CENTRE NATIONAL D'ÉTUDES SPATIALES

CU6 Spectroscopic Processing
2nd Workshop 12-13 Oct 2006

Work status on the definition of the CU6 System Architecture (WP 602-10000)

prepared by Anne JEAN-ANTOINE PICCOLO (CNES)

Scope

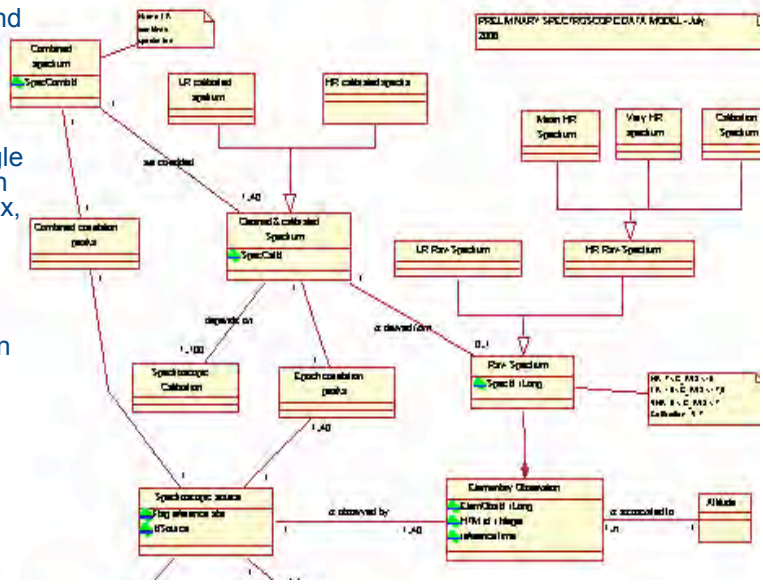
- **Work status on the WP 602-10000 for each task :**
  - **Task 1 : CU6 Functional analysis for spectroscopic processing (see Frédéric Thévenin 's talk)** ***IN PROGRESS***
    To be done : detail in depth the half-yearly processing and specially the multi transit analysis function (SADT diagrams and textual description)
  - **Task 2 : Software Requirement Specification phase at CU6 level** ***IN PROGRESS*** **(Focus on spectroscopic data structure studies and first definition of the spectroscopic data model in this talk)**
  - **Task 3 : Interface definition (*STARTING for external interfaces with the SOC*)**
  - **Task 4 : CPU (*NOT STARTED*) & storage resource assessment (*STARTING*)**
  - **Task 5 : Software Design phase at CU6 level (*NOT STARTED*)**
  - **Task 6 : Technical studies on data access layer (DAL), data management and centralized/distributed architectures (see François Jocteur 's talk about the current technical studies)** ***IN PROGRESS***

Task 2 : Spectroscopic data structure definition : objectives and context

- **Define the main spectroscopic data structures that shall be make persistent in the most appropriate storage space (either in a Relational Data Base Management System or on flat files or mixed solution) in order to :**
  - *Derive a significant data model (or class model) that could be the most appropriate for spectroscopic processing and where the main classes or objects will be identified in term of attributes, main methods, number of instances, … and also their relationships (number of tuples).*
- **Define the main access modes on spectroscopic data in order to :**
  - *derive a relevant and unique data access layer (DAL), to assess access mode performance in term of transaction number, transaction volume, access type, … and according to that, choice the best mode to access to data - testing for instance if it'd rather extracting data from data base or flat files before job execution or step by step inside the job.*
- **Task inputs :**
  - CU6 functional analysis in the current revision (end of July 2006) – analysing data flows identified between functions,
  - external ICD (from ESAC team – Excel file from Lennard) – analysing data structures described into the document and
  - information from David Katz (Many thanks !)
- **Task outputs :**
  - Requirement specifications in term of data description and storage volume
  - Data management requirements in input for the architecture studies (see : status on task 6).
- **Activities on-going : update the Software Requirement Specification document for the CU6 spectroscopic processing system and publish it on the CU6 Wiki for comments and improvements.**

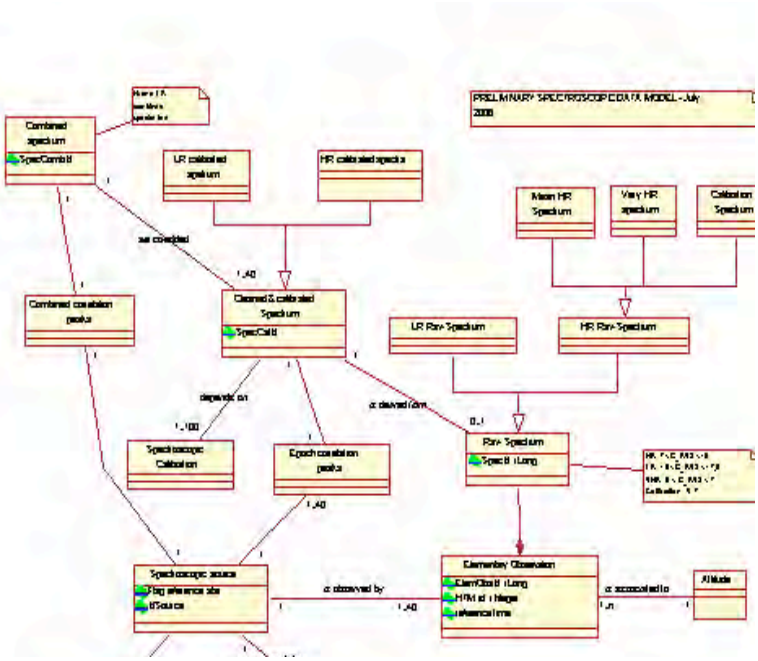
Task 2 : Design the spectroscopic data model by gathering main data classes (1/2)

- "Spectroscopic source" is a source observed with a spectroscopic point of view and defined with all parameters derived from spectroscopic processing (RV, Vsini, …) and associated to their other astrophysical parameters needed in input of some spectroscopic processing ;
- "Elementary Observation" stands for a single transit observation for a given source (or an observed object) defined in term of RVF flux, number of samples, sample values … see output from IDT processing ;
- "Raw spectrum" composes an elementary observation specialized into Low Resolution Spectrum or High Resolution Spectrum (which can be specialized themselves into Mean HR, Very HR or Calibration spectra) according to the CCD resolution ;
- "Cleaned and calibrated spectrum" output from "spectra extraction processing" ;
- "Combined spectrum" output from "multi transit analysis processing" ;



Task 2 : Design the spectroscopic data model by gathering main data classes (2/2)

- **"Spectroscopic Calibration" stands for RVS calibration parameters applied to calibrate observed spectra ;**
- **"Epoch correlation peak(s)" as peaks detected from a single transit spectrum ;**
- **"Combined correlation peak(s)" as peaks detected from a combined spectrum ;**
- **Others : "Attitude data", "Photometric magnitudes by epoch ", "Auxiliary data" as "Synthetic spectra", "Templates" and so on …**
- *TO BE CONFIRMED AND COMPLETED thanks to your algorithm specification and design.*



Task 2 : Detail on the data analysis : Sizing and estimations

- **We are currently assuming ~40 transits (or single observations) during the 5 years of the mission**
- **The limiting magnitude currently foreseen is G_RVS~17. There are ~336 10^6 stars brighter than G_RVS~17**
- **The transition between LR and HR mode is at magnitude G_RVS=10 (to be confirmed)**
- **All faint stars (G_RVS > 10) will be transmitted to the ground in LR mode (even if they have been observed in mixed mode)**
- **There are ~2 10^6 stars brighter than G_RVS=10 for a total of 1139 * 1 samples (1035 * 1 samples each (i.e. AL * AC) plus some samples for the measure of the background around the stars assumed as 10%).**
- **The number of stars observed in LR mode will be ~334 10^6. The LR windows are made of 345 * 1 samples plus ~10% background samples for a total of 380 * 1 samples.**
- **The very bright stars 6 <= V <= 7 (here the magnitudes are V magnitudes and not G_RVS magnitudes) will be observed in full 2-dim sampling. 1035 * 10 samples plus maybe an additional 10% for background measurement for a total of 1139 * 10 samples.**
- **The calibration windows will be made of 1139 * 10 samples (including the background samples).**

Task 2 : for instance … the description of the "Epoch correlation peaks" data set

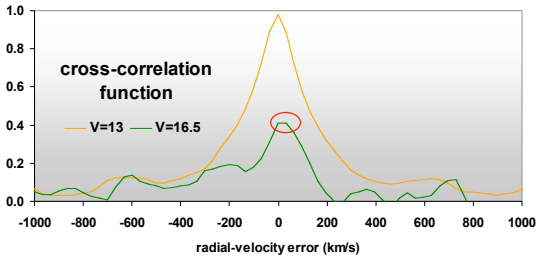
Data structure name		Bytes per logical record	Number of logical records (5 yr)	Total MB (5 yr)	Benchmark data v olume (MB)
Epoch correlation peaks		64340	336 000 000,00	20 616 760,25	
Data classification	**Name**	**Type**	**Assumed / Typical value**	**Typical number of values (multiplicity)**	**Bytes per logical record**	**Remark**
solution identifier	idSol	int		1	4	
creation time	cDate	int		1	4	
source identifier	id	long		1	8	
number of observations (transits)	nObs	int	40	1	4	
flags for the RVS transit	flagRvs	int		40	160	
velocity units	uniVel	float	200	40	160	From -1000 km/s to +1000 km/s with a step of 10 km/s
epoch velocity values	valEpPeakVel	double		8000	64000	assumed type

To be integrated in the SRS doc at CU6 level and put on Gaia Wiki

Nota : for each data structure, we assumed :

- the number of objects observed during the mission,
- the number of transits, the number of computed values,
- …

Total storage space over 5 years of mission in case of persistent data : 42 TB assumed.



Task 2 : typical access modes in spectroscopic processing

■ Extraction of spectra (each day / each six months) : will clean and calibrate raw spectra observed over some short interval of time.

To model the contaminating sources, we will have to search neighboring objects.

The data access may therefore be like this:

- ➢ Main sources : typical access by observation time.
- ➢ Contaminating sources : typical access by space references (HTM reference has be chosen here) and by source identifier.

■ Calibration of the RVS (SGIS each six months) : contains three steps :

- Characterization of the sources
  - ➢ typical access by magnitude (all sources brighter than TBD) and by source (all observations of the source maybe over some restricted interval of time).
- Identification of the sources suited as calibrators
  - ➢ typical access by magnitude (all sources brighter than TBD) and by source identifier.
- Calibration of the RVS
  - ➢ access by time (appropriate stars per calibration interval of time) and by properties (all sources qualified as appropriate calibrators).

NB : The access will not be very different for the daily calibration (if implemented) and for the 6-months SGIS.

■ Single transit analysis: performed on the source of the day, therefore the main access will probably be

➢ by observation time

■ Multiple transit analysis:

■ ➢ The access will mainly be by source identifier (analysing all the epoch spectra / epoch RV of a given source).

Conclusion

- This analysis is just a starting point and has to be confirmed and completed by the software module specifications and design in each WP.
- Some questions to the WP managers and developers will be put in the next weeks to help us in the DB prototype implementation.

Thanks for your attention