Serving a simulated Gaia Catalogue from ESAC/GAP

William O'Mullane

Gaia Science Operations Centre
European Space Astronomy Centre
European Space Agency

GREAT workshop Feb 29rd- March 2nd 2012,
University of Barcelona, Spain

Gaia Archive Preperation (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.

Gaia Archive Preparation

- Monthly telecons held and minuted - although not very active allows information trade.
- Topical meeting on technology May 24 in ESAC at least one follow telecon
- First plenary June 23 2011 after GREAT Brussels.
- Requirements Review Nov 29 2011- Cambridge.
- Work areas remain
  - Management - WOM + Luri, Walton
  - Requirements - Malapert
  - Architecture and Development - O'Mullane + Tapiador
  - Data Validation - Arenou, Matteo
  - Documentation - Van Leeuwen
  - Operations and Support - Mercier
  - Outreach and Academics - Luri
- A lot of work in GENIUS can be reused for proposal
  https://gaia.am.ub.es/Twiki/bin/view/GENIUS/

Precursor Archive(s)

- first Plato requested dataset now available from ESAC servers
- http://www.rssd.esa.int/index.php?project=GAIA&page=GPA
- provided by Satellite Archive Team at ESAC
- This provides TAP like access - limited by back end
- Other DBMS being looked at (GreenPlum, Cache) - for TAP back end.
- Hadoop seems a possible approach - experimenting with this for GREAT (tomorrow)
- experimenting with some data centres to see CANFAR

Use Cases

Everyone agrees 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)

The GAP list is here :
http://great.ast.cam.ac.uk/Greatwiki/GaiaDataAccess

These were analysed in Cambridge in November 2011 - document still pending.
Comments to GST members or Nigel Hambly

GAP - Simulation Data

- Have GUMS10.0 from CU2 http://gaia.am.ub.es/GUMS-10/
  - A&A paper on GUMS out Robin et al. (2012).
  - XL-028 for detailed stats
  - Made available to GAP members from ESAC via scp/rsynch
  - CDS quickly put in Vizier - awaiting ok to make public.
  - Used with Hadoop in Leiden school
- Now have GOG $G < 17$
  - Will make available once get ok from CU2
  - Will use with Hadoop here in this workshop tomorrow!

Current GAP Timeline

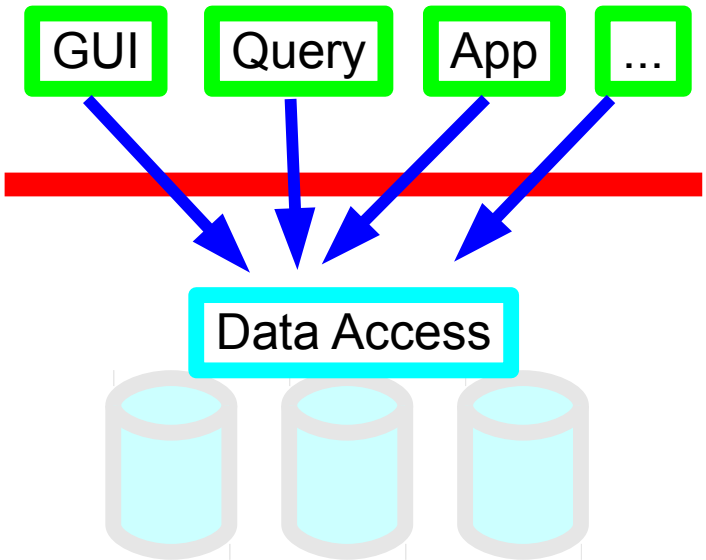
- Tentative plan

Now	Topical meetings - try HADOOP. CANFAR , discussions etc.
Feb 2012	Agree release schedule internaly (WOM-066) - PS to make public version
Sept 2012	Commence response writing - agree chapters perhaps small meeting
Nov 2012	First complete draft of response
Nov 2012	Announcement of opportunity (short)
Jan 2013	Hopefully successful negotiations and acceptance
Feb 2013	Preliminary set up work.
Aug 2013	(Launch) - Start regular CU9 work.

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

General Concept



e.g. TAP

Virtual Observatory
Table Access Protocol

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 O'Mullane (WOM-033) will be updated to Open Archives Initiative (OAI) architecture.
More interesting ideas are in O'Mullane (WOM-057)

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)
- 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 Gaudet et al. (2010) 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 .. WWT now offers ADS links.

Acronyms I

The following table has been generated from the on-line Gaia acronym list:

Acronym	Description
AO	Announcement of Opportunity
API	Application Programming Interface
CU	Coordination Unit (in DPAC)
DBMS	DataBase Management System
DPAC	Data Processing and Analysis Consortium
DPC	Data Processing Centre
FAST	Fundamental Astronomy by Space Techniques (Hipparcos)
FITS	Flexible Image Transport System
GAP	Gaia Archive Preparations (DPAC WG)
GREAT	Gaia Research for European Astronomy Training
IT	Information Technology
KM	Model Factor
OAI	Open Archives Initiative
OF	Object Feature (source packet)
SDSS	Sloan Digital Sky Survey
SQL	Structured Query Language
SRS	Software Requirements Specification
TAP	Table Access Protocol
TOC	Table of Contents
VM	Virtual Machine
VO	Virtual Observatory

References I

Gaudet, S., Hill, N., Armstrong, P., et al., 2010, In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 7740 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ADS Link

[XL-028], Luri, X., Isasi, Y., Borrachero, R., et al., 2012, *Gaia Universe Model Statistics - version 10*, GAIA-C2-TN-UB-XL-028, URL http://www.rssd.esa.int/llink/livelink/open/3110371

[WOM-033], O'Mullane, W., 2009, *Gaia Catalogue and Archive Software Requirements and Specification*, GAIA-C9-SD-ESAC-WOM-033, URL http://www.rssd.esa.int/llink/livelink/open/2907710

[WOM-057], O'Mullane, W., 2011, *Blue skies and clouds, archives of the future*, GAIA-C9-TN-ESAC-WOM-057, URL http://www.rssd.esa.int/llink/livelink/open/3072045

[WOM-066], O'Mullane, W., van Leeuwen, F., 2012, *Release scenarios for the Gaia archive*, GAIA-TN-PL-ESAC-WOM-066, URL http://www.rssd.esa.int/llink/livelink/open/3111396

Robin, A.C., Luri, X., Reylé, C., et al., 2012, ArXiv e-prints, ADS Link