoken">en</cen:language>
    </Description>
  </Description>
</RDF>
```

It might even be possible to "RDF bubblewrap" any EN 15907-structured XML, for adding LoD (similar to the above example), but without hindering 1:1 compatibility with its non-RDF version.

Resources

- RDF Page (W3C)[67]
- RDF Syntax Grammaer (W3C)[68]
- XML RDF (W3C)[69]
- SKOS Reference Schema 2004[70]
- ORE (Object Reuse and Exchange) Specifications and User Guides[71]

[67] https://www.w3.org/RDF/ (18.12.2019)
[68] https://www.w3.org/TR/rdf-syntax-grammar/ (18.12.2019)
[69] https://www.w3schools.com/XML/xml_rdf.asp (18.12.2019)
[70] https://www.w3.org/2009/08/skos-reference/skos.rdf (18.12.219)
[71] http://www.openarchives.org/ore/1.0/toc (18.12.2019)

DIF-XML (filmportal.de Schema)

Although DIF-XML (as used by DIF on filmportal.de as export format) is closely aligned with EN 15907 we decided against using it in VHH for the following reason:

- Its data and structure resemble EFG, yet EFG seems to be more suitable for international (multilingual) use and interoperability:
  - DIF-XML contains German terms for XML elements/attributes.
  - It uses Umlauts in element/attribute names.
  - Umlauts in XML elements/attributes appear in mixed forms: unescaped and escaped. (It should be either-or, yet non-ASCII characters are to be avoided in general for this purpose.)

    Examples:

    - aufgeführt
    - Auffuehrung (instead of "Aufführung")

Resources

- DIF-XML on filmstandards.org[72]

[72] http://www.filmstandards.org/schemas/de-dif/zf-fw-view-1.5/ (31.12.2019)

PBCore (Public Broadcasting Metadata Dictionary)

"PBCore is a way to organize information about audiovisual content. PBCore records can easily be shared, allowing information about media assets and collections to be exchanged between organizations and media systems." What is PBCore?[73]

PBCore is well documented and also well structured, but except for a subset (e.g for technical metadata), it doesn't seem to have advantages over EFG for the cataloging of filmographic works.

Advantages

Tech-Metadata

PBCore offers fields that no other schema reviewed here provides; for example specific fields for technical metadata apparently derived from similar fields in EBUCore. See: PBCore-RDF_data_modeling worksheet_updated.xlsx[74]

For example:

PBCore field	EBUCore field
essenceTrackType	ebucore:trackType
essenceTrackIdentifier	ebucore:trackName
essenceTrackStandard	ebucore:hasStandard
essenceTrackEncoding	ebucore:hasEncodingFormat
essenceTrackDataRate	ebucore:bitrate
essenceTrackFrameRate	ebucore:frameRate
essenceTrackPlaybackSpeed @unitsofMeasure="frames per second"	ebucore:inchesPerSecond
essenceTrackPlaybackSpeed @unitsofMeasure="frames per second"	ebucore:framesPerSecond
essenceTrackSamplingRate	ebucore:sampleRate
essenceTrackBitDepth	ebucore:bitDepth
essenceTrackFrameSize (combined with Frame Height)	ebucore:width

[73] https://pbcore.org/what-is-pbcore (18.12.2019)
[74] https://pbcore.org/assets/downloads/PBCore-RDF_data_modeling worksheet_updated.xlsx (18.12.219)

essenceTrackFrameSize (combined with Frame Width)	ebucore:height
essenceTrackAspectRatio	ebucore:aspectRatio
essenceTrackTimeStart	ebucore:start
essenceTrackDuration	ebucore:duration

Exchange with Broadcasters

PBCore could be a suitable format for exporting EFG/EN 15907 structured data when exchanging data with broadcasters.

Linked Open Data (LoD)

The "ref" attribute allows referring to an online resource for each field. This is suitable for implementing/supporting LoD.

Possible Disadvantages

Contributor = Agent

PBCore uses the term "Contributor" instead of "Agent" (which seems a more common term used in e.g. EN 15907, PREMIS, EFG, etc.). Although this is cosmetic, it might be a source of confusion when collaborating or using code/schemata from different projects.

Similar thing might be "pbCoreInstantiation" which seems to depict a "Manifestation".

The "pbcore" prefix

One cosmetic thing though might be its XML-prefix "pbCore" for all node names, which seems a bit superfluous – especially for PBCore-only documents and considering that there are XML-namespaces.

Resources

- PBCore Website[75]
- PBCore Elements (hierarchical)[76]

[75] https://pbcore.org/ (18.12.2019)
[76] https://pbcore.org/elements/element-hierarchy (18.12.2019)

DC (Dublin Core)

We propose to use DC terms for fields wherever reasonable in the VHH-EFG Schema instead of creating near duplicates, unless there is a good reason not to do so.

The fields are:

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

Examples are:

- "Work.Title" instead of Work.Label
- "Work.Language" instead of Work.Lang

EAC-CPF (Encoded Archival Context for Corporate Bodies, Persons, and Families)

While EAC-CPF appears suitable for specific use cases, we recommend restricting it to a suitable subset of fields instead of supporting the whole standard, as this might be too excessive.

Possible Disadvantages

Translated fields

Existing implementation examples (XML) show that several fields are used translated into the local language, but without reference to a common identifier for this vocabulary. This may cause problems for seamless interoperability and machine-readability.

In the XSD however, an attribute is available to define a "vocabularySource" with a corresponding URI.

Examples are:

```
German:
  <placeRole>Geburtsort</placeRole>
  <placeRole>Sterbeort</placeRole>

English:
  <placeRole>birth</placeRole>
  <placeRole>residence</placeRole>

French:
  <part localType="nom">Henrard</part>
  <part ocalType="prenom">Roger</part>
  <part localType="dates_biographiques">1900-1975</part>
```

References

- LoC Encoded Archival Description Website[77]
- EAC website of Staatsbibliothek Berlin[78]
- EAC XSD Schema 2010 (Staatsbibliothek Berlin) 2018-12-01[79]
- EAC-CPF Examples (Staatsbibliothek Berlin)[80]

[77] http://www.loc.gov/ead/ (18.12.2019)
[78] https://eac.staatsbibliothek-berlin.de/ (18.12.2019)
[79] https://eac.staatsbibliothek-berlin.de/schema/cpf.xsd (18.12.2019)
[80] https://eac.staatsbibliothek-berlin.de/examples-for-the-eac-cpf-schema-2010/ (18.12.2019)

IIIF (International Image Interoperability Framework)

EFG/CEN are designed and used for filmographic data, but a different data model and mechanism is required for describing content.

For VHH, traditional text-only time-based annotation (technically similar to closed captions) is insufficient: a structured text schema must be defined which allows content description for different scholarly demands, yet be machine readable and interoperable.

It is desirable to use existing standards, especially ones that are designed for- and used in the preservation domain and which are technically well-designed.

IIIF, originally designed for paper/image materials, was recently extended to support time-based media too (API v3.0).

The related standards and definitions as used by IIIF are documented here:

- International Image Interoperability Framework (IIIF) API V3.0[81]
- Web Annotation Data Model (W3C)[82]
- DCAT: Data Catalog Vocabulary (W3C)[83]

In IIIF, multiple so called *Manifests*[84] can be defined and used, existing in parallel to each other. This allows to annotate the same material differently, for example for different scholarly projects or use-cases, without cluttering the annotation metadata or user interface.

The *Manifests* are defined as JSON data, and online examples use pre-stored .json files. It is however more likely that in practice an API will translate between the data stored in the database, generate these JSON Manifests on demand and serve them to the annotation UI, as well as translate them back to the database when editing or adding new annotations.

IIIF specs even allow to handle granular access to different media sources, based on rights-management or similar access restrictions, which might be valuable for certain materials within the VHH projects that are not to be released as public-domain due to personal rights.

Relevant IIIF parts for VHH:

- Manifest[85]
- Range[86]

[81] https://iiif.io/api/presentation/3.0 (18.12.2019)
[82] https://www.w3.org/TR/annotation-model/ (18.12.2019)
[83] https://www.w3.org/TR/vocab-dcat/ (18.12.2019)
[84] https://iiif.io/api/presentation/3.0/#52-manifest (18.12.2019)
[85] https://iiif.io/api/presentation/3.0/#52-manifest (18.12.2019)
[86] https://iiif.io/api/presentation/3.0/#54-range (18.12.2019)

The following other features of IIIF are also of interest:

- label
- thumbnail
- behavior:thumbnail-nav

Resource Types

The following resource types are currently supported by IIIF. This list has been sorted in order of priority for VHH:

- **Video:** Moving images, with or without accompanying audio, such as might be rendered with a <video> HTML tag
- **Sound:** Auditory resources primarily intended to be heard, such as might be rendered with an <audio> HTML tag
- **Image:** Two dimensional visual resources primarily intended to be seen, such as might be rendered with an <img> HTML tag
- **Text:** Resources primarily intended to be read

These types are also supported, but not used in VHH:

- **Dataset:** Data not intended to be rendered to humans directly
- **Model:** A three (or more) dimensional model intended to be interacted with by humans

Time-based Annotation

Not only can IIIF be used to define annotations of image-regions within text documents (which it was originally designed for), but now it also allows to define time-base annotations.

Examples

- German Parliament Transcript[87]

Classifying objects

W3C Annotation Vocabulary: classifying[88]

"The motivation for when the user intends to that classify the Target as something"

[87] https://tomcrane.github.io/bbctextav/iiif/ID19409900-transcript.json (18.12.2019)
[88] https://www.w3.org/TR/annotation-vocab/#classifying (18.12.2019)

This can be used to explicitly classify objects, etc in the content. For example: audio.music, audio.noise, written words, objects (person, animal, plant, thing, etc).

Since IIIF uses W3C's Open Annotation specification as a subset, this seems to already be available.

Ranges

IIIF defines ranges in audiovisual material, as well as within text documents:

"Ranges are used to represent structure within an object beyond the default order of the Canvases in the items property of the Manifest, such as newspaper sections or articles, chapters within a book, or movements within a piece of music. Ranges can include Canvases, parts of Canvases, or other Ranges, creating a tree structure like a table of contents."

– IIIF Presentation API 3.0[89]

IIIF uses the W3C Web Annotation Data Model[90] including IIIF-specific Open/Web Annotation Extensions[91] to define positions/ranges within text documents as well as AV media.

There are different types of Selectors[92] depending on which media type is to be annotated. This provides a common method to refer to certain positions/ranges not only in audiovisual media, but also relate to pages, paragraph or quotes within text documents.

Timeline

IIIF "navDate" feature may be a suitable candidate for this.

"A date that clients may use for navigation purposes when presenting the resource to the user in a date-based user interface, such as a calendar or timeline."

- IIIF Presentation API 3.0[93]

IIIF Search API

By making use of the IIIF data model structure, it is also possible to make use of existing designs and implementations for issuing complex search queries over the content description data.

---

[89] https://iiif.io/api/presentation/3.0/ (18.12.2019)
[90] https://www.w3.org/TR/annotation-model/ (18.12.2019)
[91] https://iiif.io/api/annex/openannotation/ (18.12.2019)
[92] https://www.w3.org/TR/annotation-model/#selectors (18.12.2019)
[93] https://iiif.io/api/presentation/3.0/ (18.12.2019)

The document describing this is: IIIF Content Search API 1.0[94].

[94] https://iiif.io/api/search/1.0/ (18.12.2019)

Appendix C. Glossary of abbreviations used in this document

API	Application programming interface
CC	Creative Commons
CDH	Cinematography of the Holocaust (Cinematography des Holocaust)
CEN	European Committee for Standardization
CERCEC	Centre D'études des Mondes Russe, Caucasien Et Centre-européen
DC	Dublin Core
DCAT	Data Catalogue Vocabulary
DIF (DFF)	Deutsches Filminstitut & Filmmuseum
DFMR	Digital Film Master Repository
DIF-XML	Export format provided by DIF (DFF)
EAC-CPF	Encoded Archival Context for Corporate Bodies, Persons, and Families
EBU	European Broadcasting Union
EDM	Europeana Data Model
EFG	European Film Gateway
EN15907	European metadata standard for the description of cinematographic works
FHIs	film heritage institutions
FIAF	International Federation of Film Archives
FRBR	Functional Requirements for Bibliographic Records
HUJI	The Hebrew University of Jerusalem
IIIF	International Image Interoperability Framework
IPR	Intellectual property rights
IRI	Internationalized Resource Identifier
ISAN	International Standard Audio-visual Number
JLU	Justus-Liebig-Universität Gießen
JSON	JavaScript Object Notation
LBI	Ludwig Boltzmann Institute for Digital History
LD	Linked Data
LoD	Linked open Data
MM	Mauthausen Memorial
MMSI	Media Management and Search Infrastructure
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OFM	Österreichisches Filmmuseum
ORE	Object Reuse and Exchange
PBCore	Public Broadcasting Metadata Dictionary
PREMIS	Preservation Metadata: Implementation Strategies
RDA	Resource Description and Access

RDF	Resource Description Framework
RDFS	RDF Vocabulary Description Language
SBG	Stiftung Bayerische Gedenkstätten
SKOS	Simple Knowledge Organization System
SNG	Stiftung Niedersächsische Gedenkstätten
TUW	Technische Universität Wien
UBremen	Universität Bremen
URI	Uniform Resource Identifier
URN	Uniform Resource Name
UUID	Universally unique identifier
VHH	Visual History of the Holocaust
W3C	World Wide Web Consortium
XSD	XML Schema Definition
ZDB / DIF-ZDB	DIF metadata schema