

## Global Digital Format Registry (GDFR)

### Classification

Version 1.0.5  
 Status: PUBLISHED  
 Issued 2007-11-09

## 1 Introduction

The Global Digital Format Registry (GDFR) provides 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.

Classification is an important component of the format representation information managed by the GDFR network.

## 2 Classification

The GDFR uses a system of faceted classification. Each format registered in the GDFR may be associated with a set of classification entries. Each entry is specified in terms of a facet type (or name) and value. Entries are represented notationally in the case-insensitive form: *type:value*. The defined facets are:

```
genre
role
composition
form
constraint
basis
domain
transform
```

The `genre` and `role` facets are required in a GDFR classification; all others are optional.

```
EXAMPLE   TIFF (Tagged image file format)   genre:still-image
                                                role:family
                                                composition:container-wrapper
                                                form:binary
```

```
EXAMPLE   TIFF 6.0                       genre:still-image
                                                role:file-format
                                                composition:container-wrapper
                                                form:binary
                                                basis:sampled
```

```
EXAMPLE   MXF (Material exchange format)  genre:generic
```

		<code>subsidiary-genre:moving-image</code> <code>subsidiary-genre:sound</code> <code>role:file-format</code> <code>composition:container-bundle</code>
EXAMPLE	SVG (Scalable vector graphics)	<code>genre:still-image</code> <code>role:file-format</code> <code>form:text</code> <code>basis:symbolic</code>
EXAMPLE	ISO 8859-1	<code>genre:text</code> <code>role:file-format</code> <code>constraint:unstructured</code>
EXAMPLE	Word 2002	<code>genre:text</code> <code>subsidiary-genre:still-image</code> <code>role:file-format</code> <code>constraint:structured</code>
EXAMPLE	ZIP	<code>genre:aggregate</code> <code>role:file-format</code> <code>composition:container-bundle</code> <code>transform:compression</code>
EXAMPLE	Jar (Java archive)	<code>genre:executable</code> <code>role:file-format</code> <code>composition:container-bundle</code> <code>transform:compression</code>

## 2.1 Genre facet

The `genre` facet defines the main classes found in the GDFR classification system. It is intended to indicate broadly the type of content associated with a format. The defined genres are:

```
genre:aggregate
genre:any
genre:database
genre:dataset
genre:executable
genre:model
genre:moving-image
genre:other
genre:presentation
genre:sound
genre:spreadsheet
genre:still-image
genre:text
```

It is recognized that the boundaries between these genre categories are imprecise, elastic, and permeable. However, the intention of the `genre` facet is to be practical and to provide broad indication of the primary high-level content associated with formats, as would be generally understood by the users of those formats.

Since many formats permit the aggregation of content drawn from multiple genres, the `genre` facet has a `subsidiary-genre` sub-type. The three permitted use cases are:

- A single `genre` entry
- A single primary `genre` entry accompanied by one or more `subsidiary-genre` entries
- One or more `genre` entries, none of which is of greater primacy than the others

EXAMPLE    ASCII                            `genre:text`

EXAMPLE    PDF 1.7                                `genre:text`  
    `subsidiary-genre:still-image`  
    `subsidiary-genre:moving-image`  
    `subsidiary-genre:sound`

EXAMPLE    MXF                                    `genre:moving-image`  
    `genre:sound`

NOTE        Since the containership capability of MXF is fairly generic, it could alternatively be characterized as `genre:aggregate`. However, since it is primarily associated with the management and delivery of time-based content, the characterization given above is preferable.

### 2.1.1 Genre values

The `aggregate` genre indicates a format that represents aggregations of arbitrary content drawn from multiple content genre categories.

EXAMPLE    Tar (Tape archive)                      `genre:aggregate`

EXAMPLE    ZIP                                      `genre:aggregate`

EXAMPLE    Open XML                                `genre:aggregate`  
    `subsidiary-genre:text`  
    `subsidiary-genre:still-image`  
    `subsidiary-genre:presentation`  
    `subsidiary-genre:spreadsheet`

The `any` genre indicates a format that can represent content from any single arbitrary genre.

Example    Gzip                                      `genre:any`

The `database` genre indicates a format that represents DBMS (hierarchical, object-oriented, relational, post-relational, etc.) content.

EXAMPLE    DBF (dBASE/Xbase)                      `genre:database`

The `dataset` genre indicates a format that represents non-DBMS data.

EXAMPLE    NetCDF (Network common data form)    `genre:dataset`  
    `form:binary`

## Global Digital Format Registry (GDFR)

EXAMPLE	SAS XPORT	genre:dataset form:binary
EXAMPLE	FITS (Flexible image transport system)	genre:dataset subsidiary-genre:still-image form:binary domain:astronomy
EXAMPLE	CSV (Comma separated value)	genre:text subsidiary-genre:dataset form:text

The `executable` genre indicates a format that represents interpreted or compiled executable content.

EXAMPLE	ELF-32	genre:executable basis:binary
EXAMPLE	Java byte code	genre:executable form:binary
EXAMPLE	Perl	genre:executable form:text

The `model` genre indicates a format that represents two- or three-dimensional geometric models.

EXAMPLE	IGES (Initial graphics exchange specification)	genre:model
---------	--	-------------

The `moving-image` genre indicates a format that represents dynamic (i.e., time-based) visual content.

EXAMPLE	AVI (Audio video interleave)	genre:moving-image
EXAMPLE	MPEG-4	genre:moving-image

The `other` genre indicates a format that represents content not well characterized by any other pre-defined genre.

The `presentation` genre indicates a format that represents interactive presentation content.

EXAMPLE	PowerPoint	genre:presentation
---------	------------	--------------------

The `sound` genre indicates a format that represents dynamic (i.e., time-based) auditory content.

EXAMPLE	MIDI (Musical instrument digital interface)	genre:sound basis:symbolic
EXAMPLE	WAVE	genre:sound basis:sampled

The `spreadsheet` genre indicates a format that represents a spreadsheet.

EXAMPLE	Excel	genre:spreadsheet
---------	-------	-------------------

		form:binary
	SpreadsheetML	genre:spreadsheet
		form:text

The `still-image` genre indicates a format that represents static (i.e., non-time-based) visual content.

EXAMPLE	JPEG	genre:still-image
		form:binary
		basis:sampled

EXAMPLE	SVG	genre:still-image
		form:text
		basis:symbolic

The `text` genre indicates a format that represents textual content.

EXAMPLE	Big-5	genre:text
		form:text
		constraint:unstructured
EXAMPLE	PDF 1.7	genre:text
		subsidiary-genre:still-image
		form:binary
		constraint:structured
EXAMPLE	SGML	genre:text
		form:text
		constraint:structured

The `genre` facet is required in a GDFR classification.

## 2.2 Role facet

The `role` facet indicates the ontological role of the classified format. The defined roles are:

```
role:family
role:file-format
role:encoding
role:serialization
```

The `family` role indicates a classified entity that is a *general class* of formats with common familial features.

The `file-format` role indicates a classified entity that is a *specific* format most usefully considered as independent file.

**NOTE** The specific formats that make up a familial class will most often exist within a network of versioning, extension, restriction, and/or affinity relationships to each other.

The `encoding` role indicates a classified entity usefully considered as a bit stream component of an encompassing file.

**NOTE** The `encoding` role corresponds to the Syntactic (*CE*) and Serialized Byte Stream (*BE*) Encodings defined by the GDFR format model.

The `serialization` role indicates a classified entity most usefully considered as a serialization algorithm.

**NOTE** The `serialization` role corresponds to the Serialized Bytestream Encoding (*BE*) defined by the GDFR format model.

As with the `genre` facet, the boundaries between these categories are imprecise, elastic, and permeable. Once again, however, the intention of the `role` facet is to be practical and to provide broad indication of the primary high-level ontological function associated with formats, as would be generally understood by the users of those formats.

<b>EXAMPLE</b>	GIF	<code>genre:still-image</code> <code>role:family</code>
	GIF 87a	<code>genre:still-image</code> <code>role:file-format</code>
	GIF 89a	<code>genre:still-image</code> <code>role:file-format</code>
	LZW (Liv-Zempel-Welch)	<code>genre:still-image</code> <code>role:encoding</code> <code>transform:compression</code>

**NOTE** The GIF family includes two specific formats, GIF 87a and GIF 89a, both of which make use of the LZW compression algorithm to encode their embedded raster image bit streams.

<b>EXAMPLE</b>	JFIF (JPEG file interchange format)	<code>genre:still-image</code> <code>role:file-format</code>
	JPEG DCT (Discrete cosine transform)	<code>genre:still-image</code> <code>role:encoding</code>

**NOTE** The JFIF format can use the DCT compression algorithm to encode its embedded raster image bit stream.

<b>EXAMPLE</b>	BWF (Broadcast WAVE format)	<code>genre:sound</code> <code>role:file-format</code>
	LPCM (Linear pulse code modulation)	<code>genre:sound</code> <code>role:encoding</code>

**NOTE** The BWF format must use LPCM sampling to define its embedded audio bit stream.

<b>EXAMPLE</b>	XML 1.1	<code>genre:text</code> <code>role:file-format</code>
	UTF-8	<code>genre:text</code> <code>role:serialization</code>

NOTE The XML 1.1 format can use UTF-8 for its serialization.

The `role` facet is required but not repeatable.

## 2.3 Composition facet

The `composition` facet indicates the compositional nature of the classified format. The defined compositions are:

```
composition:unitary
composition:container-bundle
composition:container-wrapper
```

The `unitary` composition indicates that the format is most usefully considered as a single atomistic entity.

The `container-bundle` composition indicates a format most usefully considered as an aggregation of multiple independently-formatted bit streams without significant metadata defining the descriptive, technical, and structural characteristics of the constituent bit streams or the relationships between them. Bundle formats are typically used for convenient packaging for distribution or storage.

Note Bundle formats will generally be classified with the `genre:aggregate` genre.

The `container-wrapper` composition indicates a format most usefully considered as an aggregation of multiple independently-formatted bit streams and significant metadata defining the descriptive, technical, and structural characteristics of the constituent bit streams and the relationships between them.

EXAMPLE	US-ASCII	<code>genre:text</code> <code>composition:unitary</code>
EXAMPLE	Tar	<code>genre:aggregate</code> <code>composition:container-bundle</code>
EXAMPLE	QuickTime	<code>genre:moving-image</code> <code>subsidiary-genre:sound</code> <code>composition:container-wrapper</code>

The `composition` facet is optional and not repeatable. If unspecified, the assumed default value is `composition:unitary`.

## 2.4 Form facet

The `form` facet indicates the nature of the format encoding. The defined forms are:

```
form:binary
form:text
```

The `binary` form indicates a format that primarily encodes its content in binary form, dependent upon a format-specific rendering application for human consumption.

The `text` form indicates a format that primarily encodes its content in textual form, susceptible to human

consumption using a non-format-specific text rendering application.

EXAMPLE	PBM (Portable bit map)	genre:still-image form:text
	JPEG 2000	genre:still-image form:binary
EXAMPLE	LaTeX	genre:text form:text
	WordPerfect	genre:text form:binary

The `form` facet is optional and not repeatable. If unspecified, no default value is assumed.

## 2.5 Constraint facet

The `constraint` facet indicates the organizational nature of the classified format. The defined constraints are:

```
constraint:structured  
constraint:unstructured
```

The `structured` constraint indicates a format most usefully considered as requiring a highly constrained organizational structure.

The `unstructured` constraint indicates a format most usefully considered as allowing a loosely constrained organizational structure.

The `constraint` facet is generally applied to `genre:text` formats.

EXAMPLE	SGML	genre:text constraint:structured
EXAMPLE	ISO 8859-1	genre:text constraint:unstructured

The `constraint` facet is optional and not repeatable. If unspecified, no default value is assumed.

## 2.6 Basis facet

The `basis` facet indicates the nature of the representation of content used by the format. The defined bases are:

```
basis:sampled  
basis:symbolic
```

The `sampled` basis indicates a format that primarily represents its content as sample values.

The `symbolic` basis indicates a format that primarily represents its content as notational values.

## Global Digital Format Registry (GDFR)

EXAMPLE	PNG (Portable network graphics)	genre:still-image basis:sampled
EXAMPLE	PostScript	genre:still-image genre:text basis:symbolic
EXAMPLE	AIFF (Audio Interchange File Format)	genre:sound basis:sampled
EXAMPLE	MIDI	genre:sound basis:symbolic

The `basis` facet is optional and not repeatable. If unspecified, no default value is assumed.

### 2.7 Domain facet

The `domain` facet indicates an intellectual domain in which the format is commonly used. The defined domains are:

```
domain:astronomy
domain:cad-cam
domain:gis
domain:web-archive
```

EXAMPLE	DXF	genre:model domain:cad-cam
EXAMPLE	FITS (Flexible image transport system)	genre:dataset subsidiary-genre:still-image domain:astronomy
EXAMPLE	GeoTIFF	genre:still-image domain:gis
EXAMPLE	WARC (Web archive)	genre:aggregate domain:web-archive

The `domain` facet is optional and repeatable. If unspecified, no default value is assumed.

### 2.8 Transform facet

The `transform` facet indicates for a format that defines a transformation on the form of content.

```
transform:compression
transform:encryption
transform:message-digest
```

The `transform:compression` and `transform:encryption` values imply a reversible transformation, while `transform:message-digest` implies a one-way transformation. In general, a format designated with the `transform` facet will also have the encoding role.

## Global Digital Format Registry (GDFR)

EXAMPLE	Gzip	genre:any role:encoding transform:compression
EXAMPLE	RSA-4	genre:any role:encoding transform:encryption
EXAMPLE	MD5	genre:any role:encoding transform:message-digest

The `transform` facet is optional and repeatable. If unspecified, no default value is assumed.

### References

- Bately, Sue, *Classification in Theory and Practice* (Oxford: Chandos, 2005).
- Clausen, Lars R., *Handling File Formats* (Århus: State and University Library; Copenhagen: Royal Library, 2004) <<http://netarchive.dk/publikationer/FileFormats-2004.pdf>>.
- Global Digital Format Registry (GDFR): Format Model and Relationships*.
- IANA, *MIME Media Types*, 2007-03-06 <<http://www.iana.org/assignments/media-types/>>
- Library of Congress, *Formats, Evaluation Factors, and Relationships*, 2007-03-07 <[http://www.digitalpreservation.gov/formats/intro/format\\_eval\\_rel.shtml](http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml)>.
- Marcella, Rita and Robert Newton, *A New Manual of Classification* (Aldershot: Gower, 1994).