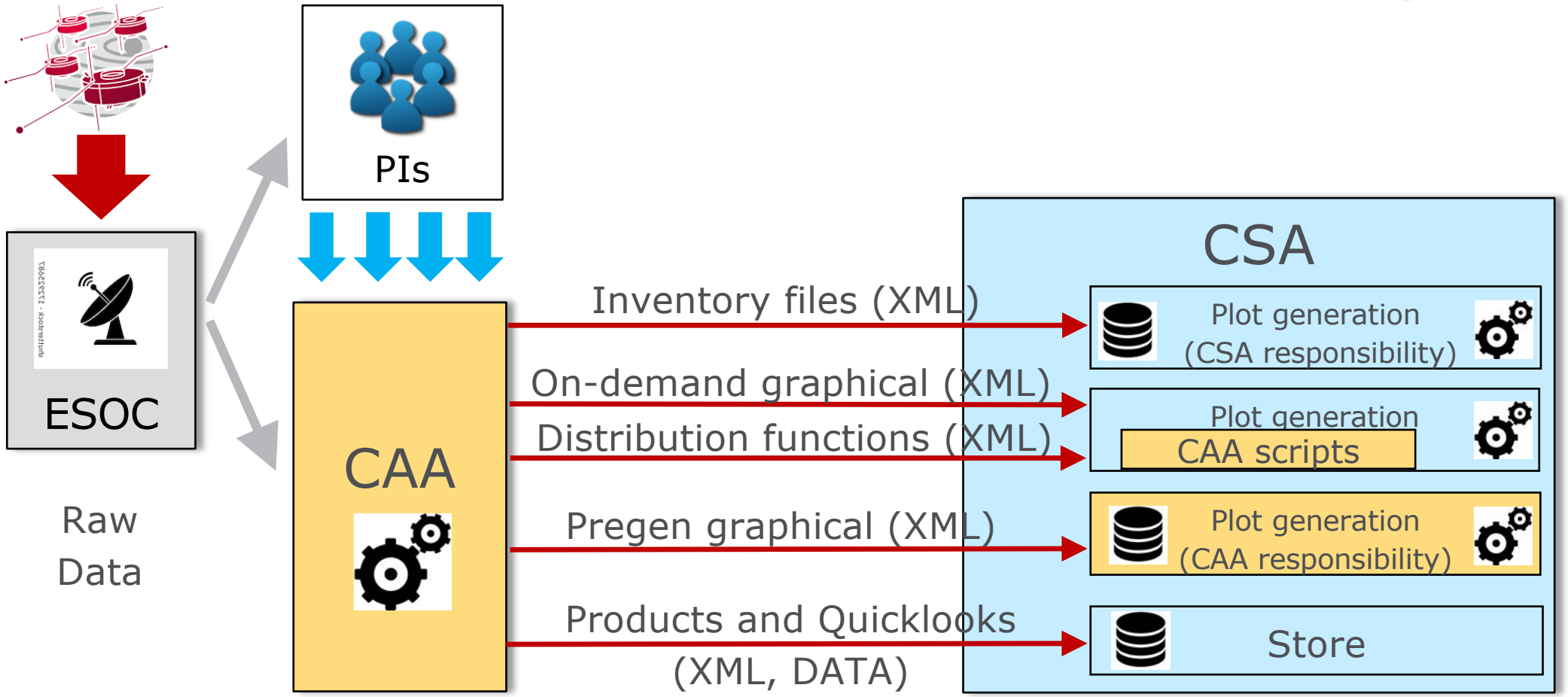
CSA data storage and concatenation

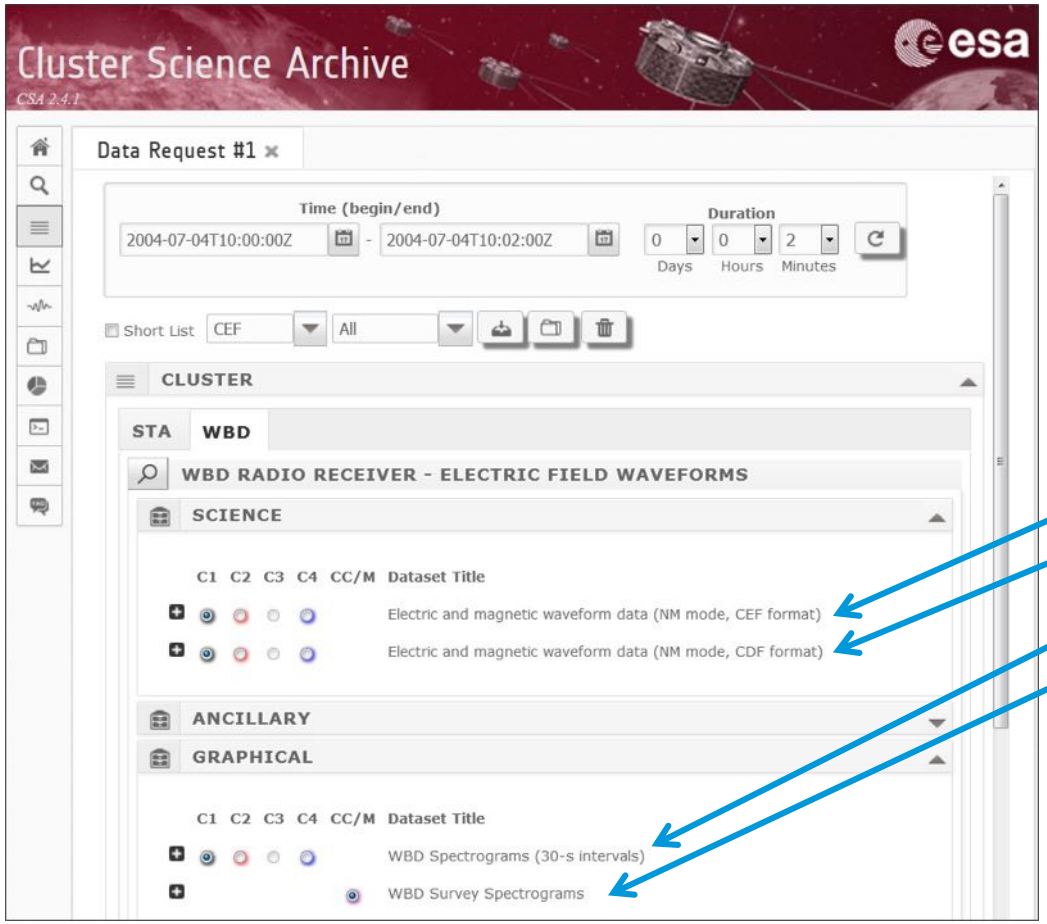
Beatriz Martinez
ESDC, European Space Astronomy Centre, ESA, Spain

IHDEA meeting, 16-18 Oct. 2019,
GSFC, Maryland, USA

Cluster Context Overview



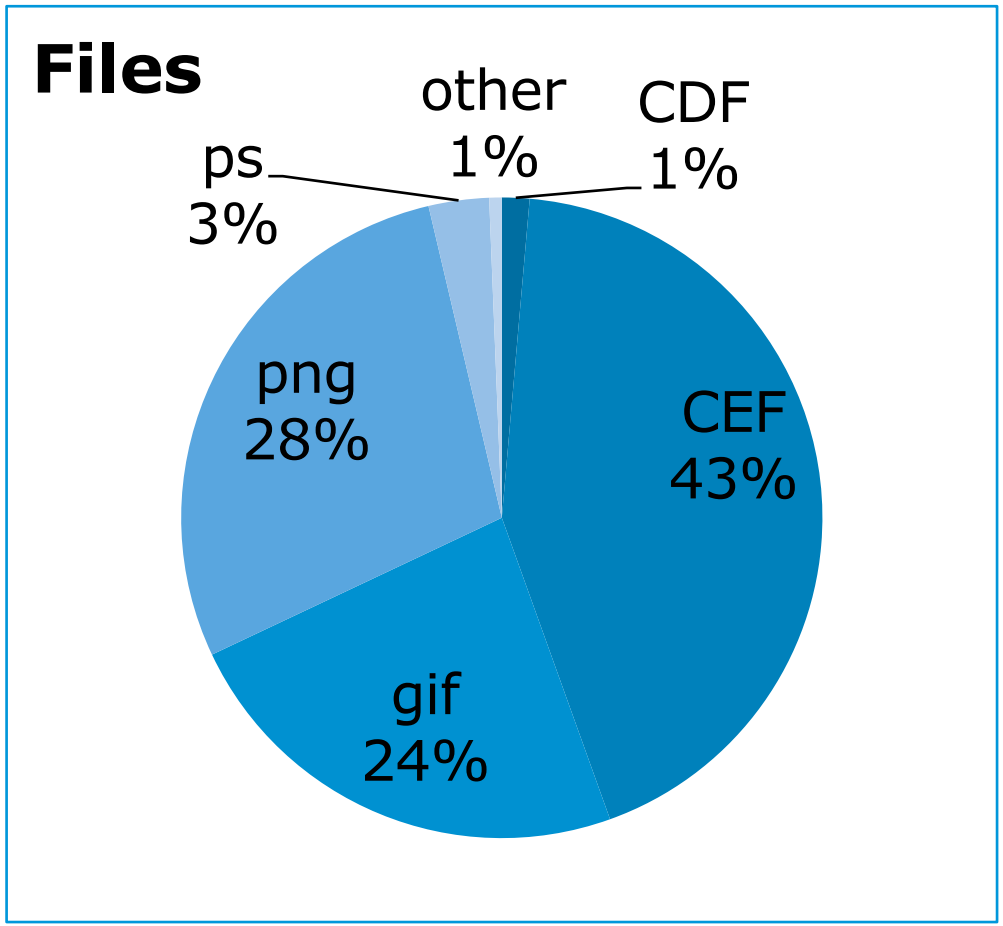
Cluster Science Archive (CSA) data



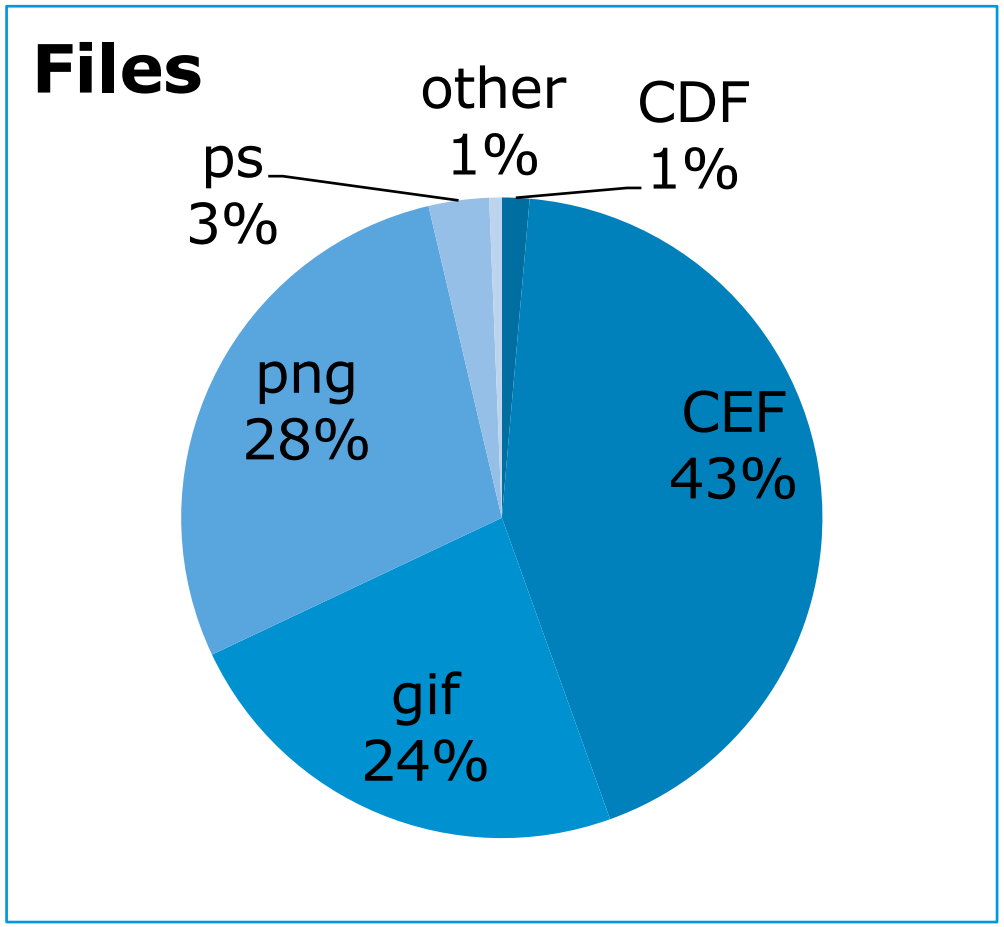
https://csa.esac.esa.int/csa-web/

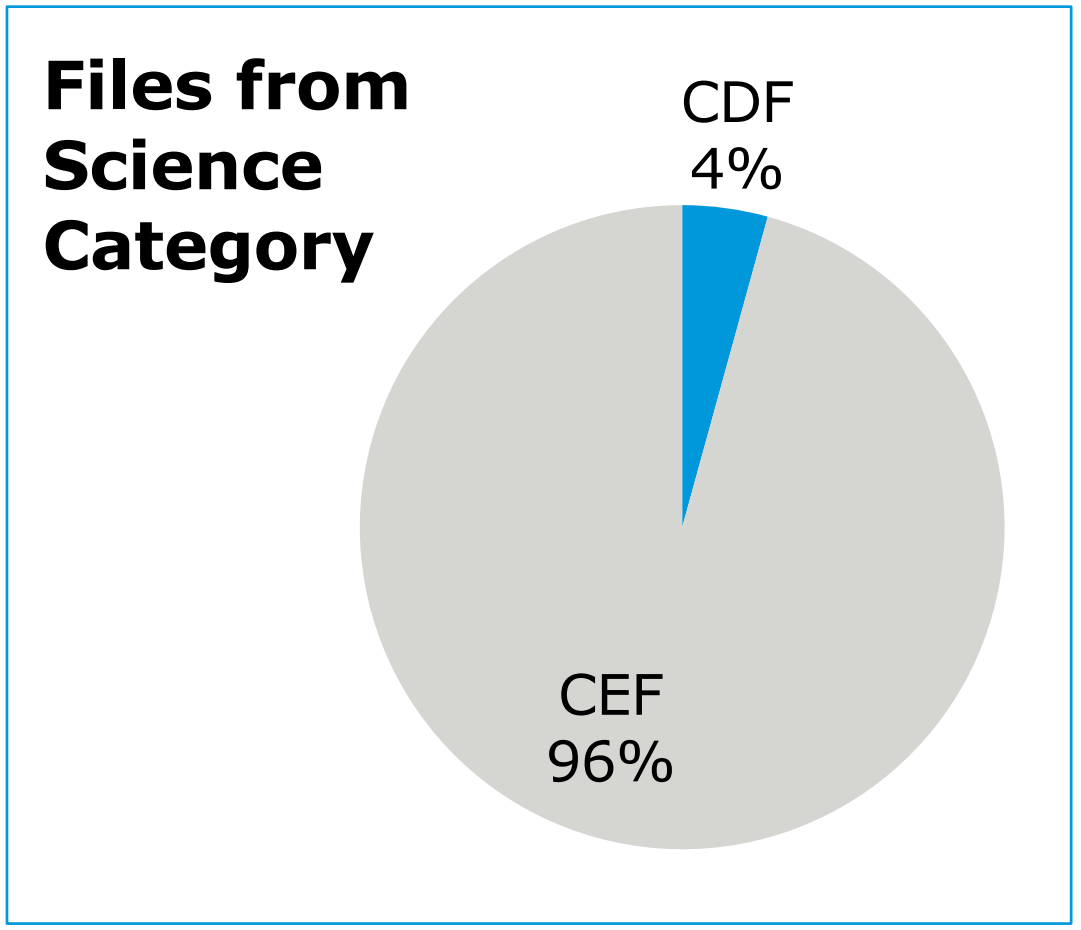
- Cluster data are transferred to CSA with associated metadata files (in XML), that describe them.
- Different data file formats:
  - CEF
  - CDF
  - gif
  - png
  - …
- > 17 millions of data files, ~ 117 TB

Format ratio (#files) at CSA



Format ratio (#files) at CSA





CEF – Cluster Exchange Format

- "**A single ASCII format data file syntax was recommended** by the Cluster Science Data System (CSDS) Archive Task Group **for the exchange of science data between instrument teams**. This format was intended as an exchange format **to allow translation between the several native data formats used by science tools and data bases within the Cluster community**. [...] It will also **facilitate delivery of science products** to future scientists **without access to specific software** in use at the time of data archival."
- Advantages:
  - ASCII format: easily readable and software independent
  - It handles sub-millisecond time resolution (WBD in ns, 50 µs accuracy), EPOCH16 only appeared in 2005 (EPOCH: ms accuracy)
  - Expanded the length of variable and attribute name from 64 to 256 (e.g. some PEACE variables longer than 64)
  - Detached header/metadata information
  - Control of the Metadata dictionary

CEF structure

HEADER

- File specific metadata
- Mission metadata
- Spacecraft metadata
- Experiment metadata
- Instrument metadata
- Dataset metadata
  - Variable(s) metadata

DATA

!! Data is a list of records, where the values are separated by “comma”, and dates are in ISO format.

```
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!                   File Metadata                      !!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
FILE_NAME = "C4_CP_FGM_SPIN__20040704_175500_20040704_180500_V140305.cef"
FILE_FORMAT_VERSION  = "CEF-2.0"
END_OF_RECORD_MARKER = "$"
!
START_META  = LOGICAL_FILE_ID
   ENTRY    =   "C4_CP_FGM_SPIN__20040704_175500_20040704_180500_V140305"
END_META    = LOGICAL_FILE_ID
!
START_META  = VERSION_NUMBER
   ENTRY    =   140305
END_META    = VERSION_NUMBER
!
....
!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!                 Global Metadata                      !!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

START_META = MISSION
    ENTRY  = "Cluster"
END_META   = MISSION

START_META      = MISSION_TIME_SPAN
    VALUE_TYPE = ISO_TIME_RANGE
    ENTRY      = 2000-07-16T00:00:00Z/2026-08-22T00:00:00Z
END_META       = MISSION_TIME_SPAN

...

START_META     =   OBSERVATORY
    ENTRY      =   "Cluster-4"
END_META       =   OBSERVATORY

...

START_META     =   EXPERIMENT
   ENTRY       =   "FGM"
END_META       =   EXPERIMENT

...
START_META     =   DATASET_ID
   ENTRY       =   "C4_CP_FGM_SPIN"
END_META       =   DATASET_ID

...
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!                      Variables                       !!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
START_VARIABLE  =  time_tags__C4_CP_FGM_SPIN
...
END_VARIABLE    =  time_tags__C4_CP_FGM_SPIN
START_VARIABLE  =  half_interval__C4_CP_FGM_SPIN
...
END_VARIABLE    =  half_interval__C4_CP_FGM_SPIN
START_VARIABLE  =  B_vec_xyz_gse__C4_CP_FGM_SPIN
...
END_VARIABLE    =  B_vec_xyz_gse__C4_CP_FGM_SPIN
START_VARIABLE  =  B_mag__C4_CP_FGM_SPIN
...
END_VARIABLE    =  B_mag__C4_CP_FGM_SPIN
START_VARIABLE  =  sc_pos_xyz_gse__C4_CP_FGM_SPIN
...
END_VARIABLE    =  sc_pos_xyz_gse__C4_CP_FGM_SPIN
START_VARIABLE  =  range__C4_CP_FGM_SPIN
...
END_VARIABLE    =  range__C4_CP_FGM_SPIN
START_VARIABLE  =  tm__C4_CP_FGM_SPIN
...
END_VARIABLE    =  tm__C4_CP_FGM_SPIN
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!                        Data                          !!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
! White space between entries (for readability) is safe but not needed! !
! Time field must be in ISO form

DATA_UNTIL = EOF
2004-07-04T17:55:03.460Z, 2, -211.301, -442.226, -83.166,
        497.120, 16601.7, 18335.1, 15196.6, 4, 22  $
2004-07-04T17:55:07.594Z, 2, -211.464, -442.113, -83.701,
        497.179, 16603.6, 18324.4, 15212.8, 4, 22  $
2004-07-04T17:55:11.728Z, 2, -211.688, -441.949, -84.236,
        497.219, 16605.5, 18313.7, 15229.0, 4, 22  $
...
2004-07-04T17:56:50.948Z, 2, -216.500, -438.649, -97.058,
        498.704, 16649.7, 18056.0, 15615.8, 4, 22  $
...
2004-07-04T18:04:54.643Z, 2, -241.176, -412.802, -155.982,
        502.894, 16827.4, 16759.4, 17465.1, 4, 22  $
2004-07-04T18:04:58.778Z, 2, -241.573, -412.453, -156.510,
        502.961, 16828.7, 16748.1, 17480.7, 4, 22  $
!RECORDS= 145
```

CEF – detached storage

- Metadata information is stored in separated files with extension CEH (Cluster Exchange Header).

  Ex.: CL_CH_FGM_EXP.ceh (FGM experiment description)

- Data files (CEF) stored contain data records and reference to the header files.
- Stored data files coverage is decided by Principal Investigators (PIs).
- Versioning only apply to data files, following an increasing number approach.

  Ex.: C4_CP_FGM_SPIN__20040704_173516_20040707_024233_V10.cef

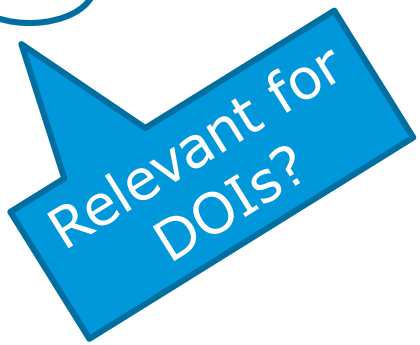
CEF – detached storage

- Metadata information is stored in separated files with extension CEH (Cluster Exchange Header).

  Ex.: CL_CH_FGM_EXP.ceh (FGM experiment description)

- Data files (CEF) stored contain data records and reference to the header files.
- Stored data files coverage is decided by Principal Investigators (PIs).
- Versioning only apply to data files, following an increasing number approach.

  Ex.: C4_CP_FGM_SPIN__20040704_173516_20040707_024233_V10.cef



CEF – detached storage

- Metadata information is stored in separated files with extension CEH (Cluster Exchange Header).

  Ex.: CL_CH_FGM_EXP.ceh (FGM experiment description)

- Data files (CEF) stored contain data records and reference to the header files.
- Stored data files coverage is decided by Principal Investigators (PIs).
- Versioning only apply to data files, following an increasing number approach.

  Ex.: C4_CP_FGM_SPIN__20040704_173516_20040707_024233_V10.cef

⇒ **Changes** on metadata information (ex. PI email) and on data (ex., recalibration) are **decoupled**.

⇒ **Concatenation required** before delivering to users, since data files stored are not valid CEF files.

Examples of CEH

```
!  CL_CH_MISSION.ceh
!  Global mission metadata provided by the CAA.
!
START_META = MISSION
    ENTRY  = "Cluster"
END_META   = MISSION

START_META     = MISSION_TIME_SPAN
    VALUE_TYPE = ISO_TIME_RANGE
    ENTRY      = 2000-07-16T00:00:00Z/2026-08-22T00:00:00Z
END_META       = MISSION_TIME_SPAN

START_META = MISSION_AGENCY
    ENTRY  = "ESA"
END_META   = MISSION_AGENCY

START_META = MISSION_DESCRIPTION
    ENTRY  = "The aim of the Cluster mission is to study small-scale structures of the magnetosphere "
    ENTRY  = "and its environment in three dimensions. To achieve this, Cluster is constituted of four "
    ENTRY  = "identical spacecraft that will flight in a tetrahedral configuration. The separation
distances "
    ENTRY  = "between the spacecraft will be varied between ~40 km and 10 000 km, according to the "
    ENTRY  = "key scientific regions."
END_META   = MISSION_DESCRIPTION

START_META = MISSION_KEY_PERSONNEL
    ENTRY  = "Philippe Escoubet>Philippe.Escoubet@esa.int >Cluster Project Scientist"
END_META   = MISSION_KEY_PERSONNEL

START_META = MISSION_REFERENCES
    ENTRY  = "The Cluster and Phoenix Missions>Cluster project and instrument teams>Space Sci. Rev. 79,
Nos. 1-2, 1997"
END_META   =  MISSION_REFERENCES

START_META = MISSION_REGION
    ENTRY  = "Solar_Wind"
    ENTRY  = "Bow_Shock"
    ENTRY  = "Magnetosheath"
    ENTRY  = "Magnetopause"
    ENTRY  = "Magnetosphere"
    ENTRY  = "Magnetotail"
    ENTRY  = "Polar_Cap"
    ENTRY  = "Auroral_Region"
    ENTRY  = "Cusp"
    ENTRY  = "Radiation_Belt"
    ENTRY  = "Plasmasphere"
END_META   = MISSION_REGION

START_META = MISSION_CAVEATS
 ENTRY = "*CL"
END_META = MISSION_CAVEATS
```

```
! CL_CH_FGM_EXP.ceh
! EXPERIMENT metadata
!
START_META     =   EXPERIMENT
   ENTRY       =   "FGM"
END_META       =   EXPERIMENT
!
! Description of the experiment
!
START_META     =   EXPERIMENT_DESCRIPTION
   ENTRY       =   "Each Cluster spacecraft carries an identical FGM instrument (Fluxgate Magnetometer) to "
   ENTRY       =   "measure the DC magnetic field vector. Each instrument, in turn, consists of two triaxial"
   ENTRY       =   "fluxgate magnetometers and an onboard data processing unit."
   ENTRY       =   "The instrument samples the magnetic field at a cadence of 22 Hz"
   ENTRY       =   "(67 Hz in Burst mode). In order to minimise the magnetic "
   ENTRY       =   "background of the spacecraft, one of the magnetometer sensors"
   ENTRY       =   " (the outboard, or OB sensor) is located at the end of one"
   ENTRY       =   " of the two 5 m radial booms of the spacecraft, the other "
   ENTRY       =   "(the inboard, or IB sensor) at 1.5 m inboard from the end "
   ENTRY       =   "of the boom. Since the start of the scientific operations "
   ENTRY       =   "on February 1, 2001, only the outboard sensor on each "
   ENTRY       =   "satellite has been used."
END_META       =   EXPERIMENT_DESCRIPTION
!
! Name and coordinates of the PI, and possible earlier PIs
!
START_META     =   INVESTIGATOR_COORDINATES
   ENTRY       =   "Chris Carr>PI>c.m.carr@imperial.ac.uk"
END_META       =   INVESTIGATOR_COORDINATES
!
! List of standard reference documents for the experiment
!
START_META     =   EXPERIMENT_REFERENCES
   ENTRY       =   "*CL_CD_CAA_FGM_ICD_V60.pdf"
   ENTRY       =   "*CL_CD_FGM_USERMAN.pdf"
   ENTRY       =   "http://www.sp.ph.ic.ac.uk/Cluster/"
END_META       =   EXPERIMENT_REFERENCES
!
! Name, role and coordinates of experiment key personnel
!
START_META     =   EXPERIMENT_KEY_PERSONNEL
   ENTRY       =   "Chris Carr>PI>c.m.carr@imperial.ac.uk"
END_META       =   EXPERIMENT_KEY_PERSONNEL
!
! Miscellaneous information concerning the experiment
!
START_META     =   EXPERIMENT_CAVEATS
   ENTRY       =   "*CL_CQ_FGM_CAVF.txt"
END_META       =   EXPERIMENT_CAVEATS
```

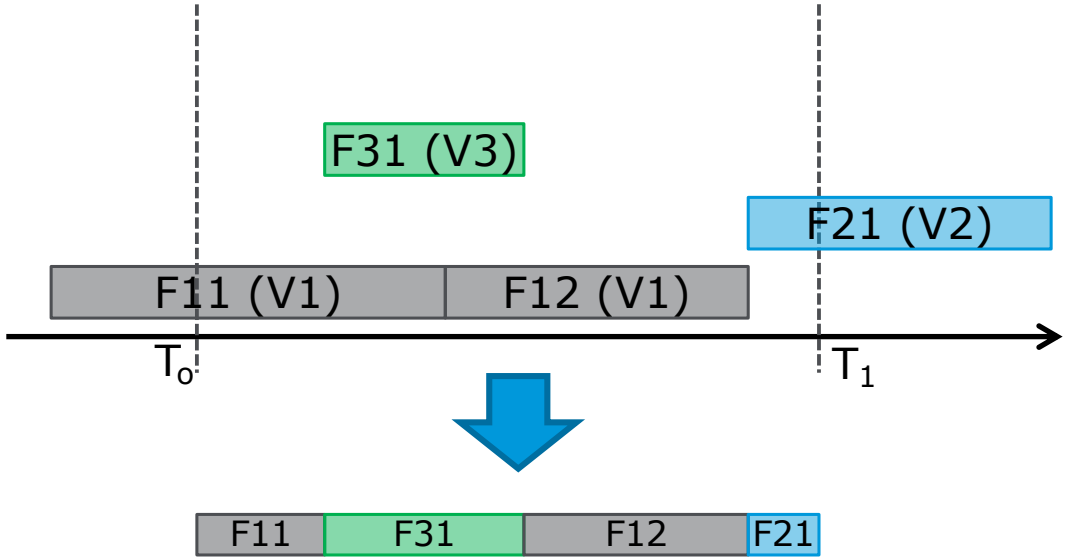
Example of CEF delivered by PI

```
!+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
! Production date: 2015-11-06 16:14:39
! Cluster/RAPID high resolution data in Cluster Exchange Format
!-------------------------------------------------------------|
FILE_NAME            = "C4_CP_RAP_ESPCT6__20040704_V02.cef"
FILE_FORMAT_VERSION  = "CEF-2.0"
END_OF_RECORD_MARKER = "$"

!================================
! Mission Level Meta Data
!================================
INCLUDE        =   "CL_CH_MISSION.ceh"
!================================
! Observatory Level Meta Data
!================================
INCLUDE        =   "C4_CH_OBS.ceh"
!================================
! Experiment Level Meta Data
!================================
INCLUDE        =   "CL_CH_RAP_EXP.ceh"
!================================
! Instrument Level Meta Data
!================================
INCLUDE        =   "C4_CH_RAP_INST.ceh"
INCLUDE        =   "C4_CH_RAP_ESPCT6.ceh"

!================================
! File Level Meta Data
!================================
START_META     =   LOGICAL_FILE_ID
   ENTRY       =   "C4_CP_RAP_ESPCT6__20040704_V02"
END_META       =   LOGICAL_FILE_ID
START_META     =   VERSION_NUMBER
   VALUE_TYPE  =   INT
   ENTRY       =   02
END_META       =   VERSION_NUMBER
START_META     =   FILE_TYPE
   ENTRY       =   "cef"
END_META       =   FILE_TYPE
START_META     =   METADATA_TYPE
   ENTRY       =   "CAA"
END_META       =   METADATA_TYPE
START_META     =   METADATA_VERSION
   ENTRY       =   "2.0"
END_META       =   METADATA_VERSION
START_META     =   FILE_CAVEATS
   ENTRY       =   "Release V011 of RAPID Data"
   ENTRY       =   "RAPID Processing software: MSF2SCI V9.1_20151106"
   ENTRY       =   "Spike removal: DESPIKE_SCI V20151027; sd=5.0, ratio=50., gaps>60.s, recs=+-5, At jumps"
   ENTRY       =   "IES SC heater noise (hatchet) removal: OFF"
   ENTRY       =   "Electron calibration file: RAP_IES_C4_V332.CAL_20140707"
   ENTRY       =   "SCI to CEF Packaging software: SCI_TRANS [LNX-20151105]"
   ENTRY       =   "With configuration: SCI2CAA [V011 20150801]"
   ENTRY       =   "Production Date: 2015-11-06T15:31:44Z"
   ENTRY       =   "RAPID Data produced with best-effort general calibration files."
   ENTRY       =   "Revised ion calibrations, version 3"
   ENTRY       =   "Revised electron calibrations, version 3"
   ENTRY       =   "Number of noise spikes removed: 0"
END_META       =   FILE_CAVEATS
START_META     =   FILE_TIME_SPAN
   VALUE_TYPE  =   ISO_TIME_RANGE
   ENTRY       =   2004-07-04T00:00:00.000Z/2004-07-04T23:59:59.999Z
END_META       =   FILE_TIME_SPAN
START_META     =   GENERATION_DATE
   VALUE_TYPE  =   ISO_TIME
   ENTRY       =   2015-11-06T15:31:44Z
END_META       =   GENERATION_DATE
START_META     =   DATASET_VERSION
   ENTRY       =   "3120"
END_META       =   DATASET_VERSION

!================================================
! Data
!================================================
DATA_UNTIL = "End_of_File"
 2004-07-04T00:00:03.117Z,   2.0670,
  1.65E+01, 3.97E+00, 1.85E+00, 1.98E+00, 8.46E-01, 5.95E-01,
  5.51E+00, 2.34E+00, 1.27E+00, 1.24E+00, 4.07E-01, 2.92E-01,
 3 $
 2004-07-04T00:00:07.251Z,   2.0670,
  4.26E+00, 4.06E+00, 2.72E+00, 0.00E+00, 4.06E-01, 4.39E-01,
  2.62E+00, 2.34E+00, 1.56E+00, 0.00E+00, 2.88E-01, 2.53E-01,
 3 $
 2004-07-04T00:00:11.385Z,   2.0670,
  0.00E+00, 3.06E+00, 3.62E+00, 7.16E-01, 2.04E-01, 5.83E-01,
  0.00E+00, 1.91E+00, 1.80E+00, 7.16E-01, 2.04E-01, 2.92E-01,
 3 $
 2004-07-04T00:00:15.519Z,   2.0670,
  8.87E+00, 4.06E+00, 8.99E-01, 2.15E+00, 8.13E-01, 5.84E-01,
  4.39E+00, 2.33E+00, 8.99E-01, 1.24E+00, 4.07E-01, 2.92E-01,
 3 $
 2004-07-04T00:00:19.653Z,   2.0670,
  4.71E+00, 1.35E+00, 4.44E+00, 2.17E+00, 6.11E-01, 4.41E-01,
  3.00E+00, 1.35E+00, 2.01E+00, 1.25E+00, 3.52E-01, 2.52E-01,
 3 $
...
 2004-07-04T23:59:57.892Z,   2.0670,
  5.35E+01, 1.30E+01, 7.30E+00, 2.19E+00, 4.11E-01, 2.93E-01,
  1.08E+01, 4.04E+00, 2.37E+00, 1.25E+00, 2.88E-01, 2.07E-01,
 3 $
End_of_File
```

Header includes

Delivery to user - Concatenation

- File to deliver is created on-the-fly after user selection of Dataset and Time Interval.
- If the Time Interval covers several original files with different versions, the highest version is selected per fraction:



Delivery to user - Concatenation

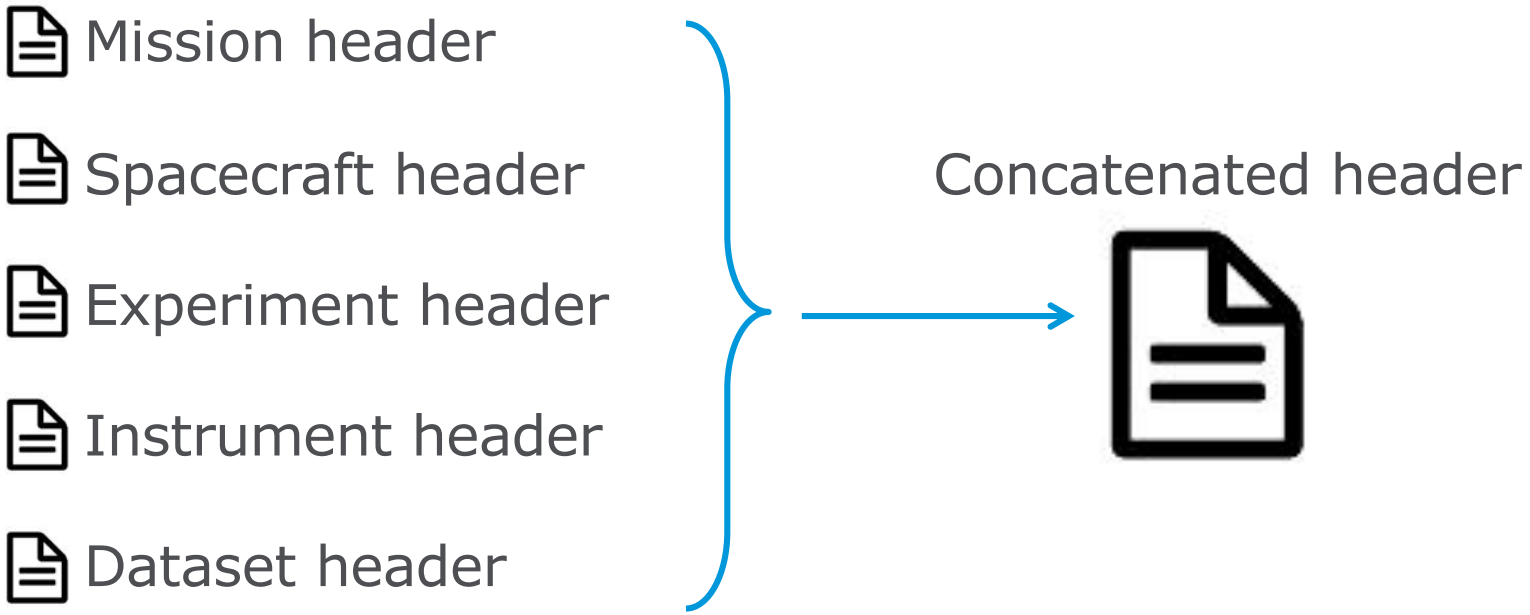
- File to deliver naming convention is:

```
<dataset id>__<start time>_<end time>_V<new version>.cef
```

Where:

- <start time> and <end time> are user selection in ‘yymmdd_HHmmss’ format
- <new version> is the date of the most recently ingested file in ‘yymmdd’ format.

Concatenation: build header of common metadata



Concatenation: aggregate data content

Example: F11 | F31 | F12 | F21

F11: from 'T_0' till start of F31

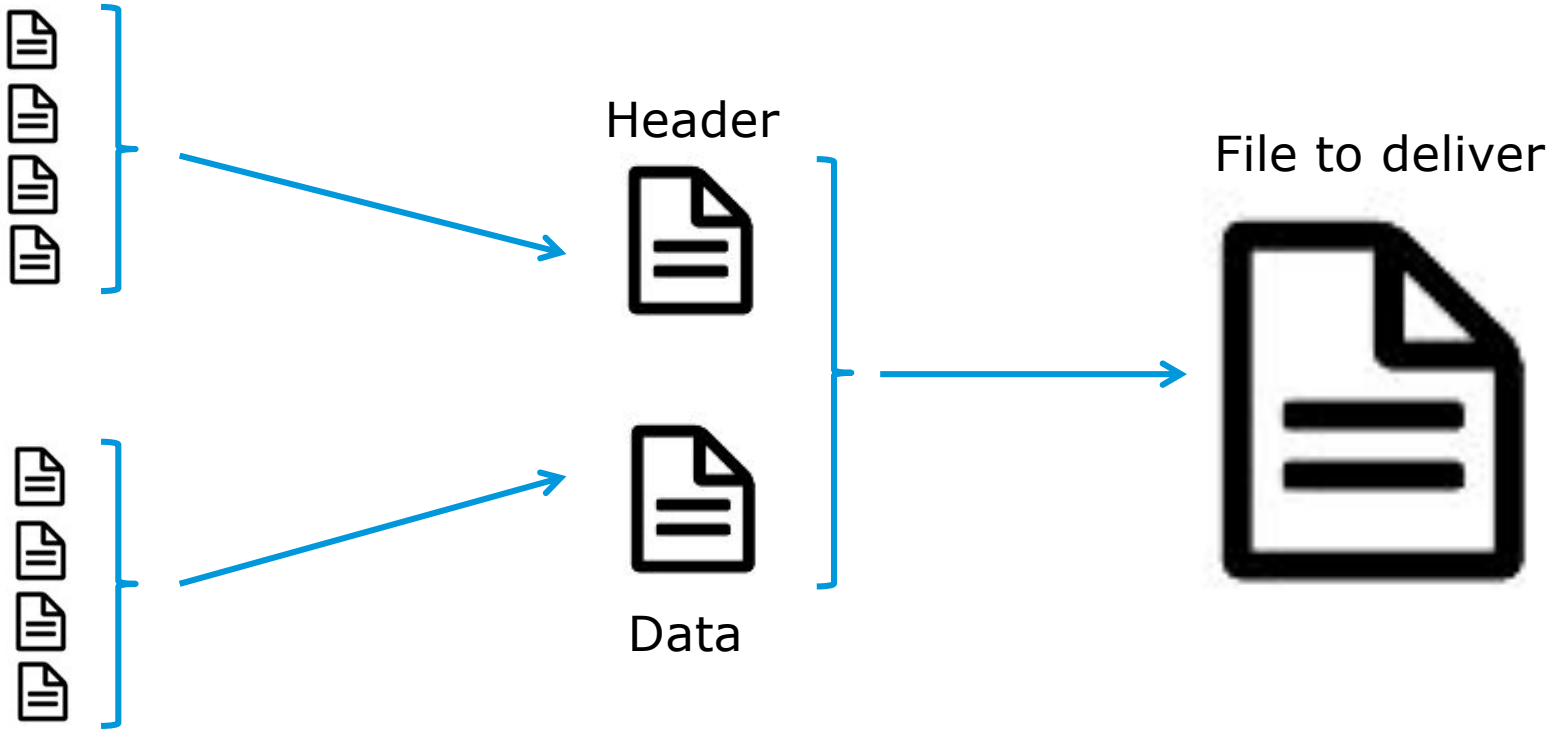
F31: from start F31 till end F31

F12: from end F31 till start F21

F21: from start F21 till 'T_1'

Concatenated data

Concatenation: attach Header and Data content



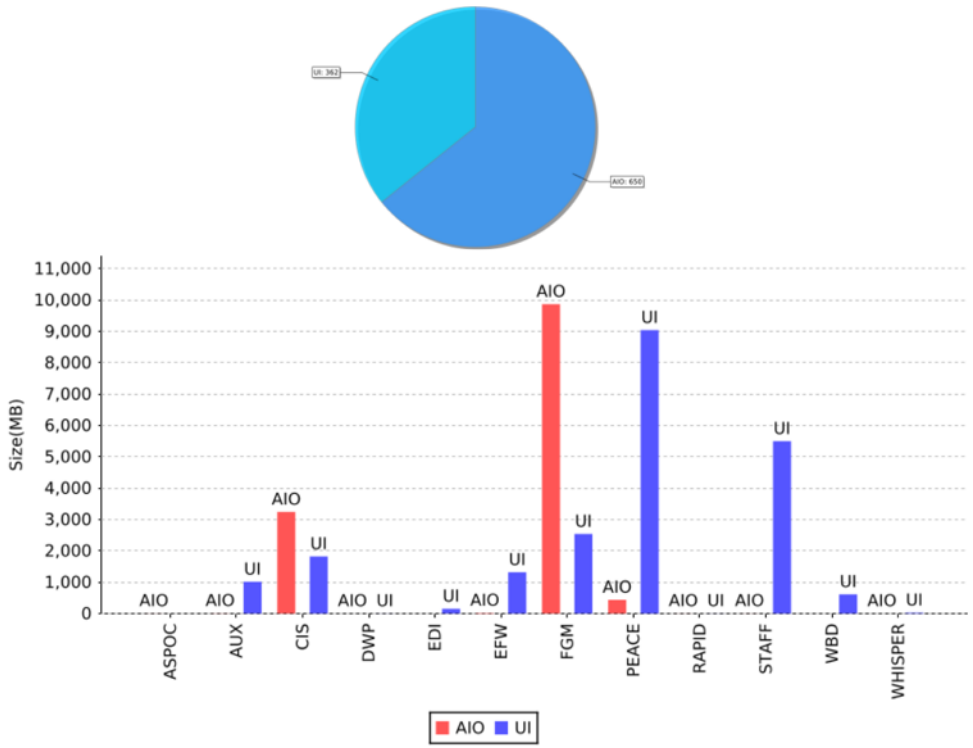
CEF files - Delivery format

Once the file is concatenated, it is delivered to the user:

- In CEF format (by default)
- In CDF format (plain conversion with SPARTA, previously Qtran)
- In CDF ISTP format (advance conversion with SPARTA)

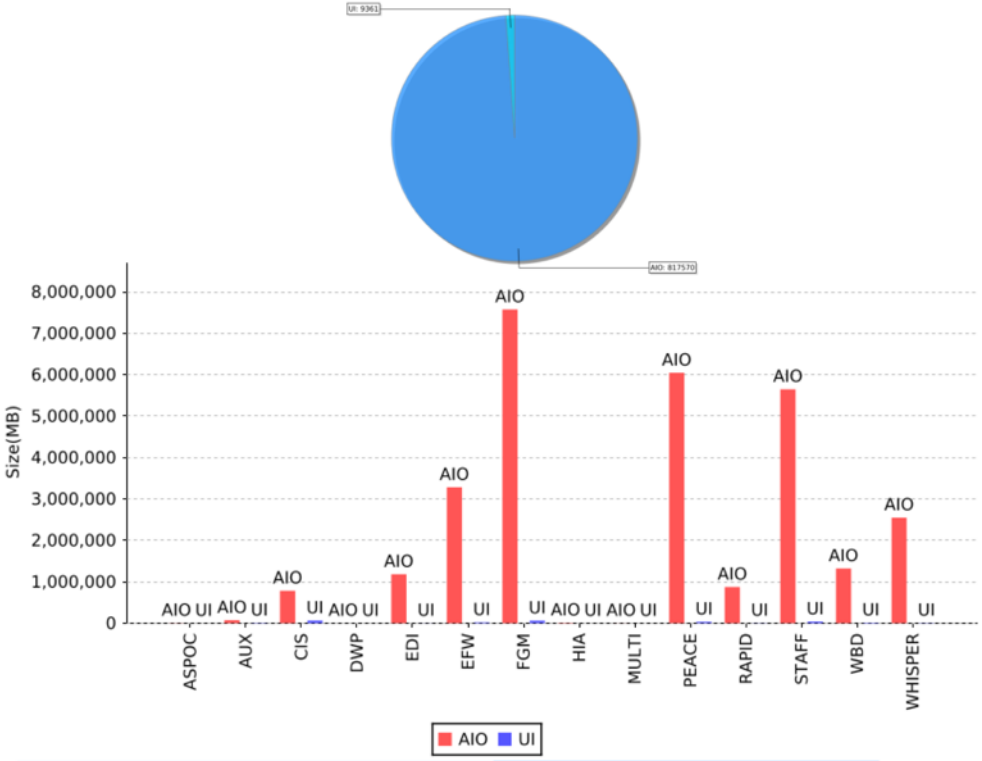
Download statistics per interface

Since 1 Oct. 2019



TEAM	AIO (MB)	UI (MB)	Total (MB)	Total (GB)
ALL	12938.2	20980.0	33918.2	33.1

Since 1 Jan. 2019



TEAM	AIO (MB)	UI (MB)	Total (MB)	Total (GB)
ALL	27982953.	232316.1	28215269.9	27554.0

CSA REST-api & HAPI

Through the Archive Inter-Operability subsystem (AIO), CSA archive offers a scripting interface similar to HAPI:

- CEF streaming ~ HAPI “data” endpoint

https://csa.esac.esa.int/csa/aio/html/streamingrequests.shtml

Ex.:

```
https://csa.esac.esa.int/csa/aio/streaming-
action?DATASET_ID=C3_CQ_DWP_INST&START_DATE=2007-01-15T18:00:00Z&END_DATE=2007-01-
16T00:00:00Z
```

CSA REST-api & HAPI (II)

- Metadata request ~ HAPI "catalog" endpoint

https://csa.esac.esa.int/csa/aio/html/metadatarequests.shtml

Ex.:

```
https://csa.esac.esa.int/csa/aio/metadata-action?
SELECTED_FIELDS=DATASET.DATASET_ID,
DATASET.TITLE&RESOURCE_CLASS=DATASET&EXPERIMENT.NAME=CIS&RETURN_TYPE=JSON
```

- Data Header request ~ HAPI "info" endpoint
  (Retrieves dataset headers in XML format)

https://csa.esac.esa.int/csa/aio/html/datarequests.shtml#HeaderRequests

Ex.:

```
https://csa.esac.esa.int/csa/aio/product-
action?RETRIEVALTYPE=HEADER&DATASET_ID=C*_CIS-CODIF_RPA_*
```

CSA REST-api working with IDL

Cluster data can be accessed through the AIO from:

- IDL https://csa.esac.esa.int/csa/aio/html/other_clients.shtml#idl

```
;define URL host and path
  csa_host = 'csa.esac.esa.int'
  csa_product_path = 'csa/aio/product-action'

;construct URL query from supplied parameters and keywords
  csa_product_query = ''
  for i = 0, n_elements(csa_dataset_id) -1 do begin
            csa_product_query=csa_product_query+'DATASET_ID='+csa_dataset_id[i]+'&'
  endfor
  csa_product_query =
csa_product_query+'&START_DATE='+csa_start_date+'&END_DATE='+csa_end_date+'&NON_BROWSER`
...
  csa_product_response = csa_product_obj->get(filename='csa_buffer.dat')
  csa_product_obj->getproperty, response_header=csa_product_header
...
```

CSA REST-api working with Python and Matlab

- Python https://csa.esac.esa.int/csa/aio/html/other_clients.shtml#python

```
...
myurl = 'https://csa.esac.esa.int/csa/aio/product-action'
query_specs = {'DATASET_ID': 'C1_CP_FGM_SPIN',
               'START_DATE': '2003-03-03T12:00:00Z', END_DATE': '2003-03-04T12:00:00Z',
               'DELIVERY_FORMAT': 'CEF','NON_BROWSER': '1',
               'DELIVERY_INTERVAL': 'houxrly',  'CSACOOKIE': }
download(myurl, query_specs, '20160616test.tar.gz')
...
```

- Matlab https://csa.esac.esa.int/csa/aio/html/other_clients.shtml#matlab

```
URL = 'https://csa.esac.esa.int/csa/aio/product-action';
fileName=tempname;
gzFileName = [fileName '.gz'];
options = weboptions('RequestMethod', 'get', 'Timeout', Inf);
tgzFileName = websave(gzFileName, URL, 'DATASET_ID', 'C1_CP_PEA_PITCH_SPIN_DPFlux', ...
    'START_DATE', '2008-04-24T21:40:00Z', 'END_DATE', '2008-04-25T22:10:00Z', ...
    'DELIVERY_FORMAT', 'CDF', 'NON_BROWSER', '1', 'DELIVERY_INTERVAL', 'HOURLY', ...
    'CSACOOKIE', <csacookie>, options);
...
```

Future Interoperability/Interfaces

- Simple Application Messaging Protocol (SAMP, from IVOA) implementation

Will allow send CEF/CDF files to SAMP enabled applications (Autoplot, AMDA,…)

- Table Access Protocol (TAP, from IVOA) server

Will allow to query and retrieve Cluster data through standard IVOA services, and eventually EPN-TAP to interact with Planetary services.

- Heliophysics Application Programmer's Interface (HAPI) server

Will allow scripting access from HAPI compliant clients

- Phython wrapper interface

Will allow to contribute to Sunpy and Heliopy for common access from python clients.

Thank you