METS and the CIDOC CRM – a Comparison

Martin Doerr

February 2011

Acknowledgments

The work was commissioned and financed by Cultural Heritage Imaging (http://www.c-h-i.org) with majority funding from the US Institute of Museum and Library Services' (IMLS) National leadership Grant Program (Award Number LG-25-06-0107-06). Additional funding was provided by charitable contributions to Cultural Heritage Imaging.

1 Introduction

This report compares the functional roles of the CIDOC CRM and METS (Metadata Encoding and Transmission Standard) with respect to metadata encoding and the management of interoperability and information integration in heterogeneous, distributed Digital Library environments. It investigates the ways in which both standards can be optimally used in combination. The report addresses managers and practitioners to support decisions about the deployment of these standards and to serve as a guideline for effective use of both standards.

This report is based on the *CIDOC CRM version 5.0.2*[1], *<METS> Metadata Encoding and Transmission Standard: Primer and Reference Manual version 1.6 Revised*[2], *©2010 Digital Library Federation*, and the METS Schema version 1.9[3], as well as associated information. This report does not intend to replace the reading of those documents, but to complement the respective documentation of both standards from a functional, integrated perspective. In order to talk about data structures we use the terminology known from RDFS and XML as appropriate. We use the terms *data structure* as a generalization of *database schema* and other ways to organize digital data formally.

In order to introduce this report's major concepts, we first discuss the role of metadata and characterize the most prominent approaches currently in use or under discussion. In section 2 and 3 we describe in detail the functional roles of METS and the CRM respectively, and clarify the differences between those and other standards. Since both are designed a priori for completely different functions, we describe in section 4 how they can be used in a complementary way in practical applications, and how possible semantic conflicts can be avoided.

1.1 The role of Metadata

In this report we consider data structures relevant to the following processes and activities in the life-cycle of digital objects (see also [1]):

a) *Content creation*, such as authoring, documentation, digitization of physical items, scientific measurements or experiments.
b) *Metadata creation*, as part of the content creation process or afterwards, partially automatically or by manual documentation, also comprising content packaging or reformatting.
c) *Ingestion* into a repository for public or community access, also comprising the distribution of content to multiple locations.
d) *Indexing* contents for accessing it in a repository or digital library, by installing metadata elements as elements of a database schema for querying a database system.
e) *Harvesting* of metadata from multiple repositories in order to create homogeneous indices of distributed materials.

---

[1] This version has been submitted to ISO for the revision of ISO21127:2006 due in 2011. See http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.2.pdf
[2] http://www.loc.gov/standards/mets/, http://www.loc.gov/standards/mets/METSPrimerRevised.pdf
[3] http://www.loc.gov/standards/mets/mets.xsd

f) *Extraction and migration* of contents and metadata from one repository to another
g) *Preservation* of contents by comprehensive documentation, monitoring accessibility of formats and dependencies, issuing migration and content duplication.

In these contexts, metadata play a key role. Metadata are data about data. They are created to describe information objects kept by digital libraries or other repositories or collections. A successful era of content retrieval with powerful search engines is converging towards its technological limits. Richer metadata are now regarded as the key to improve the still unsatisfactory access methods to the knowledge hidden in ever-growing digital repositories. But metadata are far more than "finding aids". Their basic functions are to enable:

1. *Identification,* by listing identifiers, characteristics and essential properties.
2. *Use*, by describing format, structure and encoding so that the information can be accessed, decoded and presented as intended by their creators. This also includes legal regulations and licenses.
3. *Adequate interpretation,* by describing provenance, context of creation and intended use. This includes information to understand the relevance, authenticity and reliability of information objects.
4. *Finding and selection,* by summarizing relevant content features, such as subjects, references, relationships, key-words and key propositions, but also anything under item 1., 2. and 3.
5. *Preservation and future use,* by describing on which related or background information and tools the future use may depend. It includes information to control the integrity of information objects, but also anything under item 1., 2. and 3.

Beginning with the natural languages, our data and metadata, even the same facts or statements can come in thousands of different formats and encodings,. The hope is to overcome this interoperability problem by metadata everyone understands, and which contain the information needed to deal with the heterogeneous contents adequately. Therefore many organizations engage in defining metadata standards, i.e., shared data structures[4] with commonly known meanings, to foster interoperability at least on the metadata schema level. So, hundreds of metadata schemata emerged instead of one.

The metadata heterogeneity has different causes. For instance, there are metadata definitions separately dealing with selected functions and goals as the ones described above, instead of dealing with all of them in a coherent way [1,2]. This causes ambiguities where to document facts that serve multiple functions, such as if provenance and rights assessment metadata are separated. On the other side, there are clearly independent functions, such as packaging and identification, which we may find interconnected in some metadata. Further, local organizations may have gone their local ways, before standards appeared or ignoring standards. Different application sectors may encounter different requirements not served by standards, or standards comprising all sectors may become to complex to be understood. This hinders access to data relevant for other sectors. To make the situation even more

[4] Frequently called "metadata vocabularies" or even ontologies, a misleading terminology.

complex, metadata are by themselves data. So there exist "metametadata" and so forth. The metadata themselves may be used as information contents about things that have happened in the real world in their own right, for example, metadata about the processing history of digital data.

It is quite popular to regard metadata structures as "vocabularies" from different communities, as if they were natural languages, and not engineering constructs. If this were true, the elements differing from one metadata structure and to another would either have complementary meaning or be equivalent, and a simple "translation" of element names would resolve this form of heterogeneity. However, there are forms of heterogeneity that originate in functional and conceptual differences that are more difficult to address. Therefore, we are now in the situation that we need yet another complex mechanism to overcome metadata heterogeneity. Currently, the so-called Application Profiles (SCHEMAS project[5]), Dublin Core[6], METS and the CIDOC CRM [3,4] can be regarded as the most prominent attempts to overcome metadata heterogeneity. Very recently, the ORE model[7] has attracted significant attention. Each of these approaches is based on a completely different paradigm, solves different problems and serves different functions. We present them here in short:

1.1.1 Flat metadata structures and Dublin Core

We call here a data structure a "flat list of attributes", or simply "flat", if each data value is connected by one direct property (field, element, link) with the described root object. "Application profiles" aggregate flat lists of metadata properties from different applications, in the assumption that applications may need more or less of these direct properties to which the possible organization of information is restricted. The aggregation aims at combining metadata structures designed for different functions that may be simultaneously needed in one application, by merging the identical elements and aggregating the others into an extended flat list of properties. The method ignores the fact, that information may a) not be adequately represented by flat lists and b) properties may have overlapping meaning. Therefore we regard this approach as a naïve oversimplification of the more general problem of schema integration that has already been dealt with extensively in the literature of the 1980's and 1990's. In particular it requires the knowledge representation mechanisms of specialization and generalization. See for instance the work of D. Calvanese, M. Lenzerini and others [5]. The problem can more adequately be solved by the combination mechanisms that XML Schema or RDF Schema foresee based on namespaces, but in general it is even more complicated, and may require complex transformation algorithms.

The Dublin Core Metadata Element Set (DC) is a particular flat list of attributes designed as *finding aids*. It serves primarily to increase the precision of finding things (function 3. above) in a digital repository by a set of simple properties. Finding is regarded as one of the core functions a digital repository should support. The particular choice of the DC properties exhibits a library bias: for instance the name of

---

[5] SCHEMAS Registry, http://www.schemas-forum.org/registry/

[6] http://dublincore.org/

[7] Open Archives Initiative Object Reuse and Exchange, ORE Specification - Abstract Data Model 2 June 2008

a *publisher* was regarded as more important than the place of creation. It simplifies more complex structures of the things described: for instance, the property *creator* is not combined with the property *date* to describe a creation event, but the meaning of *date* remains ambiguous. This is the price of a flat data structure. Selecting "core" attributes means selecting the most frequently used/asked ones, rather than the most generic ones, i.e., the ones that would cover as many more specialized properties as possible. Therefore, for all the other properties, there exists a series of application profiles containing Dublin Core as a part, i.e., as "core".

If used as finding aids, one can argue that the flattening reduces precision but preserves recall, which is per se not bad. If the same information however should be used for other purposes, the flattening corrupts meaning significantly, and is a great obstacle to effective information integration. It should not be taken as a documentation format, except for very elementary digital objects. At least, Dublin Core is widely supported as a standard and constitutes a great progress over keyword-based-only access. The self-imposed restriction to flat metadata in large parts of the Digital Library community is hard to understand, given the sophistication of modern database engines on one side, and the rather complicated workarounds to recover from the shortcomings of flat metadata, which are characteristic for many Dublin Core extensions (see section 3).

1.1.2 METS, OAI-ORE and CIDOC CRM

METS (Metadata Encoding and Transmission Standard), a Digital Library Federation initiative, attempts to provide an XML document format for encoding metadata necessary for both management of digital library objects within a repository and exchange of such objects between repositories (or between repositories and their users). Its primary function is the *packaging* of complex digital objects, consisting of multiple parts associated with multiple metadata about the whole and its parts, into *one self-contained document*, so that it can be preserved as one integral unit, and be unpacked and understood at another place. It addresses the description of physical objects only indirectly through digital representations. It primarily addresses format and structure. All other metadata are only represented as parts of the overall document without analyzing their content. No other standard currently deals with the packaging of complex digital objects. As such, it should be complementary to all other metadata structures. In practice, it contains more information than necessary for this particular function, which causes some overlaps with other formats.

The Open Archives Initiative Object Reuse and Exchange model (**OAI-ORE)** [8] is a " data model for the description and exchange of aggregations of Web resources, named Aggregations. OAI-ORE introduces the notion of Resource Maps that describe an Aggregation. A Resource Map describes an Aggregation: it asserts the finite set of constituent resources (the Aggregated Resources) of the Aggregation, and it can express types and relationships pertaining to the Aggregation and its Aggregated Resources. This data model conforms to the concepts defined in the Architecture of the World Wide Web. The ORE Model can be implemented in a variety of

---

[8] (ORE Specification - Abstract Data Model, 17nd October 2008, http://www.openarchives.org/ore/1.0/datamodel)

serialization formats." and "A Resource Map asserts a set of RDF triples expressing information about an Aggregation, its constituent Aggregated Resources, metadata about the Aggregation and Resource Map, and other Relationships. The RDF Graph that is manifested by the triples asserted by a Resource Map MUST conform to a number of restrictions."

The ORE Model is a very simple schema, and deliberately lacks all kinds of administrational data. It describes a simple abstract semantic structure similar to that underlying METS, but does not aim at creating a container to transport the aggregated resources. Resources are referred only, rather than included. As such, it can be seen as a blueprint or part of the schema of a digital repository or metadata repository dealing with the resource aggregation. Metadata may appear in the form of Dublin Core or *any* RDF Schema. McDonough (2009) [6] describes "Aligning METS with the OAI-ORE Data Model". Basically, he concludes that not all METS features transform easily into ORE, but that METS could be restricted to conform with ORE. This appears to us as quite natural, because METS makes use of the physical containment of content for packaging and internal reference characteristic for XML, which have no equivalents in RDF. This paper does not investigate transforming ORE data into METS.

The actual RDF Schema provided by ORE [9] declares among others a class "ore"AggregatedResource", which is nothing else than a Resource that happened to be aggregated, and hence is redundant with the respective property "ore:aggregates" This feature causes problem when someone wants to integrate the ORE Schema with other ontologies. It sounds intuitive but we regard it as bad modeling.

The **CIDOC CRM** is an ontology developed by working groups of the International Committee for Documentation (CIDOC) of the International Council of Museums. It was accepted as ISO standard ISO21127:2006. It was developed by generalizing over the most prominent documentation formats in museums and archives. It describes the core concepts behind these formats by interpreting these data structures as human conceptualizations of how the things documented in an information system are thought to be related in the real world. The purpose is to have a common language to *mediate* or *translate* between different data formats or to merge complementary data with respect to their intended meaning, the most general mechanisms to deal with data structure heterogeneity. In the manner of formal ontologies, it interprets data as a semantic network of interconnected facts about the world. In the narrower sense, it is not a metadata structure by itself, but compatible metadata structures can be derived from it, even if the metadata structure may just be a one-to-one encoding of the ontology in RDFS or OWL (these would already be two different metadata structures!). In contrast to Dublin Core, it aims at *covering* most of the properties in metadata structures by adequate generalizations rather than just finding the most frequent ones. However, the generalizations it provides are limited in order not to compromise the clarity of semantics necessary to make relevant inferences. For instance, it does not accept the genericity of Dublin Core "date" [4]. Also, in contrast to many metadata standards, it does not prescribe what to document, but brings into a *homogeneous, integrated form* what *has already been documented*. As such, it is complementary to many metadata standards.

[9] http://www.openarchives.org/ore/terms

1.1.3 OAI PMH

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability. *Data Providers* are repositories that expose structured metadata via OAI-PMH. *Service Providers* then make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP[10] in order to communicate with the repositories to collect metadata. As minimal requirement, DC metadata are requested, but the repository software may create DC metadata on the fly from other, internal formats. Therefore it does not require primary documentation in DC format. Repositories may expose other metadata formats than DC for harvesting via OAI-PMH. The minimal record structure for returning metadata to the harvester functionally overlaps a bit with METS.

In this report, we compare in detail the functional roles of the CIDOC CRM and METS with respect to metadata encoding and management for interoperability and information integration in heterogeneous, distributed Digital Library environments. In particular, we investigate the ways both standards can optimally be used in combination in an effective implementation.

2 The scope and potential of METS

2.1 Scope

"Maintaining a library of digital objects of necessity requires maintaining metadata about those objects. The metadata necessary for successful management and use of digital objects is both more extensive than and different from the metadata used for managing collections of printed works and other physical materials. While a library may record descriptive metadata regarding a book in its collection, the book will not dissolve into a series of unconnected pages if the library fails to record structural metadata regarding the book's organization, nor will scholars be unable to evaluate the book's worth if the library fails to note that the book was produced using a Ryobi offset press. The same cannot be said for a digital version of the same book. Without structural metadata, the page image or text files comprising the digital work are of little use, and without technical metadata regarding the digitization process, scholars may be unsure of how accurate a reflection of the original the digital version provides. For internal management purposes, a library must have access to appropriate technical metadata in order to periodically refresh and migrate the data, ensuring the durability of valuable resources.

The Making of America II project (MOA2) attempted to address these issues in part by providing an encoding format for descriptive, administrative, and structural metadata for textual and image-based works. METS, a Digital Library Federation initiative, attempts to build upon the work of MOA2 and provide an XML document format for encoding metadata necessary for both management of digital library objects within a repository and exchange of such objects between repositories (or between repositories and their users). Depending on its use, a METS document could be used in the role of Submission Information Package (SIP), Archival Information

[10] http://www.openarchives.org/pmh/

Package (AIP), or Dissemination Information Package (DIP) within the Open Archival Information System (OAIS) Reference Model."[11]

Throughout the METS documentation, an initial bias towards digitizing paper documents is apparent. The consequences of generalizing over this scenario seem not to be consistently applied to all METS definitions.

2.2 The METS document parts.

The idea of METS is to define an aggregate of content and metadata that form a coherent whole by some criteria and to package it into a single file. The great achievement of METS is to describe such aggregates with complex interrelations in a *self-contained document*. METS is basically neutral to the metadata schema to be used, but some exceptions are discussed below. Virtually all data structure elements are optional, and unfortunately many are functionally redundant. Therefore a guide to good practice is necessary for any application of METS. METS' splitting of the metadata into five separate functional units may be cause of unnecessary inconsistencies in wider applications, particularly in e-science.

In this section, we discuss the design of the parts of a METS document. We refer to version 1.9 of the METS Schema, which contains some unsubstantial but useful extensions with respect to the version 1.6 still described in the revised METS Primer and Reference Manual. The reader is kindly asked to look up the respective parts in the original documents and to consult the extracts in the section 7.2 of this report for more details.

2.2.1 METS Header

The METS Header contains metadata about the creation of the METS document itself ("metametadata"). It describes the identifier, last modification date, creation date and agents contributing to the creation or having other roles. This violates the neutrality in other parts of METS to the metadata schema employed, but a link to an ADMID (administrative metadata section identifier) in the METS Header actually allows for using other metadata formats. Curiously enough, even the identifier for the whole document (OBJID) is not required (see also [6]). In the METS Schema version 1.9, yet another identifier for the whole document was introduced: "The metsDocument identifier element <metsDocumentID> allows a unique identifier to be assigned to the METS document itself. This may be different from the OBJID attribute value in the root <mets> element, which uniquely identifies the entire digital object represented by the METS document." This suggests that the METS container may be regarded as different from the contained objects taken as one whole. We regard this distinction as mandatory.

The <agent> "ROLE" attribute lacks definition of a temporal validity. It is not clear, if the "ROLE" is meant to be maintained during the document's life-cycle. Further, exchange of documents over the Internet requires that an agent should at least be qualified by a location, in order to be identified later. In the CIDOC CRM view, dates and agents are exclusively related via events to objects, and belong to the object

[11] METS: An Overview & Tutorial, http://www.loc.gov/standards/mets/METSOverview.v2.html

history. In this respect, we regard the METS header information as possibly insufficient. The 'last modification date' together with an identifier should probably be seen as an extended identification mechanism for a document version rather than as description of a historical event. The latter conforms to the OAI-PMH record header.

2.2.2 Metadata Sections.

The metadata sections (<dmdSec>, <amdSec>) break down into a list of distinct elements:

1. **Descriptive Metadata** (<dmdSec>) are described as those needed for "discovery". "The descriptive metadata section may point to descriptive metadata external to the METS document (e.g., a MARC record in an OPAC or an EAD finding aid maintained on a WWW server), or contain internally embedded descriptive metadata, or both. Multiple instances of both external and internal descriptive metadata may be included in the descriptive metadata section." Here the user expects to find the finding aids or other metadata used for harvesting.
2. The **Administrative Metadata** Section (<amdSec>) "contains the administrative metadata pertaining to the digital object, its component data files and any original source material from which the digital object is derived." As with descriptive metadata, administrative metadata may be either external to the METS document, or encoded internally. The metadata must be divided into:
    - 2.1 **Technical Metadata** (<techMD>) "may contain information regarding a file's creation, format, and use characteristics."
    - 2.2 **Rights Metadata** (<rightsMD>) are "used to record information about copyright and licensing."
    - 2.3 **Source Metadata (**<sourceMD>) are **"**used for associating descriptive and administrative metadata about the source format or media of the digital object being described by the METS document." It "would contain information regarding the original source" (This definition shows the bias of METS towards digitization of physical documents)
    - 2.4 **Digital Provenance Metadata** (<digiprovMD>) "can be used to record any preservation-related actions taken on the various files which comprise a digital object (e.g., those subsequent to the initial digitization of the files such as transformation or migrations) or, in the case of born digital materials, the files' creation."

The distinct feature of METS is to be able to embed a set of different standard metadata formats and even any user-defined one. This gives METS a great potential as mechanism for syntactic interoperability of complex digital objects.

In contrast to the others, descriptive metadata are restricted in use: They pertain only to the internal "<file>" and " " elements. Why not to the <fileGrp> elements? So, if multiple files want to share a "description", a element has to be introduced uniting them. "Administrative metadata" apply to <dmdSec>, header and the four parts of the <amdSec> listed above. So the administrative metadata can describe administrative metadata, which is a good thing. The reasons for the specific

logic of applying differently the kinds of metadata to parts of a METS document are not explained in the METS documents, and not obvious.

2.2.3 The content sections

We regard as a distinct achievement of METS the definition of the four content sections. It enables content to be packaged and described as self-contained transportation and preservation units.

3. **File Section** - The file section lists all files containing content which comprise the electronic versions of the digital object. <file> elements may be grouped within <fileGrp> or <file>elements, to provide for subdividing the files by object version.
4. **Structural Map** - The structural map is the heart of a METS document. It outlines a hierarchical structure for the digital library object, and links the elements of that structure to content files and metadata that pertain to each element.
5. **Structural Links** - The Structural Links section of METS allows METS creators to record the existence of hyperlinks between nodes in the hierarchy outlined in the Structural Map. This is of particular value in using METS to archive Websites.
6. **Behavior** - A behavior section can be used to associate executable behaviors with content in the METS object. Each behavior within a behavior section has an interface definition element that represents an abstract definition of the set of behaviors represented by a particular behavior section. Each behavior also has a mechanism element which identifies a module of executable code that implements and runs the behaviors defined abstractly by the interface definition. It is not clear why the "behavior" is not regarded as metadata, rather important one for Digital Preservation.

Particularly important are the mechanisms to identify a file internally and to connect its internal identifier with external identifiers. So, content internally expanded can be referred to without creating ambiguity or confusion about external copies and identifiers. The structural map has the distinct capability to organize the same content by multiple overlapping principles, another important feature. Sequencing of content parts, such as "pages", is however only supported by numbering. This is somehow limited, because in general sequencing might be represented by directed graphs (to represent "guided tours" etc.). A more general solution would be linking.

Somehow limited is the Structural Links section, because it deals only with binary links (for instance, some types of "XLinks" can described n-ary relations). There is a link title, which might be used to describe typed hyperlinks. However, more detailed linking might be described in metadata, but the METS document does not indicate in which of the five different sections it should be stored.

2.3 Critical consideration of METS elements

The METS schema generally has a very clear structure. In the documentation, motivation is given for all parts based on a running example, but the user misses a more rigorous and more generalized justification of its particular elements. We

describe here our view of some problems we find in the system of XML attributes and the division of metadata into five semantic sections.

2.3.1 Attribute problems

Names and applicability of some attributes in different elements appear unnecessarily heterogeneous and partially inconsistent:

a) A *creation date* is referred in three different ways, and not for all elements: CREATED, VERSDAT, CREATEDATE (<metsHdr> has a "createdate"; sections, files are "created", a but a <fileGrp> is "versioned" and not "created"; a is neither versioned nor created). We regard that these three forms have identical semantics. As they appear together with identifiers, there should be a consistent use of a "last modified date" logic for identification. See the <datestamp> element in OAI-PMH.xsd (the datestamp -- the date of creation, modification or deletion of the record for the purpose of selective harvesting.) All other use of *date information* should be in the metadata section(s).

b) CONTENTIDS and OWNERID: A <file> has an owner-provided id "OWNERID", but *no owner*, but the METS header is created by an "agent". <fptr> and in the structmap have "CONTENTIDS, whereas the <file> does not. There should be a richer, consistent representation of identifiers in use in the real world, each one optionally connected to an agent.

c) GROUPID: This additional grouping mechanism has no other semantics than the special cases in the document that motivate its use. There is no possibility to describe in any way what such a group means. We would suggest to *drop completely* or extend this feature.

d) RECORDSTATUS, STATUS: It is neither clear why some elements have a status and some have not, nor why a document should have a status, and why there are two different attributes for it. We regard "status" as an internal workflow feature, which should be implemented by a respective workflow manager, and not be dropped here and there into the document. If the document should be self-contained, only a general classification of the whole METS document makes sense. RECORDSTATUS however may map to the OAI-PMH.xsd element <status>! Both standards could be better harmonized in this point. An adequate mapping to the CRM would capture the relationship between the versions of the same document with different "STATUS", or classify the creation processes accordingly.

e) ORDER is competing with SEQ: why are there two different attributes?

f) SEQ, ORDER provide a linear, numbering order. There are cases of more complex ordering not captured by this model. A "n-m" link indicating the "next element" would be more flexible.

g) Note a change of interpretation from METS version 1.6 to 1.9: "The metsDocument identifier element <metsDocumentID> allows a unique identifier to be assigned to the METS document **itself**. This may be different from the OBJID attribute value in the root <mets> element, which uniquely identifies the entire digital object **represented** by the METS document."

2.3.2 Use of the metadata sections

The content for these kinds of metadata sections is based on functional distinctions and the distinction becomes mandatory by the different assignment of these metadata kinds to content types. But metadata contents cannot be separated by function. The same information may serve different functions. Therefore, in general, either information is redundantly described in different sections, or arbitrarily aggregated only into the one or another section. For instance, should be the writing and writer of a born-digital text (as this report) be documented in the "description metadata", since I would expect it to be registered in a DC record for "discovery", and "discovery" metadata should be in the "description metadata" following METS documentation. Or should it be in the digital provenance metadata, as the METS definition explicitly suggests, or both? On the other side, the documentation excludes the initial digitization from Digital Provenance. For us, the whole history of a digital object basically forms one connected graph of events and intermediate products, and spans over all the parts of the (<amdSec>. How can digital provenance be separated from "source metadata", and why should the "original source" not be subject of "description"?

Metadata guidelines may make such distinctions as a kind of subject list in order to prescribe when documentation is complete, and to structure natural language text. But when formal metadata elements are used, the different topics cannot be confused. These distinctions may have made sense for digitizing paper documents, but seem neither to be necessary for managing the metadata, nor can they be easily generalized. Even the rights metadata may overlap with digital provenance, since provenance may entail inheritance of rights. Whereas METS is in most parts quite "tolerant", using optional and alternative representations, the separation into five metadata sections is not generalized, so it cannot be overcome.

The would rather suggest the following distinctions, which are based on a causal-historical consideration: We would reserve "descriptive metadata" (<dmdSec>) for finding aids made for harvesting, and, if the digital material is about a physical item in any sense, the metadata of this item. There is a distinct event of digitizing, measuring or documenting a physical thing or feature, which starts the digital life-cycle. We would suggest using the <digiprov> element for this and all subsequent events in the digital life-cycle. This implies the reference to physical items, but not their properties and history up to the capture event, which could be in described in the <dmdSec>.

Note however, that the chain of indirection of representation can even be deeper: The Science Museum of London keeps a modern digital photo of a model of Columbus' Santa Maria made in the late 1940ies, but described the original Santa Maria from the 15th century in the metadata. Note also, that for instance data measuring ocean temperatures represent in a different sense the ocean than an image of the Mona Lisa represents the Mona Lisa. There may be a need to express these modalities, and there should be a unique place to it.

Rights metadata are distinct in the sense that they pertain to the future, i.e, what can, should or should not be done with a thing. Even though they root in the provenance, we understand that one may record ***consequences*** of the provenance and local legislation in a separate section, even though it may imply redundant documentation.

However, we could not find a good criterion to separate <sourceMD> and <techMD> from the <digiprovMD> or <dmdSec>. Alternatively, it may be much wiser to drop all those distinctions, because they ***conflict*** *with the metadata format neutrality* METS claims. So METS could be simpler, clearer, and the job to distinguish semantic of metadata parts is a question of the employed metadata schemata.

We suspect that behind these distinctions is a general information modeling problem: Many data formats are designed as a kind of questionnaire, such as "document the object, its source, its provenance, the associated rights etc." This should better be achieved with a guideline, rather than with distinct data elements. Other distinctions originate in confusing data structure with data presentation to the user. It is the job of style sheets and other software to take data structures for presentation apart into suitable units. A data format for information management should be determined by criteria of efficiency of automated management and genericity, and not ease of writing a style-sheet. Redundancy or fuzzy definitions as we encounter in this case make automated maintenance very difficult.

2.4 Potential of METS

Since METS deals with content packaging, we regard its role as enabling syntactic interoperability for information exchange, in contrast to semantic interoperability [1]. As such, it is without a competitor.

METS *manages heterogeneity of semantic metadata formats but does not resolve it.* Therefore it does not solve semantic heterogeneity. This is not a negative point. It should just be clear. A standard is the better the more concise its function is.

Information integration is supported up to the identity of the content granularity METS supports. With this focus on syntactic interoperability, METS closes a gap in the metadata landscape. No other standard describes how physical content is transported together with its descriptions. No other standard describes how the binary integrity of such an aggregation can be described. Suitable semantic metadata formats can complement METS to complete interoperability beyond the transport level. **Dublin Core** is just one option, as finding aids in the <dmdSec>.

The potential of METS can be associated with three functions:

a) As a means for a producer to submit a complete complex data set to a repository (SIP) or to transport for migration or preservation to another repository (AIP).
b) As a dissemination information package (DIP).
c) As a database schema to index and manage digital contents.

We see the potential of METS particularly in function a) and c). Applying function a), Content should be physically included in the METS container as much as possible, rather than linked. Any link to external content causes a dependency which may be broken after being received, in the near or far future. It is particularly important to describe alternative identifiers and locations of content, such that a receiver can trace and merge content again in his database if he receives multiple packages containing identical files, of if there are other copies of some content part in the world. Further,

all historical identification information helps for deciding authenticity or recovering otherwise lost content. The CHECKSUM attributes are very important for that sake. For digital preservation, it is vital to monitor the availability of all external links that cannot be avoided, in particular the "behaviors", i.e., the availability of S/W to interpret the content. It is easy to extract all external references from a METS document to implement such a function.

It may be wise to store a METS document as a whole in back-up repositories or back-up media for digital preservation, even though it may contain redundant data with respect to other METS documents, just to reduce the chance than any data loss corrupts the coherence of the documented unit. This use might be even more relevant than that of a self-contained submission package. Redundancy across different agents can also be helpful to assess authenticity.

In contrast to that, applying function b), one may choose to link to all content that is assumed to be available to the receiver at the time of sending the package. As a DIP, METS could also be used in **OAI-PMH harvesting**. In this case, no content would be expected to be physically transported. The METS header attributes LASTMODDATE, RECORDSTATUS and OBJID match with the OAI-PMH.xsd fields <status>, <datestamp>, <identifier>. The whole METS document could be transported as a community specific metadata format, as a "document map".

One problem of the current METS metadata structure is, that there is no clear place where to find finding aids for the METS document as a whole, such as a **Dublin Core** record. Perhaps it should be the <dmdSec> linked from the first element in the <structMap>, or a <dmdSec> without any link to it?

In both functions, the receiver would "unpack" the METS document and map its elements to the database structures his/her repository foresees to provide detailed access and management to such a complex object. Characteristically, it would enable browsing through hierarchical content organization, hyperlinks and *enable queries on semantic metadata elements* etc. This has to be done with suitable *ingest tools*. To have a standard input format such as METS is a great thing.

With suitable XML tools, even the METS document itself can be seen as a rudimentary database. Native XML databases would even allow for implementing the METS schema right away as database schema. So, in principle, METS could also be used as database schema, i.e. the function c) above. We however maintain that METS is **not suited** for this use, because consistent management of redundant content and support of the workflow of administrative processes needs other and more generic structures. We would recommend to drop all STATUS information from the METS schema. They could easily be overlaid with an appropriate XML schema in another namespace. It is not helpful to overload very relevant functions with half-hearted attempts of functionality others can better provide.

Revision of the separation of the metadata sections as suggested above would be helpful, and would give METS even more power and a clearer layering of content composition (syntax) and metadata (semantics).

The ORE model may be a candidate as a core database schema of "aggregation" and a competing model as DIP format:

2.4.1 METS and ORE

The ORE model is a very simple schema defined in RDF. This means, it forms a semantic network of assertions about resources. It *does not* deal with physically transporting content. The basic unit is the "Resource Map". As with METS, resources representing content are hierarchically "aggregated". In addition to METS, arbitrary graphs denoting sequences of resources can be formed (using "proxy" nodes). Further to the definition of the structure of the aggregation(s), The Resource Map describes the aggregation by a set of RDF statements that play the role of metadata about the aggregation. This corresponds to a union of all metadata sections in a METS document. As with METS, the Resource Map has metadata that correspond to the METS header.

This set of statement is, to say so, the "content" of the Resource Map, which can be implemented as an RDF Named Graph, a new RDF construct. The problem of semantic networks was the lack of an efficient reification construct, i.e., the possibility to express that a number of statements are "from" some source. It is only possible to link identifiers of objects to the source, not the statements. In a nested container model such as METS, this problem does not exist. The Named Graph mechanism now solves this problem and assigns to a set of RDF statements an identity, such that its provenance can be represented even though all statements from all contributors form one connected network in an RDF triple store. It basically closes the gap between a document view and a view of the integrated knowledge from all documents in a database.

Via so-called "proxy-nodes", arbitrary statements can be made that hold only relative to an aggregation. This is particularly useful in order to describe archival aggregation of objects that may have belonged or belong to other aggregations or wholes as well. However, the mechanism is relatively data intensive, and to our opinion should not be used to replace Named Graphs in order to register the source of a statement. METS has no such problem, because XML elements are by definition local.

In a database, information must be analyzed into its elements in order to be manageable. Therefore a semantic network view is superior to a document view in many respects, but reification and management of local units of knowledge remained for long time a problem. Since about 2008, scalable "RDF Triple Store" databases exist that provide an effective "Named Graph" or "context" mechanism, which basically solves to problem of updating a semantic network with larger units of knowledge. *Therefore ORE could be seen as a core database schema to store METS objects*. In [6] a mapping of METS to ORE is described, and we can imagine future implementations of ORE-based schemata to provide effective database functions for handling METS documents. Even though the ORE model can be seen as an interesting attempt to model rigorously a basic structure as the one underlying METS, mature database implementations will still need some time to appear.

The title of ORE refers to "reuse and exchange", but it seems the exchange function is limited to information about objects, not the objects themselves. XML encodings of

ORE instance data may evolve to efficient DIP packages, as “document maps”, even though ORE currently makes no provisions for packaging physical content nor for particular “submission” metadata. In our view, ORE might be a core model of aggregation for a metadata repository and METS a transport and exchange format, which both can go hand in hand, but are not effective to replace each other despite other claims in their documentation. Moreover, both METS and ORE are not concerned with semantic interoperability of metadata beyond the aggregation aspect.

3 The scope and potential of the CRM

The CIDOC CRM is a core ontology designed for schema integration of metadata, or in other words, for semantic interoperability of data structures, in contrast to handling content objects. It concentrates on the global relationships that may connect information elements in different information systems, so that they can be merged into coherent units of knowledge, and does not concern the terminology that typically appears as data in metadata formats.

3.1 History and scope

The CIDOC CRM is a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. It was developed by interdisciplinary teams of experts, coming from fields as diverse as computer science, archaeology, museum curation, history of arts, natural history, library science, physics and philosophy, under the aegis of the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). It started ‘bottom up’, by re-engineering and integrating the semantic contents of the most relevant and widely used database schemata and documentation structures from all kinds of museum disciplines, archives and, more recently, also from libraries.

The very first schema analyzed in 1996, the CIDOC Relational Data Model with more than 400 tables [7], was reduced in 1996 to a model of about 50 classes and 60 properties, with a far wider applicability than the original schema. Now, the model contains 86 classes and 137 properties, representing generically the semantics of hundreds of schemata. The development team applied strict principles to admit only concepts that serve *the function of global information integration*, and other, more philosophical restrictions about the kinds of discourse to be supported (for more detail see [4]).

The application of these principles was successful in two ways. First, the model became very compact without compromising adequacy. Second, the more schemata were analyzed, the fewer changes were needed in the model (see version history of the CIDOC CRM[12]). This experience convinced CIDOC in 2000 to begin the ISO standardization process and the model was accepted in Sept. 2006 as ISO21127:2006. The current version 5.0 has been submitted to ISO for the due revision of the standard in 2011.

The ABC Harmony model, a competitive core model developed independently by the digital library and multimedia communities, was harmonized with the CRM in 2001

[12] See http://cidoc.ics.forth.gr

[8], enriching the CRM with some abstractions of material and immaterial things. Between 2003 and 2009, the model of library concepts maintained by the International Federation of Library Associations (IFLA), the FRBR model [9], has been formulated as a specialization[13] of the CRM [10,11]. It required minor adaptations of the CRM itself, in order not to compromise the genuine library conceptualization. This ease of convergence, even with models from new domains, is an encouraging evidence that the CRM captures nearly generic concepts far beyond its originally limited scope (there has been reuse of the CRM in several other domains as well).

Three ideas are central to the CRM:

a) The relationship between entities and the identifiers that are used to refer to the entities, and the ambiguity of reference, are part of the historical reality that has to be documented rather than be resolved in advance. Therefore, the CRM distinguishes nodes representing real-world items from nodes representing names per se.
b) Types and classification systems are not only a means of structuring information about reality from an external point of view, but also part of the historical reality itself in the sense of human inventions. Similarly, all documentation is seen as part of a reality, and may be described coherently together with the documented content.
c) A characteristic way to analyse the past is to divide it up into discrete events. The past as it is documented can be formulated as events involving "Persistent Items" (continuants or endurants) , both material (Caesar the Roman, Lucy the hominid) and immaterial (The Empire, Hominid). Material and immaterial items have the potential to be present in events. Immaterial items are regarded to be present through physical information carriers.

From this point of view, a picture emerges of history as a network of lifelines of persistent items in space-time that meet each other in events (fig.1). This abstraction turns out to be extraordinary powerful. Many intuitive relationships are analyzed in terms of events, such as "has creator" or "has origin". With a minimal schema, there arise a surprising wealth of inferences and any event can be described by the CRM. For instance:

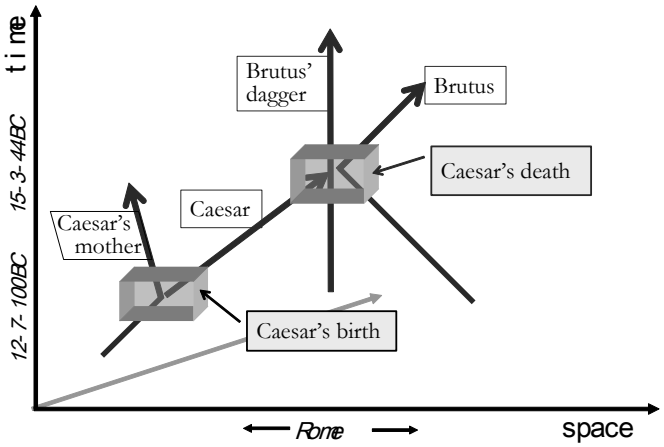


Figure 1: Historical events as meetings of things and people

[13] http://cidoc.ics.forth.gr/frbr_inro.html

Complex genetic family relations can be represented by birth events from a father and a mother. Chronologies can be justified by causal ordering of events [12]. Experimental knowledge in the sciences is gained by actual human experiments, which are carried out by individuals and teams of researchers in space/time and can be documented as events, etc. In all metadata stored in libraries, including digital libraries, is an embedded historical perspective that can be described by events. Recently, the Europeana Digital Library has adopted this core event model [13].

From its practical scope, i.e., the data structures that served as empirical base for its development, the CIDOC model inherits a focus on the material history of discrete, mesoscopic things, whereas the harmonization with FRBR introduced the notion of intellectual derivation. However, the CRM does not analyze particular internal structures of objects beyond a generic part-whole relationship. This is a field for specializations. The CRM is also not concerned with planning the future, as for instance engineering data or exhibition planning do. This aspect of reality does not appear in metadata.

3.2 Potential of the CRM

The CRM helps resolving semantic heterogeneity of metadata formats but does not manage it.

In contrast to ER models and other traditional data structures, an ontology describes the world referred to by data and data structures in an information system, rather than being a data structure itself. It describes how the different things, concepts and processes in a "domain of discourse" can be related. Since the ontology is described in a formal or objective way, it can be used to discuss which information elements a system should have, and how they should be connected, in order to create an effective information system that allows for managing a specific task. Thus the ontology may be more detailed than it is necessary for a particular information system. This richness provides a basis for deciding what the consequences are of neglecting parts of the possible information. These consequences can be formulated in terms of the questions the resulting system will be able to answer or not.

On the other hand, an ontology does not prescribe particular values or any completeness of documented data. Therefore, if the CRM is to be used for primary documentation, for instance, in RDF syntax, it should be complemented by prescriptions of necessary or preferred content. Alternatively, the prescriptions might be formulated via an XML document structure used to capture the primary data, which can afterwards be transformed into a CRM compatible form, such as the LIDO[14] format. Such prescriptions are highly disciplined and application specific, and therefore not part of the CRM specifications.

Further, an ontology is arranged in hierarchies or levels of generalisation. This allows for recognising optimal simplifications of seemingly unrelated information elements. Based on a suitable selection of CRM concepts, one can implement very simple

[14] http://www.**lido**-schema.org

information systems that still represent all the key features [3,13]. Furthermore, one can use the CRM in order to compare two different information systems and decide which is more effective for a particular task by examining the questions they can answer.

The CRM is a "core ontology". It provides the base schema for the integrated information that can support the evaluation of historical records and scientific observation for building hypotheses about individual facts about the past and for induction of hypotheses about categorical behavior. In these areas, there is a distinct level of mesoscopic entities observed and handled by people. We directly handle only things that can be touched. We do not handle bacteria, but we handle microscopes, samples and datasets. This is the key to define a "core" level of basic concepts and relationships, and to separate this level from specialized terminology and detailed object structures.

Based on these considerations, we distinguish four kinds of use of the CRM:

- As a virtual schema for transforming data from one metadata format to another (so-called "mapping)
- As a virtual or materialized global schema in information integration systems (mediators, data warehouses etc.). Note, that the object-oriented formalism underlying RDF/OWL is the only one known that allows for integrating multiple conceptual models into one, while still preserving the identity of the constituent parts.
- As an intellectual guide to develop good data structures, typically extensions and simplifications of the model combined with prescriptions about the desired content, such as schemata for describing inscriptions and transcriptions, art conservation activities, bibliographic data, empirical provenance of digital data, etc.
- As a formal core schema and framework to define sets of compatible metadata schemata.

In this paper, we are particularly interested in the last function. Immediate access compatibility of two different metadata formats is achieved, if the data are given in a metadata format which is a specialization in the object-oriented sense of the one by which we query (i.e., more detailed). Then, all queries and functions developed for the more general one apply to the more specialized one as well.

For a detailed description of the kinds of compatibility with the CIDOC CRM see section "Compatibility with the CRM" in http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.2.pdf. The document distinguishes the ability of a data structure to be source or target of a loss-less automated data transformation and the ability of an information system to ingest and then query adequately data in certain formats. We repeat here four key definitions:

1. "**A data structure is *export-compatible* with the CRM** if it is possible to transform any data from this data structure into a CRM-compatible form *without loss of meaning*. Implicit concepts may be present in elements of the data structure that are not supported by the CRM. As long as these concepts can be encoded as instances of E55 Type (i.e. as terminology) and attached

unambiguously to their respective data items with suitable properties, the data structure is *still regarded as* export compatible."

2. **"A data structure is *import-compatible* with the CRM** if it is possible to automatically transform any data from a CRM-compatible form into this data structure *without loss of meaning*, simply on the basis of knowledge about the data structure elements being used. This implies that a data record transformed into this data structure from a CRM-compatible form can be transformed back into the CRM-compatible form *without loss of meaning*." Obviously, the METS metadata sections are CRM import compatible in a trivial way.

3. **"An information system is *import-compatible* with the CRM** if it is possible to import data encoded in a CRM-compatible form and to access the data in a manner equivalent to and homogeneous with all generic data of this system that fall under the same concepts. This capability is considered as the normal kind of CRM compatibility for *integrated access systems* that physically copy source data in a *data warehouse* style (materialized access systems)."

4. **"An information system is *access-compatible* with the CRM** if it is possible to access the user data in the information system by querying with CRM classes and properties so that the meaning of the answers to the queries corresponds to the query terms used. It is not regarded as a reduction of compatibility if access is limited to data deemed to be exchanged."

For this purpose, the CRM Special Interest Group has developed a detailed data structure, an **RDF Schema in XML encoding** which comprises all CRM concepts and properties. There is also a definition of a minimal CRM subset for "partial import compatibility". A respective OWL form is also available. Both can be used as metadata schema. RDF/OWL is not a necessity, one can also use nested XML structures that uniquely map to the CRM, such as LIDO. Further, there is a proposal of a format encoding a minimal set of concepts, the **CRM Core metadata element set**[15]. It is an XML DTD of only 18 elements that is CRM export compatible, when suitable "type" and "role" values are chosen. It is also Dublin Core compatible.

Minimal compatibility with the CRM requires representation of events. In terms of data structures, this means that **at least one** level of nesting or indirection is required. A field like Dublin Core "dc:date" can be interpreted as the date of an unknown event the object "was present at". By virtue of this interpretation, it can be imported into a CRM compatible form.

Note, that **no** single data structure is identical with the CRM, but they are all applications of the CRM. The CRM is not a data schema. Even an RDFS encoding must make system-dependent interpretations of "primitive values" such as time expressions, which restrict the general ontological-scientific understanding of time. Other non-ontological features of RDF are the confusion of the item itself with its URI and the lack of an implicit inverse of any property.

Depending on the requirements, users may develop or choose different metadata formats by specialization or selection or use a CRM compatible database schema in a

[15] http://www.cidoc-crm.org/working_editions_cidoc.html

repository to index content, and still enjoy access and/or import compatibility, given suitable management S/W is in place. This flexibility is the unique result of choosing one underlying ontology (here the CRM) in an encoding-neutral form, rather than a data structures, like RDFS, XML or whatever, albeit that many people now call RDFS files "ontologies". During its life-cycle, the CRM has "survived" many KR encoding languages, without any need to change its form in order to be compatible with any of them.

3.2.1 CRM, Dublin Core, ORE, EDM and OAI-PMH

The **Dublin Core Metadata Element Set** is a finding aid of frequent associations people may make to look for content - content they typically know that it exists. It has only loose connections to an underlying reality, in particular, because the "dc:date" field is underspecified, and no place can be associated with a creation. These associations can easily be produced from a CRM compatible form. Therefore, the CRM can be used as intermediate representation to map to Dublin Core, if necessary. Dublin Core records can be mapped to the CRM, but in case of multiple dates and agents it is undecidable how these information elements would combine in the real world. This is a weakness of Dublin Core, which should not be mistaken as documentation format.

CRM Core provides a far more powerful "core form", because it preserves an elementary event structure, being only slightly more complex than DC, but extremely small compared to the "full" CRM. It could replace Dublin Core in most applications and is more appropriate for cultural-historical material and scientific observations. The more specialized fields of Dublin Core, such as Format, or Source, could be easily added as specializations to CRM Core, if regarded necessary. This is slightly more complex, but is much more ontologically precise.

ORE can be seen as a complete (small) subset of the CRM, except for the sequencing links for document parts. The CRM currently regards sequencing of content as an interpretation of identifiers, exactly as the current treatment in METS. Representation of sequence in the form of links implies a part-of relation, which is also represented in the CRM. A named graph can be seen as specialization of *E73 Information Object*. In this sense, the ORE model can be seen intellectually as a specialization of the CRM, and its encoding in RDF as a respective application.

On the other hand, if a serialization of ORE is used as a DIP, *any CRM compatible metadata* can be transported with it. (A previous version of ORE foresaw only "flat" metadata. It was not possible to use this version to describe CRM metadata). But, missing implementations of Named Graphs currently hinders its use.

The **EDM** model [13] is the core schema foreseen for the 2011 release of the Europeana Digital Library, a huge metadata repository for the European cultural heritage. It is an "integrated access system" in the sense of the CRM. The EDM integrates concepts from ORE, DC and core concepts from the CRM, and generalizes even over all of these metadata standards with some new properties for access purposes. Even though the schema is very small, it will be "access compatible" with

the CRM. It allows for querying a basic event representation, and the repository will allow to ingest and preserve original CRM metadata in RDF encoding.

Any particular CRM compatible metadata encoding can be chosen as community metadata format of **OAI-PMH**, which implies the possibility to create a Dublin Core projection of such metadata, the minimal compatibility requirement of OAI-PMH.

4 Possibilities of combined use of METS and CRM

As we alluded to above, METS manages heterogeneity of semantic metadata formats but does not resolve it, and the CRM helps resolving semantic heterogeneity of metadata formats but does not manage it. Both are complementary, and can be used together in two ways, depending on the kind of use: A METS document may contain metadata in CRM compatible form, or the information content of a METS document may be represented in CRM compatible form.

4.1 CRM within METS

For the task of creating self-contained Submission Information Package (SIP) or a Dissemination Information Package (DIP), METS appears currently to be one of the best solutions. The CRM does not transport content and content structure information and handle identifiers inside such packages. The CRM addresses semantic interoperability of metadata. CRM-compatible metadata formats, be it selections or specializations of CRM concepts, currently provide the highest quality and flexibility of semantic interoperability.

Besides a list of well-known metadata formats that METS lists and endorses explicitly, i.e., “MARC, MODS, EAD, DC, NISOIMG, LC-AV, VRA, TEIHDR, DDI, FGDC, LOM, PREMIS, TEXTMD, METSRIGHTS, NAP”, METS foresees metadata to be in any other XML format (“OTHERMDTYPE”). Here we can specify a CRM compatible format, such as the XML encoded RDF Schema http://www.cidoc-crm.org/rdfs/cidoc_crm_v5.0.1.rdfs.

CRM compatible metadata form a contiguous network of statements expanding history, context and structure of things, rather than aggregates of attribute lists. As discussed in section 2.3.2, such metadata may be foreseen by METS to appear in different sections in a METS document, but then there will be redundancy between the contents of the different sections. In order to reduce the redundancy and merge metadata on ingest, we suggest restricting CRM metadata to the Descriptive Metadata Section and the <digiprovMD> element, following the rules described in section 2.3.2, and to leave <techMD>, <sourceMD> empty. The content intended for the latter should instead be expressed in the Descriptive Metadata Section and the <digiprovMD> element.. If the various date attributes (see section 2.3.1) in METS are used, another merging of metadata facts will be necessary during the ingest process.

The CRM does not analyze rights in detail, as this has to do with the future, and the future falls out of the scope of the CRM. In the framework of the European funded project CASPAR, elaborate CRM compatible rights metadata in RDF have been

proposed [14], but other formats may be used for the rights descriptions. Possible overlaps with digital provenance data should be avoided or harmonized.

Since CRM compatible metadata form a contiguous network of statements, they may describe complex relationships between the content parts of a METS document, such as those between input, output and processing parameter sets in scientific workflow, or cultural-historical transfer of ideas etc. For that purpose it will be useful to make suitable conventions how a URI referred to as subject or object in an RDF triple may refer to the internal IDs of contents in a METS document. CRM compatible metadata are particularly powerful to declare the connection of content with multiple identifiers used by different agents in different contexts.

It is possible to use multiple CRM compatible metadata formats in the same document, and METS can express such a fact, but normally one should try to exploit the integration capabilities of CRM compatible formats to merge such metadata into one larger common format.

If the receiver of such an SIP or DIP has the suitable repository software, he/she can make the content of the metadata accessible as one merged network of knowledge, which can be based on a CRM-compatible schema (see below). From such a merged state, content and metadata can be rearranged, reaggregated or annotated. In this way, one can make the optimal use of CRM compatibility. Besides that, it is advisable to preserve a self-contained version of the received documents in a physical or virtual form, which might be the original METS encoding or a repository-internal equivalent of it, an AIP in the proper sense.

4.2 METS under the CRM

From another point of view, everything that METS adds to content can be regarded as facts about the content in the sense that the content items are objects in our universe of discourse and the real world. The partial mapping of METS to the CRM described in the following section demonstrates that the CRM can, relative to its level of abstraction, more or less completely describe the meaning of a METS document.

There are two areas that the CRM does not particularly analyze:

a) “behavior”: From a CRM point of view, compatibility with tools is a question of classification, which the CRM describes in a generic way. It is feasible, and not difficult, to describe all METS elements adequately as a specialization of the CRM. This view can be useful to develop a generic repository schema, in order to provide full access to all knowledge expressed in metadata, even the container information.
b) “Locations” in an Information Object: METS attempts to define a generic <area> element, and some other location notions. It is however far from genericity. For instance, it cannot capture 3D or 4D (such as movie segments in time and image) areas. A location notion on an Information Object depends on a virtual space, which can be defined in quite different, mutually incompatible forms. The varying interpretations of HTML “range” are a good example for that. Therefore the CRM does not have a location concept in Information Objects, but models two different notions of parts, which are much less unambiguous, a semantic component, such a paragraph or a tune in

a song, and a syntactic, such as a cluster of bytes or characters. “Areas” in the sense of METS can then be identified in the CRM with the respective denoted parts. Nevertheless, a CRM based repository may choose to extend the CRM with “areas”.

Since the CRM has been developed by generalizing over many data structures and database schemata, it is particularly suited to be expanded into a *repository schema* to manage integrated access to the semantics of the metadata and the general content structures.
In order to develop a full-fledged repository schema from the CRM, one would need to add:

- Specializations about document structures relevant to repository functions, such as update, browsing, display etc.
- Structures for epistemological functions, such as co-reference, reification and units of knowledge, which apply uniformly to all CRM constructs and are mainly a technical problem. Therefore the CRM does not deal with it as a matter of functional purity.
- Workflow structures, i.e., management processes and states, which are also outside the scope of the CRM.

METS contains details and sporadic references to some of the above functions, but not in any systematic way detailed enough to make it a management schema. ORE contains a restricted proposal for reification, but misses the co-reference problem. ORE and METS do not have any model how metadata elements intellectually connect to the “resources” the model deals with as objects of discourse.

Only the CRM provides a historically valid view of how metadata information and content are coherent parts of the same universe of discourse. For instance, in a scientific experiment, images of the instrumental set-up may be part of the data package itself. A scientific publication may contain many details about its making. A library or museum activity on an object may be integral part of the object history, i.e. things like making a METS document may appear as digital provenance in the next step. However, a standard repository model with satisfactory epistemological functionality to deal with these issues is still to be developed.

4.3 Mappings METS to the CRM

In this section, we describe an abbreviated mapping of elements and selected attributes of Appendix B of the METS Primer and Reference Manual version 1.6 to the CIDOC CRM, version 4.2.5, which is listed in this report in the Appendix in section 7.2.

A document structure such as METS can be mapped in two ways to the CRM: as an object or as an equivalent. As an object, a METS document is a document (“E31 Document” in the CRM, and all its elements are information objects, part of the overall document. This view is correct, but does not tell us what the meaning of these elements is. Some elements in an XML document have no other meaning. Those are the containers for a particular kind of information, without being such information. To

this category belong the different “sections” (<dmdSec>) etc. They are all mapped to the same class.

For all other elements we do not refer to this “trivial” mapping, but the mapping to the respective equivalent. For instance, for the element <agent> we refer the mapping to “E39 Actor”, and not to “E31 Document”, even though the <agent> element is a part of the overall document.

We do not map the internal XML identification attribute “ID”, because in a CRM representation all elements would be globally identified by a key or URI. In the section 7.2 we refer by “*” to this mapping.

We provide a mapping for all types, key elements and all of their attributes. For some of the leave elements we do not provide a detailed mapping of their attributes. We assess however that all METS elements can be mapped to the CRM, i.e., all elements and attributes correspond to a suitable abstraction in the CRM, and all nesting of elements and attributes corresponds to property paths in the CRM. In order to capture adequately the more specialized notions, a particular vocabulary of respective METS types has to be used. Instances of such a vocabulary are referred to in double quotes, such as: E31 Document. *P2F has type:“METS:metsType”*.

Starting with this mapping, an algorithm could be defined that transforms a METS document automatically into an equivalent RDF or OWL instance.

5 Conclusions

In this report, we have compared the functional roles of the CIDOC CRM and METS (Metadata Encoding and Transmission Standard) with respect to metadata encoding and the management of interoperability and information integration in heterogeneous, distributed Digital Library environments. We have also described the ways in which both standards can optimally be used in combination, depending on the situation and requirements.
We find METS particularly strong in the role of Submission Information Package syntax and in the preservation of identity of composite digital objects (AIP). The CRM is particularly strong in the field of flexible definition of semantically interoperable metadata and as a blueprint for the semantic part of a repository schema. Together, both make an ideal pair with multiple, complementary roles.

Under the experience of this comparison and the analysis of wider application cases than those that seem to be initially envisaged by METS, such as data from scientific observation, we have suggested some improvements to METS.

Dublin Core metadata play a peripheral role in this scenario. If someone wants them, they can be handled by METS and can be generated from CRM compatible metadata. As a core format for documentation or a repository schema they appear to be inappropriate. As finding aids they do a very good job for documents. For cultural objects, better core formats, such as LIDO, can be found.

The ORE model is a small but interesting abstraction, which could show ways how to simplify METS, but is far from being a full repository model or metadata exchange model.

Finally, METS and CRM compatible metadata can be operated with OAI-PMH.

6 Acknowledgement

This study was sponsored by Cultural Heritage Imaging, California, USA.

7 References

[1] PATEL, M., KOCH, T., DOERR, M., TSINARAKI, C., GIOLDASIS, N., GOLUB, K., AND TUDHOPE, D. 2005. Semantic Interoperability in Digital Library Systems, DELOS Network of Excellence on Digital Libraries – deliverable 5.3.1, June 2005. http://delos-wp5.ukoln.ac.uk/project-outcomes/SI-in-DLs/ (accessed July 2008).

[2] MANJULA PATEL, MARTIN WHITE, NIKOLAOS MOURKOUSSIS, KRZYSZTOF WALCZAK, RAFAL WOJCIECHOWSKI, JACEK CHMIELEWSKI. 2005. Metadata Requirements for Digital Museum Environments. In: *International Journal of Digital Libraries* 5(3) May 2005, Special Issue on the Digital Museum.

[3] DOERR, M., & IORIZZO, D. 2008. The dream of a global knowledge network-A new approach. *Journal for Computing and Cultural Heritage, ACM*, New York, NY, USA, 1 (1), 1-23.

[4] Doerr, M. 2003. The CIDOC CRM - An Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, 24(3).

[5] CALVANESE, D., GIACOMO, G., LENZERINI, M., NARDI, D. AND ROSATI, R. 1998. Description Logic Framework for Information Integration. In *Proceedings of the 6th International Conference on the Principles of Knowledge Representation and Reasoning* (KR'98), 2-13

[6] MCDONOUGH, JEROME P. 2009. Aligning METS with the OAI-ORE Data Model. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, Austin, TX, June 15-19, 2009. New York: Association for Computing Machinery.

[7] REED, P-A. 1995. CIDOC Relational Data Model, A Guide. *http://www.willpowerinfo.myby.co.uk/cidoc/model/relational.model/datamodel.pdf*, (accessed July, 2008)

[8] Doerr M., Hunter, J., and Lagoze C. 2003. Towards a Core Ontology for Information Integration. *Journal of Digital Information, Volume 4 Issue 1Article No. 169*.

[9] P. LeBoeuf Ed. 2005. *Functional Requirements for Bibliographic Records (FRBR): Hype or Cure-All?*. Haworth Press, Inc, ISBN: 0789027984.

[10] DOERR, .M, AND LEBOEUF, P. 2006. Modelling Intellectual Processes: The FRBR - CRM Harmonization. In *Conference Proceedings of ICOM-CIDOC annual meeting, Museum of World Culture,* Gothenburg, Sweden, 10-14 September 2006, ISBN 91-85222-12-7.

[11] DOERR, .M, LEBOEUF, P. AND BEKIARI, C. 2008. $FRBR_{OO}$, a Conceptual Model for Performing Arts. To appear in: *Conference Proceedings of ICOM-CIDOC annual meeting, Museum of World Culture,* Gothenburg, Sweden, 14-18 September 2008.

[12] Doerr, M., Plexousakis, D., Kopaka, K., AND Bekiari, C. 2004. Supporting Chronological Reasoning in Archaeology. In *Proceedings of Computer Applications and Quantitative Methods in Archaeology Conference, CAA2004,* Prato, Italy, 13-17 April, 2004. At http://www.ics.forth.gr/isl/publications/paperlink/caa2004_supporting_chronological_reasoning.pdf, accessed Nov.16, 2006.

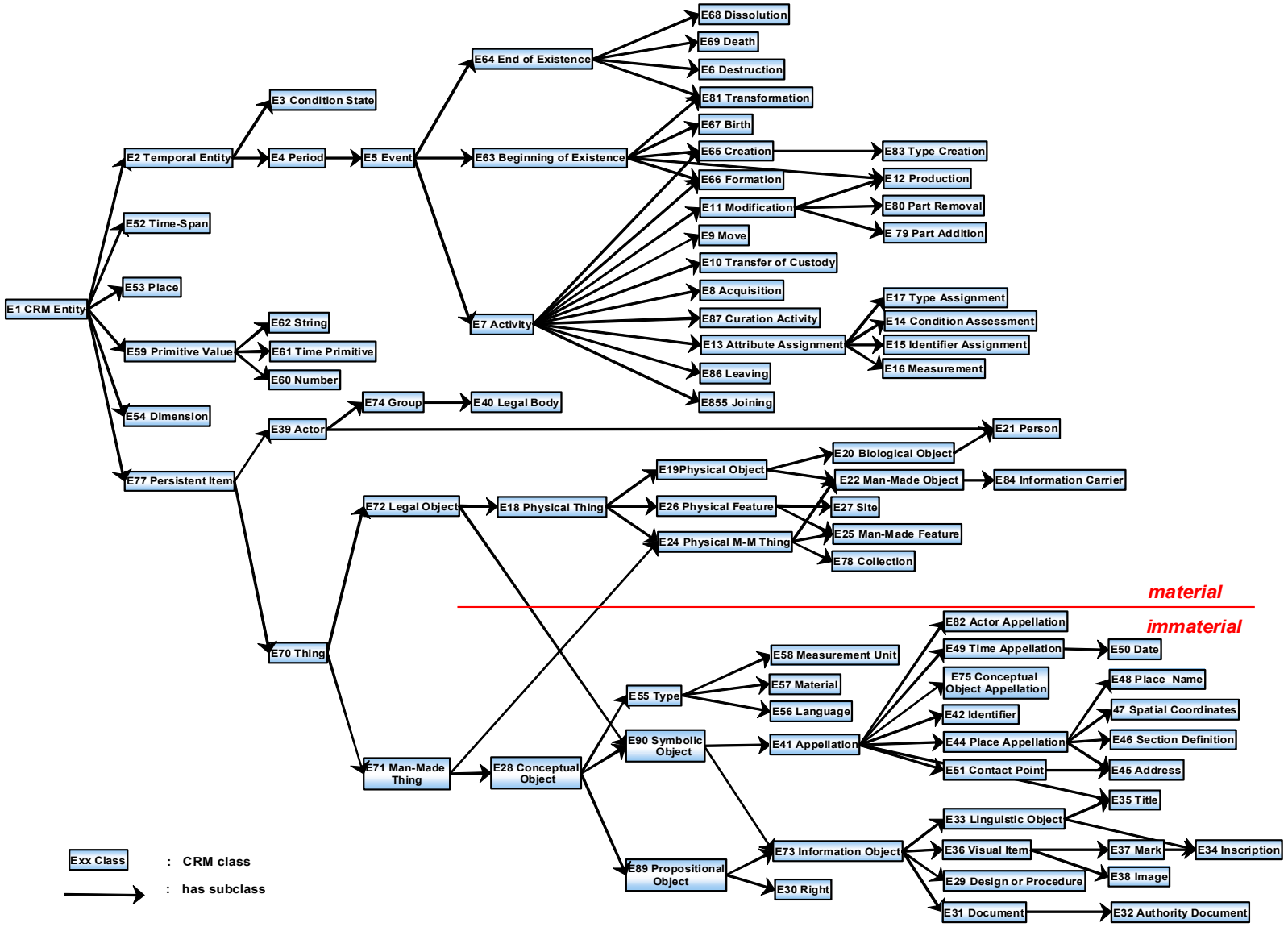
[13] DOERR, M. GRADMANN, S. HENNICKE, S. ISAAC, A. MEGHINI, M. VAN DE SOMPEL, H. 2010. The Europeana data model. In: Proceedings of IFLA 2010, World Library and Information Congress: 76th IFLA General Conference and Assembly, 10-15 August 2010, Gothenburg, Sweden, http://www.ifla.org/files/hq/papers/ifla76/149-doerr-en.pdf

[14] PRANDONI, C. VALENTINI, M. AND DOERR, M. 2009. Formalising a Model for Digital Rights Clearance. In Proceedings of the 13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 327–338, 2009.

8 Appendix

8.1 A – CRM class hierarchies

We present here the CRM class hierarchy version 4.2.5 in graphical form. The essence of the CRM are however the properties. A complete set of graphical representations of the CRM properties can be found in: http://cidoc.ics.forth.gr/comprehensive_intro.html.



8.2 B – METS –CRM mapping tables

Mapping of elements and selected attributes of METS XML Schema version 1.9 to the CIDOC CRM, version 5.0.2, following the format of Appendix B of the METS Primer and Reference Manual version 1.6.

TABLE 1: ELEMENT, ATTRIBUTE AND COMPLEX TYPE TABLE

COMPLEX TYPE	CRM Equivalent	ELEMENTS OF THIS TYPE	ATTRIBUTES and CRM equivalents		MAY CONTAIN	
<amdSecType>	E31 Document. *P2F has type: "METS:amdSecType"*	<amdSec>	ID	*	<techMD> <rightsMD> <sourceMD> <digiprovMD>	E31 Document[amdSec]. *P148 has component:* E31 Document[<techMD]. etc.
<areaType>	E75 Conceptual Object Appellation. *P2F has type: "METS:areaType"* *P106F is composed of:* E75 Conceptual Object Appellation. *P2F has type: "METS:areaType"* Instances of this class **contain** a composite identifier of a "file area" and may be **a surrogate for** a "file area". These elements further imply the following useful information: E73 Information Object[*X1*]. *P106F is composed of:* E90 Symbolic Object[*X2*], E73 Information Object[*X1*]. *p1F is identified by*: E75	<area>	ID "A unique identifier for the <area> element"	E75 Conceptual Object Appellation[<area>]. *p1F is identified by:* E75 Conceptual Object Appellation[ID]*. **Mapping comment:** It is not clear, if ID is used to point to the <area> declaration or meant to designate the described segment itself (*X2*) or both: E90 Symbolic Object[*X2*].*p1F is identified by*: E75 Conceptual Object Appellation [ID] **Mapping comment:** This can be ambiguous if two area declarations declare the same area.		
			FILEID SHAPE COORDS BEGIN END BETYPE EXTENT EXTYPE	E75 Conceptual Object Appellation. *P106F is composed of:* E90 Symbolic Object Instances of this class contain a **composite identifier** for an arbitrary file segment ("*X2*") of a file ("*X1*"), built of XML attributes that can be interpreted as a procedure to locate the		

	Conceptual Object Appellation [FILEID], E90 Symbolic Object[*X2*].*p1F is identified by*: E75 Conceptual Object Appellation [<area>]. X1 is the file denoted by FILEID and X2 the segment identified by the instance of the <area> element.			corresponding segment. **Mapping comment:** The part (*X2*) of the referred file (*X1*) is identified by a tuple of begin [BEGIN] and endpoints [END] or image coordinates [COORD], which can all be interpreted as identifiers of document parts (E75), a system of E55 Types SHAPE, BETYPE, EXTYPE], and an E54 Dimension [EXTENT].[16]		
			CONTENTIDS ADMID	E90 Symbolic Object[*X2*]. *P1F is identified by:* E75 Conceptual Object Appellation [CONTENTIDS] E31 Document. *P2F has type:* *"METS:mdSecType"* *P1F is identified by:* E75 Conceptual Object Appellation [ADMID] *P67F refers to:* E90 Symbolic Object[*X2*]		
<behaviorSecType>	E31 Document. *P2F has type:* *"METS:behaviorSecType"*	<behaviorSec>	ID LABEL	* E31 Document. *P1F has note:* E62 String [LABEL]	<behavior> <behaviorSec>	E31 Document [<behaviorSec>]. *P148 has component:* E31 Document [<behaviorSec>]. E31 Document [<behaviorSec>]. *P148 has component:* E29 Design or Procedure [<behaviour>]
			CREATED	E31 Document.*P94B was created by*: E65 Creation. *P4F has time-span*: E52 Time-Span **Mapping comment:** METS does not foresee to refine a< behaviorSec> by admin metadata. Why?		
<behaviorType>	E29 Design or Procedure. *P2F has type:*	<behavior>	ID LABEL	* E29 Design or Procedure[*X1*].	<interfaceDef> <mechanism>	E29 Design or Procedure[*X1*]. *P69 is*

[16] This analysis of the constituents is not useful in order to create a CRM equivalent of a METS document, because they make sense only with the interpretation mechanism of this identifier defined by METS. It makes only sense if more information about the area should be inferred.

	"METS:behaviorType" "Behavior" has to do with what can be done in the future. Internal analysis of such designs or procedures is not part of the current scope of the CRM. From point of the CRM, this should be regarded as a detailed description of a compatibility type of things that can be run/shown by the respective methods. In addition, we can regard the behavior as a description of the object (refers to/ has note). "Behaviors" should be standardized, and not be regarded as individual metadata. We denote an instance of <behaviour> as *"X1"*		BTYPE	*P1F has note:* E62 String [LABEL], *P2F has type:* E55 Type [BTYPE]		*associated with*: E29 Design or Procedure [<interfaceDef>] E29 Design or Procedure[*X1*]. *P69 is associated with*: E29 Design or Procedure [<mechanism>]
			CREATED	E29 Design or Procedure. *P94B was created by*: E65 Creation. *P4F has time-span*: E52 Time-Span [CREATED]		
			STRUCTID	E29 Design or Procedure[*X1*]. *P67F refers to:* E73 Information Object [STRUCTID] **AND/OR:** E73 Information Object [STRUCTID]. *P2F has Type:* E55 Type [METSbehaviourType:ID]. *P1 has note:* [<behaviour>]		
			GROUPID ADMID "administrative metadata sections within the METS document that pertain to the given behavior."	E29 Design or Procedure[*X1*]. *P69 is associated with*: E29 Design or Procedure [GROUPID] E31 Document. *P2F has type:* *"METS:mdSecType"* *P1F is identified by:* E75 Conceptual Object Appellation [ADMID] *P67F refers to:* E29 Design or Procedure[*X1*]		
<divType>	E73 Information Object. *P2F has type:* E55 Type [TYPE]		ID xlink:label CONTENTIDS ORDER ORDERLABEL LABEL used, for example, to identify a to an end user **Mapping comment:** Altogether a rich set of identifiers. It is hard to understand, why those and no others, i.e. why identification has not been parameterized in	* E73 Information Object [*X1*]. *P1F is identified by:* E75 Conceptual Object Appellation [xlink:label]* E73 Information Object [*X1*]. *P1F is identified by:* E75 Conceptual Object Appellation [CONTENTIDS]* E73 Information Object [*X1*]. *P1F is identified by:* E75 Conceptual Object Appellation [ORDER]. *p2 has type:* "METS:ORDER"	 <mptr> <fptr> The <fptr>/<mptr> element "represents digital content that manifests its parent element."	E73 Information Object. [ ]. *P148 has component:* E73 Information Object. [ ]. <fptr> and <mptr> are internal identification mechanisms. They may contain pointers to actual carriers. we identify in the mapping ,<fptr> and

			a generic way.	E73 Information Object [*X1*]. *P1F is identified by:* E75 Conceptual Object Appellation [ORDERLABEL]. *p2 has type:* "METS:ORDERLABEL" E73 Information Object [*X1*]. *P1F has note:* E62 String [LABEL]		<mptr>
			TYPE	E73 Information Object. *P2F has type:* E55 Type [TYPE]		
			DMDID ADMID	E31 Document. *P2F has type:* *"METS:mdSecType"* *P1F is identified by:* E75 Conceptual Object Appellation [ADMID] *P67F refers to:* E73 Information Object [*X1*] E31 Document. *P2F has type:* *"METS:mdSecType"* *P1F is identified by:* E75 Conceptual Object Appellation [DMDID] *P67F refers to:* E73 Information Object [*X1*]		
<fileGrpType> Why does the FileGrp not point to a dmdSec, whereas a and <file> does?	E73 Information Object. *P2F has type:* *"METS:fileGrpType"* We denote an instance of fileGrpType as *"X1"* **Mapping comment**: The "fileGrp" element is competing with the structmap. There shouldn't be two alternative organization principles.to describe data object hierarchies	<fileGrp>	ID	*	<file> <fileGrp>	E31 Document [<fileGrp>]. *P148F has component:* E73 Information Object. [<fileGrp>], *P148F has component:* E73 Information Object. [<file>]
			VERSDATE **Mapping comment:** Why this date should denote a "version" is hard to understand. There seems to be unnecessary heterogeneity. Why is a separate date necessary at all? In the CRM, we regard any version as a new object.	E73 Information Object [*X1*]. *P94B was created by*: E65 Creation. *P4F has time-span*: E52 Time-Span [VERSDATE]		
			ADMID USE	E31 Document. *P2F has type:* *"METS:mdSecType"*		

				P1F is identified by: E75 Conceptual Object Appellation [ADMID] *P67F refers to:* E73 Information Object [*X1*] E73 Information Object [*X1*]. P103 was intended for (was intention of): E55 Type [USE].		
<fileSecType>	E31 Document. *P2F has type:* *"METS:fileSecType"*	<fileSec>	ID	*	<fileGrp> <file>	E31 Document [<fileSec>]. *P148F has component:* E73 Information Object. [<fileGrp>], *P148F has component:* E73 Information Object. [<file>]
<fileType>	E73 Information Object. *P2F has type:* E55 Type[MIMETYPE] We denote an instance of fileType as *"X1"*	<file>	ID OWNERID why not "contentids"? SEQ	* E73 Information Object [*X1*]. *P1F is identified by:* E75 Conceptual Object Appellation [OWNERID] E73 Information Object [*X1*]. *P1F is identified by:* E75 Conceptual Object Appellation [SEQ]. *p2 has type:* "METS:SEQ"	<file> <transformFile> <FLocat> <FContent> <stream>	E73 Information Object. [<file>]. *P148F has component:* E73 Information Object. [<file>]. E29 Design or Procedure [<transformFile]. *P67 refers to:* E73 Information Object. [<file>]. <FLocat> <FContent> <stream> are internal identification mechanisms. They may contain pointers to actual carriers. we
			MIMETYPE USE	E73 Information Object[*X1*]. *P2F has type:* E55 Type[MIMETYPE] E73 Information Object [*X1*]. P103 was intended for (was intention of): E55 Type [USE].		
			CREATED	E73 Information Object [*X1*]. *P94B was created by*: E65 Creation. *P4F has time-span*: E52 Time-Span [CREATED]		

			SIZE CHECKSUM CHECKSUMTYPE	E73 Information Object [*X1*]. *P43F has dimension:* E54 Dimension [SIZE]. *p2 has type:* "Number of Bytes" *P43F has dimension:* E54 Dimension [SIZE]. *p2 has type:* E55 Type [CHECKSUMTYPE].		identify in the mapping ,<fptr> and <mptr>
			ADMID DMDID GROUPID	E31 Document. *P2F has type:* *"METS:mdSecType"* *P1F is identified by:* E75 Conceptual Object Appellation [ADMID] *P67F refers to:* E73 Information Object [*X1*] E31 Document. *P2F has type:* *"METS:mdSecType"* *P1F is identified by:* E75 Conceptual Object Appellation [DMDID] *P67F refers to:* E73 Information Object [*X1*] E73 Information Object [*X1*]. *P148B is component of:* E73 Information Object[GROUPID].		
			BEGIN END BETYPE version 1.9 added BEGIN, END and BETYPE attributes to the <file> and <stream> elements for specifying the location of a nested file or a stream within it's parent file.	E73 Information Object[*X1*]. *P106F is composed of:* E90 Symbolic Object[*X2*], E73 Information Object[*X1*]. *p1F is identified by*: E75 Conceptual Object Appellation [FILEID], E90 Symbolic Object[*X2*].*p1F is identified by*: E75 Conceptual Object Appellation []. where X2 is the segment identified by the BEGIN,END,BETYPE attributes..		

<mdSecType>	E31 Document. *P2F has type:* *"METS:dmdSec"/* *"METS:techMD"/* *"METS:rightsMD"/* *"METS:sourceMD"/* *"METS:digiprovMD"* We denote an instance of mdSecType as *"X1"*	<dmdSec> <techMD> <rightsMD> <sourceMD> <digiprovMD>	ID	*	<mdRef> <mdWrap>	<mdRef> <mdWrap> **Mapping comment:** these are identification mechanisms for the metadata content itself, internal (wrapped) or external. The CRM would refer to it via its ID ("*X1*") wherever it might be, point to a physical carrier of it (URL in "href"), or expand the content in a *"P3F has note"* in case of <mdWrap>
			CREATED	E31 Document[*X1*]. *P94B was created by*: E65 Creation. *P4F has time-span*: E52 Time-Span [CREATED]		
			GROUPID ADMID "An attribute that provides values for administrative metadata elements which apply to the current descriptive or administrative metadata." **Mapping comment:** Note that ADMID is recursive, an <admSec> can be described by an <admSec>!	E31 Document[*X1*] *P148B is component of:* E31 Document[GROUPID] E31 Document. *P2F has type:* *"METS:mdSecType"* *P1F is identified by:* E75 Conceptual Object Appellation [ADMID] *P67F refers to:* E31 Document[*X1*].		
			STATUS "Use to indicate the status of this metadata (e.g., superseded, current, etc.)"	E31 Document[*X1*]. *P3 has note*: E62 String **Mapping comment:** This is workflow information, not functional in isolation.		
<metsType>	E31 Document[*X1*]. *P2F has type:* *"METS:metsType"* We denote an instance of metsType as *"X1"*	<mets>	ID OBJID "This identifier is used to tag the entire METS object to external systems, in contrast with the ID identifier" LABEL	* E31 Document[*X1*]. *P1F is identified by:* E75 Conceptual Object Appellation [OBJID], **Mapping comment:** In version 1.9: "OBJID uniquely identifies the entire digital object represented by the METS document" in contrast to the METS document itself! This creates a different mapping. See <metsHdr>. *P1F has note:* E62 String [LABEL]	<metsHdr> <dmdSec> <amdSec> <fileSec> <structMap> <structLink> <behaviorSec>	E31 Document [*X1*]. *P148F has component:* E31 Document [<metsHdr>]. etc.
			TYPE PROFILE "Indicates to which of the registered profile(s) the METS document conforms."	E31 Document [*X1*]. *P2F has type:* E55 Type[TYPE], E31 Document [*X1*]. *P2F has type:* E55 Type[PROFILE]		
<parType>	E31 Document[<par>]. *P2F has type:*	<par>	ID	*	<area> <seq>	E73 Information Object [<par>].

	"METS:parType"					*P106F is composed of:* E90 Symbolic Object [<area>].
<objectType>	E73 Information Object. *P2F has type:* "METS:interfaceDef"/ "METS:mechanism"	<interfaceDef> <mechanism>	ID LABEL	* E73 Information Object. *P1F has note:* E62 String [LABEL]		
			LOCTYPE OTHERLOCTYPE attributeGroup ref: xlink:simpleLink A URL is regarded as identifying a section on a physical machine that carries the information.	Depending on LOCTYPE: E73 Information Object. *P1F is identified by:* E75 Conceptual Object Appellation [xlink:href] (e.g., a DOI) OR E73 Information Object. *P128B is carried by:* E24 Physical Man-Made Thing. *P1F is identified by:* E42 Identifier [xlink:href]. *P2F has type: "URL"*		
<seqType>	E73 Information Object. *P2F has type:* "METS:seqType"	<seq>	ID	*	<area> <par>	E73 Information Object. [<seq>]. *P106F is composed of:* E90 Symbolic Object [<area>]. E73 Information Object. [<seq>]. *P106F is composed of:* E73 Information Object [<par>].
<structLinkType>	E31 Document. *P2F has type:* *"METS: structLinkType"*	<structLink>	ID	*	<smLink>	E73 Information Object. [<structLink>]. *P148F has component:* E73 Information Object. [<smLink>]. This is a reification construct. the <smLink> links are collected in the <sctructLink>
<structMapType>	E31 Document	<structMap>	ID	*		E31 Document.

	[<structMap>]. *P2F has type:* *"METS: structMapType"*		LABEL "Describes the <structMap> to viewers of the METS document" TYPE "kind of organization principle of the structure"	E31 Document[<structMap>]. *P1F has note:* E62 String [LABEL], E31 Document[<structMap>]. *P2F has type:* E55 Type [TYPE]		[<structMap>]. *P148F has component:* E73 Information Object. [ ].

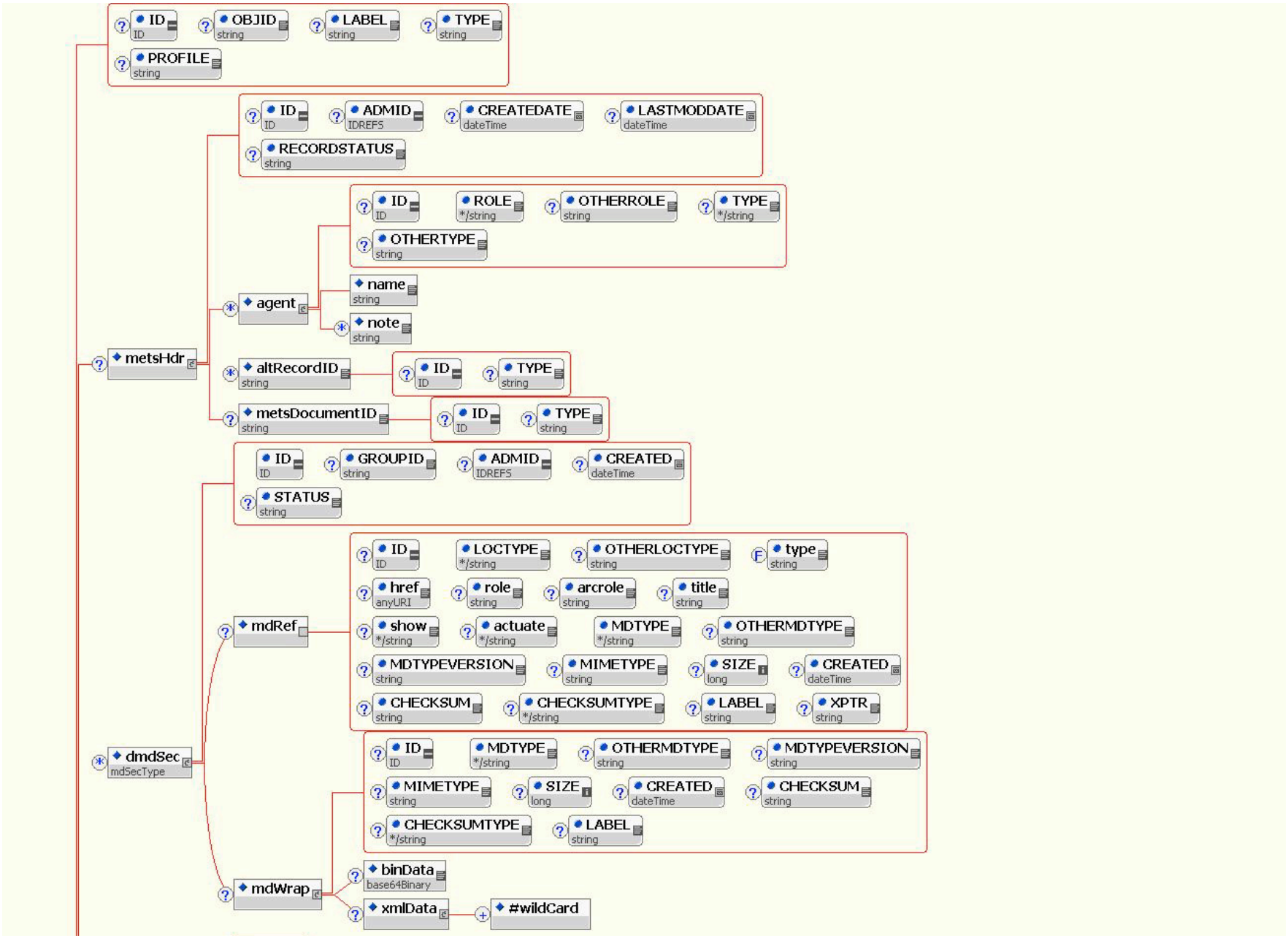
***** The attribute "ID" could be mapped to the CRM as "E73 Information Object. *P1F is identified by*: E75 Conceptual Object Appellation [ID]. Since it serves internal identification, it should not be mapped to the CRM. Rather, the links resulting from referring to the ID should be instantiated, and the ID should be used to compose an internal identifier (or URI in case of RDF networks) for the CRM instance corresponding to this information object.

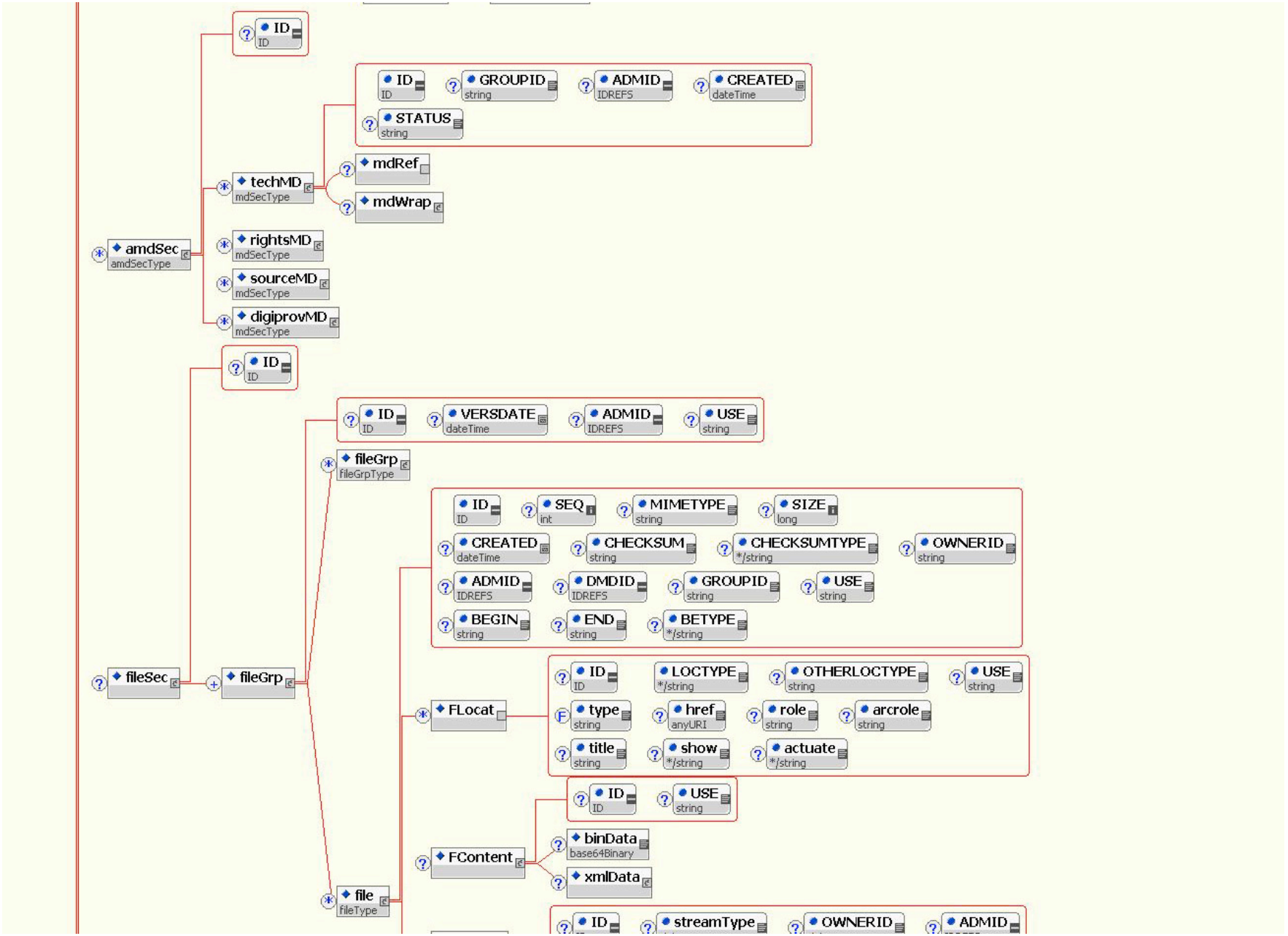
TABLE 2: Other Mapped ELEMENTS

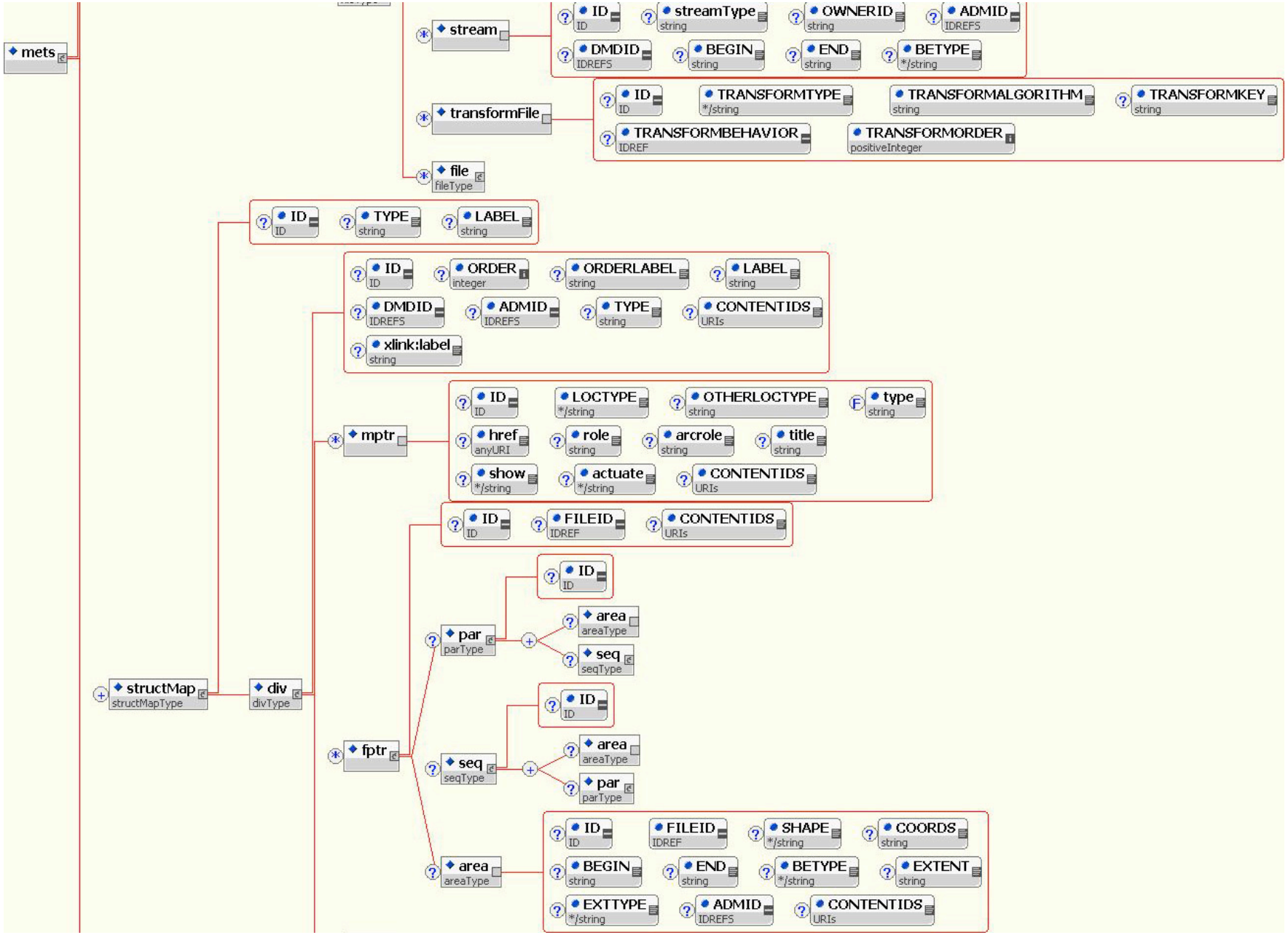
ELEMENT	CRM Equivalent	MAY CONTAIN	CRM Path	ATTRIBUTES and CRM equivalents		CONTAINED WITHIN	MIN/MAX
<agent>	E39 Actor	<name> <note>	E39 Actor. P131F is identified by: E82 Actor Appellation[<name>], *P3 has note:* E62 String [<note>]	ID	*	<metsHdr>	0/•
				TYPE OTHERTYPE	E39 Actor. *P2F has type:*E55 Type[TYPE]		
				ROLE OTHERROLE	E31 Document[<mets>]. *P94B was created by*: E65 Creation. *P14F carried out by(P14.1 in the role of: E55Type[ROLE])*: E39 Actor [<agent>]		
<metsHdr>	E31 Document. *P2F has type:* *"METS:* metsHdr*"* The header is mapped to a set of activities on the mets document as a whole. This view as document is not needed.	<agent> <altRecordID> <metsDocumentID> "The metsDocument identifier element <metsDocumentID> allows a unique identifier to be assigned to the METS document itself. This may be different from the OBJID attribute value in the root <mets> element, which uniquely identifies the entire digital object represented by the METS document."	E31 Document[<mets>]. *P94B was created by*: E65 Creation. *P14F carried out by*: E39 Actor [<agent>] **IF <metsDocumentID> is not used:** E31 Document[<mets>]. *P1F is identified by:* E75 Conceptual Object Appellation[<altRecordID>] **IF <metsDocumentID> is used:** E31 Document[<mets>]. *P1F is identified by:* E75 Conceptual Object Appellation[<metsDocumentID>] E31 Document [<mets>]. *P148F has component:* E31 Document [<mets OBJID>]. *P1F is identified by:* E75 Conceptual Object Appellation[OBJID] E31 Document [<mets>]. *P148F has component:* E31 Document [<mets OBJID>]. *P1F is*	ID	*	<mets>	0/1
				CREATEDATE	E31 Document[<mets>]. *P94B was created by*: E65 Creation. *P4F has time-span*: E52 Time-Span [CREATEDATE]		
				LASTMODDATE In the CRM, modification is the creation of a new version, using the previous version.	E31 Document[<mets>]. *P94B was created by*: E65 Creation. *P4F has time-span*: E52 Time-Span [LASTMODDATE], *P16F used specific object:* E31 Document, *P2 has type: "Modification"*.		
				ADMID "An attribute that	E31 Document[<mets>].		

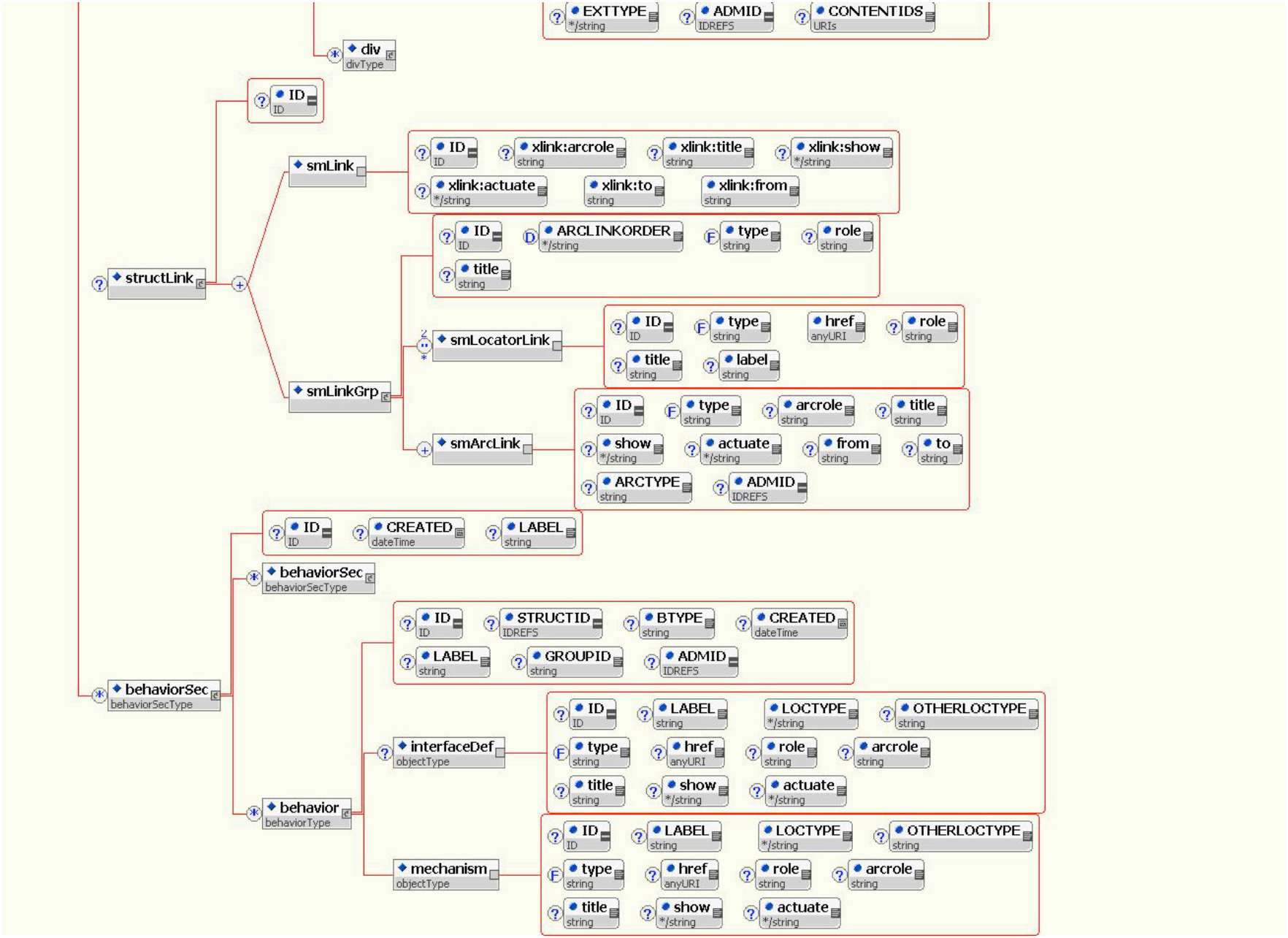
			identified by: E75 Conceptual Object Appellation[<altRecordID>] **Mapping comment:** The idea is, that the object "represented by METS" is "incorporated" in the METS wrapper. Instead of P148F one may use "incorporates" from FRBRoo.	provides values for administrative metadata elements which apply to the current descriptive or administrative metadata."	*P1F is identified by:* E75 Conceptual Object Appellation [ADMID] *P67F refers to:* E31 Document[*X1*].		
				RECORDSTATUS	E31 Document[<mets>]. *P3 has note*: E62 String **Mapping comment:** This is workflow information, not functional in isolation.		

8.3 The METS schema in graphical form




ID
ID
GROUPID
string
ADMID
IDREFS
CREATED
dateTime
STATUS
string
techMD
mdSecType
mdRef
mdWrap
amdSec
amdSecType
rightsMD
mdSecType
sourceMD
mdSecType
digiprovMD
mdSecType
ID
ID
VERSDATE
dateTime
USE
string
fileGrp
fileGrpType
SEQ
int
MIMETYPE
SIZE
long
CHECKSUM
CHECKSUMTYPE
OWNERID
DMDID
GROUPID
BEGIN
END
BETYPE
fileSec
FLocat
LOCTYPE
OTHERLOCTYPE
type
href
anyURI
role
arcrole
title
show
actuate
FContent
binData
base64Binary
xmlData
file
fileType
streamType


mets
stream
ID
ID
streamType
string
OWNERID
string
ADMID
IDREFS
DMDID
IDREFS
BEGIN
string
END
string
BETYPE
*/string
transformFile
ID
ID
TRANSFORMTYPE
*/string
TRANSFORMALGORITHM
string
TRANSFORMKEY
string
TRANSFORMBEHAVIOR
IDREF
TRANSFORMORDER
positiveInteger
file
fileType
ID
ID
TYPE
string
LABEL
string
ID
ID
ORDER
integer
ORDERLABEL
string
LABEL
string
DMDID
IDREFS
ADMID
IDREFS
TYPE
string
CONTENTIDS
URIs
xlink:label
string
mptr
ID
ID
LOCTYPE
*/string
OTHERLOCTYPE
string
type
string
href
anyURI
role
string
arcrole
string
title
string
show
*/string
actuate
*/string
CONTENTIDS
URIs
structMap
structMapType
div
divType
fptr
ID
ID
FILEID
IDREF
CONTENTIDS
URIs
par
parType
ID
ID
area
areaType
seq
seqType
seq
seqType
ID
ID
area
areaType
par
parType
area
areaType
ID
ID
FILEID
IDREF
SHAPE
*/string
COORDS
string
BEGIN
string
END
string
BETYPE
*/string
EXTENT
string
EXTTYPE
*/string
ADMID
IDREFS
CONTENTIDS
URIs


EXTTYPE
*/string
ADMID
IDREFS
CONTENTIDS
URIs
div
divType
ID
ID
structLink
smLink
ID
ID
xlink:arcrole
string
xlink:title
string
xlink:show
*/string
xlink:actuate
*/string
xlink:to
string
xlink:from
string
ID
ID
ARCLINKORDER
*/string
type
string
role
string
title
string
smLocatorLink
ID
ID
type
string
href
anyURI
role
string
title
string
label
string
smLinkGrp
smArcLink
ID
ID
type
string
arcrole
string
title
string
show
*/string
actuate
*/string
from
string
to
string
ARCTYPE
string
ADMID
IDREFS
behaviorSec
behaviorSecType
ID
ID
CREATED
dateTime
LABEL
string
behaviorSec
behaviorSecType
ID
ID
STRUCTID
IDREFS
BTYPE
string
CREATED
dateTime
LABEL
string
GROUPID
string
ADMID
IDREFS
behavior
behaviorType
interfaceDef
objectType
ID
ID
LABEL
string
LOCTYPE
*/string
OTHERLOCTYPE
string
type
string
href
anyURI
role
string
arcrole
string
title
string
show
*/string
actuate
*/string
mechanism
objectType
ID
ID
LABEL
string
LOCTYPE
*/string
OTHERLOCTYPE
string
type
string
href
anyURI
role
string
arcrole
string
title
string
show
*/string
actuate
*/string

8.4 The OAI-PMH XML Schema in graphical form