

GDPDM Blobs

1. Specification for BLOBs

Byte 1-3: Bit 1-24: Version of the compression algorithm (3 bytes ASCII string, null padded). This version: 001. All data will be stored in big-endian format.
Byte 4: Bit 25-32: Data type (1 byte ASCII string)
Byte 5-8: Bit 33-64: Number of Sites (SNP/Indel positions) (4 bytes, unsigned long integer in big endian)
Byte 9-18: Bit 65-144: Genome Version (10 bytes, null padded ASCII string)
Byte 19-43: Bit 145-344: Chromosome Name (25 bytes, null padded ASCII string)
Byte 44-47: Bit 345-376: Start Position (4 bytes, unsigned long integer in big endian). Smallest Absolute Physical Position.
Byte 48-51: Bit 377-408: End Position (4 bytes, unsigned long integer in big endian). Largest Absolute Physical Position
Byte 52-201: Bit 409-1608: Accession Name (150 bytes) (null padded ASCII string)
Byte 202-205: Bit 1609-1640: Data element length in bits (4 bytes, unsigned long integer in big endian)
Byte 206-207: Bit 1641-1656: Blob class (2 bytes, unsigned short integer in big endian)
Byte 208-357: Bit 1657-2856: Class specific fields (150 bytes, null padded ASCII string)
Remaining bytes through byte 1024 are unused for unknown header information. All unused bytes will be null bytes (0x00).
Byte 1025 - end: Bit 8193 - end: values. the length of each data element is defined in the "data element length in bits" field (byte 202 to 205).

2. Supported data types and data type code

Code 1: GDPDM genotyping values (4 bits for each data element), derived from IUPAC

0x0	A	A/A
0x1	C	C/C
0x2	G	G/G
0x3	T	T/T
0x4	R	A/G
0x5	Y	C/T
0x6	S	G/C
0x7	W	A/T
0x8	K	G/T
0x9	M	A/C
0xA	B	+/+ (insertion homoz)
0xB	D	0/0 (indel het.)
0xC	H	Reserved

0xD	V	Reserved
0xE	N	any base
0xF	-	-/- (deletion homoz.)

Code 2: Long Integer (4 bytes)

Code 5: string (variable length, length of string is defined as number of bits in the data_element_length field of the header)

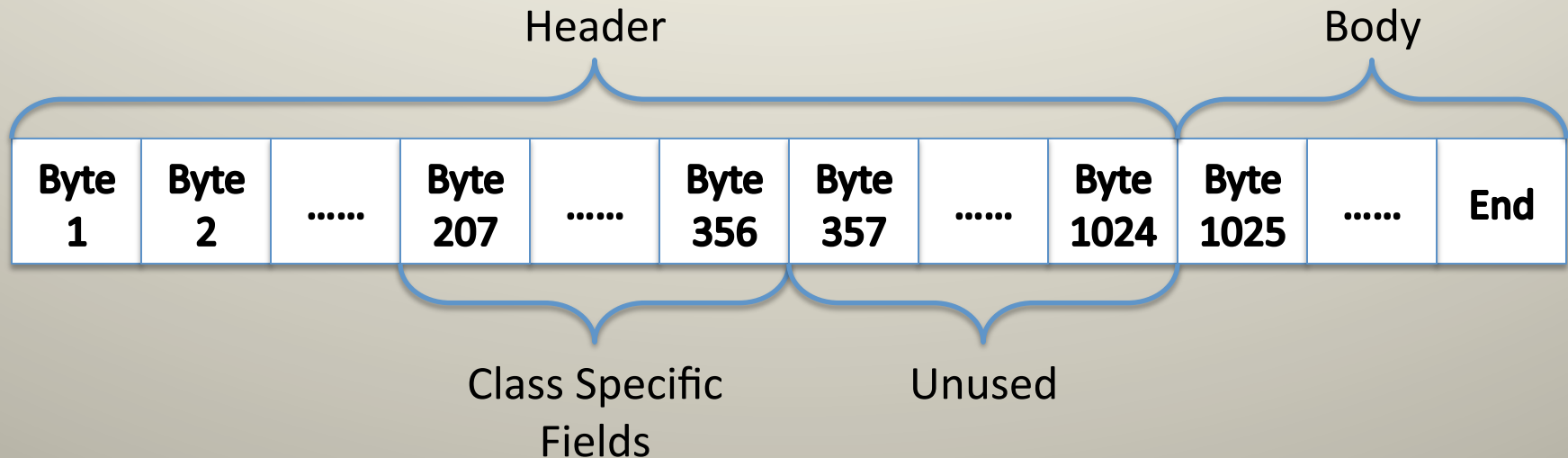
Code 8: Single precision floating point value (4 bytes)

Code 9: bit (1 bit)

3. Blob class

GDPDM Table.Column	Data type code	Blob class code	Class specific fields	stand-alone file extension	comments
div_allele.binary_value	1	1	null	bc01	Accession name required
div_allele_assay.binary_position	2	2	null	bc02	
div_allele_assay.binary_annotation	2	6	null	bc06	
div_allele_assay.binary_id	5	5	null	bc05	
cdv_g2p_study.p_value_blob	8	8	Trait <tab> germplasm_set	bc08	
cdv_g2p_study.effect_blob	8	9	Trait <tab> germplasm_set	bc09	
cdv_g2p_study.bpp_blob	8	10	Trait <tab> germplasm_set	bc10	
div_allele.binary_imputation	9	11	null	bc11	

BLOB Structure



- Header contains all information relevant to unpacking the BLOB
- Class Specific Fields used when information specific to a BLOB class is required
- Body contains all data, storage format determined in Header