Project no. 033104

MultiMatch

Technology-enhanced Learning and Access to Cultural Heritage
Instrument: Specific Targeted Research Project
FP6-2005-IST-5

D2.2.2 Metadata schema and mapping evaluation and revision

Start Date of Project: 01 May 2006
Duration: 30 Months

Netherlands Institute for Sound and Vision

Final version

Document Information

Deliverable number:	D2.2.2
Deliverable title:	Metadata schema and mapping – evaluation and revision
Due date of deliverable:	February 2008
Actual date of deliverable:	February 22 2008
Author(s):	Neil Ireson, Martha Larson, Johan Oomen, Viola Rondeboom, Hanneke Smulders.
Participant(s):	ISTI-CNR, Netherlands Institute for Sound and Vision, University of Sheffield, University of Amsterdam
Workpackage:	WP2
Workpackage title:	Content Selection and Preparation
Workpackage leader:	Netherlands Institute for Sound and Vision
Est. person months:	7.9
Dissemination Level:	PU
Version:	Final
Keywords	[semantic web, metadata, metadata schemas, controlled vocabularies, knowledge representation, data model, interoperability, OAI]

History of Versions

Version	Date	Status	Authors and Partners	Description/Approval Level
Breakdown	2007-11-19	Draft	Hanneke Smulders (Sound and Vision)	Proposal for the action plan for the evaluation and the planning.
0.1	2007-12-10	Draft	Neil Ireson (USFD), Hanneke Smulders, Martha Larson (UvA)	Draft prepared after the Pisa meeting, December 3rd , and after the distribution of the questionnaire among the developers for planning purposes.
0.2	2008-01-11	Outline	Hanneke Smulders	Proposed Table of Contents.
0.3	2008-01-22	New outline	Hanneke Smulders	New proposal for Table of Contents + change log information added.
0.4	2008-01-23	Extended version	Martha Larson	Adding sufficient information on the extra entities (section 3) and on the labels for classification and linking (section 4).
0.5	2008-01-25	Draft	Hanneke Smulders	Draft version indicating what the content should be in the other sections. Mapping rules nearly complete.
0.6	2008-01-30	Draft	Hanneke Smulders and Neil Ireson	First version of chapters 1-4.
1.0	2008-02-22	Final version	Neil Ireson, Martha Larson, Johan Oomen (Sound and Vision), Viola Rondeboom (Sound and Vision, Hanneke Smulders	Input from the metadata working group processed, summary, lay-out and spelling.

Abstract

The main objective of Deliverable 2.2 is to describe the knowledge representation framework to be adopted in MultiMatch. The deliverable is released in two versions: a preliminary version (D2.2.1) and a final version (D2.2.2) – this document. D2.2.2 describes the evaluation of the metadata model and proposes a slightly updated version and a new set of mapping rules to exploit the data model and its relations fully.

Executive Summary

The main objective of Deliverable 2.2 is to describe the knowledge representation framework to be adopted in MultiMatch. The deliverable is released in two versions: a preliminary version at month 10 (D2.2.1) and a final version at month 25 (D2.2.2). D2.2.1 presented a MultiMatch data model and functional metadata schema, influenced by current metadata schemes, primarily Dublin Core (DCMI Metadata Terms) and the reference model CIDOC-CRM. Three main factors were taken into account: user requirements ("top-down"), conceptual suitability ("bottom-up") and the support of interoperability. The original metadata, provided by the Cultural Heritage (CH) partners as well as by UVA, were then mapped to the MultiMatch schema. This second and final version evaluates, revises and further extends this schema.

The bottom-up factor – the need to represent the concepts that are present in the data – appeared the main factor for revision, since:

- The original representation proved not to be able to express all the concepts that will be present in the PT2 data. See D.2.3.1, section 2.2.4 for a description of the extra PT2 content.
- PT2 content will be semantically enriched to produce automatically generated metadata. This activitiy is done in WP4.

Evaluation results

Considering the expressiveness of the MultiMatch schema, the schema was able to express all the concepts that are present in the data of PT1. However, the complexity of the schema slows down the system and this needs to be considered. As for the mapping, all content information could be transformed into the MultiMatch schema. The study showed only few minor mapping errors, which were not caused by the schema but by lines missing in the transformation script and the deflation into Dublin Core. Finally, some metadata elements remained unpopulated because semantic enrichment was planned for PT2. In order to represent new (dynamic) data and the semantic annotation process for PT2, the schema, as expected, needs extending. These extensions were decided upon during a technical meeting and several Skype discussions.

Also, the CH partners decided to provide training data to the WP4 partners to support the semantic annotation process.

Metadata schema for PT2

Key changes in the metadata schema for PT2 are:

1. New elements added, specifically two element groups to represent dynamic data (Feed and FeedItem) and some extra new elements that were missed.
2. Several attibutes are added to several elements in order to specify them more fully or to indicate relevant knowledge about (the values of) those metadata elements.
3. The min/max occurrences for many elements have been changed to correctly represent the element concept. In some cases where multiple values are present it is possible to indicate the "preferred" value.
4. Many elements have attached the following attributes: source, classifier, reliability and manual.
5. The Catalogue entity is removed from the data model.
6. Better representation of elements related to multilinguality.
7. At the semantic level the new schema is more consistent in the application of element names.

After processing all the proposed revisions, the conclusion is that the new schema is able to express the functionality of PT2 as well as all the concepts that are present in the data of PT2 and the major concepts that are relevant in the CH domain. The schema meets the direct needs of the PT2 developers and it is adaptable in MILOS. As the CH domain is a complex domain, it remains a complex schema which probably provides here and there more options then the PT2 content and functionality will demand, to keep options open. During the field trials the added value of the new schema and the new mapping will be tested and documented.

Mapping rules for PT2

The thorough study of the mapping of the proprietary schemas of the CH partners to the new MultiMatch metadata schema resulted in a new set mapping rules for PT2. These new mapping rules:

- Correct the minor mapping errors of the PT1 transformation process.
- Avoid information loss via the addition of specified grammar in several metadata elements in the XML documents that will be provided for the PT2 transformation process.
- Populate 43 elements in PT2 newly or in an enhanced way.

Table of Contents

1 Introduction

The main objective of Deliverable 2.2 is to describe the knowledge representation framework to be adopted in MultiMatch. The deliverable is released in two versions: a preliminary version at month 10 (D2.2.1) and this final version at month 25 (D2.2.2).

1.1 Context of this deliverable

D2.2.1 presented a MultiMatch data model and functional metadata schema. Both to guide the development and enable interoperability the model was influenced by current metadata schemes, primarily Dublin Core (DCMI Metadata Terms) and the reference model CIDOC-CRM. The MultiMatch metadata schema was mapped to Dublin Core Metadata Element Set (deflating the schema for exchange purposes), DCMI Terms and to the reference model CIDOC-CRM, which is believed to become increasingly important for knowledge representation and interoperability in the Cultural Heritage domain.

The original metadata, provided by the Cultural Heritage (CH) partners as well as by UVA, were mapped to the MultiMatch schema. These metadata sets only partially populates the MultiMatch schema as the intention of the project is to augment the metadata using automatic semantic annotation processes.

Full documentation on the metadata schema for PT1 is available at:
http://www.dcs.shef.ac.uk/~nsi/xsd1.1/default.html

The intention was to revise and further extend this schema as work progresses and feedback is available together with the experimental results of the first prototype (PT1). For example, the metadata schema is augmented in order to be able to monitor a new functionality in the second prototype (PT2). Subject labels, and possibly other elements, will be populated as a result of automatic semantic analysis and classification. Section 3.3 explains this new functionality and its consequences for the schema.

This deliverable presents the methodology of augmenting the metadata schema (section 3), as well as the results of this evaluation (section 2), namely:

- A new version of the metadata schema on a conceptual level (section 3 and Appendix 1).
- A new version of the mapping rules for the CH partners (section 4).

1.2 Methodology

To start with, version 1.1 is evaluated on its expressiveness and the feedback is responded to.

The original representation was the result of three main factors (see D.2.2.1, section 2), namely:

- The need to meet the specification of the user requirements.
- The need to represent the concepts that are present in the data.
- The need for interoperability.

Evaluating the original representation, all of these factors were taken into account anew. However, the bottom-up factor, appeared the main factor for revision (of the schema as well as of the mapping), as there were no changes in the user requirements or in the need for interoperability.

For PT2 extra content is provided. See D.2.3.1, section 2.2.4 for a description of the additional content. The original representation proved not to be able to express all the concepts that will be present in the PT2 data. Especially the newly added set of cultural heritage news feeds and RSS identified from authoritative cultural heritage sites and cultural sections of newspapers asked for a revision of the data model and schema.

There is also a need to extend the expressiveness of the schema due to PT2 exploiting semantic annotation processes. The extensions relate both to representing procedural information and the discovery of information which although not in CH partner's data, or derived from CIDOC CRM, is available from other domain resources (such as ULAN and TGN).

The changes to the CH partner's data and metadata schema, and evaluation of the PT1 mapping process has necessitated a thorough study of the mapping of the proprietary schemas of the content providers to the MultiMatch metadata schema, to ensure the success of this process in PT2.

To conclude, the methodology for augmenting the metadata schema was driven by the need to include new data items or concepts (feeds) and by two new factors:

- The need to make sure that the revision had a minimal impact on the other system components, and that the system overall can be agnostic to the changes.
- The introduction of the semantic enrichment that produces automatically generated metadata.

Details on the evaluation and its results are reported in the following section.

2 Evaluating the use of the metadata schema and the mapping of the metadata

2.1 Performing the evaluation

The evaluation was performed in multiple steps:

- Analysis of the data to determine the degree to which mapping was successful.
- A thorough study of the mapping of the proprietary schemas of the content providers to the MultiMatch metadata schema of PT1 was performed to find out if the provided content asked for a revision of the original representation.[1]
- Evaluation of the system by MultiMatch partners.
- Study commentd and feedback in JIRA, the Issue Tracking system used in MultiMatch.
- A questionnaire for developers was constructed by UvA and Sound and Vision which listed all metadata issues that might be of importance for the development of PT2, sometimes with worked out proposals. This working document ('MultiMatch - Questionnaire for developers on metadata and PT2 – 20071205') structured the discussion in Sheffield (December 10th) and gave the developers the possibility to provide input in the metadata schema revision process. The outcome of this discussion is processed in the several deliverables of WP 4 as well as in the following sections of this deliverable.
- Discussion in terms of Semantic Annotation Processes (Skype, Pisa (December 3rd), Sheffield (December 10th). A developers meeting in Sheffield (December 10th) discussed the enhancements required for the newly added dynamic data (entities Feed and FeedItem, see section 3.2) and for the semantic enrichment, a new functionality in PT2 (see section 3.3 and 4).

2.2 Results of evaluation

Expressiveness of MultiMatch schema

Initially, version 1.1 of the metadata schema is evaluated in terms of its expressiveness, which showed that the model was able to express all the concepts that are present in the data of PT1, as well as the major concepts that are relevant in the CH domain.

The feedback from the developers and the content providers, a.o. via the JIRA issue tracking system, was taken into account. None of the reported problems were related to the schema itself or its expressiveness. However there was a comment that the complexity of the schema slows the system due to the time taken the marshall/unmarshall the schema from/to XML. This comment emphasises that care must be taken when implementing the use of the schema, to ensure that only the necessary information is communicated, making the complexity of the XML a function of the query rather than the expressiveness of the data model.

- [1] The actual transformation script for PT1 is: MMtransformer1.1.xslt and available from the docstore.

Mapping of metadata

Considering the mapping process, all information from the content providers could be transformed into the MultiMatch knowledge representation as far as the metadata schema was concerned.

The study of the mapping of the proprietary schemas of the content providers to the MultiMatch metadata schema of PT1 showed only few minor mapping errors.
The following lines were accidentally missing in the transformation script:

- The line to ingest the value of <luogo_scatto> from the Alinari metadata into Creation_Location.
- The line to ingest the value of <title_alternative> from the Sound and Vision metadata after the value of <title> into Creation_Title.
- The line to ingest the value of <language> from the Sound and Vision metadata into Creation_Language.

Analysis of the data showed that of all the descriptive elements these four were the most heavily populated: Creation_Title; Creation_Subject; Actor_Name and Creation_Location. These search options were therefore provided in the user interface of PT1 as fielded search options in ‘advanced search’.

The main reasons why metadata elements were not populated in PT1 are:

- The values are not available in the catalogues of the CH partners.
  Noticeable examples:
  - In the Alinari metadata not all of the photographs have a value for the Creation_Description element. This is understandable as in the Alinari metadata the Title contains a short description of what can be seen in the photo, because photos most of the time do not have an official title of their own.
  - The metadata of all the content providers focusses on the creation. As a consequence values for the Actor.Creator are often not available, especially the values for Actor_Subject and Actor_Description, but occassionaly the values for Actor_DateOfBirth as well.
- The values are available in the catalogues of the CH partners, but due to the deflation into Dublin Core, this information got lost during the transformation process.
  Notable examples:
  - In the Alinari and Cervantes metadata the values for Actor_DateOfBirth and Actor_DateOfDeath are sometimes available in the same element as the creator name. This caused loss of information during the ingestion, as the value of Creator_Name was ingested as a whole and the two date elements concerned stayed empty.
  - In the Alinari and Sound and Vision metadata several types of subject indexing are deflated into only one MultiMatch element, namely Creation_Subject. This way the MultiMatch system cannot easily distinguish between keywords that are person names, geographic names or names for period/style.
- The semantic enrichment was only planned for PT2, so at a later stage in the project.

The transformation script for PT1 is also able to process the 6 The European Library (TEL) collections that are transformed into MultiMatch Creation records. This conversion is described in D2.3.1, section 3, page 17-19. A new export of the TEL metadata will be provided in March 2008. Where necessary the mapping rules will be adapted to the new schema. Mapping of date from TEL and OAI data to the schema will be done in the context of Task 2.3.

The conclusion of this part of the evaluation is that the mapping errors were not caused by the schema. The mapping process involves exporting the metadata into Dublin Core. The thus converted data was ingested into MultiMatch. There appeared some loss of information due to the ingestion of the rich catalogue descriptions via a conversion to DC, which is a deflation in itself. It was decided to avoid information loss by adding specified grammar into the values of several DC elements. Especially into <dc:subject> and <dc:creator> to be able to make use of the several types of subject indexing and of the available birth and death dates of Creators. Semantic enrichment of Actor.Creator records in MultiMatch will further be processed via the use of ULAN and the extra biographical information that Sound and Vision has provided on 500 Dutch broadcast celebrities.

The mapping rules for PT2 were changed to:

- Correct the minor mapping errors.
- Avoid information loss.
- Populate more metadata elements than in PT1.

See section 4 for an overview of the corrections to the mapping rules.

Extensions need to represent semantic annotation process and new data

A number of discussions, both over Skype and in a Pisa (December 3rd) and a Sheffield (December 10th) meeting, took place to discuss the next step for the augmentation of the knowledge representation.

In the Pisa discussion it was decided that the Subject element needs a language property for the subject terms that are not person names. Person names in Creation_Subject should point to a ULAN record in which all translations of the name should be present.

Furthermore, the CH partners decided to provide the following training data to the WP4 partners to support the semantic annotation process:

- For training Creator: type/genre (UvA):
  Sound and Vision will provide their controlled list of actor roles.
- For training Subject: periods/styles (UvA) :
  Alinari will provide their controlled list ‘periodo_stile’.
  Cervantes will provide UDC subject codes with Spanish and English explanations.
  Sound and Vision will provide their complete thesaurus in SKOS format. Terms to indicate periods or styles are only in there if there are TV/radio programs about it in the collection. These terms will have to be derived from the controlled list with subjects and from the controlled list with so called “own names” (not being Person names or Locations).
  Alinari and Sound and Vision annotate their data with the PT2 themes provided by Alinari.
- For training Subject: Place/Time (USFD):
  Alinari will provide some associated locations to which the image is related.
  Cervantes will provide birth/death dates.
  Sound and Vision will provide a separate collection with biographies about 500 broadcast celebrities from values might be derived for the following Actor elements: SourceIdentifier; Nationality; Subject; Description; Gender; Date of birth; Date of death; Place of birth and Place of death..

During the Sheffield meeting and several Skype discussions the extensions to the metadata schema, needed to represent the semantic annotation process and the new data (namely dynamic data) were decided upon. See sections 3.2, 3.3 and 4 for further details on these extensions.

Conclusion

The PT1 metadata schema was able to express all the concepts that are present in the data of PT1, as well as the major concepts that are relevant in the CH domain. As expected the schema needed extending for the data and the functionality of PT2.

The CH partner metadata mapping and the semantic extraction needs to be carefully performed to preserve information.

3 Metadata Schema for PT2

This Section gives the reasoning behind the changes and additions to the PT1 metadata schema. See Appendix 1 for the complete listing of the metadata elements of the PT2 metadata schema per (sub) entity. Full documentation on the PT2 metadata schema is available at http://www.dcs.shef.ac.uk/~nsi/xsd2.0alpha/default.html.[2]

3.1 Main revisions

Key changes in the metadata schema for PT2 are:

1. New elements are added, specifically two groups of elements to represent dynamic data (Feed and FeedItem) and some extra new elements that were missed during the first round. See section 3.2 and 3.4.1.
2. Several attibutes are added to several elements in order to specify them more fully or to indicate relevant knowledge about (the values of) those metadata elements. The added attributes are grouped into five basic types. See further section 3.4.2.
3. The min/max occurrences for many elements have been changed to correctly represent the element concept. In some cases where multiple values are present it is possible to indicate the "preferred" value, for example the value to be displayed in the UI. See section 3.4.1.
4. Many elements have attached attributes which allow the definition of the source of the element value (Source), the process to derive the value (Classifier), the perceived reliability of the value (Reliability) and whether the value was manually (or automatically) derived (Manual). See section 3.3.
5. The entity for Exchibition Catalogues is removed from the data model. See Figure 1, section 3.2.
6. Better representation of elements related to multilinguality.
   - In PT2 more elements than in PT1 are of the LinguisticType. In PT1 already all Description elements and all Title elements had a language property. For PT2 also all Subject elements are typed linguistic.
   - In the MultiMatch system, some elements have translations that are human generated and supplied by the content providers. In other cases, translations are generated automatically. It is necessary to keep these two apart. Therefor the Translated from attribute was added to the Linguistic Type.
7. At the semantic level the new schema is more consistent in the application of element names. For example, the Identifier element has the same meaning in all its appearances. Also, all elements where language can play a role now have language properties.

All changes are further explained in the sections below.

Conclusion of the discussion on the metadata schema for PT1 was to maintain it, even though not every entity is in use or will be in use. It was decided to keep the options open when they were likely to be needed sometime in the future of providing digital CH information.

An exception was made with removing the Exhibition Catalogue entity from the schema. This entity, representing (printed) catalogues produced by curators and published alongside exhibitions, is only sporadically used and considered too redundant to maintain it in the schema.

However, it appeared necessary to extend the metadata schema, as for PT2 feed-based Web data will be provided. This data will include conventional text-based feeds such as news feeds from museum websites and also podcasts.

[2] Note that the documentation for Version 1.1 is available at http://www.dcs.shef.ac.uk/~nsi/xsd1.1/default.html.

1.

3.2 Changes to handle Dynamic Data

The datamodel is extended with two extra entities to handle the dynamic data content, namely Feed and FeedItem.

Feed (extra type of Collection)

This entity is a derivation of Website, so the Feed elements contain all Website elements plus some extra FeedInformation elements (Type, Status, Engine and Query). It was decided to make Feed a derivative of Collection because Feeds consist of individual feed items and each feed item (for example, an individual news item, or an individual podcast) is a document in its own right. It was not possible to represent a Feed using the WebSite element, since a Feed is dynamic, meaning the new information is added to the feed over time.

A Feed entity must therefore carry information about the Status of its content, i.e., whether or not the content has a short shelf life (e.g., news) or can be maintained indefinitely in the system (e.g., documentary podcasts). Making Feed a separate element from WebSite also has the beneficial side effect that the UI can chose to display a Feed in a different way than it displays a WebSite. The Type element contains a label that indicates whether the feed is a podcast, vodcast or conventional feed.

The other FeedInformation elements (Engine and Query) are used to retain further information about the origin of the feed. Some feeds result from a query being submitted to a search engine. The original query contains important information about the subject of the feed, which is retained for the case that it may be useful in the workflow.

Further information about how feeds were gathered for PT2 and mapped to the MultiMatch metadata schema will be available in D4.3 "Focused Crawling component and documentation."

FeedItem (item of a Feed)

This entity is a derivation of Web page, in the sense that it has a Feed Homepage. Further FeedItem elements contain the FeedInformation elements plus a link to a possible enclosure. The FeedInformation elements of a Feed items are copies of the FeedInformation elements of the parent feed. They are copied to the feed item in order for the system to be able to easily treat either the parent feed or the individual feed items as the basic unit of retrieval.

FeedItem cannot be conflated with webpage it requires this extra information. Moreover, separating the two also gives the UI the choice of treating them as separate types of results, as was the case for Website vs. Feed. Further information about how feed items were gathered for PT2 and mapped to the MultiMatch metadata schema will be available in D4.3 "Focused Crawling component and documentation."

See Figure 1 for a basic graphical representation of the new datamodel.

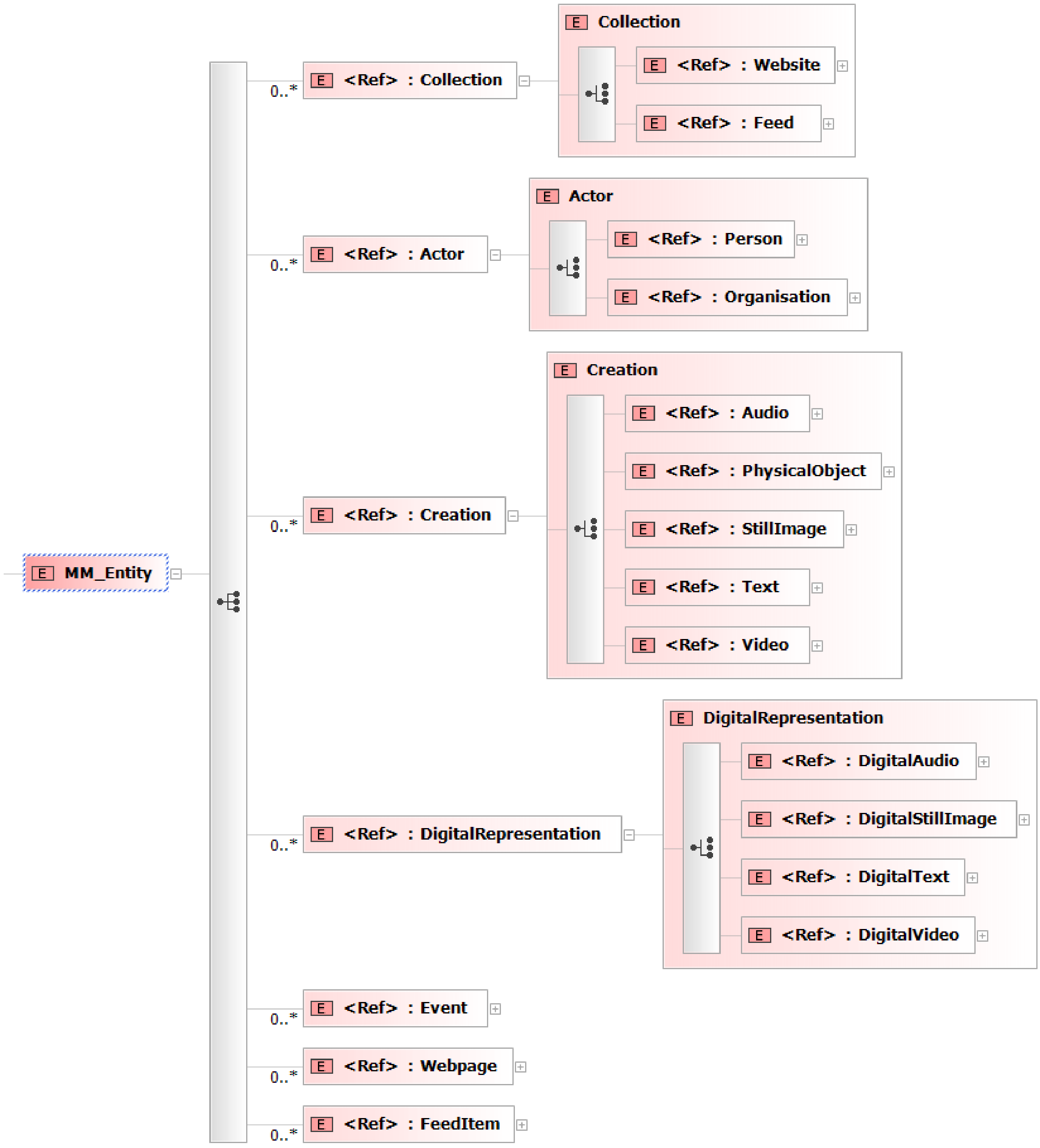


Figure 1 Data model for PT2 in a basic graphical representation.

3.3 Changes to handle Semantic Annotation Process

In PT2 automatic semantic analysis and classification create subject labels (or class labels) for MultiMatch entities. As noted in D2.1, there are numerious vocabularies in use at CH institutions throughout the world. For example, within the consortium, Alinari and Sound and Vision have their own proprietary vocabularies and BVMC uses the UDC standard[3].

Creating a mapping between controlled vocabularies (and thus using one controlled vocabulary) is outside the scope of the project, seeking semantic similarities between values are used as an alternative approach. This work is mainly undertaken in Task 4.6 "Semantic analysis and classification" and the details of these algorithms will be the subject of D4.4. Additionally, documents in the MultiMatch system will be grouped using knowledge discovery approaches, the domain of Task 4.7 "Cross media / cross collection knowledge discovery", cf., D4.5.

The MultiMatch metadata standard for PT2 must necessarily have the capacity to represent subject labels and inter-entity links. These considerations constitute one of the three major factors influencing the PT2 metadata schema development work as listed in D2.3.1 "Selected and harmonized content: metadata", where it is described as "the need to represent the concepts that are present in the data: the metadata is expected to consider the ability to adequately represent the concepts that are extractable from the data itself" (p. 18). For the PT2 metadata schema, it was decided to concentrate the representation of the output of semantic enrichment processes in the Subject element. Subject labels output by classifiers are stored in the Subject element. The connection between two linked entities is encoded by assigning both the same label in the Subject element.

Because the knowledge discovery task is placing heavy emphasis on linking resources to web pages, the possibility was left open to represent a webpage related to a specific entity directly in the RelatedWebpage or RelatedWebsite element of that entity.

Although the chosen approach for PT2, namely concentrating the representation of automatically generated semantic enrichment in the Subject element, is simple and straightforward, designing the structure and the syntax of the Subject element proved to be more challenging. In particular, there were three aspects that needed to be taken into consideration.

- First, the metadata schema needed to be revised in order to be able to distinct between the manually generated values and the automatically generated ones.
- Second, the Subject element had to be suitable for search. In particular, the Subject element needed to be compatible with the system and with the needs of the user interface.
- Third, the Subject element needed to be suitable to support the workflow for the training and the testing of the automatic enrichment methods. Each of these aspects will be discussed in turn.

3.3.1 The ManualAutoType

To be able to monitor the semantic annotation process the ManualAutoType is introduced. This is a set of four attributes to indicate if a value is generated manually or automatically. These attributes are associated to the Subject element and to all other metadata elements that will possibly be populated both via ingestion of the content providers data (i.e. manually generated metadata) and via semantic enrichment.

The ManualAutoType attributes are defined as follows:

- Manual: Indicates whether the value was derived from a manual process or not (i.e. some automatic process).
- Reliability: A numerical value for the confidence that the resource contains this value, between [0,1]. Generally (although not necessarily) manually derived values will have value of 1.0.

[3] http://www.udcc.org/scheme.htm

- Source: Identifies the source from which the value is taken. For example for Automatically derived values this may be a URL of the Wikipedia Article, or for Manually derived values the controlled vocabulary (MMTop40, ULAN, PT2Themes, TateGroupMovement...).
- Classifier: The name of the classifier/algorithm (automatic process, organisation or person) that populated the field concerned..

The following metadata elements in the PT2 schema are of the ManualAutoType:

Affliation; Copyright; DateOfBirth; DateOfDeath; DepictedActor; DepictedCreation; Description; DigitalRepresenation; Format; Gender; LicenseCondition; Publisher; RelatedActor; RelatedCreation; RelatedWebpage; RelatedWebsite; RightsHolder; SourceIdentifier; Subject; Tags; Title; Transcription and Type.

3.3.2 The Subject element and search requirements

The fact that subject information is consolidated in the Subject element facilitates search. One particular effect is to make the PT2 metadata standard maximally compatible with the delivery formats of content providers. Keywords and subject fields from content providers can all be mapped to the Subject element. At the same time, an extensive list of attributes ensures that no information concerning the subject label is lost. For example, information about the original controlled vocabulary of the content provider from which the subject label is drawn is carefully retained. The attributes are illustrated in the example of two subject elements in Diagram 1.

```
<subject category="creator" manual="yes" reliability="1.0"
language="spa" source="MMTop40" classifier="gold" translated-
from="eng">
Cervantes
</subject>

<subject category="creator" manual="no" reliability="0.2"
language="eng" source="BGThesaurus" classifier="SVM”>
Harry Mulisch
</subject>
```

Diagram 1: Two examples of subject elements compliant with the MultiMatch PT2 metadata schema

It is preferable to encode additional information about the subject in attributes rather than defining several different elements to contain subject-related information, since if there is only a single subject element, the system need index only a single field and the UI need only search a single field.

At the same time, it was deemed necessary to give the user interface information about the category of the subject label. The PT2 categories are: Creator, Period/Style, Place, Time, Artwork and Exhibition. The categories correspond to the major directions in which research in semantic analysis and classification is being undertaken by Task 4.6.

Although at the time of defining the metadata schema, it is not yet clear how/if the UI will use the category attribute, the metadata schema incorporates the flexibility necessary for representing this information. The same comment is also appropriate for the attribute Source, which encodes the list from which the class label originates. These lists include: BGThesaurus, PT2Themes and MMTop40 as well as ULAN. Giving the Subject element all of the attributes associated with linguistic types makes it possible to support multilingual retrieval on subjects. Additionally, it was important to be able to separate subject labels that are manual generated from subject labels that are automatically generated. This function is performed by the attribute "manual".

Labels generated by automatic classifiers will be characterized by certain, possibly quite high, levels of noise. It is a subject of research, how the system should deal with noise and how disruptive it is for users if evidence of noise is apparent in their interactions with the system. For this reason, high quality labels are carefully kept apart from automatically generated ones. Also for this reason, the attribute "reliability" encodes a score that reflects the confidence with which the classification algorithm generated the subject label. This attribute makes it possible for the UI to deal only with high quality automatically generated labels.

Finally, Subject element design is compatible with the system because the subject label is the content of the element and not the value of an attribute. In order not to lose any information about the subject, an extensive list of attributes was defined.

3.3.3 The Subject element and workflow requirements

The data in the MultiMatch PT2 system is not only used for search, it is also used for the purpose of research and development of automatic enrichment algorithms (T4.6 and T4.7). The MultiMatch metadata standard must also fulfil the requirements of the workflow needed for training and testing new classifiers as new data is added to the system. Certain attributes described above as necessary for search also play a role in supporting workflow. These attributes are source, category and reliability, which are used to keep track of the label list the classifier is drawing its subject labels from and the confidence score that it outputs. The attribute "classifier" was introduced exclusively for workflow purposes.

This attribute encodes the classification algorithm with which the label was generated. If the document has been processed for use as training or test data, the classifier attribute has the value "gold" indicating that the subject label represents the gold standard. Note that the attribute "manual" does not directly encode whether the subject label represents the gold standard. Gold standard labels are manually generated, but manually generated labels are not necessarily gold standard. In this way the system allows for less-than-perfect manual labels. Gold standard labels are checked by hand before their classifier attribute is set to "gold."

3.3.4 Further changes to handle Semantic Annotation Process

The semantic annotation process includes furthermore the enrichment of:

- Actor.Creator records with data from the Getty Union List of Artist Names (ULAN; see D2.2.1 section 6.5).
- Location data in MultiMatch with data from the Getty Thesaurus of Geographic Names (TGN; see D2.2.1 section 6.6).

ULAN will also provide a controlled vocabulary for Creator person names. Because the MultiMatch project disposes of ULAN in a semantic web version, the intention is not to add creator names in Creation records as a word, but as a link to the Actor concerned. Linking up the Actor.Creators from the CH partner collections with ULAN, at least for the top 40 artists, will provide a concrete example of how MultiMatch is intended.

In order to make full use of ULAN the metadata schema is revised as follows:

- The possibility of multiple names for an Actor was added, together with an attribute to indicate which value is the preferred one.
- The element Nationality is added to Actor.
- The element Gender is added to Actor.Person.

In order to make full use of TGN latitude and longtitude are added as an extra attribute of the Location element. See also LocationType in section 3.4.2

.

3.4 Other Changes

3.4.1 Changes to Elements

This section lists the changes in the metadata schema per sub(entity). The changes can concern:

- New element.
- New element name.
- New semantic meaning to existing element.
- Existing semantic meaning added to other existing element (Moved to …).
- Extended the min/max occurrences to correctly represent the element concept (Occurrences). This last type of change occurs most frequent.

Website Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Homepage	New element name.			(1)	(1)
Type	Occurrences.	(1)	(1)	0	unbounded
Subject	New element.			0	unbounded
Links	New element.			0	unbounded
Content	New element.			0	1

Webpage Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Type	Occurrences.	(1)	(1)	0	unbounded
URI	New element name.			(1)	(1)
Author	New element.			0	1
Subject	Occurrences.	(0)	(1)	0	unbounded
Related Website	New element.			0	unbounded
Related Webpage	New element.			0	Unbounded
Content	New element.			0	1

Actor.Person Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Type	New semantic meaning to existing element. Occurrences.	(1)	(1)	0	unbounded
Main Role	Moved to Type.	0	1		
Name	Occurrences.	(1)	(1)	0	unbounded

Source	Occurrences.	(1)	(1)	0	unbounded
SourceIdentifier	New element.			0	unbounded
Nationality	New element.			0	unbounded
Subject	New element.			0	unbounded
Gender	New element.			0	1
DateOfBirth	Occurrences.	(1)	(1)	0	1
PlaceOfBirth	Occurrences.	(1)	(1)	0	1
DateOfDeath	Occurrences.	(1)	(1)	0	1
PlaceOfDeath	Occurrences.	(1)	(1)	0	1

Actor.Organisation Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Type	New semantic meaning to existing element. Occurrences.	(1)	(1)	0	Unbounded
Main Role	Moved to Type.	0	1		
Name	Occurrences.	(1)	(1)	0	Unbounded
Source	Occurrences.	(1)	(1)	0	Unbounded
SourceIdentifier	New element.			0	Unbounded
Nationality	New element.			0	Unbounded
Subject	New element.			0	Unbounded

Creation.Audio Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Type	Occurrences.	(1)	(1)	0	Unbounded
Source	Occurrences.	(1)	(1)	0	Unbounded
SourceIdentifier	Occurrences.	0	1	0	Unbounded
Subject	Occurrences.	0	1	0	Unbounded
Transcription	New element name. New semantic meaning to existing element. Occurrences.	(1)	(1)	0	Unbounded

Creation.PhysicalObject Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs

Type	Occurrences.	(1)	(1)	0	Unbounded
Source	Occurrences.	(1)	(1)	0	Unbounded
SourceIdentifier	Occurrences.	0	1	0	Unbounded
Subject	Occurrences.	0	1	0	Unbounded

Creation.StillImage Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Type	Occurrences.	(1)	(1)	0	Unbounded
Source	Occurrences.	(1)	(1)	0	Unbounded
SourceIdentifier	Occurrences.	0	1	0	Unbounded
Subject	Occurrences.	0	1	0	Unbounded

Creation.Text Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Type	Occurrences.	(1)	(1)	0	Unbounded
Source	Occurrences.	(1)	(1)	0	Unbounded
SourceIdentifier	Occurrences.	0	1	0	Unbounded
Subject	Occurrences.	0	1	0	Unbounded

Creation.Video Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Type	Occurrences.	(1)	(1)	0	Unbounded
Source	Occurrences.	(1)	(1)	0	Unbounded
SourceIdentifier	Occurrences.	0	1	0	Unbounded
Subject	Occurrences.	0	1	0	Unbounded
Transcription	New element name. New semantic meaning to existing element. Occurrences.	(1)	(1)	0	Unbounded

DigitalAudio Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Type	Occurrences.	(1)	(1)	0	Unbounded
Streamed	New element.			0	1

AudioChannels	New element.			0	1
SamplingRate	New element.			0	1
BitRate	New element.			0	1
FrameRate	New element.			0	1

DigitalStillImage Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Type	Occurrences.	(1)	(1)	0	Unbounded
Thumbnail	New element.			0	1
Height	New element.			0	1
Width	New element.			0	1

DigitalText Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Type	Occurrences.	(1)	(1)	0	Unbounded

DigitalVideo Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Type	Occurrences.	(1)	(1)	0	Unbounded
Thumbnail	New element.			0	1
Height	New element.			0	1
Width	New element.			0	1
Streamed	New element.			0	1
AudioChannels	New element.			0	1
SamplingRate	New element.			0	1
BitRate	New element.			0	1
FrameRate	New element.			0	1

Event Elements

Name	Change	PT1 Min Occurs	PT1 Max Occurs	PT2 Min Occurs	PT2 Max Occurs
Language	New element.			1	1
Subject	New element.			0	unbounded
Related Catalogue	Removed element.				

3.4.2 Changes to Attributes

Section 3.3 describes the attributes that are added to elements to be able to handle the semantic annotation process, the ManualAutoType. In the PT2 Metadata Schema several other attibutes are added to several elements in order to specify them more fully or to indicate relevant knowledge about (the values of) that metadata element. The detailed reasons for adding these attributes are listed below. The usage of these different types of attributes is illustrated in Appendix 1.

LinguisticType

In the PT2 metadata schema more elements are of the LinguisticType. In the PT1 metadata schema already all Title elements and all Description elements had a Language attribute (The Language of the element text.) and a Translated from attribute (If applicable the original language from which this text was translated.). For PT2 also all Subject and Tags elements are of the LinguisticType.

LocationType

In order to use introduce controlled vocabulary for geographic names (applying TGN) and to be able to apply a form of geographic clustering and/or presentation in the user interface, it was necessary to add the latitude and longtitude to names of places. The following four elements in the PT2 metadata schema that are of the LocationType: ArchiveLocation; Location; PlaceOfBirth and PlaceOfDeath. This means that these four elements have a Name attribute (The name of the geographic location concerned.) and a LatLong attribute to store the latitude and longtitude of the geographic location concerned.

PreferredType

As some elements can contain more than one value, it can be necessary e.g. for the presentation in the user interface, to indicate which value is the preferred value. This Boolean Preferred attribute can be used for example the indicate which of the available values is the preferred representation of a Person's name or Creation's title. The following eight elements in the PT2 metadata schema that are of the PreferredType: Depiction; Description; Name; Nationality; RelatedWebpage; RelatedWebsite; Title and Type.

Attributes associated with Subject elements

In the PT2 metadata schema the Subject elements appeared to need specific attributes (see also section 3.3).

As described above, they needed the attributes of LinguisticType and of the ManualAutoType. To these (sets of) attributes one extra attribute is added, namely Category. This is done to support the semantic enrichment as well as the presentation of search results in the user interface.

The Category attribute indicates the category to which the subject concerned belongs (creator, period_style, place_time, artwork_name or exhibition).

To conclude, in the PT2 Metadata Schema there are five basic complex types:

1. LinguisticType.
2. LocationType.
3. ManualAutoType (described in section 3.3).
4. PreferredType.
5. SubjectType. Includes all attributes from LinguisticType and ManualAutoType plus attribute Category.

A number of combined types are also defined: ManualAutoPreferredType, ManualAutoPreferredLinguisticType, ManualAutoLinguisticType, PreferredLinguisticType.

3.5 Influence of changes on MultiMatch System

The changes in the PT2 metadata schema, summarized in section 3.1, imply that the PT1 Metadata is not valid with respect to the PT2 metadata schema (see section 3.4.1 for a detailed overview of the element changes). However the differences are mainly superficial. The influence of the changes are listed in the table below.

Change	Concerning the following elements	Consequences
Some of the elements have been removed.	Event_RelatedCatalogue DigitalRepresentation_Low-level features	All of these elements are unused in PT1, so their removal will have no functional effect.
There are some name changes to the elements.	Identifier MultiMatch into Identifier. Identifier-Source into SourceIdentifier. Plural form into Single form, e.g. Related Creators into RelatedCreator; Related Creations into RelatedCreation; Related Webpages into RelatedWebpage. Actor_Main Role into Actor_Type. Actor.Person_Affiliated Organisation into _Affiliation. DigitalRepresentation_CreationRepresented into DepictedCreation. DigitalRepresentation_ActorRepresented into DepictedActor.	This is basically done to clarify the semantics and to make the element names more consistent. These changes should require minimal changes in the current services.
There are some new elements	Actor_Nationality Actor.Person_Gender Webpage_Author Exhibition_Subject Exhibition_Language DigitalStillImage_Height DigitalStillImage_Width DigitalVideo_Height DigitalVideo_Width	While examining the metadata schema anew, it became clear that some information was missed off in PT1. The new elements should not effect the current services.
The number of occurrences has been changed for many elements.		This is the major functional change to the metadata. E.g. where there was only one Creator Name in PT1 to display, now there are possible many. Therefore the search process looks at all the Names and when the MM Creator Entity is returned the UI must select one to display (this is signified by the element with the preferred="true" attribute value).
Several attributes have been added to many of the elements (manual, classifier, source, reliability, preferred, etc.)		These attributes are added mainly for storing the semantic annotation information (see also section 3.4.2). Although using the attribute values will provide added functionality, to a large extent the system can be agnostic about these. Aside from the preferred attribute for multiple occurrences described above. The semantic annotation information can be examined for developing purposes with the these new attributes. Also, these attributes make it possible for the developers to study a specific part of MultiMatch, for example only the automatically generated data or only the Wikipedia content.

3.6 Conclusion

Finally, a complete overview of the metadata elements of the new PT2 metadata schema is listed in Appendix 1.

Per element one can see:

- The element name.
- The number of allowed occurrences (minimum and maximum).
- If applicable, the type of added attributes, e.g. ManualType.

The new mapping rules for the Cultural Heritage partners, presented in section 4, are compliant to the PT2 metadata schema.

After processing all the proposed revisions, the conclusion is that the new schema is able to express the functionality of PT2 as well as all the concepts that are present in the data of PT2 and the major concepts that are relevant in the CH domain. The schema meets the direct needs of the PT2 developers and will be implemented in MILOS.

As the CH domain is a complex domain, it remains a complex schema which probably provides here and there more options then the PT2 content and functionality will demand, to keep options open. During the field trials the added value of the new schema and the new mapping will be documented.

4 Revised mapping rules for PT2

The mapping rules drive the ingestion process via the scripts of the transformation file. These rules define the details of how to convert the provided metadata from various sources into the MultiMatch metadata schema.

Although not strictly lineair, the ingestion process consists of the following steps:

- Convert the original metadata into Dublin Core XML documents.
- Transform the metadata into the MultiMatch schema.
- Enrich the metadata during the semantic annotation/enrichment process.
- Ingest the data into MILOS.

During this whole process it will become clear (just as in PT1) in what way the schema will be actually used and what elements will be actually populated (either manually via the content providers or automatically via semantic enrichment) in PT2. The intention is that any element that is not populated with data provided by the content providers becomes a candidate for semantic enrichment, research conducted in WP4.

In the sections below the PT2 mapping rules are listed per CH partner. Every table concerns a different MultiMatch entity. The tables do not list all metadata element of each entity, they only contain the metadata elements that might be populated via the provided metadata of the CH partner concerned. Note, that 'OK' in the following tables means that the transformation script for PT1 does not need to be changed.

4.1 PT2 mapping rules for the Alinari metadata

The following mapping rules are applicable for the photographs that were selected for PT1.

Alinari Creation.StillImage metadata

MM metadata element	Mapping remarks
Identifier	OK
Source	OK, default "Alinari".
SourceIdentifier	OK, it is readable in the Identifier.
Type	OK, dc:type is default "Still Image". Further specification of this value can not be derived from the Alinari metadata.
Title	OK. The title contains the title of the work of art, plus extra information like the name of the creator and the museum or other location where the work of art is located/photographed. e.g. Lamento per la morte di Picasso, Galleria Toninelli d'Arte Moderna, Roma
Subject	The different types of subject indexing (<periodo_stile>, <tipo_opera>, <eventi>, <personaggi> and <keywords>)will be ingested with specified grammar.
Location	The values of <luogo_scatto> and <localita_raffigurata> will be ingested here.
Archive Location	OK, default "Florence, Italy".
Format	The same as the format of the Digital Representation.
Digital representation	OK.

Rights Holder	Populate default with “Contact fototeca@alinari.it for more details.”
Copyrighted	OK, default = yes
License condition	Populate default with “Contact fototeca@alinari.it for more details.”
Related Actor	OK, this link points to <fotografo> and/or to <artista>. In case of self portraits, add an extra Related Actor, namely the Actor.Person that is depicted in the photographed painting.
Related Actor –Type	Change this default text into “is created by” for the related <artista>. Default = “is photographed by” for the related <fotografo>. In case of self portraits, default = “is depicted in”.
Related Actor – Date	Populate with the value from <data opera> for the related <artista>. Populate with the value from <data> for the related <fotografo>.

Alinari Actor.Person metadata

MM metadata element	Mapping remarks
Identifier	OK
Type	OK, default = Creator.
Name	OK, value taken from <artista> Contains the name of the artist, sometimes with birth and death date. E.g. Nolde, Emil (1867-1956)
Source	OK, default = Alinari.
Related Creation	Ingest the Creation_Identifier concerned.
Related Creation_Type	Ingest default “created by”.
Related Creation_Date	Populate with the value from <data opera>.
Date of birth	These values will be provided in a recognizable way with extra specified grammar in <artista> to make the distinction between Name, Birth date (year) and Death date (year).
Date of death	These values will be provided in a recognizable way with extra specified grammar in <artista> to make the distinction between Name, Birth date (year) and Death date (year).

Alinari DigitalStillImage metadata

MM metadata element	Mapping remarks
Identifier	OK
Type	Populate with default value “Whole”.
Source	OK, default http://www.alinari.com.
SourceIdentifier	OK
Format	OK
Rights Holder	OK, default = “Alinari”.
Copyrighted	OK, default "Yes".
Depicted Actor	Ingest here the Actor_ Identifier for those photos/images of paintings that are self-portraits. It is the vice versa relation of Creation_Related Actor_type (depicted in).

Alinari metadata for the new PT2 content

Alinari has selected additional content for PT2. These records concern the e-Dotto/EDU educational content created by Alinari and relate to the themes identified in D2.3.1, for example ‘immigrations and emigrations’ or ‘La Belle époque’. Note, the newly selected images do not follow the cataloguing as provided in the past by Alinari. Additional note, the PT1 metadata was provided in three languages. The new metadata will be provided in Italian only, as the cataloguing of the newly selected photographs is done especially for this project.

The following mapping rules are only applicable for the new content that is provided by Alinari for PT2.

MM metadata element Creation.StillImage	Mapping remarks
Identifier	The MM Identifier.
Source	Ingest the value that is provided (Options: www.alinari.com; www.e-dotto.it; www.edu.alinari.it.).
SourceIdentifier	Ingest the Alinari identifier. For example “ACA-F-009629-0000-112”.
Type	dc:type is default "Still Image". Further specification of this value can not be derived from the Alinari metadata.
Title	Ingest here the values (Italian only) from <short_description>. Example 1: <short_description>Underwood and Underwood, Folla su un molo saluta gli emigranti verso l'America, 1905, Norvegia.</short_description>. Example 2: <short_description> Mario Castagneri, Tre donne con bambini sono ritratte sul ponte di una nave diretta a Buenos Aires, 12 Maggio 1926, America. </short_description>.
Subject	If available, the different types of subject indexing (<periodo_stile>, <tipo_opera>, <eventi>, <personaggi> and <keywords>) will be ingested with specified grammar. Plus the theme itself will be ingested with specified grammar. Example of encoding the PT2 themes: <subject> manual="yes" reliability="1.0" language="eng" source="PT2Themes" classifier="gold"> Archeology </subject>
Archive Location	Default "Florence, Italy".

Format	The dimensions of the physical photograph, if available; otherwise the same dimensions as for the digital representation.
Digital representation	Automatically created: the link to the DigitalStillImage representing the physical photograph.
Description	Ingest the value of <long _description>, which describes the theme involved in Italian. Example1: <long_description>Il fenomeno delle migrazioni si riferisce per lo più all'incremento demografico di una popolazione cui non corrisponde più la capacità di popolamento di un paese: si fa necessaria di conseguenza la migrazione di masse di uomini verso regioni che presentino ancora riserve da utilizzare.</long_description> Example 2 : <long_description> La storia degli Ebrei è un caso di storia di migrazione. Essa inizia con la vocazione di Abramo che da Ur dei Caldei va in Palestina ad abitare la terra promessa da Dio e poi in Egitto. Il soggiorno in terra straniera più tardi si muta per gli ebrei in oppressione dalla quale riesce a liberarli Mosé, che sul Monte Sinai riceve da Dio le tavole della Legge. A Mosé succede Giosuè. Alla fine del XIII secolo a.C. si compie la conquista della Palestina e nasce la monarchia. Alla morte di Salomone il regno si divide in due: Israele e Giudea. Con la distruzione del Regno di Giuda ad opera dei Babilonesi (598) gli ebrei sono deportati a Babilonia. Nel 135 d.C. un tentativo di rivolta contro Roma che ne erano entrati in possesso, si conclude con la distruzione di Gerusalemme; per gli Ebrei comincia da questo momento la diaspora che li disperderà nel mondo.</long_description>
Rights Holder	Default "Alinari 24 ORE".
Copyrighted	Default "Yes".
License condition	Populate default with "Contact fototeca@alinari.it for more details."
Related Actor	If available, this link points to <fotografo>.
Related Actor –Type	Default "creator" .
Related Actor – Date	If available, ingest here the value from <data>.

MM metadata element Actor.Person	**Mapping remarks**
Identifier	The MM Identifier.
Type	Default = Creator.
Name	Ingest the value from <fotografo>.
Source	Default = Alinari.
Nationality	Populated via Alinari if available, otherwise populated via ULAN if the Actor is available.
Related Creation	Ingest the Creation_Identifier concerned.
Related Creation_Type	Default "created by".
Related Creation_Date	Populate with the value from <data>.
Date of birth	If available, these values will be provided in a recognizable way with extra specified grammar in <fotografo> to make the distinction between Name, Birth date (year) and Death date (year).
Date of death	If available, these values will be provided in a recognizable way with extra specified grammar in < fotografo > to make the distinction between Name, Birth date (year) and Death date (year).

MM metadata element Digital.StillImage	**Mapping remarks**
Identifier	The MM Identifier.
Type	Default “Whole”.
Source	Default "http://www.alinari.com.; www.e-dotto.it; www.edu.alinari.itAlinari".
SourceIdentifier	Ingest the filename or number that Alinari uses to identify the file concerned.
Rights Holder	Default “Alinari 24 ORE”.
Copyrighted	Default "Yes".
Depicted Creation	The MM Identifier of the Creation (see table above) that is represented by this Digital representation.

4.2 PT2 mapping rules for the Cervantes metadata

The following mapping rules apply to the content provided by Biblioteca Virtual Miguel de Cervantes.

Cervantes Creation.Text metadata

MM metadata element	Mapping remarks
Identifier	OK
Source	Ingest the value from <dc:source>, which contains the URL of the catalogue card of this creation record.
Source Identifier	Ingest here the value from <dc:identifier>, which is the original record number from the BVMdC catalogue.
Type	Populate this element with these possible values, ingested from <dc:type>: • “Text - Biografía”: for biographies • “Text - Estudio crítico”: for critic studies • “Text - Tesis”: for thesis • “Text - Lección inaugural”: for a speech • “Text - Manuscrito”: for manuscripts • “Text - Revista”: for magazines • “Text - Mapa”: for maps This element will contain the value “Text” for all other types (books, poems or unknown types of text creations).
Title	OK
Subject	Ingest the values (all in Spanish) from <dc:subject>(c:category)</dc:subject> as well as from <dc:subject>(l:location)</dc:subject> and from <dc:subject>(s:subject)</dc:subject>.
Archive Location	Ingest default value “www.cervantesvirtual.com”.
Description	OK, ingest here <dc:description>. If there are more than 1 descriptions, they will be joined for PT2.
Digital representation	OK.This element contains the link to the MM identifier of the Digital Representation (most of the time this is a webpage with the ToC).
Related Actor	Ingest the value between “(n:” and “)”. e.g. <dc:creator>(n:Ortelius, Abraham) (b:1527)(d:1598)</dc:creator> PT2 Secondary creators, editors etc. can be ingested from <dc:contributor>(n: name).
Related Actor -Type	OK. Default value = “Creator” for the name derived from <dc:creator>. Default value = “Contributor” for the name(s) derived from <dc:contributor>.
Related Actor - Date	Ingest the value from <dc:subject>(t:time) </dc:subject>, i.e. a century or a period of years, for example (1541 – 1560).
Language	OK. Default value for the language of the creation = “spa” (see: <dc:language>).

Cervantes Actor.Person metadata

MM metadata element	Mapping remarks
Identifier	OK
Type	OK, default = “Creator”.
Name	OK
Source	OK, default = “Cervantes”.
Related Creation	OK.
Related Creation_Type	Ingest default “created by”.
Related Creation_Date	OK.
Date of birth	Ingest the value between “(b:” and “)”, if available. e.g. <dc:creator>(n:Ortelius, Abraham) (b:1527)(d:1598)</dc:creator>
Date of death	Ingest the value between “(d:” and “)”, if available. e.g. <dc:creator>(n:Ortelius, Abraham) (b:1527)(d:1598)</dc:creator>

Cervantes Web page metadata

These records concern webpages created by BVMdC about several authors and creations, of the type:

- “Web - Biografía”: for biographies
- “Web - Estudio crítico”: for critic studies
- “Web - Tesis”: for thesis
- “Web - Lección inaugural”: for a speech
- “Web - Manuscrito”: for manuscripts
- “Web - Revista”: for magazines
- “Web - Mapa”: for maps
- “Web” for all other types.

MM metadata element	Question, suggestion, remark
Identifier	OK
Type	Ingest <dc:type>, which can contain the values mentioned above this table.
Language	Default “Spanish”
URI	OK, ingest <dc:identifier> which has the URL to the content.
Date captured	Ingest the date that this page was crawled for indexing in MM.
Title	OK: <dc:title> contains the title of the actual web page.
Identifier	OK
Date - modified	OK, although <dc:date> actually contains the creation date of the web page.
Subject	Ingest the values (all in Spanish) from <dc:subject>(c:category)</dc:subject> as well as from <dc:subject>(l:location)</dc:subject> and from <dc:subject>(s:subject)</dc:subject>.
Related Actor	If available ingest the Identifier of the Actor.Person that is the author of the book concerned; otherwise possible candidate for automatic generation.
Related Creation	If available ingest the Identifier MM of the Creation that is the author of the book concerned; otherwise possible candidate for automatic generation.
Related Website	Ingest the URL of the homepage: ‘www.cervantesvirtual.com’.
Content	Automatically ingest here the link to the file with the content of this web page.

4.3 Mapping rules PT2 for the Sound and Vision metadata

The following mapping rules apply to the content provided by Netherlands Institute for Sound and Vision.

Sound and Vision Creation.Video metadata

MM metadata element	Mapping remarks
Identifier MM	OK
Source	OK
SourceIdentifier	OK: not exactly the iMMix ID, but a source number of the video file, e.g. BG_37454-out.wmv
Type	OK
Title	OK, e.g. <assets_id>1182</assets_id> <title>LIVE/LIFE (Uitz. 28-06-1979)</title> <title_alternative>Drie choreografen en een ballet</title_alternative>
Subject	Ingest the value of <subject> as it is provided. Unfortunately it is not possible to identify the different types of subject terms. Ingest the values of the PT2 themes and the Top40 artists also into Subject. Example of encoding the PT2 themes: <subject manual="yes" reliability="1.0" language="eng" source="PT2Themes" classifier="gold"> Archeology </subject> Example of encoding the Top40 artists: <subject category="creator" manual="yes" reliability="1.0" language="eng" source="MMTop40" classifier="gold"> Harry Mulisch </subject>
Location	Ingest the values of <coverage spatial>, even though this contains different types of location. If possible, maybe later the less valid terms will be filtered out manually.
Archive Location	OK
Format	OK, although formally not correct, as the format of the original creation is either an analogue format or a high-quality digital format.
Digital representation	OK
Description	OK
Copyrighted	OK, default = No.
Related Actor	The person names shall be filtered based on the thesaurus, in order to distinguish between persons and organisations or other names. The specified grammar will make sure that the same MM Identifier of the Actor_Creator will be ingested into these fields: Actor_Creator Identifier and Creation_Related Creator.
Related Actor - Type	OK
Related Actor - Date	Ingest the value of <date_issued>.
Transcript	Populate this element with the link to the related transcript file (XML file). Note, that the transcript of speech should be indexed.
Language	Ingest the value(s) of <language>.

Sound and Vision Actor.Person metadata

MM metadata element	Mapping remarks
Identifier	OK
Type	OK, default = “Creator”. If possible more specific roles will be added via semantic enrichment.
Name	OK
Source	OK
Related Creation	Attention, in PT1 there was not always the same value as in Creation_Related Actor, although it should be.
Related Creation_Type	OK, default= “Creation”. Further specification of roles is a candidate for semantic enrichment.
Related Creation_Date	Ingest the value of </date_created>.

Sound and Vision Digital.Video metadata

MM metadata element	Mapping remarks
Identifier	OK
Type	Default = “Whole”.
Source	Default = “Sound and Vision”.
SourceIdentifier	OK
Format	OK, default = “MPEG-1”.
Rights Holder	OK, default = “Sound and Vision”.
Copyrighted	OK, default = “No”.
Depicted Creation	Ingest the URN of the creation that is linked to this DigRep.

4.4 Conclusion

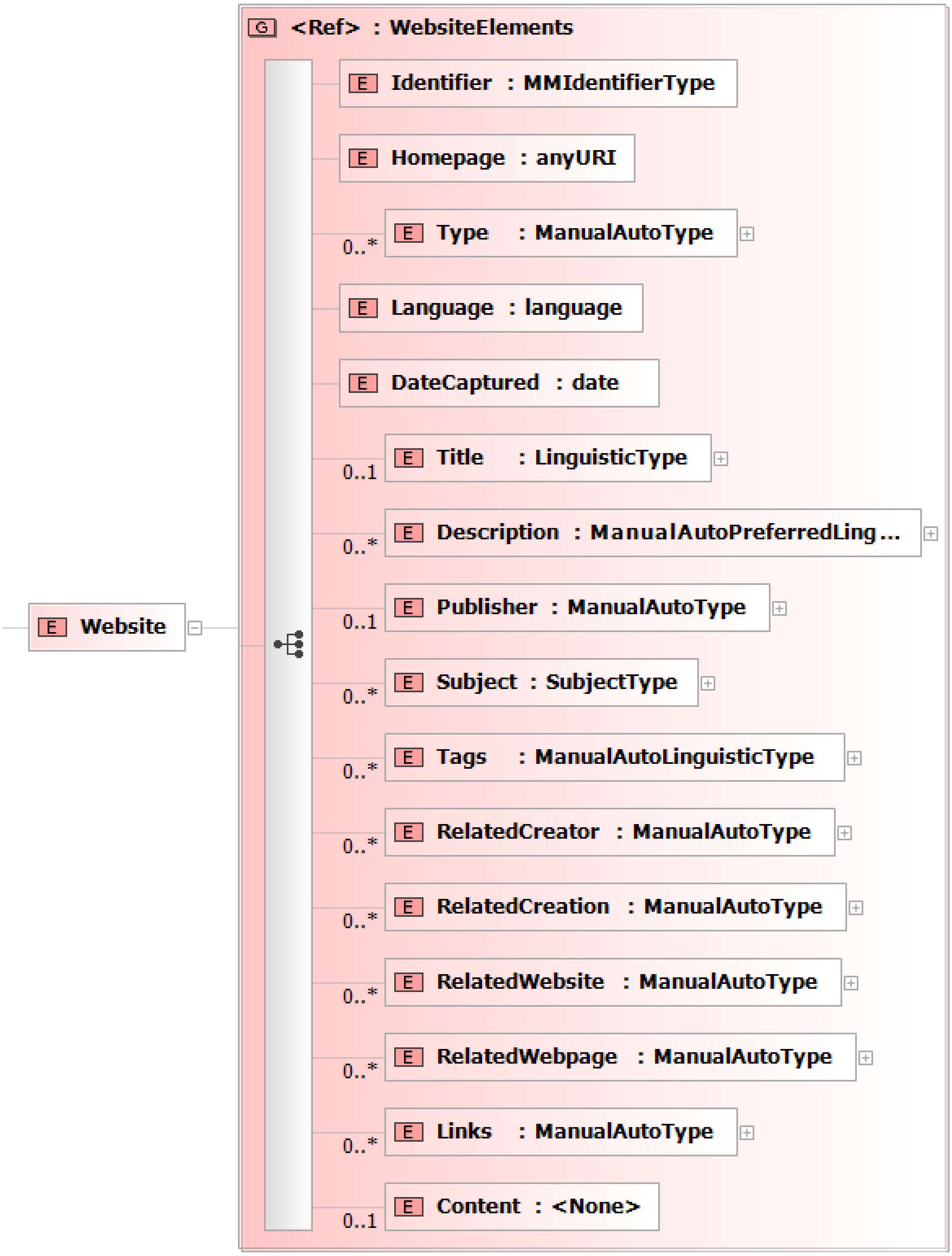
The thorough study of the mapping of the proprietary schemas of the CH partners to the new MultiMatch metadata schema resulted in the above presented new set mapping rules for PT2. These new mapping rules:

- Correct the minor mapping errors of the PT1 transformation process.
- Avoid information loss via the addition of specified grammar in several metadata elements in the XML documents that will be provided for the PT2 transformation process.
- Populate 43 elements in PT2 newly or in an enhanced way.

Appendix 1

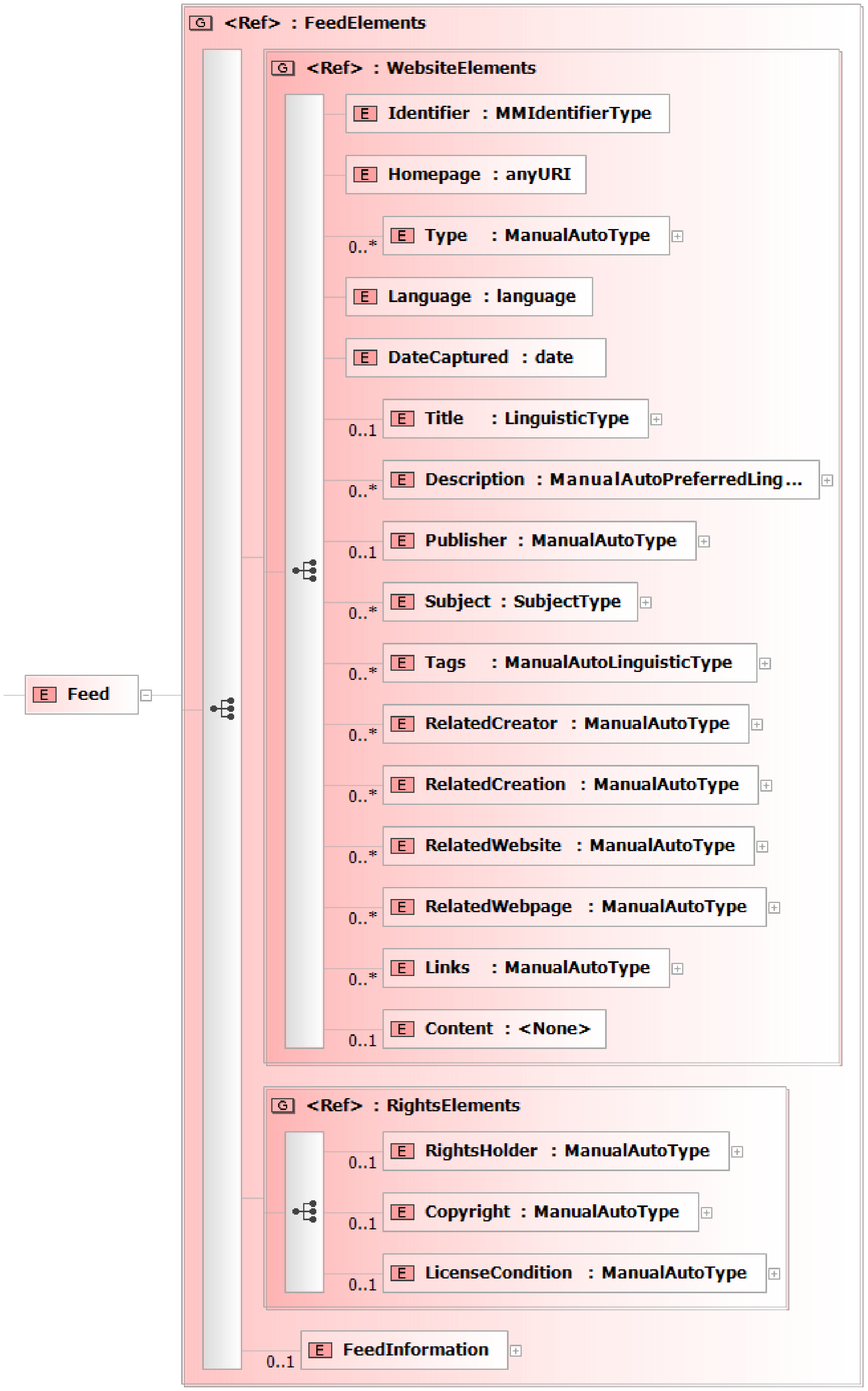
Metadata elements per (sub) entity

Website Elements


G <Ref> : WebsiteElements
E Identifier : MMIdentifierType
E Homepage : anyURI
0..* E Type : ManualAutoType
E Language : language
E DateCaptured : date
0..1 E Title : LinguisticType
0..* E Description : ManualAutoPreferredLing...
0..1 E Publisher : ManualAutoType
0..* E Subject : SubjectType
0..* E Tags : ManualAutoLinguisticType
0..* E RelatedCreator : ManualAutoType
0..* E RelatedCreation : ManualAutoType
0..* E RelatedWebsite : ManualAutoType
0..* E RelatedWebpage : ManualAutoType
0..* E Links : ManualAutoType
0..1 E Content : <None>
E Website

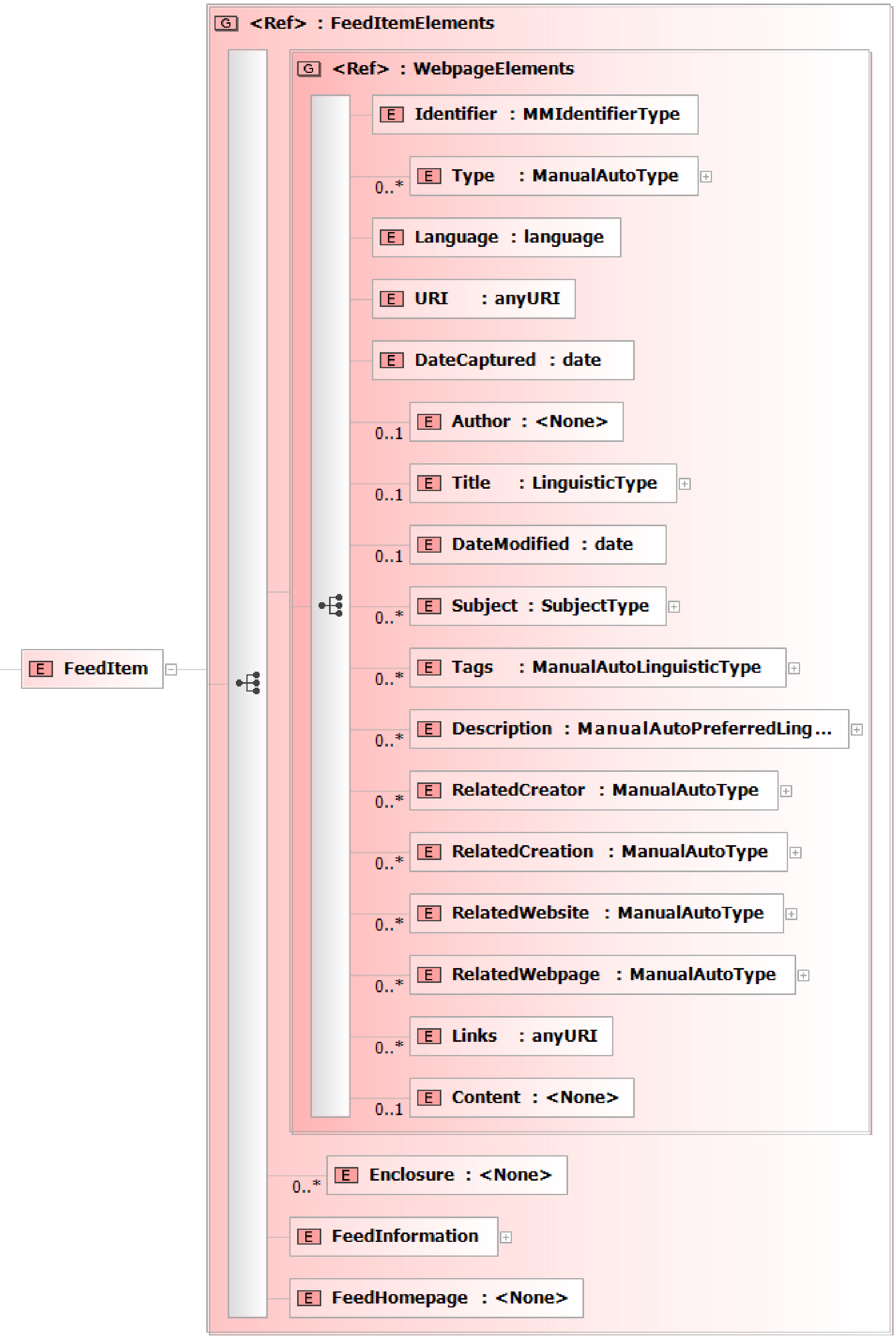
Feed Elements

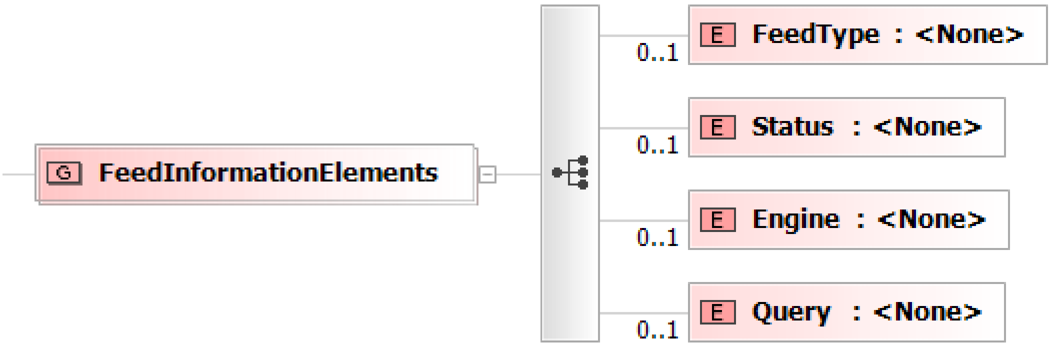
A Feed (sub-entity of Collection) contains all the elements for Website plus FeedInformationElements and RightsElements.


G <Ref> : FeedElements
G <Ref> : WebsiteElements
E Identifier : MMIdentifierType
E Homepage : anyURI
0..* E Type : ManualAutoType
E Language : language
E DateCaptured : date
0..1 E Title : LinguisticType
0..* E Description : ManualAutoPreferredLing...
0..1 E Publisher : ManualAutoType
0..* E Subject : SubjectType
0..* E Tags : ManualAutoLinguisticType
0..* E RelatedCreator : ManualAutoType
0..* E RelatedCreation : ManualAutoType
0..* E RelatedWebsite : ManualAutoType
0..* E RelatedWebpage : ManualAutoType
0..* E Links : ManualAutoType
0..1 E Content : <None>
E Feed
G <Ref> : RightsElements
0..1 E RightsHolder : ManualAutoType
0..1 E Copyright : ManualAutoType
0..1 E LicenseCondition : ManualAutoType
0..1 E FeedInformation

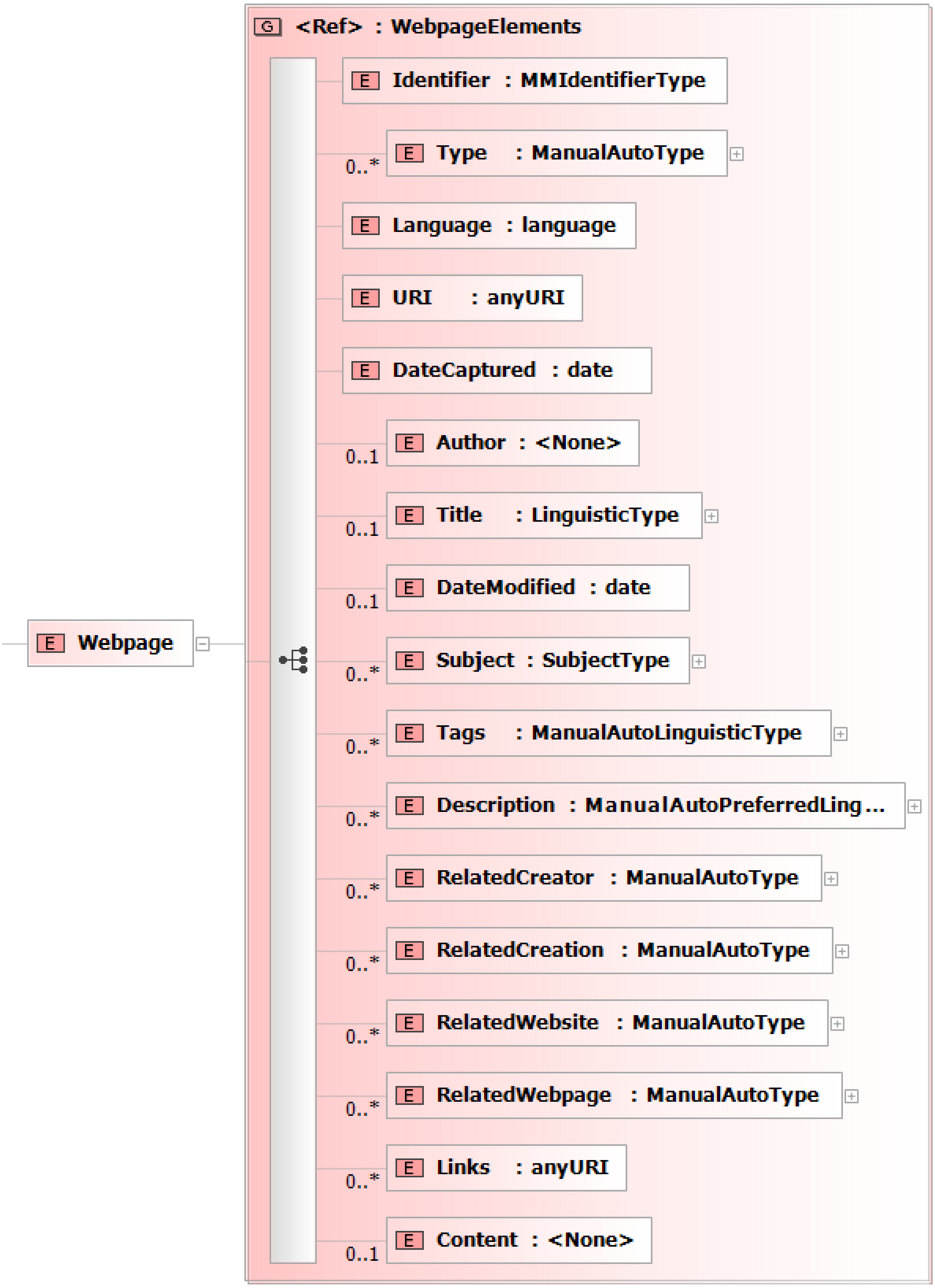
FeedItem Elements

FeedItem is a new entity which contains metadata for dynamic data feeds. This entity links to a feeds homepage.

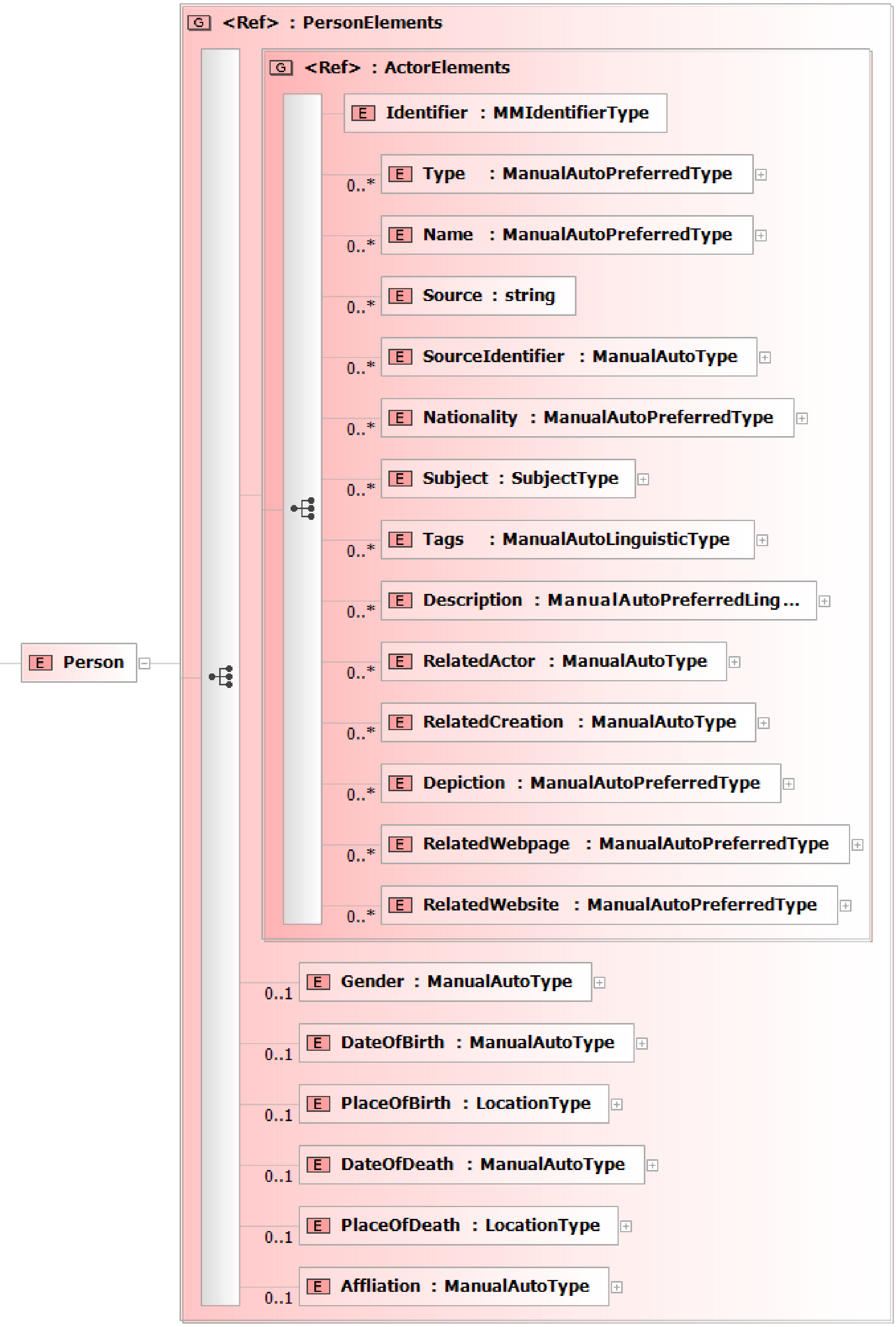

G <Ref> : FeedItemElements
G <Ref> : WebpageElements
E Identifier : MMIdentifierType
0..* E Type : ManualAutoType
E Language : language
E URI : anyURI
E DateCaptured : date
0..1 E Author : <None>
0..1 E Title : LinguisticType
0..1 E DateModified : date
0..* E Subject : SubjectType
0..* E Tags : ManualAutoLinguisticType
0..* E Description : ManualAutoPreferredLing...
0..* E RelatedCreator : ManualAutoType
0..* E RelatedCreation : ManualAutoType
0..* E RelatedWebsite : ManualAutoType
0..* E RelatedWebpage : ManualAutoType
0..* E Links : anyURI
0..1 E Content : <None>
0..* E Enclosure : <None>
E FeedInformation
E FeedHomepage : <None>
E FeedItem


G FeedInformationElements
E FeedType : <None>
0..1
E Status : <None>
0..1
E Engine : <None>
0..1
E Query : <None>
0..1

Webpage Elements

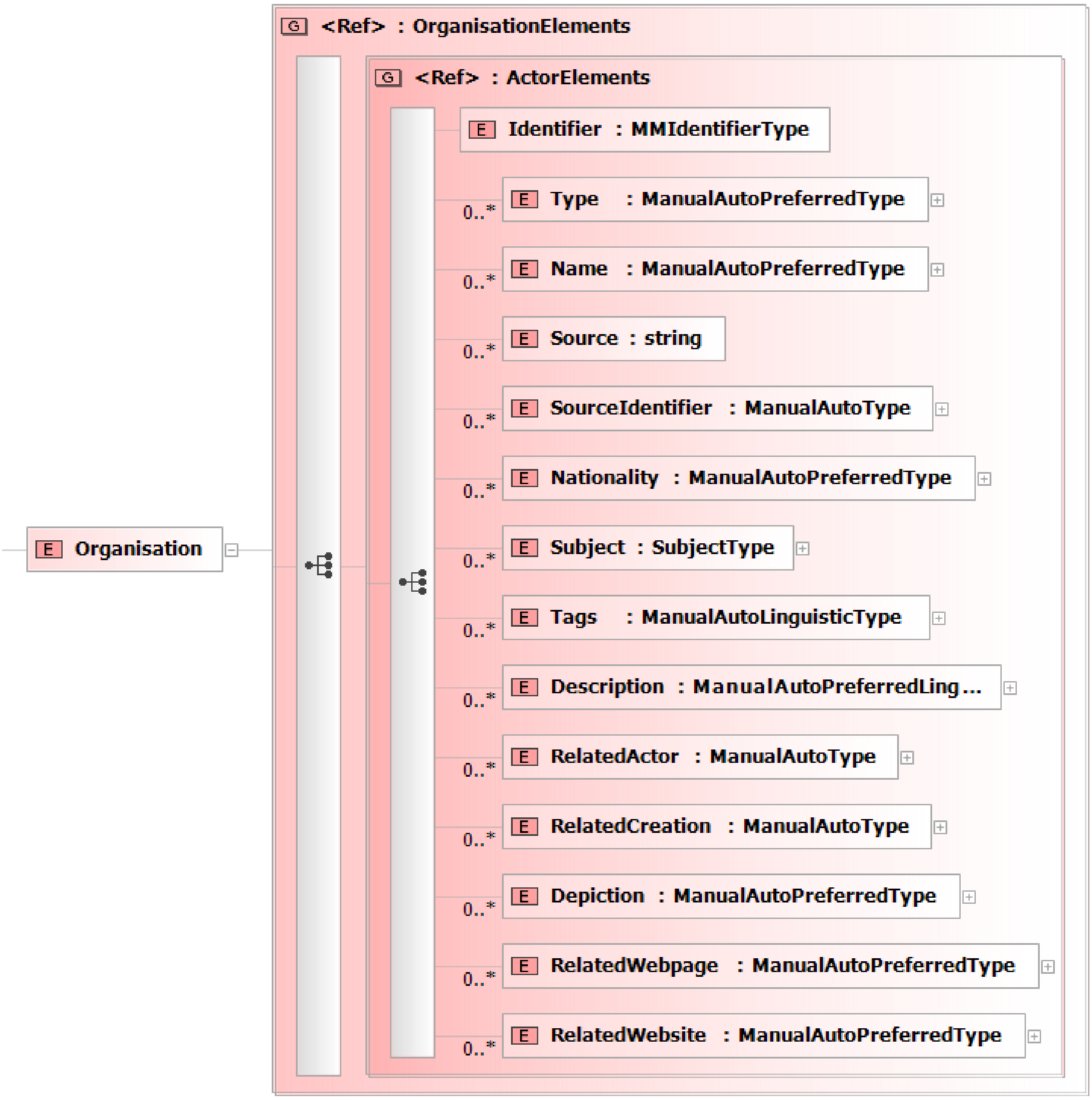

G <Ref> : WebpageElements
E Webpage
E Identifier : MMIdentifierType
0..* E Type : ManualAutoType
E Language : language
E URI : anyURI
E DateCaptured : date
0..1 E Author : <None>
0..1 E Title : LinguisticType
0..1 E DateModified : date
0..* E Subject : SubjectType
0..* E Tags : ManualAutoLinguisticType
0..* E Description : ManualAutoPreferredLing...
0..* E RelatedCreator : ManualAutoType
0..* E RelatedCreation : ManualAutoType
0..* E RelatedWebsite : ManualAutoType
0..* E RelatedWebpage : ManualAutoType
0..* E Links : anyURI
0..1 E Content : <None>

Actor.Person Elements



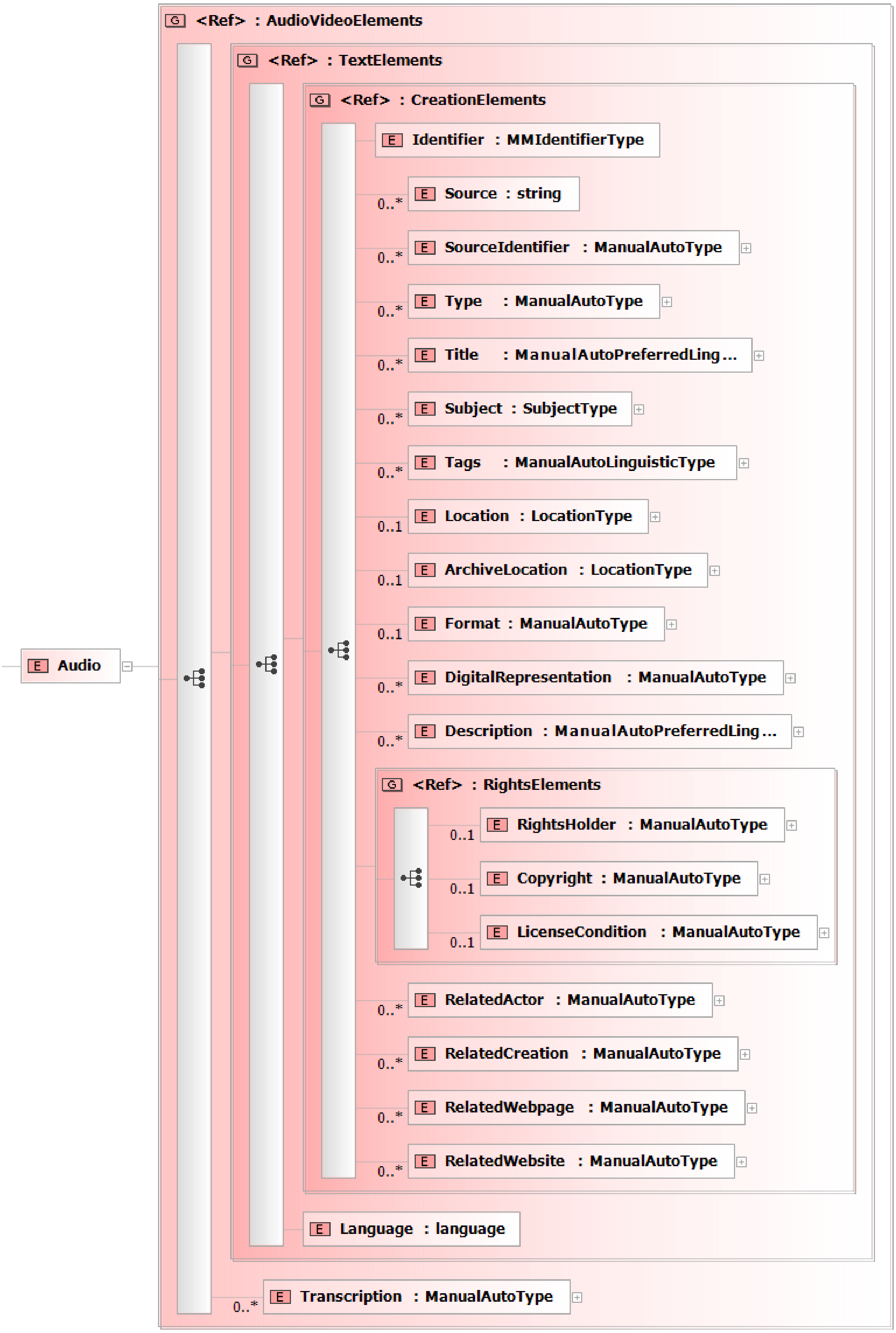
<Ref> : PersonElements
<Ref> : ActorElements
E Identifier : MMIdentifierType
0..* E Type : ManualAutoPreferredType
0..* E Name : ManualAutoPreferredType
0..* E Source : string
0..* E SourceIdentifier : ManualAutoType
0..* E Nationality : ManualAutoPreferredType
0..* E Subject : SubjectType
0..* E Tags : ManualAutoLinguisticType
0..* E Description : ManualAutoPreferredLing...
0..* E RelatedActor : ManualAutoType
0..* E RelatedCreation : ManualAutoType
0..* E Depiction : ManualAutoPreferredType
0..* E RelatedWebpage : ManualAutoPreferredType
0..* E RelatedWebsite : ManualAutoPreferredType
E Person
0..1 E Gender : ManualAutoType
0..1 E DateOfBirth : ManualAutoType
0..1 E PlaceOfBirth : LocationType
0..1 E DateOfDeath : ManualAutoType
0..1 E PlaceOfDeath : LocationType
0..1 E Affliation : ManualAutoType

Actor.Organisation Elements

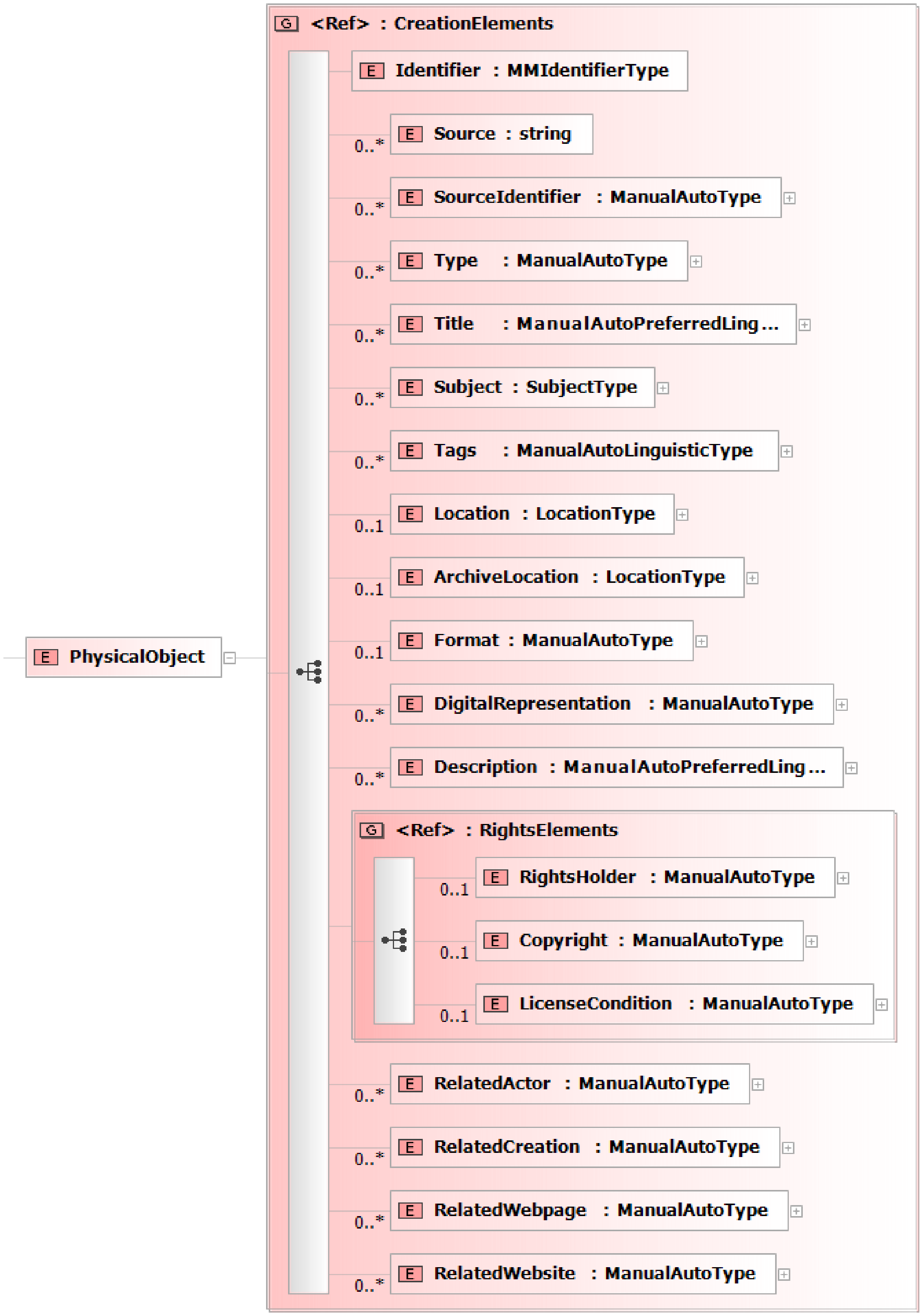


G <Ref> : OrganisationElements
G <Ref> : ActorElements
E Identifier : MMIdentifierType
0..* E Type : ManualAutoPreferredType
0..* E Name : ManualAutoPreferredType
0..* E Source : string
0..* E SourceIdentifier : ManualAutoType
0..* E Nationality : ManualAutoPreferredType
E Organisation
0..* E Subject : SubjectType
0..* E Tags : ManualAutoLinguisticType
0..* E Description : ManualAutoPreferredLing...
0..* E RelatedActor : ManualAutoType
0..* E RelatedCreation : ManualAutoType
0..* E Depiction : ManualAutoPreferredType
0..* E RelatedWebpage : ManualAutoPreferredType
0..* E RelatedWebsite : ManualAutoPreferredType

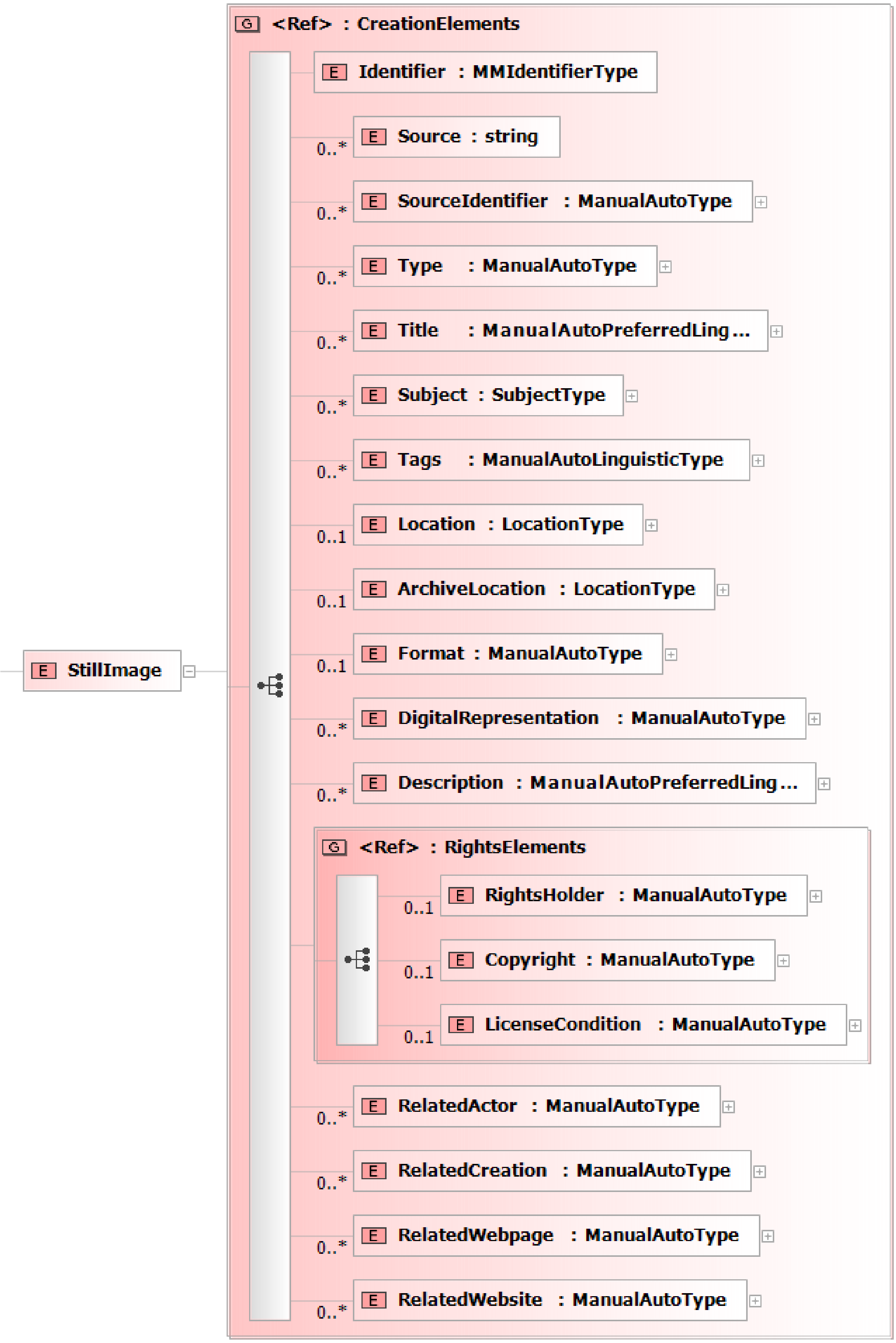
Creation.Audio Elements


G <Ref> : AudioVideoElements
G <Ref> : TextElements
G <Ref> : CreationElements
E Identifier : MMIdentifierType
0..* E Source : string
0..* E SourceIdentifier : ManualAutoType
0..* E Type : ManualAutoType
0..* E Title : ManualAutoPreferredLing...
0..* E Subject : SubjectType
0..* E Tags : ManualAutoLinguisticType
0..1 E Location : LocationType
0..1 E ArchiveLocation : LocationType
0..1 E Format : ManualAutoType
0..* E DigitalRepresentation : ManualAutoType
0..* E Description : ManualAutoPreferredLing...
G <Ref> : RightsElements
0..1 E RightsHolder : ManualAutoType
0..1 E Copyright : ManualAutoType
0..1 E LicenseCondition : ManualAutoType
0..* E RelatedActor : ManualAutoType
0..* E RelatedCreation : ManualAutoType
0..* E RelatedWebpage : ManualAutoType
0..* E RelatedWebsite : ManualAutoType
E Language : language
0..* E Transcription : ManualAutoType
E Audio

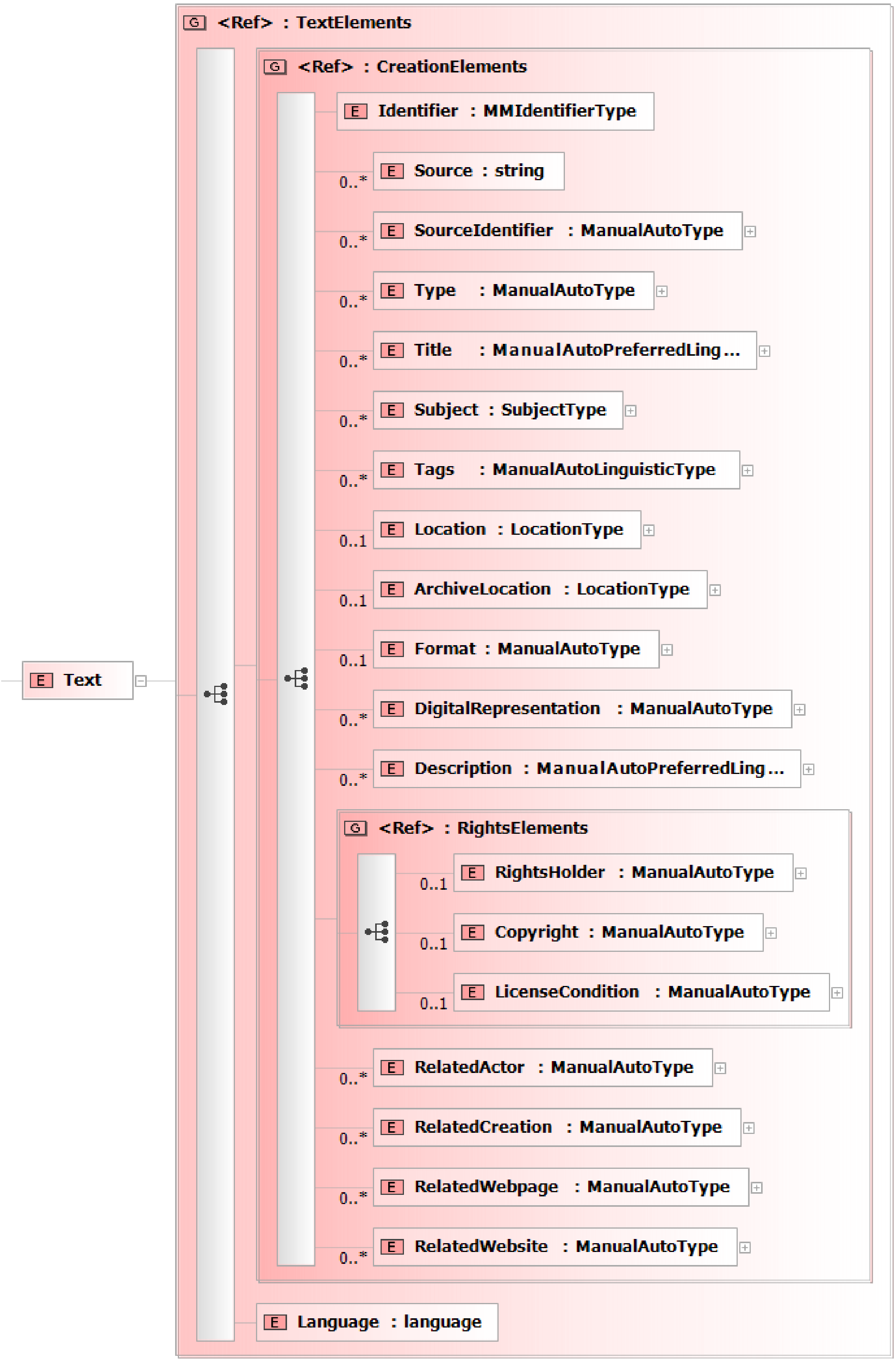
Creation.PhysicalObject Elements


PhysicalObject
<Ref> : CreationElements
Identifier : MMIdentifierType
0..* Source : string
0..* SourceIdentifier : ManualAutoType
0..* Type : ManualAutoType
0..* Title : ManualAutoPreferredLing...
0..* Subject : SubjectType
0..* Tags : ManualAutoLinguisticType
0..1 Location : LocationType
0..1 ArchiveLocation : LocationType
0..1 Format : ManualAutoType
0..* DigitalRepresentation : ManualAutoType
0..* Description : ManualAutoPreferredLing...
<Ref> : RightsElements
0..1 RightsHolder : ManualAutoType
0..1 Copyright : ManualAutoType
0..1 LicenseCondition : ManualAutoType
0..* RelatedActor : ManualAutoType
0..* RelatedCreation : ManualAutoType
0..* RelatedWebpage : ManualAutoType
0..* RelatedWebsite : ManualAutoType

Creation.StillImage Elements

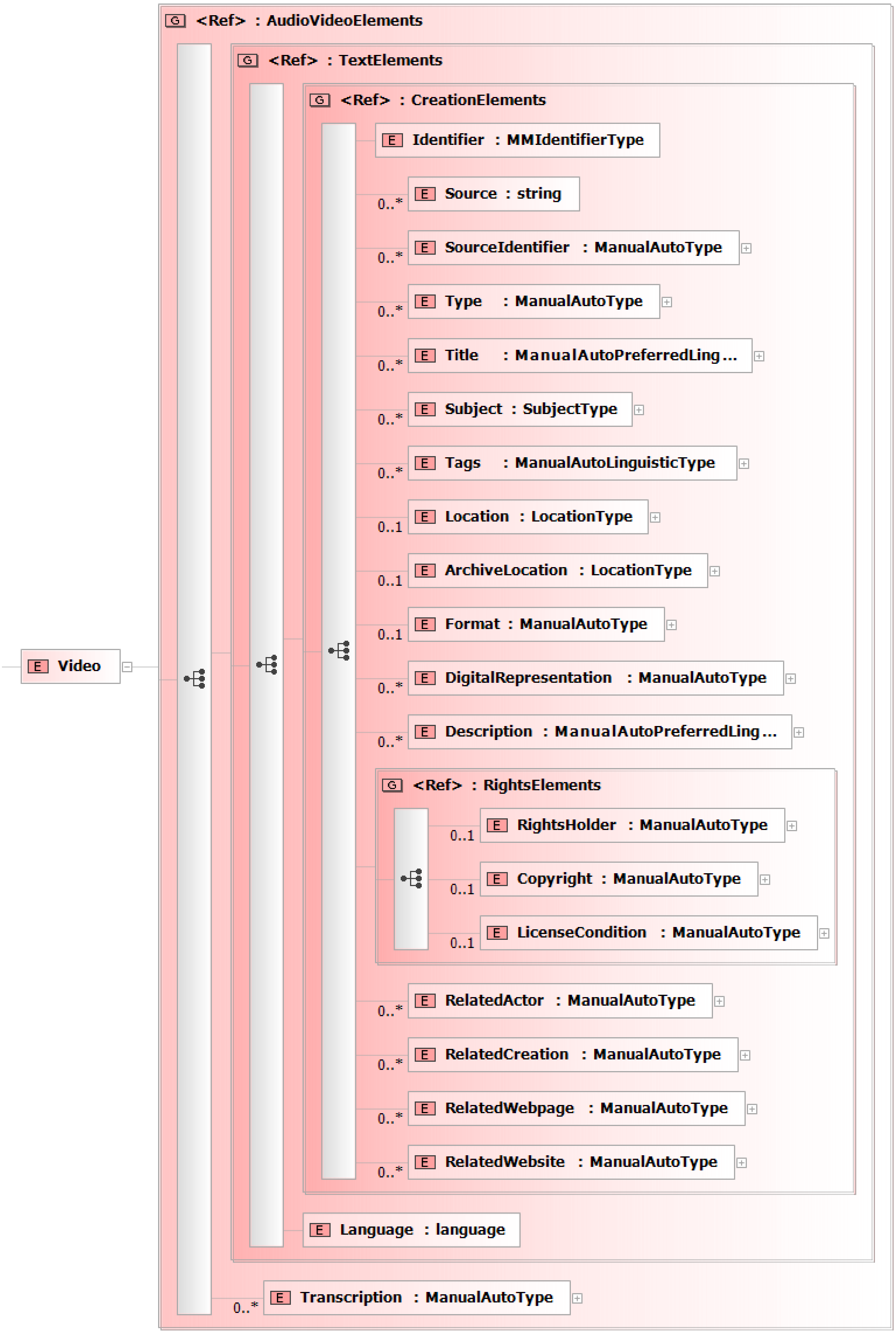

E StillImage
G <Ref> : CreationElements
E Identifier : MMIdentifierType
0..* E Source : string
0..* E SourceIdentifier : ManualAutoType
0..* E Type : ManualAutoType
0..* E Title : ManualAutoPreferredLing...
0..* E Subject : SubjectType
0..* E Tags : ManualAutoLinguisticType
0..1 E Location : LocationType
0..1 E ArchiveLocation : LocationType
0..1 E Format : ManualAutoType
0..* E DigitalRepresentation : ManualAutoType
0..* E Description : ManualAutoPreferredLing...
G <Ref> : RightsElements
0..1 E RightsHolder : ManualAutoType
0..1 E Copyright : ManualAutoType
0..1 E LicenseCondition : ManualAutoType
0..* E RelatedActor : ManualAutoType
0..* E RelatedCreation : ManualAutoType
0..* E RelatedWebpage : ManualAutoType
0..* E RelatedWebsite : ManualAutoType

Creation.Text Elements

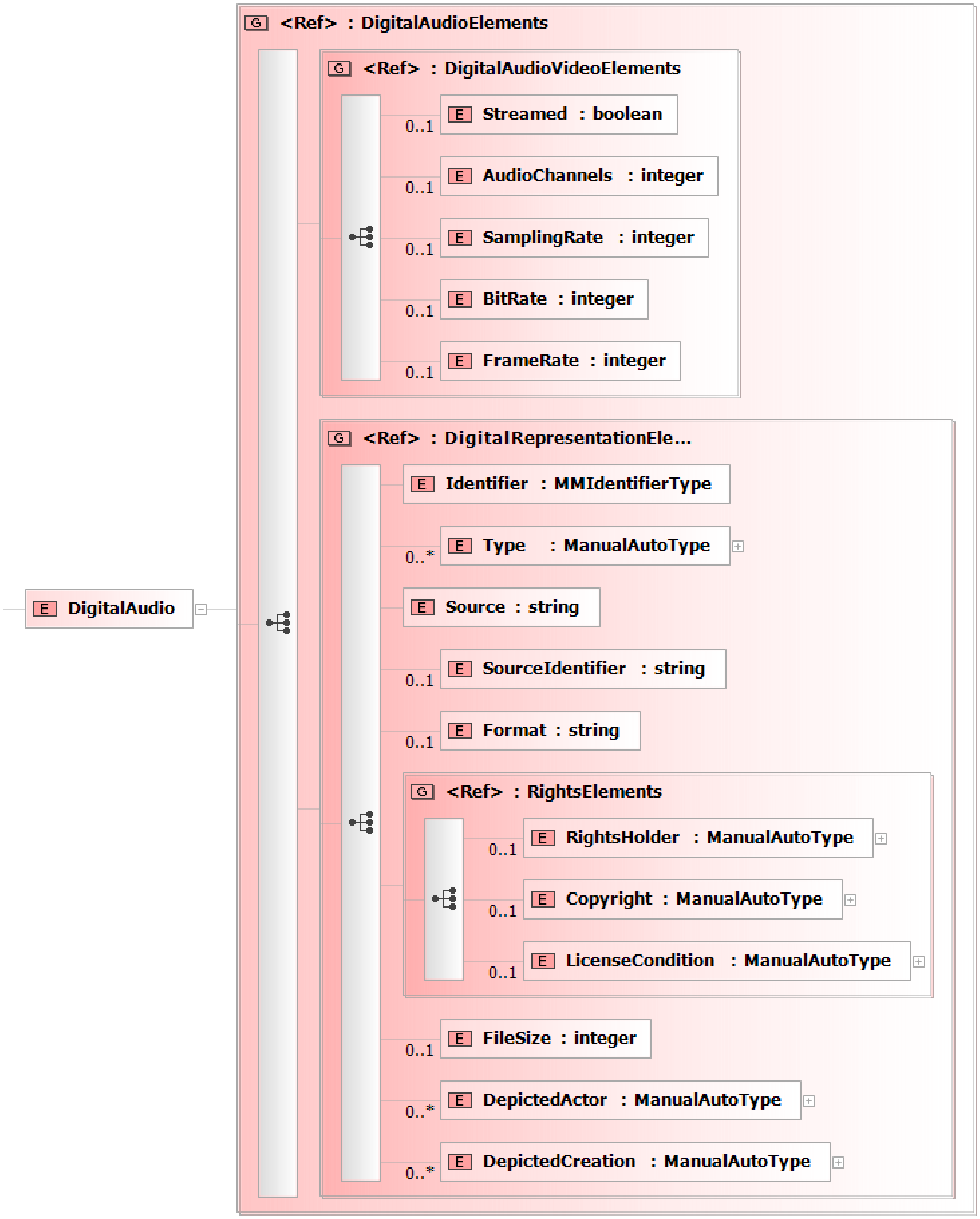


G <Ref> : TextElements
G <Ref> : CreationElements
E Identifier : MMIdentifierType
0..* E Source : string
0..* E SourceIdentifier : ManualAutoType
0..* E Type : ManualAutoType
0..* E Title : ManualAutoPreferredLing...
0..* E Subject : SubjectType
0..* E Tags : ManualAutoLinguisticType
0..1 E Location : LocationType
0..1 E ArchiveLocation : LocationType
0..1 E Format : ManualAutoType
0..* E DigitalRepresentation : ManualAutoType
0..* E Description : ManualAutoPreferredLing...
G <Ref> : RightsElements
0..1 E RightsHolder : ManualAutoType
0..1 E Copyright : ManualAutoType
0..1 E LicenseCondition : ManualAutoType
0..* E RelatedActor : ManualAutoType
0..* E RelatedCreation : ManualAutoType
0..* E RelatedWebpage : ManualAutoType
0..* E RelatedWebsite : ManualAutoType
E Language : language
E Text

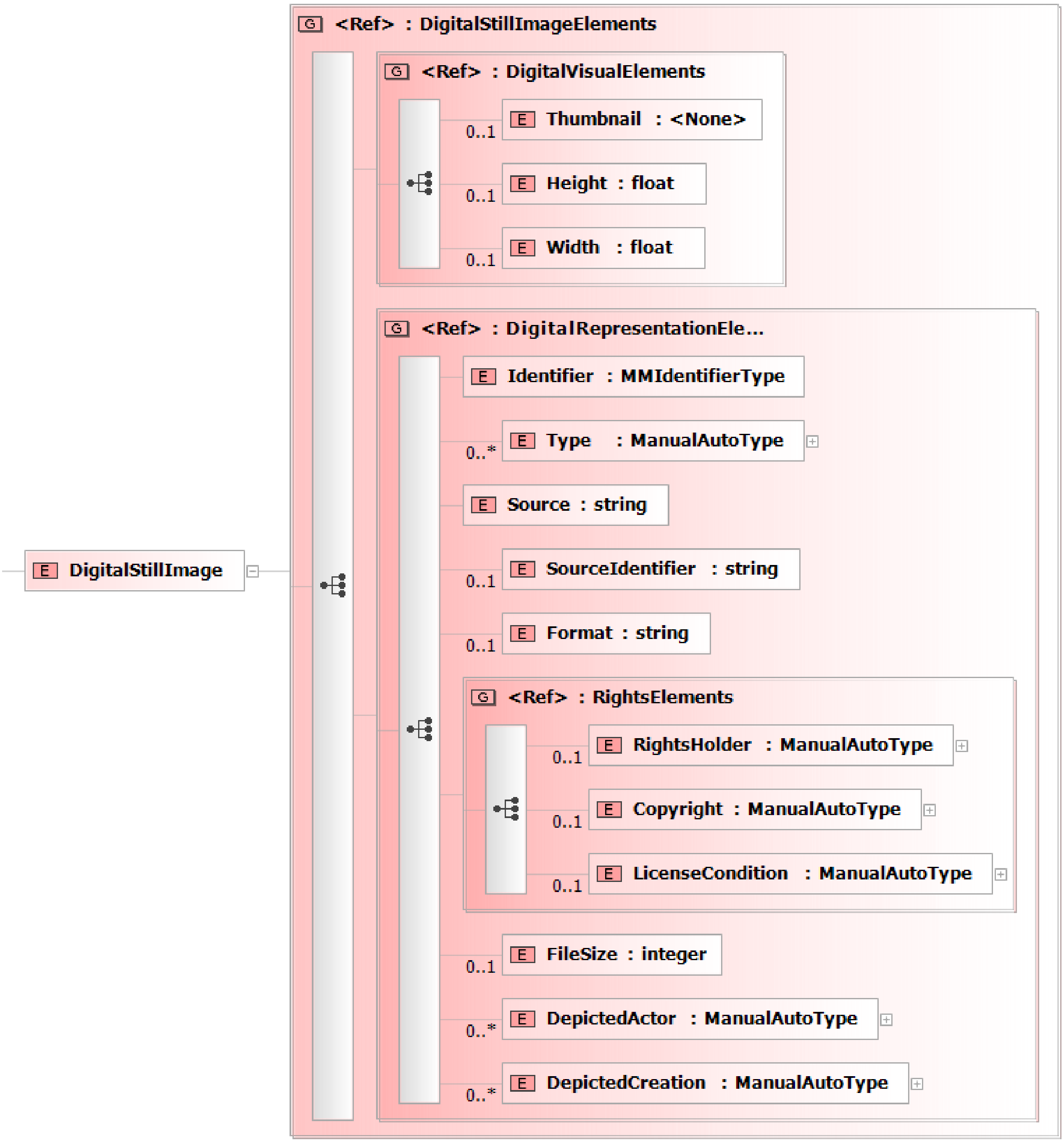
Creation.Video Elements


G <Ref> : AudioVideoElements
G <Ref> : TextElements
G <Ref> : CreationElements
E Identifier : MMIdentifierType
0..* E Source : string
0..* E SourceIdentifier : ManualAutoType
0..* E Type : ManualAutoType
0..* E Title : ManualAutoPreferredLing...
0..* E Subject : SubjectType
0..* E Tags : ManualAutoLinguisticType
0..1 E Location : LocationType
0..1 E ArchiveLocation : LocationType
0..1 E Format : ManualAutoType
0..* E DigitalRepresentation : ManualAutoType
0..* E Description : ManualAutoPreferredLing...
G <Ref> : RightsElements
0..1 E RightsHolder : ManualAutoType
0..1 E Copyright : ManualAutoType
0..1 E LicenseCondition : ManualAutoType
0..* E RelatedActor : ManualAutoType
0..* E RelatedCreation : ManualAutoType
0..* E RelatedWebpage : ManualAutoType
0..* E RelatedWebsite : ManualAutoType
E Language : language
0..* E Transcription : ManualAutoType
E Video

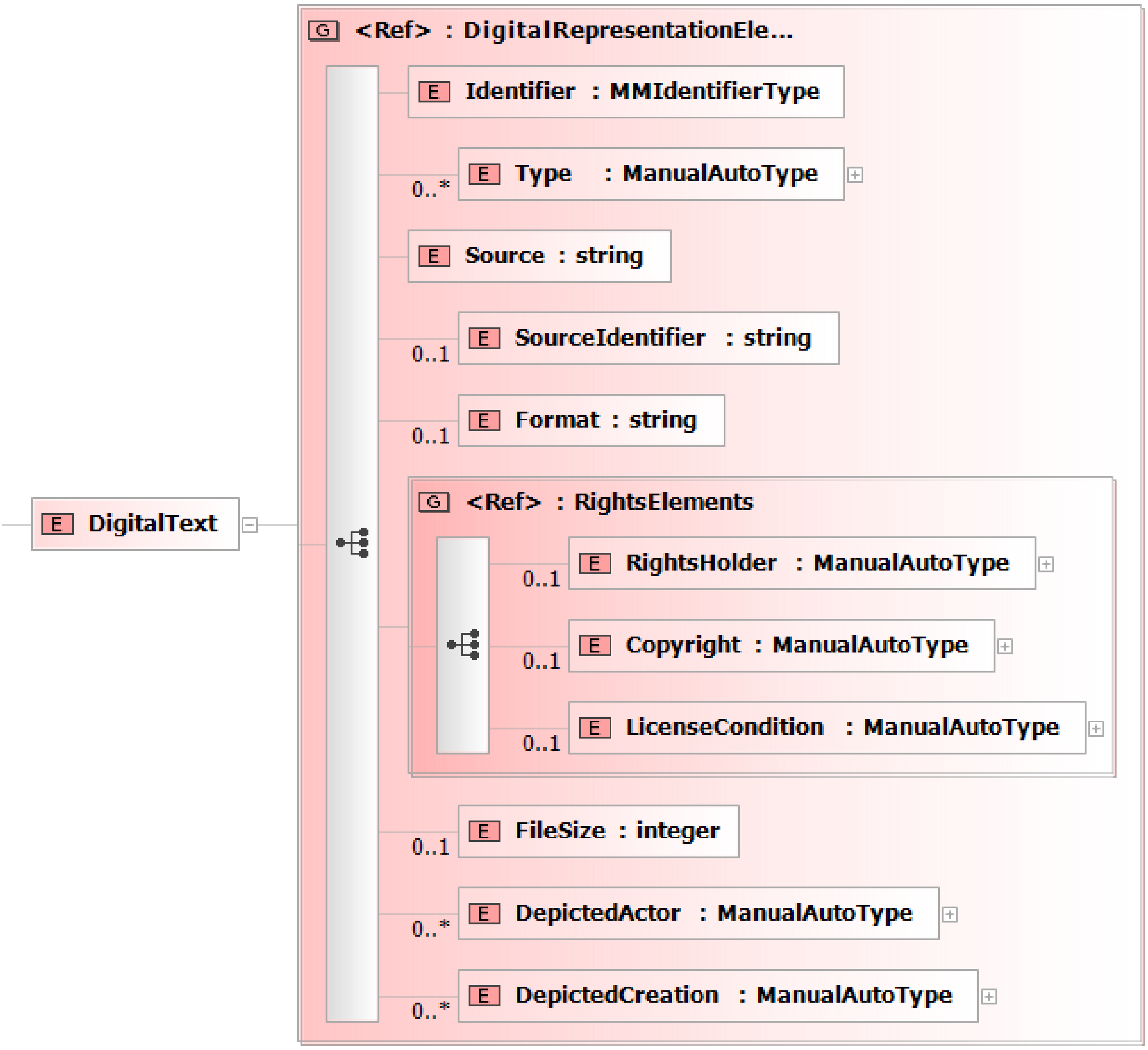
DigitalAudio Elements


E DigitalAudio
G <Ref> : DigitalAudioElements
G <Ref> : DigitalAudioVideoElements
0..1 E Streamed : boolean
0..1 E AudioChannels : integer
0..1 E SamplingRate : integer
0..1 E BitRate : integer
0..1 E FrameRate : integer
G <Ref> : DigitalRepresentationEle...
E Identifier : MMIdentifierType
0..* E Type : ManualAutoType
E Source : string
0..1 E SourceIdentifier : string
0..1 E Format : string
G <Ref> : RightsElements
0..1 E RightsHolder : ManualAutoType
0..1 E Copyright : ManualAutoType
0..1 E LicenseCondition : ManualAutoType
0..1 E FileSize : integer
0..* E DepictedActor : ManualAutoType
0..* E DepictedCreation : ManualAutoType

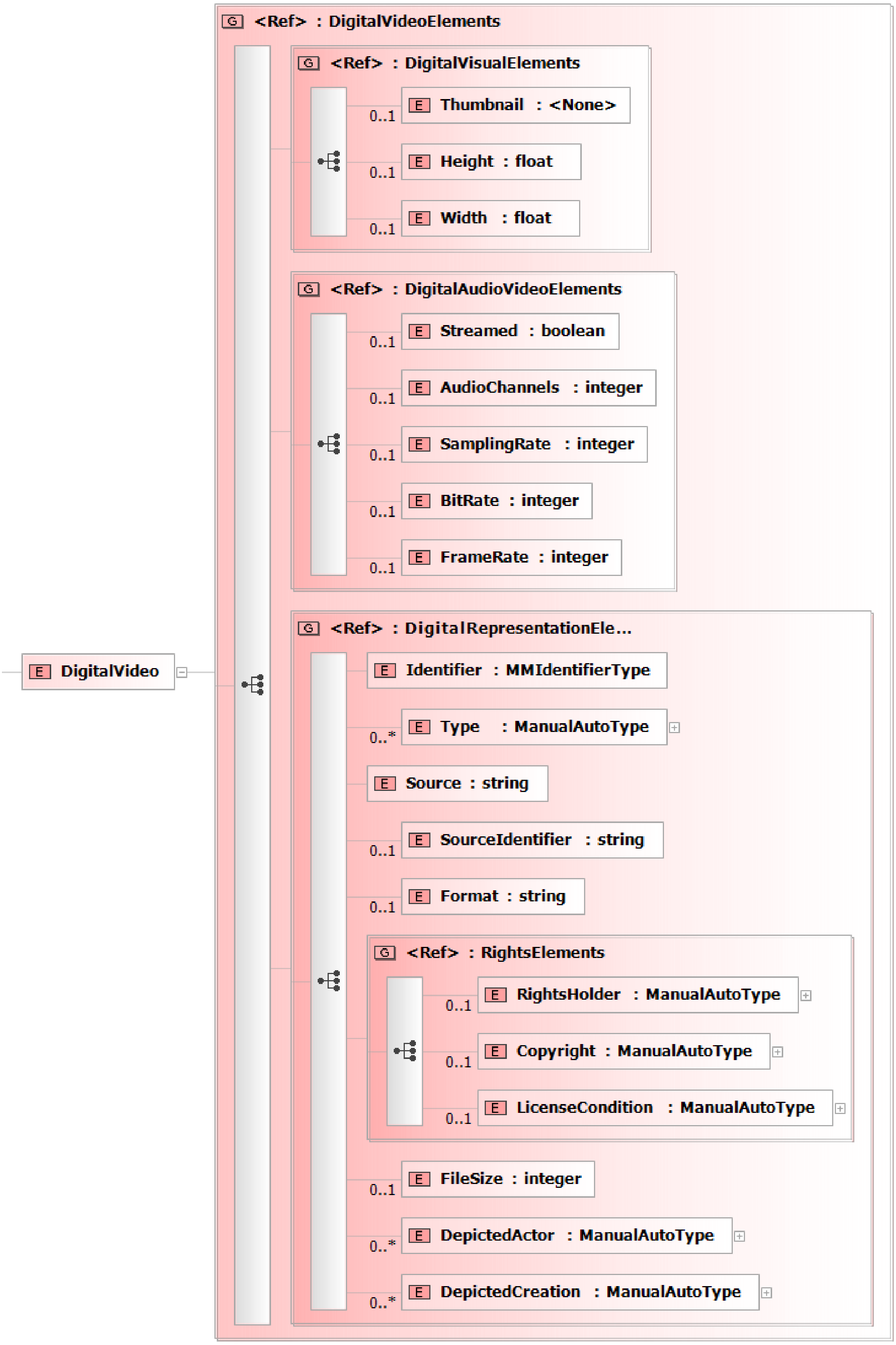
DigitalStillImage Elements


DigitalStillImage
<Ref> : DigitalStillImageElements
<Ref> : DigitalVisualElements
Thumbnail : <None>
0..1
Height : float
0..1
Width : float
0..1
<Ref> : DigitalRepresentationEle...
Identifier : MMIdentifierType
Type : ManualAutoType
0..*
Source : string
SourceIdentifier : string
0..1
Format : string
0..1
<Ref> : RightsElements
RightsHolder : ManualAutoType
0..1
Copyright : ManualAutoType
0..1
LicenseCondition : ManualAutoType
0..1
FileSize : integer
0..1
DepictedActor : ManualAutoType
0..*
DepictedCreation : ManualAutoType
0..*

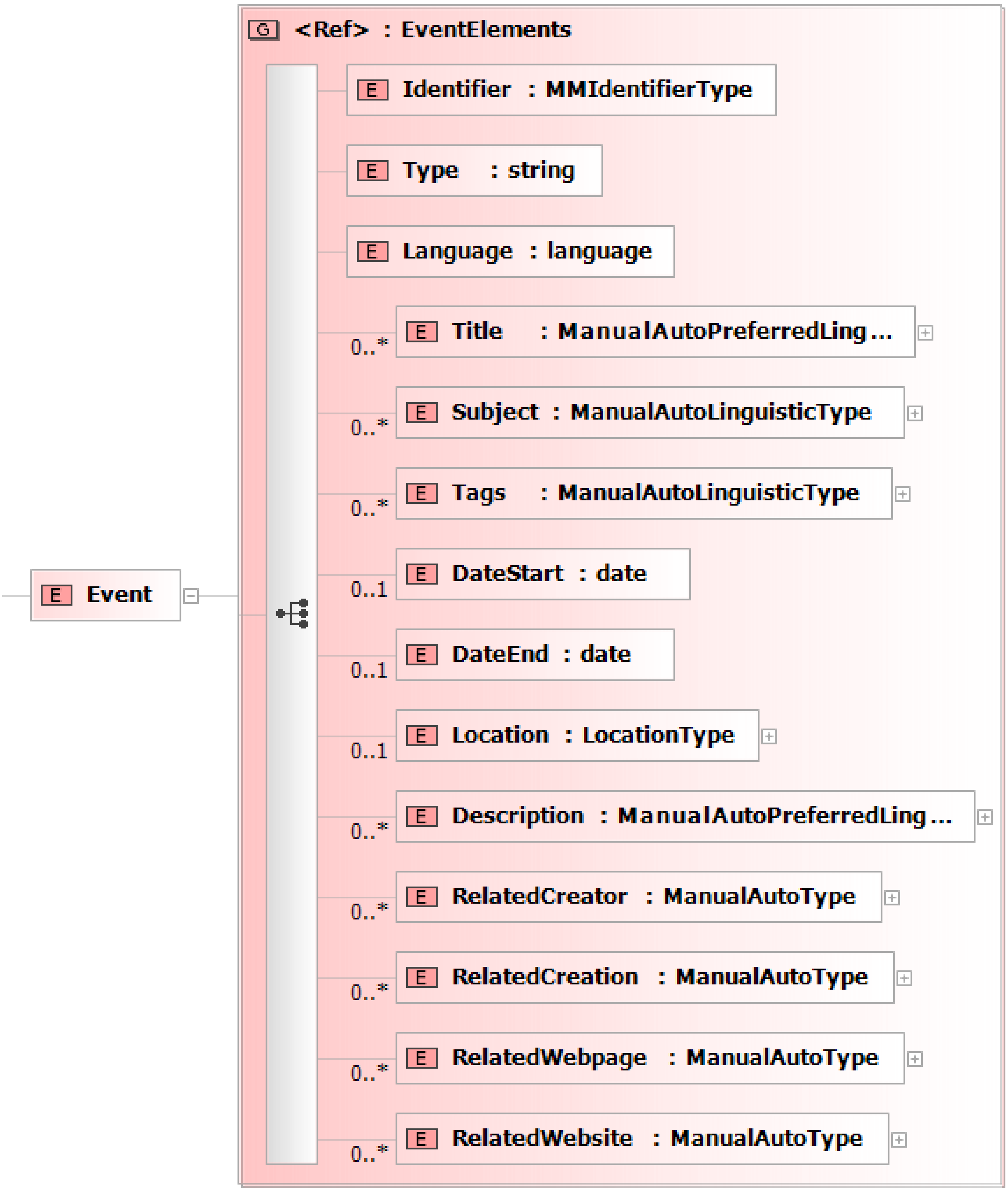
DigitalText Elements


E DigitalText
G <Ref> : DigitalRepresentationEle...
E Identifier : MMIdentifierType
0..*
E Type : ManualAutoType
E Source : string
0..1
E SourceIdentifier : string
0..1
E Format : string
G <Ref> : RightsElements
0..1
E RightsHolder : ManualAutoType
0..1
E Copyright : ManualAutoType
0..1
E LicenseCondition : ManualAutoType
0..1
E FileSize : integer
0..*
E DepictedActor : ManualAutoType
0..*
E DepictedCreation : ManualAutoType

DigitalVideo Elements


E DigitalVideo
G <Ref> : DigitalVideoElements
G <Ref> : DigitalVisualElements
E Thumbnail : <None>
0..1
E Height : float
0..1
E Width : float
0..1
G <Ref> : DigitalAudioVideoElements
E Streamed : boolean
0..1
E AudioChannels : integer
0..1
E SamplingRate : integer
0..1
E BitRate : integer
0..1
E FrameRate : integer
0..1
G <Ref> : DigitalRepresentationEle...
E Identifier : MMIdentifierType
E Type : ManualAutoType
0..*
E Source : string
E SourceIdentifier : string
0..1
E Format : string
0..1
G <Ref> : RightsElements
E RightsHolder : ManualAutoType
0..1
E Copyright : ManualAutoType
0..1
E LicenseCondition : ManualAutoType
0..1
E FileSize : integer
0..1
E DepictedActor : ManualAutoType
0..*
E DepictedCreation : ManualAutoType
0..*

Event Elements



E Event
G <Ref> : EventElements
E Identifier : MMIdentifierType
E Type : string
E Language : language
0..* E Title : ManualAutoPreferredLing...
0..* E Subject : ManualAutoLinguisticType
0..* E Tags : ManualAutoLinguisticType
0..1 E DateStart : date
0..1 E DateEnd : date
0..1 E Location : LocationType
0..* E Description : ManualAutoPreferredLing...
0..* E RelatedCreator : ManualAutoType
0..* E RelatedCreation : ManualAutoType
0..* E RelatedWebpage : ManualAutoType
0..* E RelatedWebsite : ManualAutoType