Big Data Space Science: ESA Planck & Euclid

A. Zacchei
INAF-OATs

on behalf of the Planck and Euclid Collaborations
The Planck mission

• The Planck Satellite was launched on 14 May 2009 and concluded its operations on October 2013.
• It carried on board two Instruments: LFI (Low Frequency Instrument) and HFI (High Frequency Instrument) that observed the entire sky for about 4 years in 9 different frequencies.
• Its scope, in one sentence, was to map the microwave sky with an unprecedented combination of sky coverage, frequency range, calibration accuracy, independence from systematic errors, stability and sensitivity, to provide the most precise CMB and foregrounds maps.
• Planck was a “P.I. Mission” coordinated by ESA. Two consortia lead by two P.I. had the responsibility to operate the two instruments, process the data and deliver to ESA the agreed products. ESA, through its archive system at ESAC, was then in charge of making those products available.
The Planck Data Processing

- Two Planck DPC (Data Processing centers) have been the responsible of the data analysis and both follow the same overall approach to the data reduction with specific tasks aimed to correct instrument dependent systematics;
- Process has been logically divided in three main levels:
  - Level 1 where HK and Science telemetry received from the satellite were transformed in raw timelines and stored in dedicated database with time information associated;
  - Level 2 was dedicated to synthesize the instrument information in the Instrument Model, remove the systematics, flag the data that are considered not usable, calibrate the data and finally create the maps and all associated products;
  - Level 3 was dedicated to the more scientific analysis with the responsibility to separate components into catalogues and specific astrophysical emissions, and evaluate CMB spectra and likelihood.
- Each step was internally validated and most of the DPC time was spent to cross check all the results first internally and then between instruments.
- NOTE that pipeline and entire process was developed independently at each DPC adding strong value to the cross-instrument validation;
Planck- Temperature maps

A. Zacchei "Frequency maps generation"
In Planck we had an incremental release approach.

We had three releases and we are preparing right now the final release that should serve as legacy for the incoming years.

This release contains all the data acquired, starting from raw packets till final products like CMB and foreground maps and it is correlated with all the simulation and auxiliary files that have been used to validate and assess the error bar.

Together with the data we will publish a coordinated set of scientific papers and an explanatory document that describe the data and the information necessary to “replicate” our results. Till now about 212 papers have been published by the Planck Consortia and 42 are in the editing phase to describe the last release.
Planck Products 2/3

- The Planck data set released is composed by:
  - Timelines (raw, before photometric calibration, after photometric calibration) for a total of ~ 30 TB → aimed to special Users, fits based structure NO VO compliant.
  - MAPs, for frequency, different time period → about 4000 files < 2 TB → aimed to general user, FITS based and VO compliant to be compared with other surveys.
  - CMB and Foregrounds (Dust, Synchrotron, CO etc...) maps → less than 50, few GB aimed to general user, FITS based VO compliant. Those maps represent the REAL SCIENCE OUTPUT.
  - Auxiliary products (beams, instrument parameters, noise covariance matrix etc ..) ~ 10 TB → aimed to specialists, different format not easy to use.
  - Simulations used to validate andasses the error bars, ~ 20 TB → aimed to specialists.
Planck Products 3/3

- All the product has been ingested in to the ESA – ESAC archive that is the ONLY public interface to the Planck Data.
- ESAC developed a software interface that allow the user to create own maps, correlate Planck results with others Surveys and visualize them (ESA-SKY).

A. Zacchei "Big data Space Science: ESA Planck & Euclid"
Planck LFI Polarization Amplitude at 30 GHz + Herschel footprints
Planck LFI Polarization Amplitude at 30 GHz + XMM footprints
The Euclid mission

• Euclid is a space-borne survey mission dedicated to investigate the origin of the Universe's accelerating expansion and the nature of dark energy, dark matter and gravity. Euclid will characterise the signatures of dark energy on the 3D distribution of cosmic structures. In 2012, Euclid was approved as the second Medium Class mission (M2) in the Cosmic Vision Programme.

• Euclid will be launched on May 2021 and operate (baseline) for about 7 years.

• It will carry on board two instrument VIS (Visible Imager) and NISP (Near Infrared Spectrum Photometer).

• Euclid is a ESA Mission where the Data analysis and Instrument realization are in charge to a Consortium. The Consortium has the responsibility to reach the scientific goals and to provide to ESA the agreed products. ESA, through its archive system at ESAC, will be then in charge of making those products available.
The Euclid Data Processing

• The SGS is then essentially a distributed structure.
• Euclid will produce and use a big amount of data (estimated to be at the end of the mission of the order of 100 PB). It will be then essential to avoid excessive data transfer, to develop a structure where the code will be moved instead of the data.

• A common Data Model and common infrastructure has been built → EACH Science Data Center should be able to run the same code → the sky can be divided in patches to be analyzed in parallel in different centers to minimize data transfer between SDCs.

• The Data processing pipeline in Euclid will be a series of Processing Functions: designed by the OUs (Organization Units, scientist), developed in collaboration between the OUs and SDC developers, integrated by the SDCs, and running on the SDCs infrastructure.
### Possible scenario: per-determined SDC-Sky allocation

<table>
<thead>
<tr>
<th>Science Operations Centre</th>
<th>SDC-NL</th>
<th>SDC-DE</th>
<th>SDC-FR</th>
<th>SDC-CH</th>
<th>SDC-ES</th>
<th>SDC-US</th>
<th>SDC-UK</th>
<th>SDC-FI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Data</td>
<td>Raw EXT Data</td>
<td>Raw EXT Data</td>
<td>Raw EXT Data (TBC)</td>
<td>Data Files</td>
<td>Data Files</td>
<td>Data Files</td>
<td>Data Files</td>
<td>Data Files</td>
</tr>
<tr>
<td>Level 1 Data Files</td>
<td>Data Files</td>
<td>Data Files</td>
<td>Data Files</td>
<td>Data Files</td>
<td>Data Files</td>
<td>Data Files</td>
<td>Data Files</td>
<td>Data Files</td>
</tr>
<tr>
<td>Metadata (prime)</td>
<td>Metadata (backup)</td>
<td>Metadata (backup)</td>
<td>Metadata (backup)</td>
<td>Metadata (backup)</td>
<td>Metadata (backup)</td>
<td>Metadata (backup)</td>
<td>Metadata (backup)</td>
<td>Metadata (backup)</td>
</tr>
</tbody>
</table>

*Redundant storage of the data*
The Euclid Data Processing

• Euclid will release the data to the Public with an incremental approach (three/four releases are foreseen) and will be the *legacy* galaxy catalogue. It will be then very important to build it in the MOST compatible way.

• For this reason one of the requirements is to save all the data in a VO compatible format to allow easy interoperability and cross checking with others surveys. This is true NOT only for the Euclid products but is also for the intermediate products that may become part of the formal delivery.

• This was NOT true in Planck, where we agreed to use a VO compatible structure only at the moment of creation of the public products.
Conclusions (1/2)

• Data in astronomy are not only OPEN but also PUBLIC, usually after a proprietary period of 1 – 2 years. In the case of Planck and Euclid, data are periodically RELEASED through ESAC and made therefore immediately PUBLIC.

• Compliance with the Virtual Observatory (VO) standards allow the data to follow the FAIR (Findable, Accessible, Interoperable, Reusable) paradigm. We can say that Planck data achieved VO-compliance due to an internal agreement, whereas in Euclid it is part of the requirements.
Conclusions (2/2)

• To maximize the **reusability** (FAIR) of data, it is important to offer **services** (e.g. visual inspection tools, cross-matching functionalities, etc ...) where the data are. This has been provided for Planck data.

• Due to the enormous amount of data, in case of Euclid it will be almost impossible for an external user to download them locally. This means that the ESA-ESAC archive, where Euclid data will be stored and made available, should provide services. They should be based on **cloud computing** and **EOSC** should play an important role to define rules to be followed and to provide tools.

• Data access/retrieval will be public, while services on data will be open, but may not be free for all.