Gaia Archive Preparation - Status and Plans

William O'Mullane

Gaia Science Operations Centre
European Space Astronomy Centre
European Space Agency

Red Española de Gaia Sept 21st 2011 Santillana del Mar, Spain.

GAP - Background

- DPAC officially created in answer to ESA's AO for Gaia Data Processing 2006.
- DPAC is composed of:
  - Eight Coordination Units (CU)
    - area of competence/software production
  - Six Data Processing Centres (DPC)
    - Hardware to run processing
- AO explicitly said DPAC not to include Archive/Catalogue production
  - At that time
- To be integrated in DPAC later
  - A ninth Coordination Unit CU9
  - Another AO for CU9
- GAP $\neq$ CU9 - GAP will define CU9
  - CU9 will not necessarily contain all GAP members.
  - GAP does not necessarily contain all CU9 members.
  - We must remove overlaps and fill missing parts.

What will CU9 be?

- A coordination unit within DPAC
- Following usual DPAC rules
- CU9 must:
  - DOCUMENT the data
  - develop the archive - repository of Gaia data
  - develop the access mechanism(s) to the data
  - make several data releases
  - assist the community in utilising the data
  - that probably means making software available also

Top level work areas

- Work areas for foreseeable future ..

901	Management - WOM + Luri, Walton
902	Development - WOM for now
902-10000	Requirements - Malapert
904	Data Validation - Arenou, Matteo
905	Documentation - Van Leeuwen
906	Operations and Support - Mercier
950	Outreach and Academics - Luri

- will not cover in detail mainly technical

Releases - Highly speculative

- Yes more than one - we wont make people wait till 2020. But all remains tentative
- Now have a list of *kinds* of intermediate data releases but need to agree on how many releases:
- Probably one after $\approx$ 2 years
  - Positions and some Photometry (perhaps just G mag)
  - without epoch and probably with generic error values (or formulae) for entire catalogue
- One more after 5 years
  - Full astrometry better photometry, single transit spectra?
- Then the final one 8 years after launch
  - Variable stars , Astroparams . . .
  - plus all the epoch data
- all under discussion between DPACE and GST

Releases - simulation data

- Before the real data probably good to have a simulated catalogue
- For use of GREAT researchers but also of general interest
- Currently small ($G < 16$) simulation for Plato available to GAP and DPAC (MyPortal)
- Want to make full GUMS/GOG (1.4 / 2 billion sources) available to general public.
- interface TBD ... precursor archive

Timeline

Now	GAP is set up
End 2011	Agree release scenario (including contents).
Jun 2012	Announcement of opportunity
Late 2012	Negotiations and acceptance
Late 2012	Preliminary set up work.
Launch 2013	Start regular CU9 work.
Launch+2	Initial release of data.
...	...
Launch+8	Final release of catalogue/data.

GAP - What YOU can do.

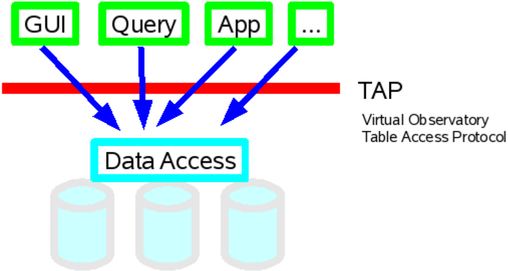
Having a list of representative scenarios which could be turned in to programs or queries is essential.
(termed 20 queries or questions - SDSS had 20 queries to test SkyServer our target is not so narrow - we will have more than 20 for sure.)

The GAP list has commenced here :
http://great.ast.cam.ac.uk/Greatwiki/GaiaDataAccess
You can sign up to edit or contact the science team with your scenario type if it is not there.

Documentation

- Probably the most important task of CU9 !
- Editorial team foreseen to set up the framework and edit documentation
- Documentation per data release
  - currently unclear how many releases but clearly we will have several phases which represent the state of progress in the reductions and the general confidence we have in the data
  - First release , just enough to support basic use of the data
    - description of tables
    - statistical tests that have been carried out on the formal errors
    - some statistics and error distribution figures probably
  - Midterm release as before plus
    - descriptions of reductions on the processing tasks (CU3, 5 & 6)
    - data descriptions on anything new (variability, classification etc).
  - Final publication
    - descriptions of all reductions and verification methods and results
    - possibly some papers on the proper use of the data
- Non negligible effort from inside each CU

Architecture - General Concept



In the simplest form we can split the archive in to applications and storage.

Between we have an agreed interface such as TAP

Hence we can work above and below the line

Current architecture in the SRS (WOM-033) will be updated to Open Archives Initiative (OAI) architecture.

User Interfaces - above the line

- Should be Public oriented i.e. no separate site for *professionals*
- Basics with no login
- Easy self-registration for advanced features
- Query language access - not just forms and check boxes..
- Multi lingual of course .
- Some sort of Sky Browser (ala SDSS, GoogleSky)
  - Visualisation is a whole area to itself - Alves
- Also probably need a more bulk access approach like Hadoop/Map Reduce (call it cloud if you want)

Interrogator ..below the line

- Whatever the interface , VO, Sky Browser etc. Need fast engine to answer queries
- Putting data in some DBMS will not be sufficient
- Tuning needed !!! *LOTS OF IT*.
- Should allow powerful queries to user (SQL and/or ADQL)
- Local space for registered users to upload data and store query results
- Extraction in multiple formats (csv, FITS, VOtable)
- VO access of course .. TAP, SIAP .
- *AND FAST !!!!!!*

Support Advanced apps ..

- Assume we will have *added value* apps
- 3D visualisation for all or part of sky (i.e. globular cluster)
  - Animated with proper motions of course
- Light curve tools
- . . . your idea here . . .
- Interrogator should provide an API to build these apps on.
- but what about a container to deploy them - will have to run in the archive.
- Then a sort of "app" store to show them off and rank them.. (CyberSka)
- this brings all sorts of problems of its own ..

Other Interesting topics ..

- Living archive (Anthony Brown)
  - Can we/should we allow additions to the Archive? e.g. improved solutions for binaries using follow observations
  - Implications for maintenance, quality and security
  - But there will be no printed catalogue so why not a new type of astronomical archive ?
-
- Archive as a model (James Binney)
  - How can we compare models of the Galaxy to Gaia Data ?
  - How can the Archive facilitate that ?
- David Hogg (http://arxiv.org/abs/0810.3851) goes one further: We should try to encode the archive in a Model

Other Interesting topics II

- Virtualized Observatory (William O'Mullane)
  - Can be a way to get big processing for complex queries
  - put data in "cloud" - use HADOOP etc to access it - researchers pay as they go
  - Virtualization may allow us to bring the computing to the data (Szalay) your code runs on VM(s) near the data.
  - Working with CANFAR on this.
- Alpha Observatory (several people)
  - Wolfram Alpha is very cool for some things
  - iKnow sifts and sorts textual information in concepts
  - Why not a natural language input box as the interface ?
  - `Plot all brown dwarfs to 10 parsec`
  - and it could offer relevant literature ..

Conclusion

- The Gaia archive will be a rich an interesting resource for astronomy
- We are currently consulting the community and bringing to correct group of together to make it usable.
- We need to have ambitious goals to start with - even if we may not achieve them all.
- We must work together on a *new* archive
- First we need to launch Gaia - and get it to L2 !!!
- Questions ?