Use Cases for the ESAC Science Exploitation and Preservation Platform (SEPP)

Christophe Arviset, Vicente Navarro, Ruben Alvarez
ESA Operations Department, ESAC, Spain

BIDS 2019, Munich, 21/02/2019
SOLAR SYSTEM EXPLORERS

- Solar Orbiter (2020)
- Proba-3 (2019)
- SMILE (2019)
- ExoMars Trace Gas Orbiter (2016)
- Juice (2023)
- Venus Express (2006-2014)
- Giotto (1985-1993)
- Double Star (1996-2004)
- SMART-1 (2003-2006)
- Chandrayaan (2008)
- Chang'e 1 (2007)
- Rosetta (2004-2016)
- Ulysses (1990-2009)

Ice Giants Mission
Concepts
In Development
Operational
Legacy
Enable maximum scientific exploitation of data sets

Enable efficient long-term preservation of data, software, and knowledge, using modern technology

Enable cost-effective archive production by integration in, and across, projects

Science Exploitation Platform
The **traditional Science Exploitation concept** is based on moving data and tools to the user, therefore transferred many times, replicated in many places, and with data exploitation taking place at users’ premises.
HOW MANY STARS WILL THERE BE IN THE SECOND GAIA DATA RELEASE?

- **position & brightness on the sky**: 1,692,919,135
- **surface temperature**: 161,497,595
- **red colour**: 1,383,551,713
- **blue colour**: 1,381,964,755
- **parallax and proper motion**: 1,331,909,727
- **radius & luminosity**: 76,956,778
- **amount of dust along the line of sight**: 87,733,672

- **14,099** Solar System objects
- **550,737** variable sources
- **7,224,631** radial velocity

The second data release of ESA's Gaia mission is scheduled for publication on 25 April 2018.
New Archive usage paradigm

Data exploitation from science community of mission data brings challenges for us

Execution of user’s code at server side
Move code to the data
Sharing data (collaborative)
State of the art technologies of big data mining
Visualization techniques

For the community:
New way of working for scientists
The fundamental principle of **SEPP’s concept** is to move the user to the data and tools. Users access a **science platform** providing the data, tools, and resources required, as opposed to downloading, replicating, and exploiting data ‘at home’. 
Top Level scenarios

Data Exploitation

- frameworks to manipulate and analyze data interactively.
- Creation and sharing of documents that contain live code, visualization, etc.

Collaborative Research env.

- user personalized storage (scratch, persistent and public mounted areas) and execution environment.
- It allows users to publish and share their assets.

Pipeline Develop. env.

- pipelines for data integration, transformation and analytics based on processing assets developed by SEPP actors.

Software Preservation

- provides on-the-fly instantiation to legacy software, through full environment or predefined processing threads.
Data Exploitation

Enable data processing where the data is, ie the Archives

Provide generic pipeline development and testing environment

Interactive data analysis Jupyter Notebook / Hub
Collaborative Research Environment

Share your data (user storage space in the platform, VOSpace)

Share your metadata (DB user space inside the archive, ...)

Publish your data through standard data protocols (ie VO)
Data Pipeline

Develop your own customized processing pipelines

Share your code through a Science App Store
Legacy Software Preservation

Software from Legacy Missions
- Mission Planning System (ie Rosetta on the 67P comet)
- Could be re-instantiated in years for future missions (ie JUICE)

Data Processing and Analysis Software
- Enable easier access to legacy software
  - On-the-fly instantiation of full system / predefined processing threads
- Reproducibility of data processing in the future (FAIR principle)
- Rescue the code and bring it to the (small) data
Early adopters, followers, observers

**Early Adopters:**
- Ready to adopt a SEPP scenario as soon as it is available
- First use cases to be implemented

**Followers:**
- Thought to adopt a SEPP scenario once positive feedback from Early Adopters is available

**Observers:**
- The need for a specific SEPP scenario is not clear today or has not explicitly been mentioned
SEPP scenarios and adoption

<table>
<thead>
<tr>
<th>SEPP Use Cases</th>
<th>SEPP Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Storage Area</td>
</tr>
<tr>
<td></td>
<td>Interactive Analysis (Jupiter)</td>
</tr>
<tr>
<td></td>
<td>Web instantiation of legacy systems</td>
</tr>
<tr>
<td></td>
<td>Execution of project/user pipelines</td>
</tr>
<tr>
<td></td>
<td>Pipeline development framework</td>
</tr>
<tr>
<td></td>
<td>Interoperability with VO</td>
</tr>
<tr>
<td></td>
<td>Science App Store</td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>PLA-CRA: Planck Legacy Archive Collaborative Research Area</td>
<td>Early adopter</td>
</tr>
<tr>
<td>BFG-PSE: BepColombo Instrument Pipelines Scheduling &amp; Execution</td>
<td>Follower</td>
</tr>
<tr>
<td>EUC-CRA: Euclid Collaborative Research Area</td>
<td>Follower</td>
</tr>
<tr>
<td>XMM-BPCA: XMM-Newton legacy Science Processing Capability Area</td>
<td>Follower</td>
</tr>
<tr>
<td>GSSC-CRA: GSSC Collaborative Research Area</td>
<td>Follower</td>
</tr>
<tr>
<td>PLT-PFIA: PLATO SOC Pipeline Framework Integration Area</td>
<td>Early adopter</td>
</tr>
<tr>
<td>EXO-MARA: ExoMars ESA Exploratory Mission Archive Research Area</td>
<td>Follower</td>
</tr>
<tr>
<td>JWST-WS: JWST Workspaces</td>
<td>Follower</td>
</tr>
<tr>
<td>GAIA-IDE: Gaia Interactive Data Exploration</td>
<td>Early adopter</td>
</tr>
<tr>
<td>GAIA-SOP: Gaia Python Script Offline Processing</td>
<td>Early adopter</td>
</tr>
<tr>
<td>GAIA-SEV: Gaia Scientific Validation Offline Processing</td>
<td>Early adopter</td>
</tr>
<tr>
<td>ERGOS: HEROS On The Fly Instrumentation</td>
<td>Observer</td>
</tr>
<tr>
<td>ESDC-CRL: ESDC Collaborative Research Lab</td>
<td>Follower</td>
</tr>
<tr>
<td>INT-QAP: Integral Offline Science Analysis Preservation</td>
<td>Early Adopter</td>
</tr>
</tbody>
</table>
Top Level scenarios

Data Exploitation
- jupyter
- R Studio
- hubble space telescope
- esasky
- gaia

Pipeline Develop. env.
- bepicolombo
- euclid
- kafka

Collaborative Research env.
- VOSpace
- euclid
- planck
- gaia
- just

Software Preservation
- docker
- exosat
- integral
- rosetta
- herschel
- lisa pathfinder
- xmm-newton
- iso
Conclusion

New missions call for a paradigm shift for science data exploitation
- From “bring the data to the user” to “bring the user to the data”
- Closer interaction between archives and data processing services

Legacy missions call for data and software long term preservation

New scientists call for collaborative research environment

Science Exploitation and Preservation Platform