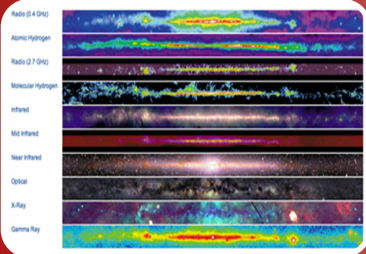# Science Archives Long-Term Strategy (2018 Edition)

Guido De Marchi
Head Scientist
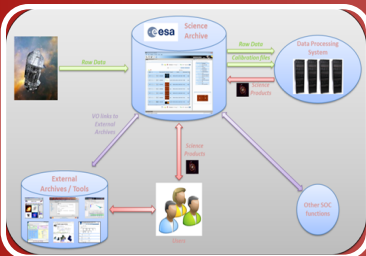ESA Science Data Centre

12.2.19

ESA Science Archives Strategy

Enable maximum *science exploitation* of data sets

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

Enable cost-effective archives, possibly through *integration in, and across, projects*

Keeping the strategy current

- *<u>Over past 15 months</u>*
  - We brainstormed on strategy, ideas, priorities
  - Discussions involved all ESA space scientists and archive software engineers (Forum), plus all Users Groups members
  - New version drafted and circulated to all stakeholders
  - Lots of comments received and included in update
  - SCI-O Executive reviewed and approved final version (2018 edition)

What has and has not changed

- Consensus that 3 pillars still stand strong
  - We want to enable best science
  - We want to preserve data, software, knowledge
  - We want efficient preservation, so when possible integrate archives with operations

- Forum and UGs: 2013 version of strategy document missed two important points
  - give more prominence to legacy data
  - expand role of community

1. Enable maximum science exploitation of datasets

1.1 Ensure data completeness

– Identify and include High Level Science Products (many H2020)

– Investigate collaborative environments (SEPP) as a framework

1.2 Improve data access and data search

– Downloads from our servers are fast, but soon data will be too big to download (Gaia, JWST?, Euclid)

– Community will need to work on data held remotely (SEPP)

1.3 Widen the users base

– For maximum exploitation of data, must promote them

– Very important to reach out to young researchers, summer schools and archival research programme

2. Efficient long-term preservation of data, software, knowledge

2.1 Preserve mission information as legacy

- Over half of our missions with data are in legacy; ESDC to preserve the most meaningful mission data (uplink, raw, processed, and calibrated, plus metadata, software, docs)

- What exactly to preserve? SOCs and communities to tell us (clear that legacy funding cannot match operations money)

2.2 Update mission data during legacy phase

- No data processing foreseen at ESA in legacy phase, but community scientists can develop more accurate versions of datasets later

- ESDC will try to ingest them in the archives after successful peer review, as advised by PS, UGs

2. Efficient long-term preservation of data, software, knowledge

2.3   Support users of legacy missions

– For legacy missions, ESDC relies on PS, contact scientists and scientists formerly involved in the missions, with links to PIs

– Possible new role: scientific curator for legacy archives, working with data for own research, takes over responsibility of science content of archives when PS retire

2.4   Archives technology

– Archives must provide data for a long time, architecture must be modular and flexible to adapt to evolving needs

– Need multi-channel approach, cannot assume one fits all

3. Archive integration in, and across, projects

3.1 Integration into projects

– Archives can belong-lived repository of mission specific data and information to preserve knowledge, but needs of SOCs and community not always overlap; must find right balance

3.2 Integration across projects

– Common framework makes powerful multi-mission, multi-instrument, multi-wavelength interfaces easier (ESASky, PSA)

– But one size does not fit all: scientific objectives of each mission are unique, flexibility is a must

Conclusions and next steps (I)

– Three pillars are solid and strong
  - Enable maximum science exploitation of data
  - Enable efficient long-term preservation of data, software, knowledge
  - Archives integration within and across projects

– Keyword for new activities is **collaboration with science community**, in particular advice from the users groups

– Keyword for new developments is **flexibility**, as needs of science community and associated technology evolve quickly

– Archives services should evolve towards facilitating this

New activities to implement

- Almost 100 individual ideas for new activities collected and discussed during brainstorming sessions

- Ideas were compared, sorted by topic, merged, and ranked

- Ranking reached through discussions with internal stakeholders and with UGs

- Activities divided in two groups
    1. Potentially high impact with limited cost
    2. Substantial development work required

1. High impact, mostly organisational work

1.1 Actively enable data exploitation

– Archival research programme: students/postdocs apply for limited funding to attend conferences and present results based on analysis of archival data; selection with UGs

– Summer schools: proactively attend and organise summer schools to highlight value of data and tools to work with them

1.2 Advertisement

– ESDC Newsletter: ongoing, reaches scientists already aware of our missions to promote new data and services

– Citizen science: with community, identify research tasks to which general public can contribute (Zooniverse project)

1.3 Assist and feedback

– Improve helpdesks; users surveys in collaboration with UGs

2. High development work

2.1 Analysis tools integrated in archives

– Data are becoming more complex and bigger; whether fashion or actual need, computing is moving online

– Community will need online data mining and analysis tools

2.2 Collaborative research environments

– Science Exploitation and Preservation Platform

– Current study to explore ways to integrate data and code, collaborative research areas (JupyterLab) to store and make available community-generated HLPSs (H2020)

2.3 Other activities

– Make satellite HK data available to help interpret (in situ) data

– Assign DOIs to data, help link to papers (starting with EO)

Conclusions and next steps (II)

– Preliminary list of new activities identified and ranked

– ESDC to update list periodically, as needs of stakeholders (internal and community) evolve

– Now proceed with implementation of the highly ranked new high-impact activities

1. Actively enable data exploitation through an **archival research programme** and summer schools for young researchers

2. Develop **a collaborative research environment**, taking into account the lessons learnt from **SEPP**, and test its functionality for the **ingestion of community HLSPs and software**

Archival Research Programme

– Original idea was to support students/postdocs attend conferences and present results based on data in archives, after their project is selected and paper submitted

– Last week we learned from ESA Science Coordination Office that this would violate agreements with national agencies: ESA must not fund research in member states

– Even travel support is considered research funding

– Archival Research Programme concept must be revised

Options available

– Annual prizes for best theses (Master and PhD level) making use of ESA data

– Support young scientists to come to ESA (ESAC/ESTEC) for 2-3 months to exploit science data in the archives (similar to existing science visitor programme)

– Other ideas possible

– Keep in mind that there must be a "deliverable to ESA" (scientist coming to ESA is considered such)

Prize for best theses

– It could be something like this, brainstorming needed…

– Annual prizes for best theses based on ESA data from universities in member states

– Students submit abstract and 3 letters

– One call a year, selection with Users Groups

– Three areas of research:
Astronomy, Heliophysics, Planetary Science

– Two prizes per area, best Master thesis and best PhD thesis

– Recipients invited to ESA to receive prize, give talk, visit sites

– Prize value ~ 1000 Euro

  • to buy books and/or software

  • to attend conference and present results

Science visitor programme

– For information this is how it works

– We welcome science visitors at ESTEC and ESAC

– Typically they stay 1-2 weeks

– Goal is to work with one of us in the faculty (also fellows) on research and a paper, give seminar

– Selection: visitors prepare simple 1 page proposal with faculty member and Faculty Council decides

– We pay travel and living costs, for longer visits co-payment

– Office space limitations: 30-40 per year

Archival research visiting scientists

- – It could be something like this, brainstorming needed…

- – Open to young scientists (students/post-docs/more?)
- – They write proposal (1500 words plus figures) addressing area of research and data to use/exploit
- – Two calls per year; selection by ESDC with Users Groups
- – Recipients come to ESA for 2-3 months to work on data
- – We pay travel (1000 Euro) and living (3,000/month) costs
- – Office space limitations: max 3 people at any given time (e.g. 2 ESAC, 1 ESTEC), so max ~15 people a year
- – Questions to be addressed
    - Open to all or only scientists based in Europe? Pros/Cons?
    - Not all areas of research covered at ESAC/ESTEC, researchers to be independent or science support from home institute needed?