

# Global Digital Format Registry (GDFR)

## Format Model and Relationships

Version: 1.0.7

Status: DRAFT

Issued: 2007-05-11

## 1 Introduction

The Global Digital Format Registry (GDFR) will provide sustainable services to store, discover, and deliver important representation information about digital formats. A format is the set of syntactic and semantic rules for serializing an abstract information model, an expression of exchangeable knowledge. The format of a digital object must be known in order to interpret the information content of that object properly. Without knowledge of its format, a digital object is merely a collection of undifferentiated bits. Thus, format typing is fundamental to the effective use, interchange, and preservation of all digitally-encoded content.

The wide diversity and rapid pace of adoption and abandonment of digital formats present an ongoing problem for long-term preservation efforts. The purpose of the GDFR is to address this concern by providing a sustainable resource for managing format-critical representation information necessary to the preservation function.

## 2 Format model

Informally, a format is a byte-wise serialization of an abstract information model. More rigorously, a format can be defined in terms of four conceptual entities:

- Information Model (*IM*) – a class of exchangeable knowledge.
- Semantic Model (*SM*) – a set of semantic information structures capable of realizing the meaning of the *IM*.
- Syntactic Model (*CM*) – a set of syntactic data units capable of expressing the *SM*.
- Serialized Bytestream (*SB*) – a sequence of bytes capable of manifesting the *CM*.

The format-specific rules governing the three-stage transformation between these entities can be defined in terms of the following conceptual encoding functions:

- Semantic Encoding (*SE*) – a mapping from the exchangeable knowledge of an *IM* to the semantic information structures of an *SM*.

$$SE : IM \rightarrow SM$$

- Syntactic Encoding (*CE*) – a mapping from the semantic structures of an *SM* to the syntactic data units of an *CM*:

$$CE : SM \rightarrow CM$$

- Serialized Bytestream Encoding (*BE*) – a mapping from the syntactic units of a *CM* to the serialized bytes of an *SB*.

$$BE : CM \rightarrow SB$$

Thus, a format *F* is the class defined by the 3-tuple,  $F = (SE, CE, BE)$ .

In practice, the formal specifications for many formats often co-mingle the rules for semantic and semantic encodings. Similarly, most processes that operate on formatted byte streams do not do so with clear demarcation between these three conceptual levels. Nevertheless, the model is useful for defining the semantics of inter-format relationships in a formal manner.

### 3 Relationships

Many digital formats exist in associative relationships with other formats. These relationships are an important component of the format representation information managed by the GDFR.

#### 3.1 Extension

The *extension* relationship is defined in terms of substitutability. Format *B* is an extension of format *A* if:

- The *SE* of *A* is a proper subset of the *SE* of *B*; and
- The *CE* of *A* is a subset of the *CE* of *B*; and
- The *BE* of *A* is a subset of the *BE* of *B*.

In other words, *B* is an extension of *A* if all instances of *A* are also instances of *B*, but not all instances of *B* are instances of *A* since the *SE* of *B* will contain additional mapping rules not found in the *SE* of *A*.

The extension relationship is *transitive*, in other words, the fact that format *C* is an extension of format *B*, which is itself an extension of format *A*, necessarily implies that format *C* is an extension of format *A*.

EXAMPLE UTF-8 is an extension of ASCII

All valid ASCII byte streams can be used in the context of any UTF-8-aware process without any loss of ASCII-enabled semantic function. Using a valid UTF-8 byte stream in the content of an ASCII-only-aware process may result in some loss of UTF-8-enabled semantic function.

EXAMPLE DNG (Digital Negative) is an extension of TIFF 6.0

#### 3.2 Restriction

The *restriction* relationship is the inverse of extension. Format *B* is a restriction of format *A* if:

- The *SE* of *A* is proper subset of the *SE* of *B*; and
- The *CE* of *B* is a subset of the *CE* of *A*; and
- The *BE* of *B* is a subset of the *BE* of *A*.

In other words, *B* is a restriction of *A* if all instances of *B* are also instances of *A*, but not all instances of *A* are instances of *B* since the *SE* of *A* will contain additional mapping rules not found in the *SE* of *B*.

The restriction relationship is transitive, in other words, the fact that format *C* is a restriction of format *B*, which is itself a restriction of format *A*, necessarily implies that format *C* is a restriction of format *A*.

EXAMPLE PDF/A-1 is a restriction of PDF 1.4

EXAMPLE Model Imaged Object Profile is a restriction of METS

NOTE The fact that format *A* is an extension of format *B* necessarily implies that format *B* is a restriction of format *A*, and vice versa. Thus, the choice of which relationships to use to define

the association is arbitrary. As a best practice, the one that is consistent with the temporal relationship of the associated formats should be used. Since PDF 1.4 was defined prior to PDF/A it makes better sense to say that PDF/A is a restriction of PDF 1.4. In other words, the selected relationship should use the temporally antecedent format as its source and the subsequent format as its target, e.g. “<antecedent-format> is a restriction of <subsequent-format>” or <subsequent-format> is an extension of <antecedent-format>.

### 3.3 Modification

Both extension and restriction are specific instances of a more general *modification* relationship. Format *B* is a subclass of format *A* if:

- The *SE* of *A* and *B* can be split into two disjoint parts  $SE_1$  and  $SE_2$  such that  $SE_1$  of *B* is a proper subset of the  $SE_1$  of *A* and the  $SE_2$  of *A* is a proper subset of the  $SE_2$  of *B*; and
- The *CE* of *A* and *B* can be split into two disjoint parts  $CE_1$  and  $CE_2$  such that  $CE_1$  of *B* is a subset of the  $CE_1$  of *A* and the  $CE_2$  of *A* is a subset of the  $CE_2$  of *B*; and
- The *BE* of *A* and *B* can be split into two disjoint parts  $BE_1$  and  $BE_2$  such that  $BE_1$  of *B* is a subset of the  $BE_1$  of *A* and the  $BE_2$  of *A* is a subset of the  $BE_2$  of *B*; and

EXAMPLE BWF (Broadcast Wave Format) is a *modification* of WAVE

BWF both extends and restricts the baseline WAVE format, defining an additional Broadcast Audio Extension (“bext”) chunk and only allowing LPCM (linear pulse code modulation) audio. Neither extension nor restriction strictly applies since there are cases where a BWF cannot be used in a WAVE context without loss of function, e.g. dependencies on the “bext” chunk, and there are cases where a WAVE cannot be used in a BWF context, e.g. non-LPCM sampling.

### 3.4 Definition

The *definition* relationship indicates the means by which a modification (or restriction or extension) relationship is expressed.

EXAMPLE NITF (News Industry Text Format) is defined in terms of XML DTD.

EXAMPLE Office Open XML is defined in terms of XML Schema.

EXAMPLE ODF (Open Document Format) is defined in terms of Relax NG.

### 3.5 Containment

The *containment* relationship indicates an encapsulation association. *Containment comes in two forms: semantic containment and serial containment.*

Format *A* *semantically* contains format *B* if:

- The *SE* of *B* is a proper subset of the *SE* of *A*.

Format *A* *serially* contains format *B* if:

- The *SE* of *B* is a proper subset of the *SE* of *A*; and
- The *CE* of *B* is a subset of the *CE* of *A*; and

- The *BE* of *B* is a subset of the *BE* of *A*.

Each of these variants of containment can be qualified with respect to its obligation:

- *Optional* containment, in which the encapsulation is permitted but not required (“can contain”)
- *Mandatory* containment, in which the encapsulation is required (“must contain”)

EXAMPLE ZIP (with compression) can semantically contain any format

EXAMPLE TAR (Tape Archive) can serially contain any format

TAR serially contains formats since it preserves the bit-level integrity of the encapsulated byte streams, while ZIP only semantically contains formats since its compression algorithm results in new semantic and serialization encodings of the original byte streams. Note that ZIP can be used without compression, in which case its containment would be serial as well.

EXAMPLE PDF/A-1 must serially contain XMP (Extensible Metadata Platform)

### 3.6 Equivalence

The *equivalence* relationship indicates that the association between formats at the level of *SE* and/or *CE*. Format *B* is *semantically* equivalent to format *A* if:

- The *SE* of format *B* is identical to the *SE* of format *A*; and
- The *CE* of format *B* is not identical to the *CE* of format *A*; and
- The *BE* of format *B* is not identical to the *BE* of format *A*.

Format *B* is *syntactically* equivalent to format *A* if:

- The *SE* of format *B* is identical to the *SE* of format *A*; and
- The *CE* of format *B* is identical to the *CE* of format *A*; and
- The *BE* of format *B* is not identical to the *BE* of format *A*.

EXAMPLE TIFF (little-endian) is syntactically equivalent to TIFF (big-endian)

EXAMPLE DXF (ASCII) is syntactically equivalent to DXF (binary)

### 3.7 Version

The *version* relationship implies a change to the baseline function of previous version of a format within a recognized “familial” context, generally indicated by product identification. Version comes in two temporal forms, *older version* and *newer version*, and can be defined in either temporal direction, e.g. “<antecedent-format> is an older version of <subsequent-format>” or “<subsequent-format> is a newer version of <antecedent-format>”.

The version relationships are complements of each other, in other words, the fact that format *B* is a newer version of format *A*, necessarily implies that format *A* is an older version of format *B*.

A newer version *may* be, but is not necessarily, an extension of the older. Similarly, an older version *may* be, but is not necessarily, a restriction of the newer.

EXAMPLE Word 6.0 is a newer version of Word 97

EXAMPLE HTML 4.0 is an [older](#) version of HTML 4.01

### 3.8 Affinity

The *affinity* relationship holds between two formats that have a significant technical resemblance to each other but that do not meet the formal requirements of any of the other relationship types.

EXAMPLE SPIFF has an affinity to JPEG

EXAMPLE Word 6.0 has an affinity Word 97

NOTE The version and affinity relationships are both subjective rather than being based on strictly definable relationships between the *SE*, *CE*, and *BE* of the associated formats.