National Polar-Orbiting Operational Environmental Satellite System (NPOESS)

HDF5 NPP Data Format

AMS 2011

Michael J. Denning,
Integrity Applications Incorporated
Richard E. Ullman,
NASA/GSFC

Introduction

HDF5 for NPP

- **Hierarchical Data Format 5 (HDF5) is the format for delivery of processed products from the National Polar-orbiting Operational Environmental Satellite System (NPOESS) Preparatory Project (NPP).**
  - HDF5 is a general purpose library and file format for storing scientific data.
  - HDF5 consists of two primary objects:
    - Dataset, which is a multidimensional array of homogeneous data elements.
    - Group, which is a container structure for organizing datasets and other groups.
  - The format offers efficient storage and I/O, including parallel I/O. HDF5 is a "free" format, with an extensive open source software library that runs on multiple platforms.
  - Data stored in HDF5 are used in many fields from computational fluid dynamics to film making.
- **Data can be stored in HDF5 in an endless variety of ways, so it is important to standardize how NPP product data are organized in HDF5.**

NPP Data Products

- **NPP Data Products**
  - Raw Data Record (RDR)
  - Sensor Data Record (SDR)
  - Temperature Data Record (TDR)
  - Intermediate Product (IP)
  - Application Related Products (ARP)
  - Environmental Data Records (EDR)
  - Geolocation Product (GEO)
  - Data Delivery Record (DDR)

- **NPP Data Products are distributed and formatted in HDF5**
  - Archived and made available to the community via the Comprehensive Large Array-data Stewardship System (CLASS), an electronic library of NOAA environmental data accessible at www.class.noaa.gov.

  - There is no “HDF-NPP” library; NPP data products have been designed using the native HDF5 library.

Documentation

- **NPP Common Data Format Control Book – External (CDFCB-X)**
  - Volume I – Overview
  - Volume II – RDR Formats
  - Volume III – SDR/TDR Formats
  - Volume IV – EDR/IP/ARP and Geolocation Formats
  - Volume V – Metadata
  - Volume VI – Ancillary Data, Auxiliary Data, Messages, and Reports
  - Volume VII – Downlink Formats (Application Packets)
  - Volume VIII – Look Up Table Formats

- **Available on the NPP project website**
  - http://jointmission.gsfc.nasa.gov/project/science-documents.html

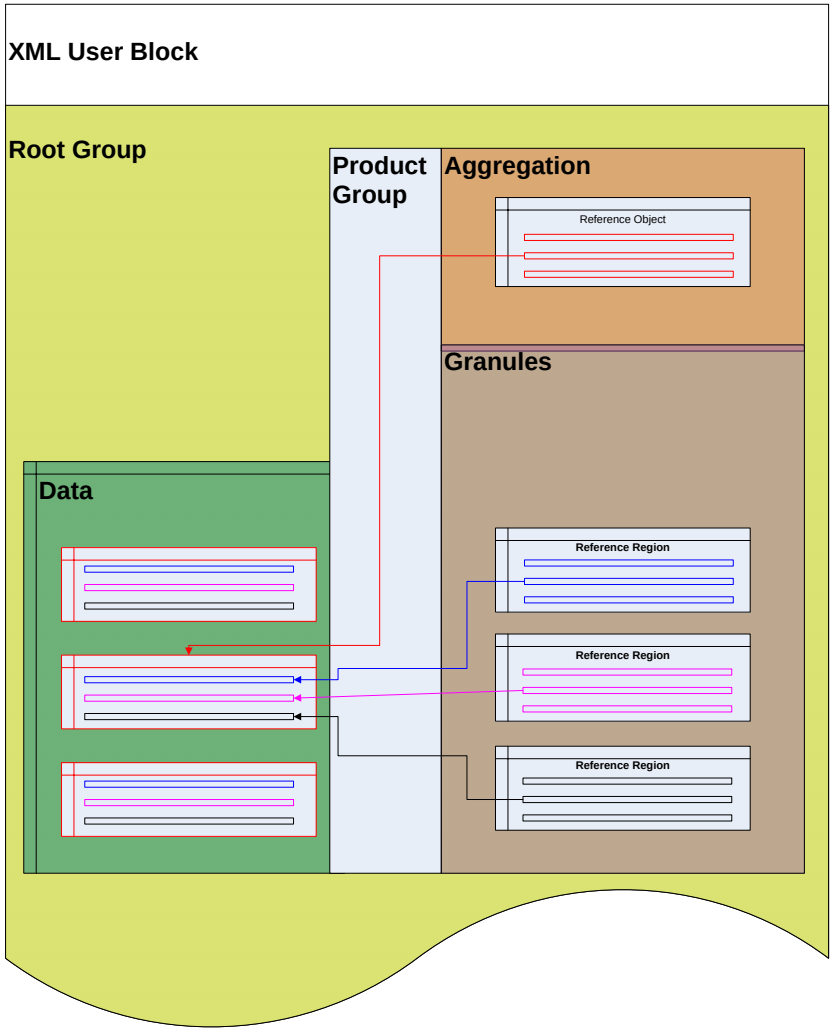
- **The CDFCB-X includes general UML models for each type of data product.**

NPP HDF5 General Overview

Data Organization

- **Data Product Granules**
  - A granule is a segment of data, with its size optimally determined to achieve maximum efficiency for an algorithm class.
  - It is associated with an integer number of sensor scans, and its definition varies for sensors and data products.
  - Gaps in granules are filled using a pre-defined “missing data” fill value.
  - Granules are represented as a set of region reference pointers to sections of the respective data set arrays.
- **Data Product Aggregations**
  - An aggregation is a grouping of homogeneous granules packaged in HDF5 covering a temporal range.
  - It may contain as few as one granule or as many as an orbit of granules.
  - It is represented as a set of object reference pointers to the various groupings of data which make up a particular data product (one for each homogenous dataset included in the granule).

NPP HDF5 Conceptual Diagram



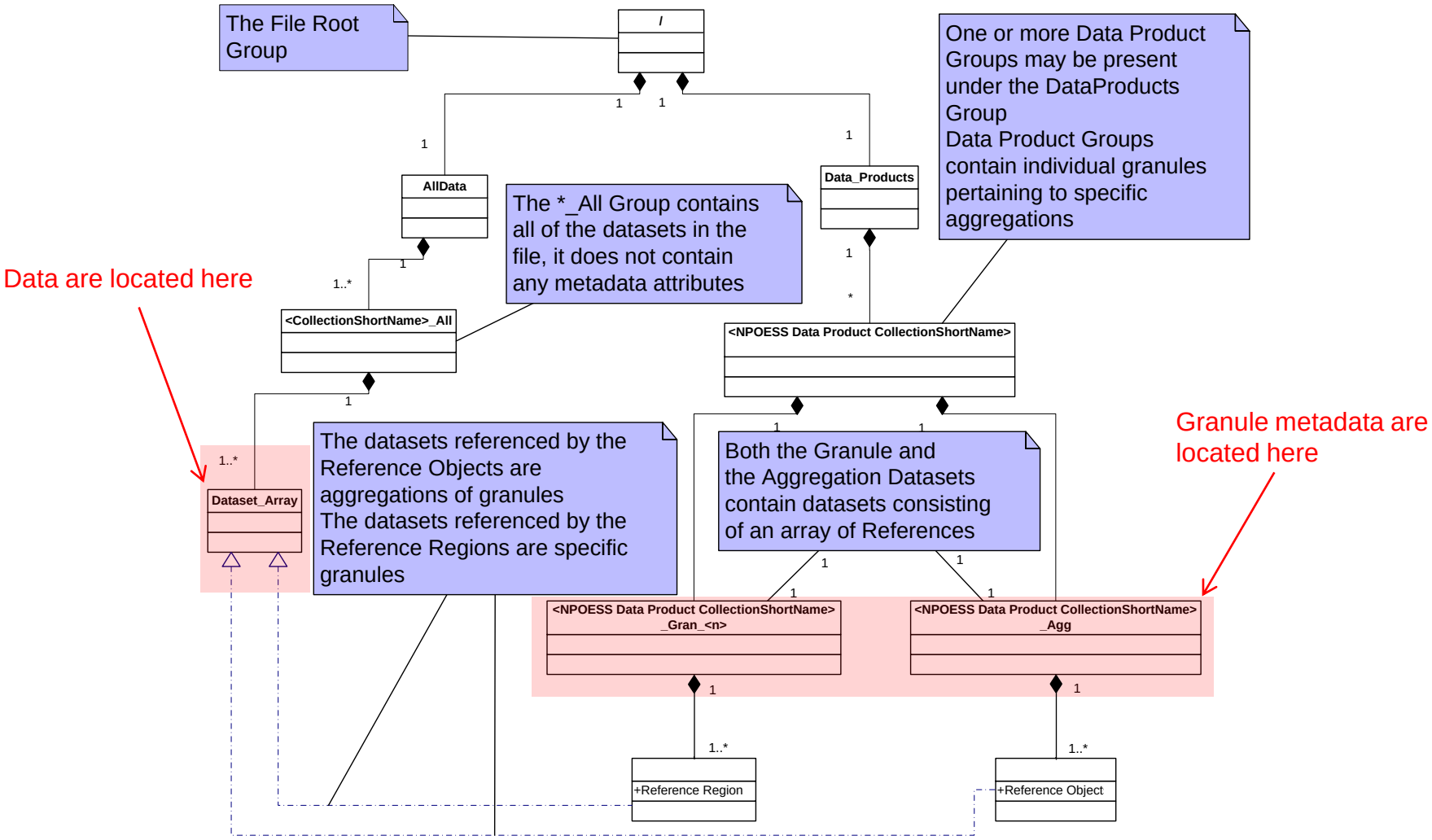
NPP HDF5 XML User Block

- **The XML user block for NPP data products provides a "quick-look" into the metadata of the associated HDF5 file.**

- **The XML user blocks are defined in the following volumes of the CDFCB-X:**
  - Volume V – Metadata
    - Contains the XML user block formats for RDRs, SDRs, TDRs, IPs, ARPs, and EDRs.
  - Volume VI – Ancillary, Auxiliary, Reports, and Messages
    - Contains the XML user block formats for the ancillary and auxiliary data files that are delivered in HDF5.

- **Example elements:**
  - Mission, Platform, and Instrument Names
  - Number of Data Products
  - Collection ShortName(s)
  - Aggregation Information
  - Timestamps

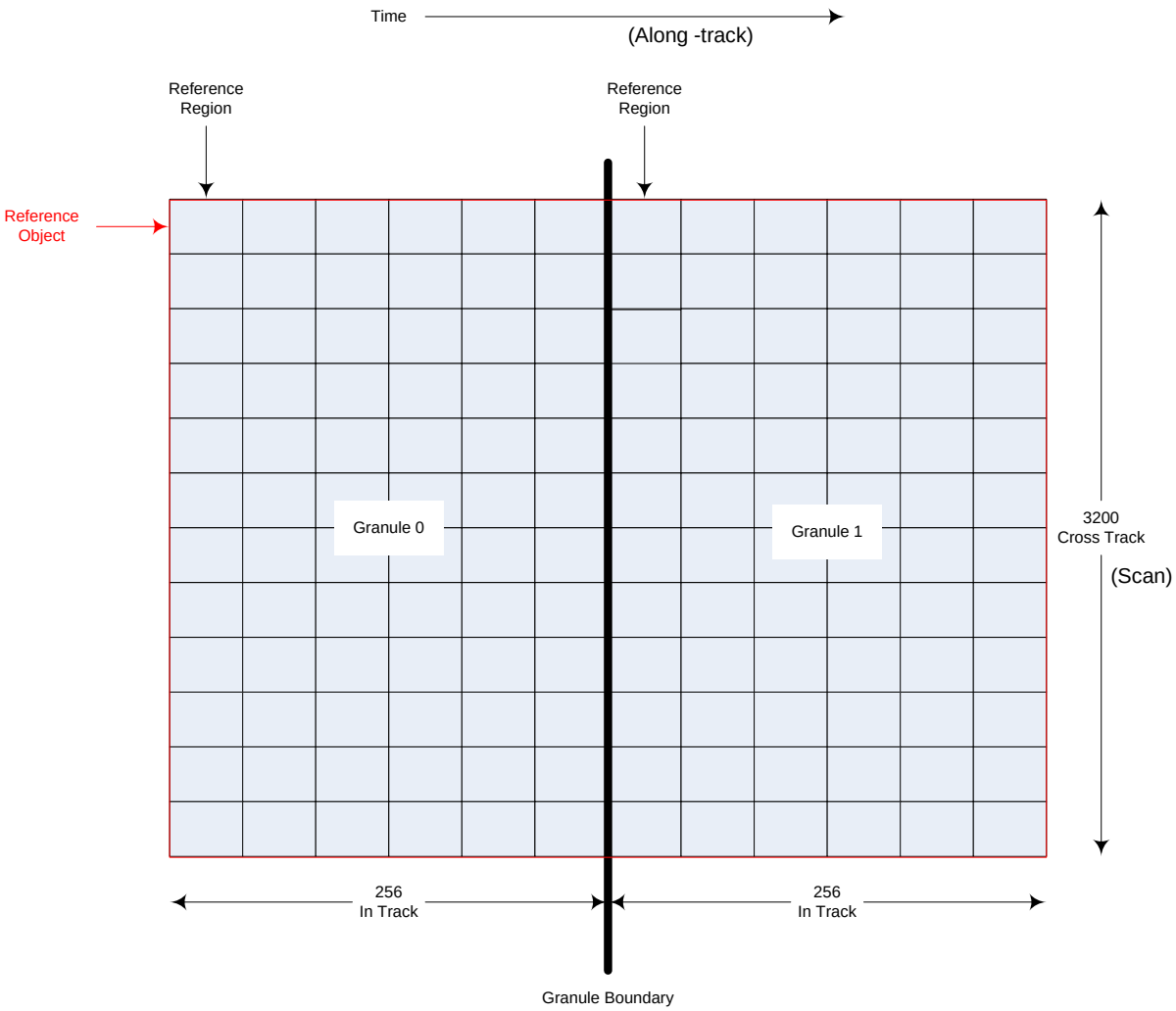
General NPP HDF5 File Structure



NPP HDF5 Metadata Locations

- **The NPP HDF5 metadata are organized hierarchically, from the top down in order to reduce duplication of information and to take advantage of the hierarchical nature of HDF5.**
  - Root Group
    - Data Products Group
      - Data Product (indicated by the specific product's identifier)
        - Product Aggregation Dataset
          - Reference Object Pointers
        - Product Granule Dataset
          - Reference Region Pointers

HDF5 Conceptual Diagram - Data



Quality Flags Overview

- **The concept is to provide consistently stored, high density, quality information about the delivered data.**
- **Quality flags are one or more consecutive bits in each byte.**
- **Quality flag arrays follow the structure of the data product.**
  - The size of the arrays are equal to or less than the size of the data to which the quality information applies (dimensions correspond to the data product arrays) .
- **Quality flags are stored in the HDF5 files as *N* number(s) of two or three dimensional, 1-byte arrays.**
  - The number of arrays is dependant on the quality flag definitions and is specific to each data product.
  - Each byte may contain multiple bit-level flags.
  - Quality flags will be ordered such that each flag is entirely contained within a single byte, occasionally resulting in a byte with reserved or meaningless bits.
  - Byte alignment is the same for every quality flag array.
    - First (left-most) bit is the LSB.

Example Product Group

An Example Product Group

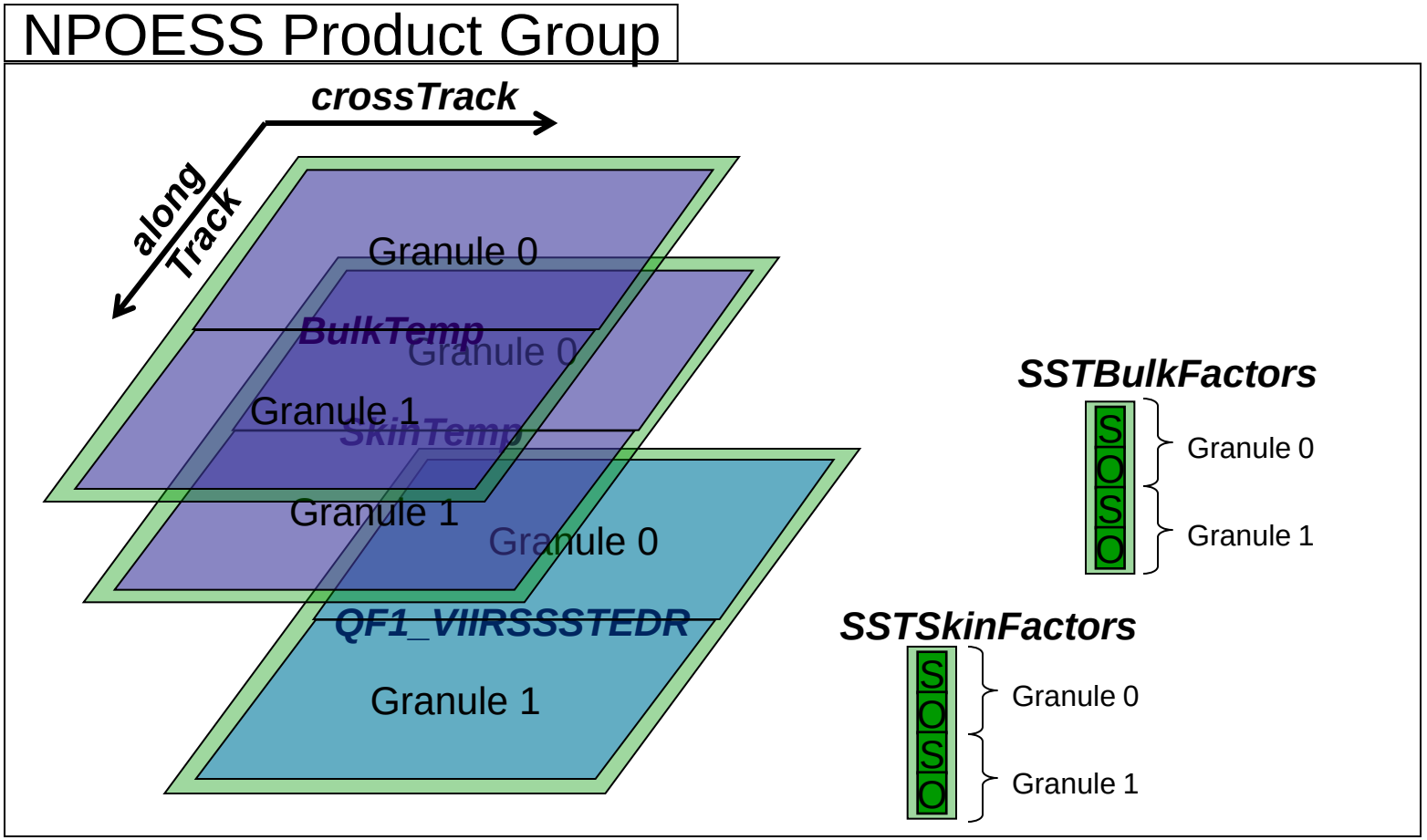
- **In this example product group:**
  - Five datasets constitute the product.
  - There are two common dimensions.
  - There are three congruent datasets.
  - Two datasets contain scale and offset values.
  - One dataset contains quality flags by element.
  - There are two granules in this aggregation.
  - Dimension “alongTrack” crosses the “granule boundary.”

Example Extracted from VIIRS Sea Surface Temperature EDR

Field Name	Description	Data Type	Dimensions	Units
BulkTemp	Sea Surface Bulk Temperature	16-bit unsigned integer	[*N**768, 3200]	Kelvin / Unitless
SkinTemp	Sea Surface Skin Temperature	16-bit unsigned integer	[*N**768, 3200]	Kelvin / Unitless
QF1_VIIRSSSTEDR	Land/Water Background	1-bit	[*N**768, 3200]	Unitless
	SST Skin Quality	2-bit		Unitless
	SST Bulk Quality	2-bit		Unitless
	Aerosol Correction	3-bit		Unitless
SSTBulkFactors	Bulk SST Scale	32-bit float	[*N**2]	Unitless
	Bulk SST Offset	32-bit float		Kelvin
SSTSkinFactors	Skin SST Scale	32-bit float	[*N**2]	Unitless
	Skin SST Offset	32-bit float		Kelvin

N is the number of granules.

Example Product Group



Dimensions

- Dimensions are defined for each field.
- Fields are related by congruency and common dimensions.
- Common dimensions are given the same name.
- One dimension crosses the granule boundary. When multiple granules are “aggregated,” the “granule boundary” dimension is extended.
- Dimension names and attributes are provided in the product profile.
- Dimensions are dependant on the temporal range.

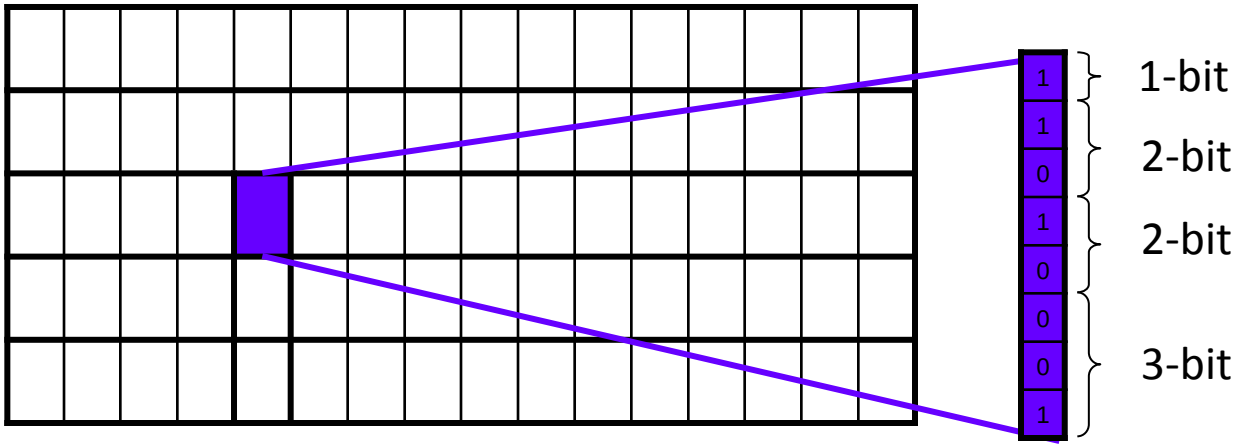
Scaled Integer Storage

- For storage efficiency, floating point data values may be stored as scaled integers.
- To regenerate the data value, the dataset element must be multiplied by a supplied scale factor and an offset added.
- The scale factor and offset are provided, one pair for each granule as a separate dataset.
- The scale and offset value are the same for all granules produced with a given version of an algorithm.

Quality Flags by Element

- Most products contain multiple indicators of quality on an element by element basis.
- Quality flags are associated by congruency (shared dimension) with a data array.
- Multiple flags of less than 8-bits are “packed” into structures aligned on 8-bit boundaries.

Quality Flags



Geolocation

- **Geolocation products are constructed using the same conventions as SDRs and EDRs.**
- **Geolocation datasets have a congruent relationship with the datasets to which they apply (same dimensions).**
- **The association between a data product and its geolocation product is determined via one of two ways.**
  - The geolocation product may be packaged as a separate product group within the same HDF5 file.
  - The name of a separate geolocation product file may be stored in the N_GEO_Ref attribute on the root HDF group.
  - The NPP Interface Data Processing Segment (IDPS) decides whether to store the geolocation information as a product group or as a separate file.

Common Geolocation Fields for VIIRS Products

Field Name	Comments	Dimensions	Units	Data Type
StartTime	since epoch 1/1/1958	[per scan or swath]	microseconds	64-bit signed integer
MidTime	since epoch 1/1/1958	[per scan or swath]	microseconds	64-bit signed integer
SCPosition	ECR coordinates	[per scan or swath]	meters	32-bit float
SCVelocity	ECR coordinates	[per scan or swath]	meters/second	32-bit float
Latitude		[per cell]	degrees	32-bit float
Longitude		[per cell]	degrees	32-bit float
SolarZenithAngle		[per cell]	degrees	32-bit float
SolarAzimuthAngle		[per cell]	degrees	32-bit float
SensorZenithAngle		[per cell]	degrees	32-bit float
SensorAzimuthAngle		[per cell]	degrees	32-bit float
Height	geoid or terrain	[per cell]	meters	32-bit float
SatelliteRange		[per cell]	meters	32-bit float

Product Profiles

- **Each data product has a corresponding XML product profile.**
  - The product profile is delivered as part of the product documentation.
  - The product profile contains metadata such as units of measure, dimension names, legend entries, etc.
  - A style sheet is also provided in order to view the profile via a web browser.

Field Attributes in the XML Product Profile (1..9)

Attribute Name	Type	Comments
DataType	String	String format is: "%d-bit %s", where %d is the number of bits and %s is one of: • signed integer • unsigned integer • floating point • <blank> (a bitfield)
Description	String	A descriptive text.
Dimension_GranuleBoundary	Set of Boolean	True (1) indicates that this dimension extends when granules are appended.
Dimension_Name	Set of string	Name match indicates that this dimension is congruent with the same dimension names in other datasets in this product group.
Field_Name	String	The name of the HDF5 dataset that contains the field values.
FillValue_Name	Set of string	
FillValue_Value	Set of number	Data type matches type of dataset.
LegendEntry_Name	Set of string	
LegendEntry_Value	Set of number	Data type matches type of dataset.

Field Attributes in the XML Product Profile (10..17)

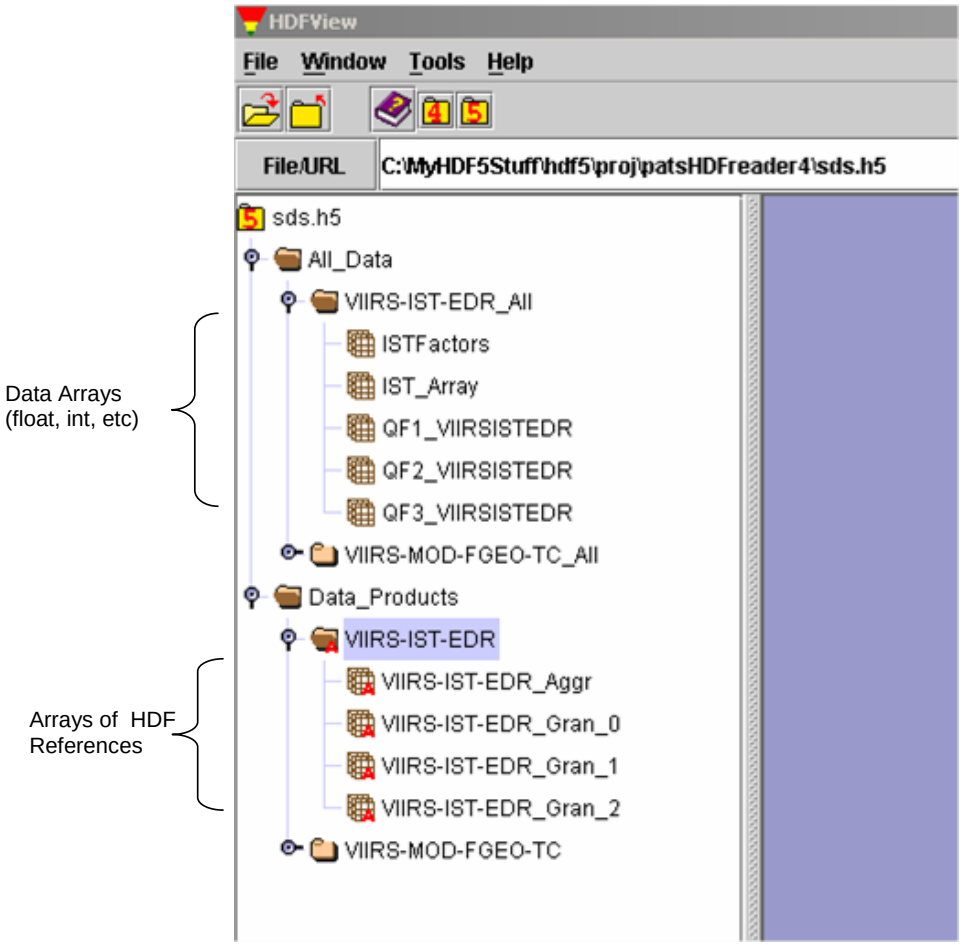
Attribute Name	Type	Comments
MeasurementUnits	String	Consistent with SI naming and Unidata's "udunits" package
NumberOfDimensions	Integer	Integer greater than zero.
NumberOfFillValues	Integer	If zero, then no FillValue_Name and FillValue_Value attributes are present. Fill Values are used for primary data fields only.
NumberOfLegendEntries	Integer	If zero, then no LegendEntry_Name and LegendEntry_Value attributes are present. Legend entries are used for quality fields only.
RangeMax	Number	Maximum expected value of field elements in the product, not just this dataset instance. Data type matches type of dataset.
RangeMin	Number	Minimum expected value of field elements in the product, not just this instance. Data type matches type of dataset.
Scaled	Boolean	True indicates that the dataset is scaled. Note that fill values are in the dataset type and so must be tested before un-scaling.
ScaleFactorName	String	The name of the HDF5 dataset that contains scaling coefficients. To un-scale the elements, first multiply the scaled element by the first element and then add the second element. If the dataset is not scaled, Scale_AttributeName will not exist.

Conclusions

- **The program will deliver official data products (RDRs, SDRs/TDRs, EDRs/ARPs/IPs) and dynamic ancillary data and auxiliary data in HDF5 files.**
  - Official deliverable data products are organized by reference objects (aggregations) which contain one or more reference regions (granules).
- **The consistent HDF5 group structure means that the organization for each product is the same as all others.**
- **HDF5 allows for flexible temporal aggregation, as granules are appended by extending dataset dimension.**
- **The format provides flexibility for a variety of end users.**
  - Straight HDF5 means there is no need for additional libraries.

Backup

VIIRS Ice Surface Temperature (IST) EDR – HDFView Screenshot



h5dump Screenshot – VIIRS Sea Surface Temperature HDF5 File

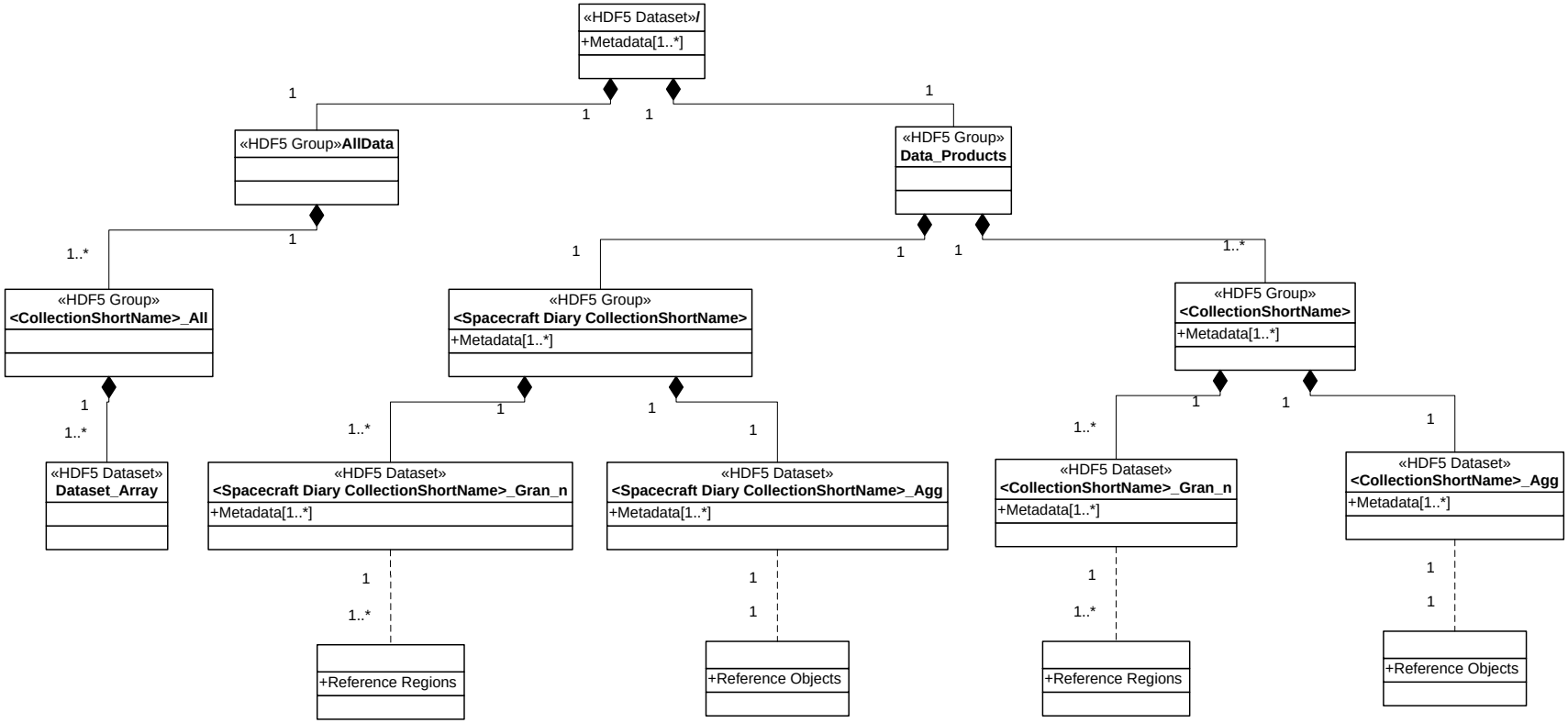
- Another way to view the arrays of references (aggregation and granule dataset arrays) is with the h5dump utility:
  - Granule:

```
Command Prompt
HDF5 "sds.h5" {
DATASET "/Data_Products/VIIRS-IST-EDR/VIIRS-IST-EDR_Gran_1" {
   DATATYPE  H5T_REFERENCE
   DATASPACE  SIMPLE { ( 1, 5 ) / ( 1, 5 ) }
   DATA {
   (0,0): DATASET 0:71284376 {(256,0,0)-(511,3199,0)},
   (0,1): DATASET 0:71287112 {(256,0,0)-(511,3199,0)},
   (0,2): DATASET 0:71287384 {(256,0,0)-(511,3199,0)},
   (0,3): DATASET 0:71287656 {(256,0,0)-(511,3199,0)},
   (0,4): DATASET 0:71287928 {(0,2)-(0,3)}
   }
```

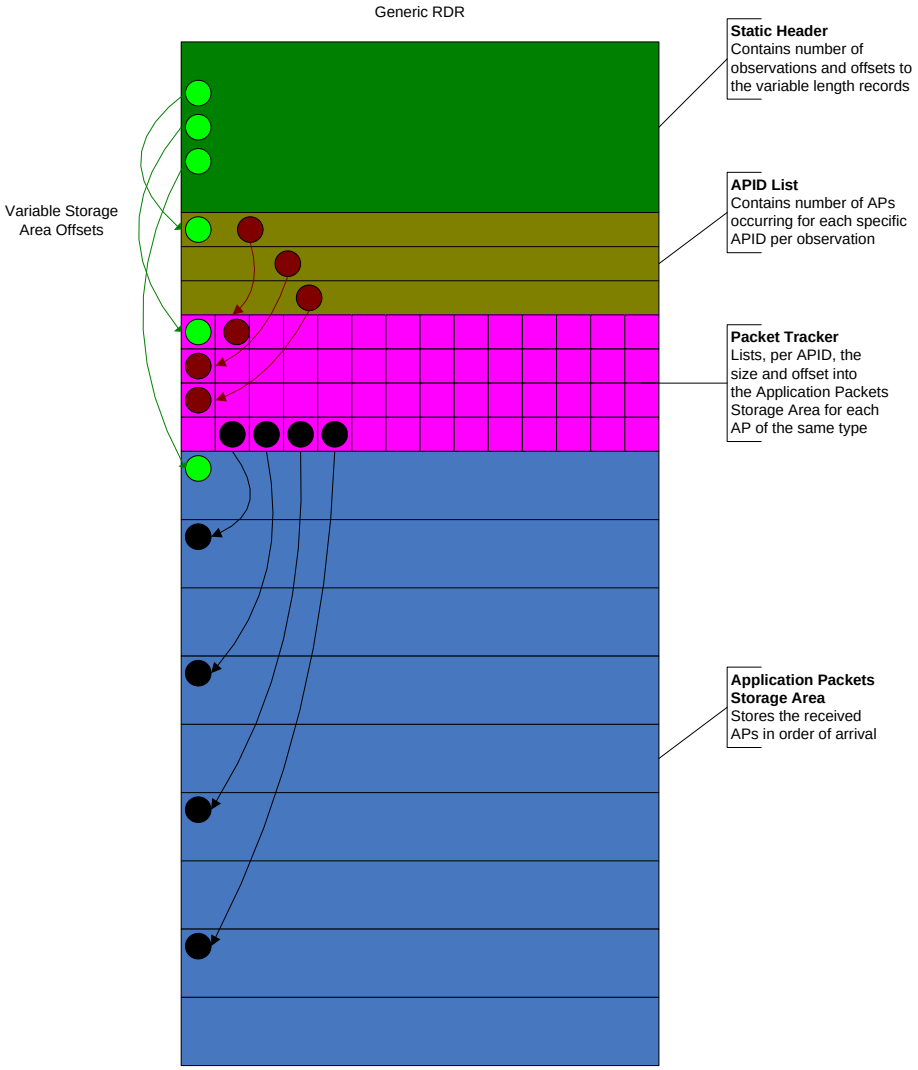
  - Aggregation:

```
Command Prompt
HDF5 "sds.h5" {
DATASET "/Data_Products/VIIRS-IST-EDR/VIIRS-IST-EDR_Aggr" {
   DATATYPE  H5T_REFERENCE
   DATASPACE  SIMPLE { ( 1, 5 ) / ( 1, 5 ) }
   DATA {
   (0,0): DATASET 0:71284376 /All_Data/VIIRS-IST-EDR_All/IST_Array ,
   (0,1): DATASET 0:71287112 /All_Data/VIIRS-IST-EDR_All/QF1_VIIRSISTEDR ,
   (0,2): DATASET 0:71287384 /All_Data/VIIRS-IST-EDR_All/QF2_VIIRSISTEDR ,
   (0,3): DATASET 0:71287656 /All_Data/VIIRS-IST-EDR_All/QF3_VIIRSISTEDR ,
   (0,4): DATASET 0:71287928 /All_Data/VIIRS-IST-EDR_All/ISTFactors
   }
```

RDR UML Model

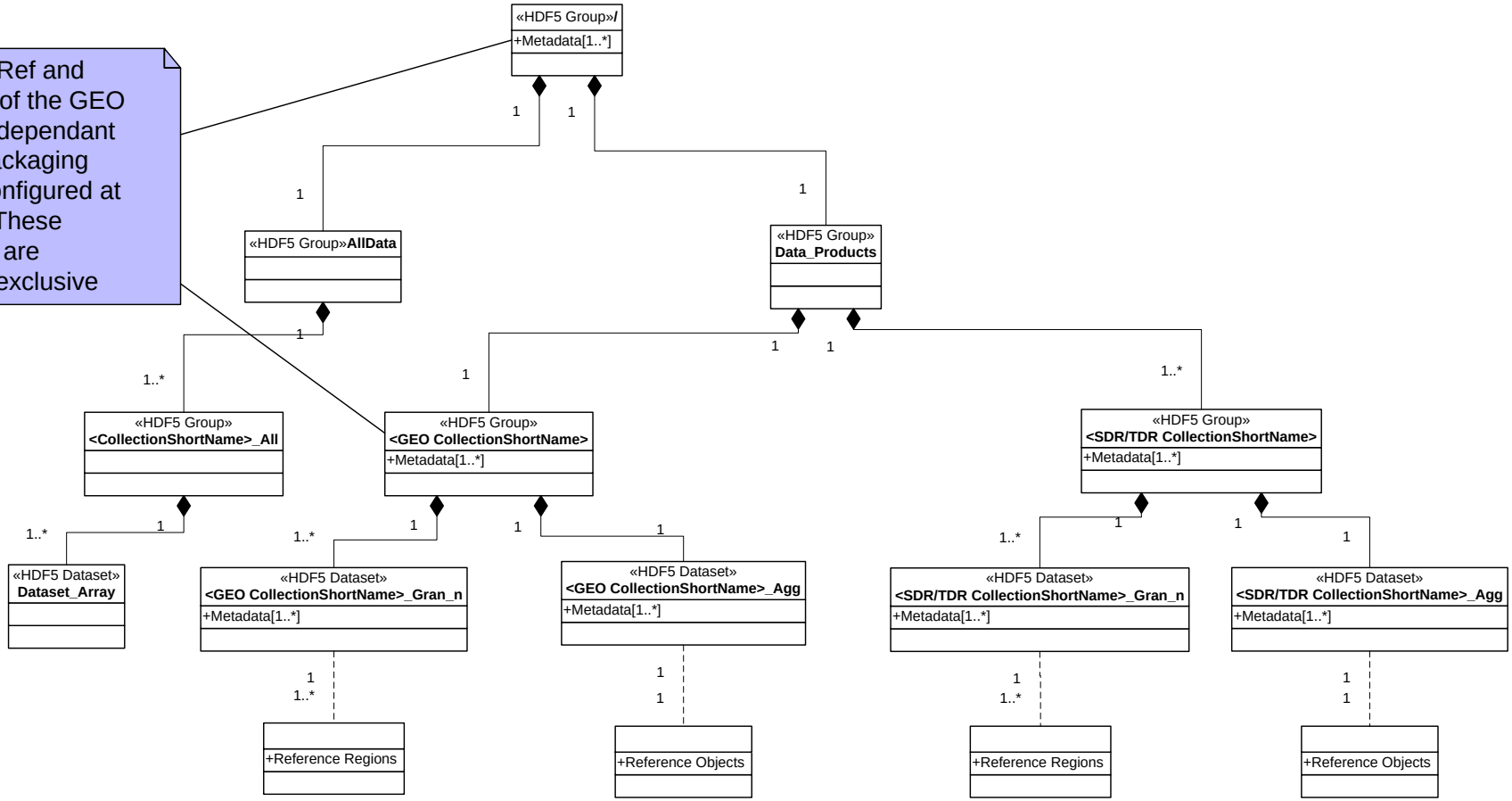


Common RDR Layout



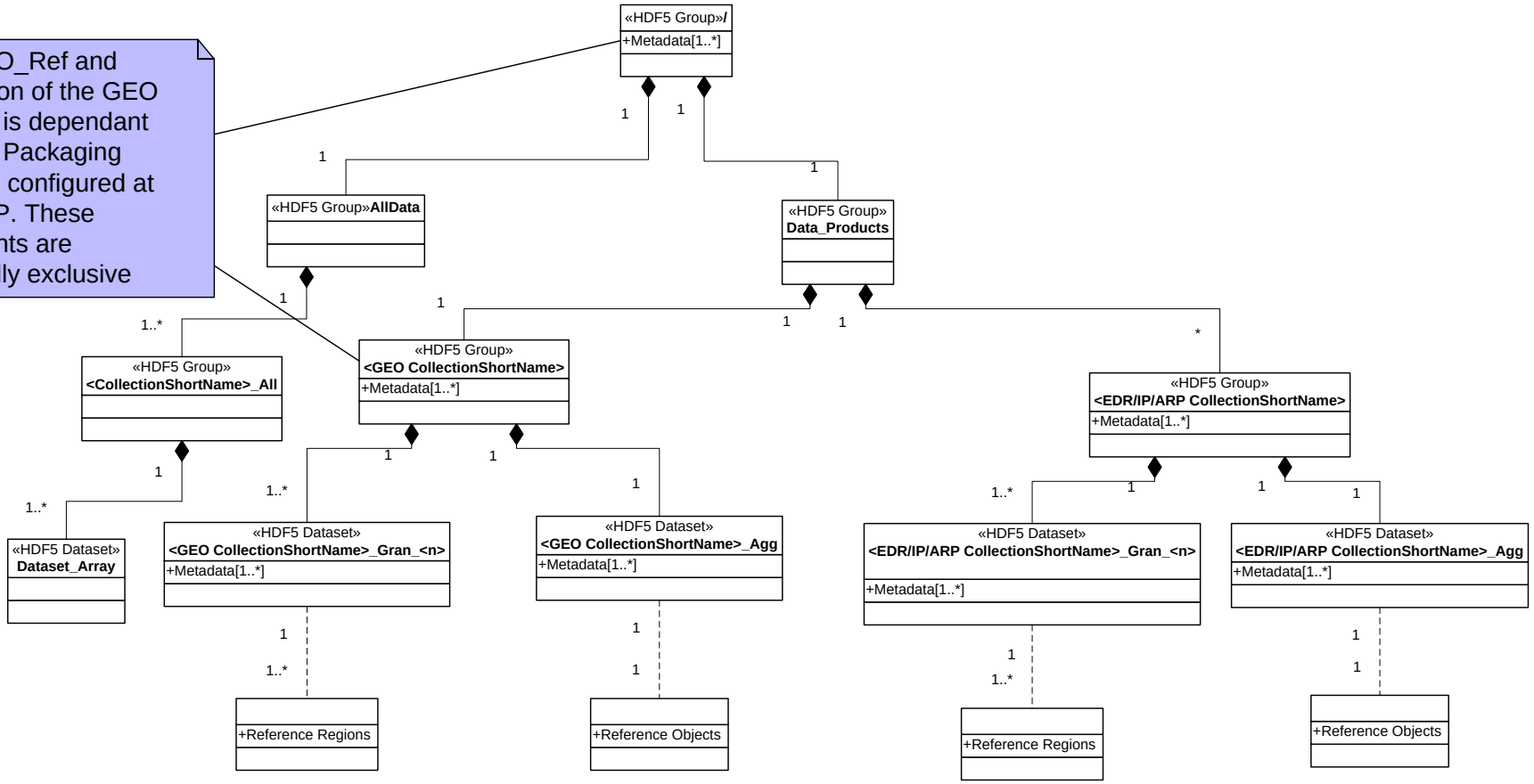
SDR/TDR UML Model

N_GEO_Ref and inclusion of the GEO Group is dependant on the Packaging Option configured at the IDP. These elements are mutually exclusive

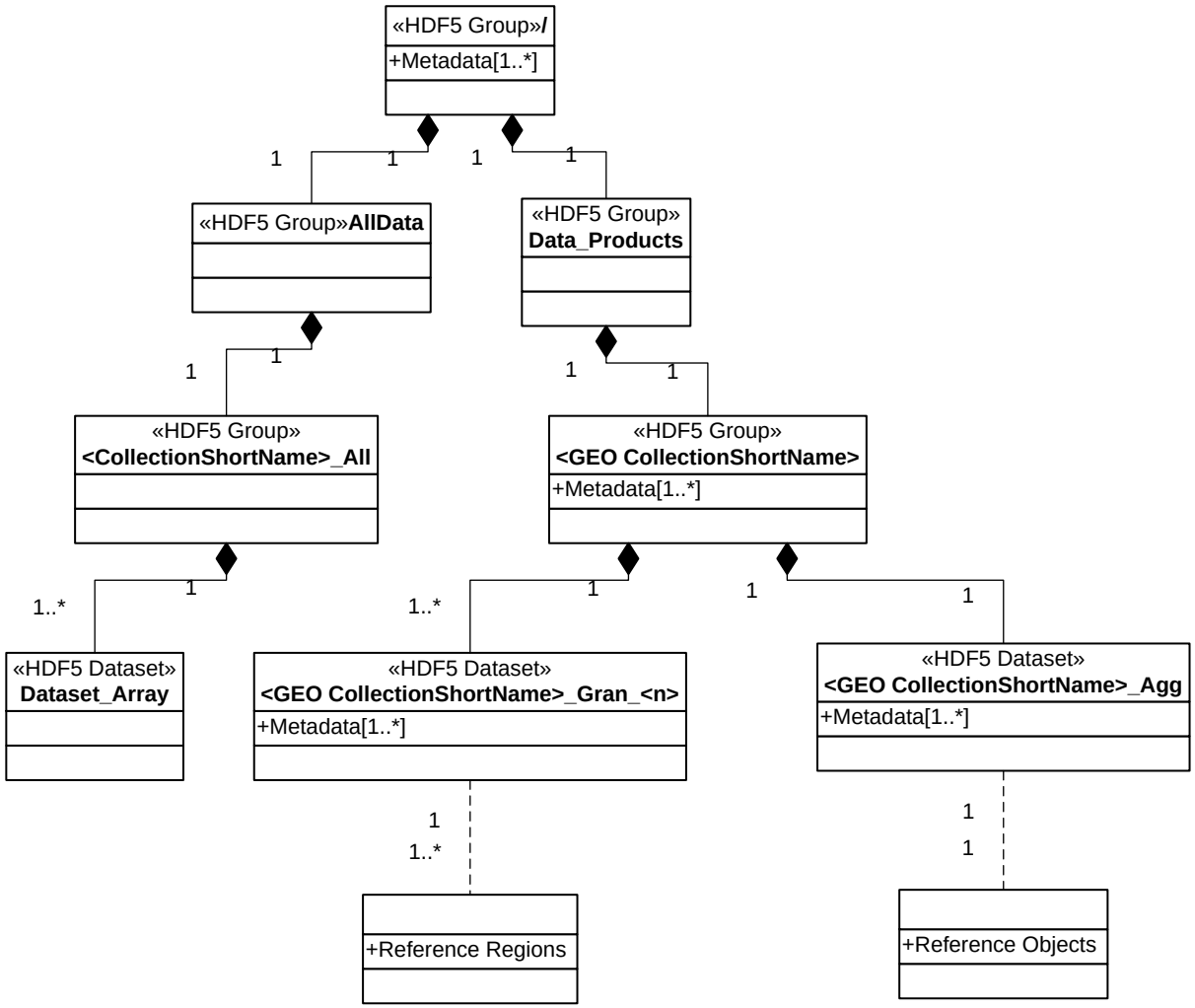


EDR UML Model

N_GEO_Ref and inclusion of the GEO Group is dependant on the Packaging Option configured at the IDP. These elements are mutually exclusive



Geolocation UML Model



Ancillary/Auxiliary UML Models