COTS, is it worth it ?

William O'Mullane, Nana Bach, Jose Hernandes, Alexander Hutton and Rosario Messineo

Science Operations Development Division
European Space Astronomy Centre
European Space Agency

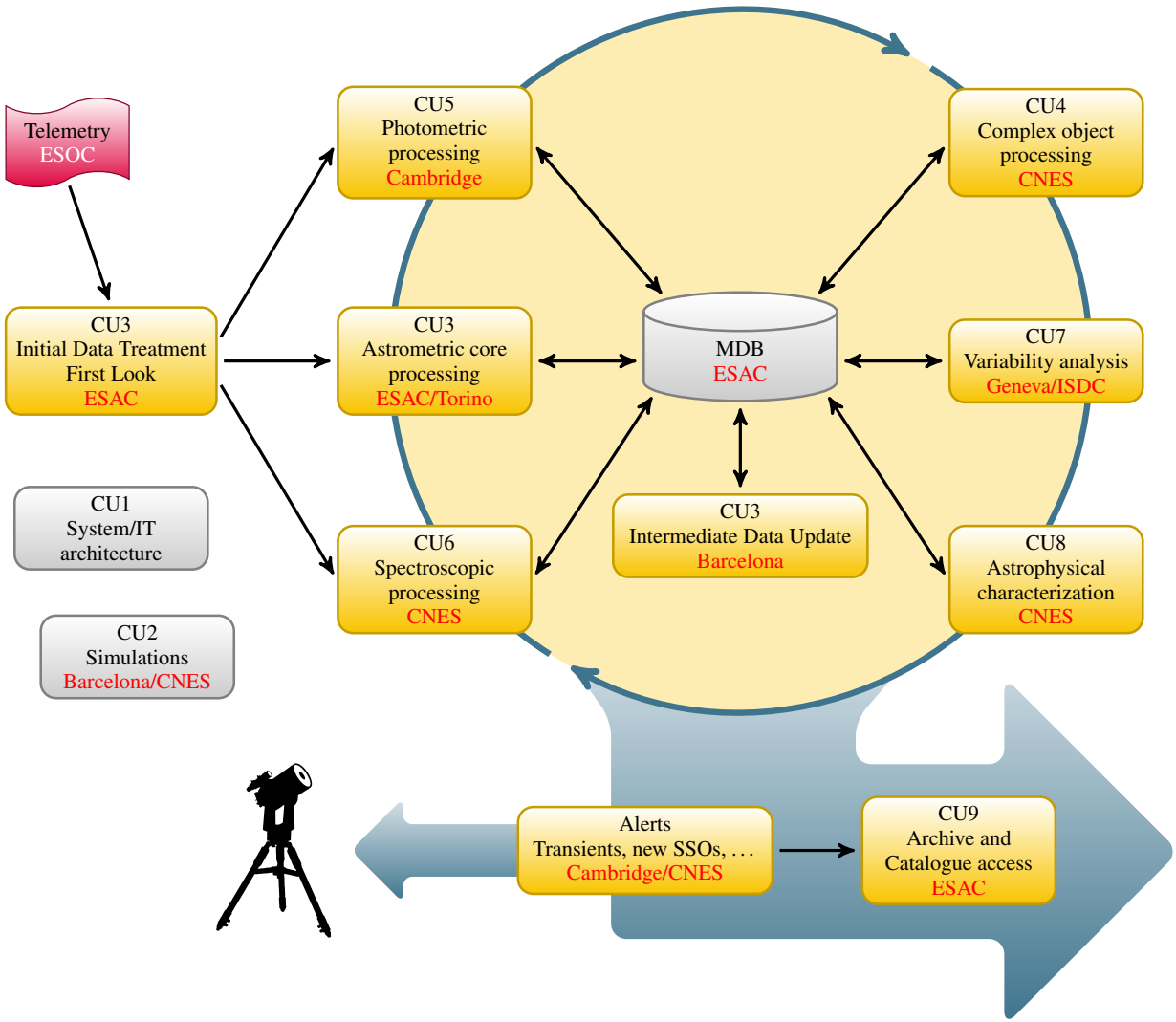
October 16-20th 2016,
ADASS XXVI,
Trieste, Italy.

esa
european space agency
agence spatiale européenne

Gaia Science Operations

- The Data Processing Ground Segment is jointly operated by ESAC and the Gaia DPAC. This is comprised of the Science Operations Centre (SOC) operated by ESA and a set of Data Processing Centres (DPCs). The SOC is at the same time a DPC (DPCE), which acts the interface between the MOC and the DPAC.
- Gaia will produce an impressive volume of raw data yielding at mission completion a telemetry data volume of roughly 100TB. Transforming the data into scientifically meaningful quantities is the task of the Data Processing Analysis Consortium (DPAC).
- DPAC is comprised of a number of Coordination Units (CUs). Each one is responsible for a well-defined part of the Gaia data processing.
- For each CU there is at least one Data Processing Center (DPC) with dedicated resources for the data processing of the CU.

Simplified processing overview

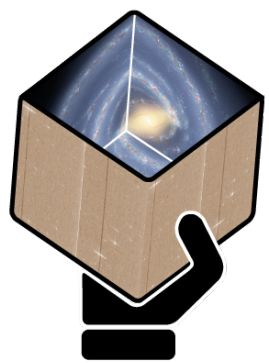


CU=Coordination Unit

Daily $30 - 60GB$ compressed

5yrs $1PB$

(see also O'Mullane et al. (2006), O'Mullane et al. (2009))



Gaia SOC/DPCE at ESAC

Hardware
- ≈100 Intel Xeon-based blades
- 1612 cores in total
- 32-128 GB RAM per blade
- ≈40 TFLOP/s
- ≈400 TB usable NetApp storage
- 10 GBit/s interconnect

Elemental satellite data
- HK (incl. attitude)
- Science (SPx, ASDx)
- Orbit
- Time couples/correlation
- IFMS

≈40 GB/d

MOC

≤8.5 Mbit/s

Ground stations
- Cebreros, Spain
- New Norcia, Australia
- Marlargüe, Argentina

SOC/DPCE

≈40 GB/d

POR

POS

SL

SCI-S

Inbox

ESTEC

DPCT (Torino)

DPCI (Cambridge)

DPCC (CNES, Toulouse)

DPCG (Geneva)

DPCB (Barcelona)

MIT → DCS → IDT → FL

≈40 GB/d

few 100 GB/d

TM Archive

Airbus D&S

≈3.5 TB/m

Main DB

IDTFLDB

IDTFLDB mirror

Payload Experts

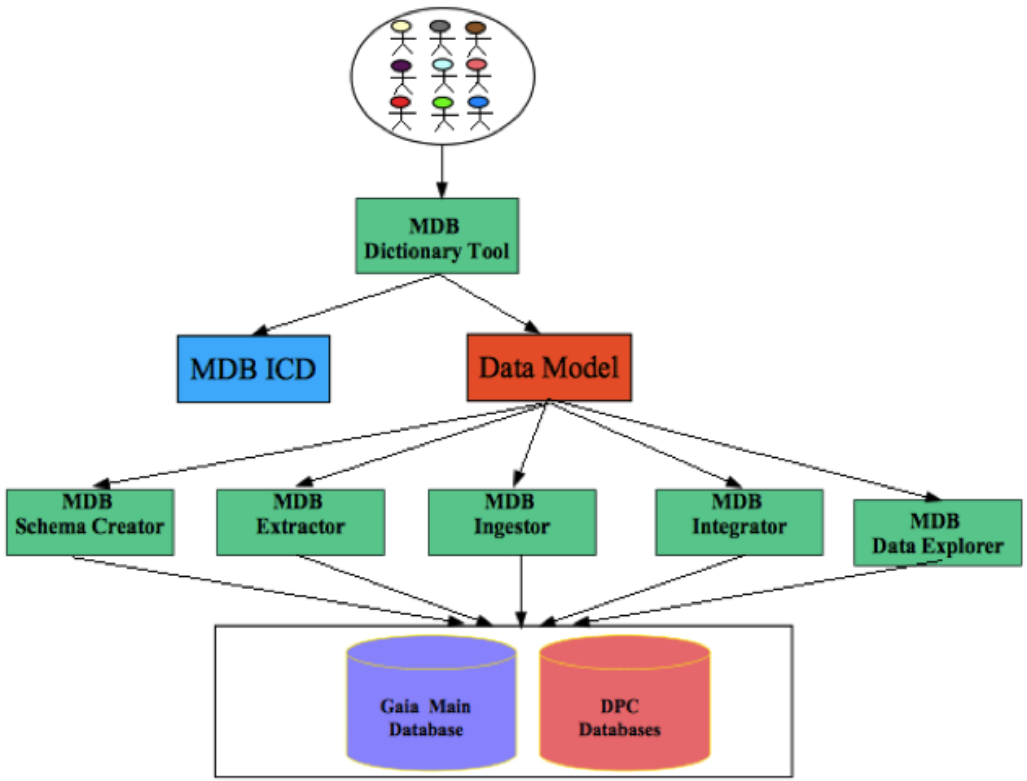
700 (200+500) TB after 5 yr

Never larger than a few TB

Out DPCE: 2.5 / In DPCx: 1.0 Gbps

Gaia Main Database (MDB)

MDB (Main Database) is the central repository of all the data produced by Gaia Scientific Processing and DPAC.
MDB Dictionary, running at DPCE, is used by all to define the Datamodel of the MDB.

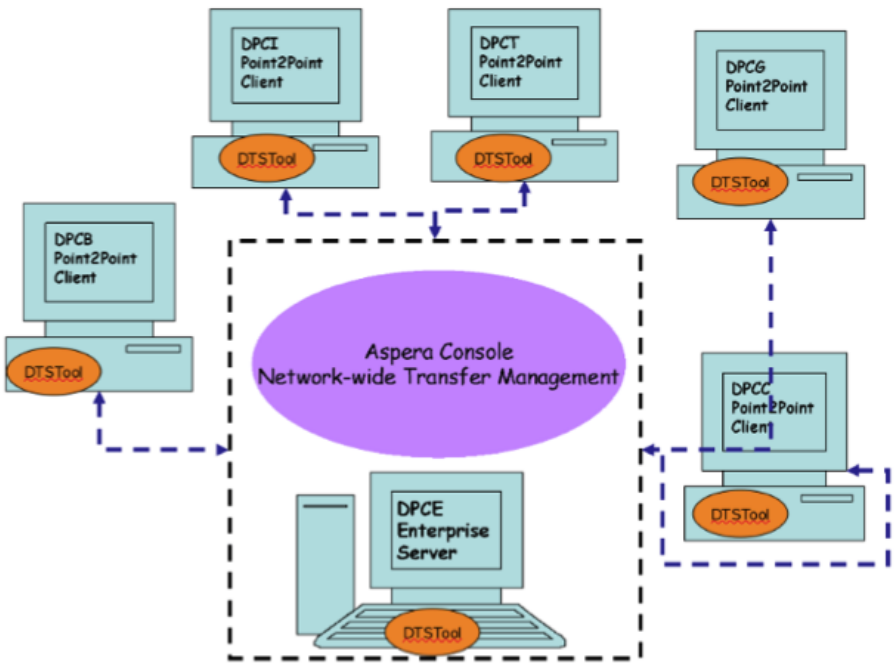


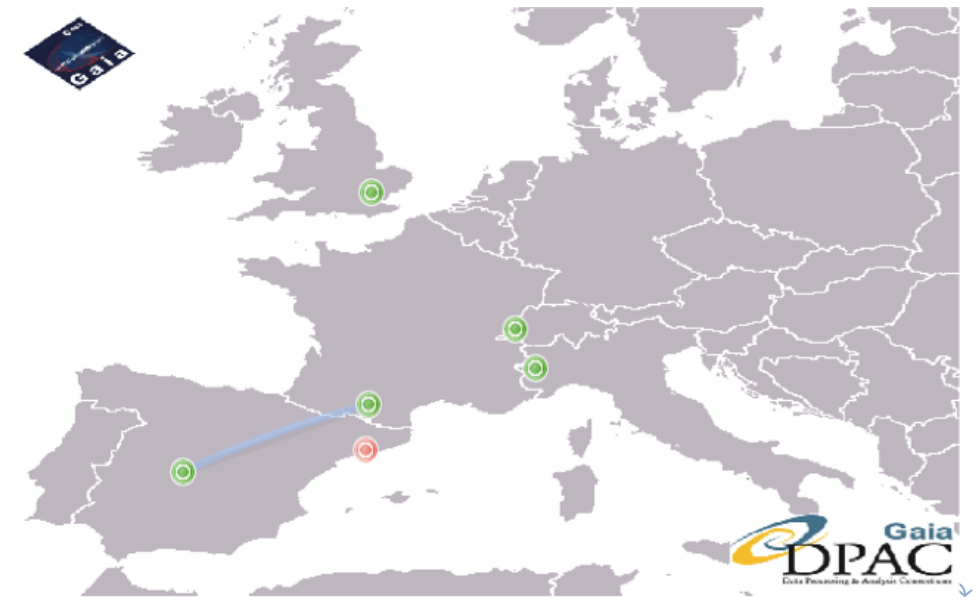
SW systems of MDB:

- MDB Database
- MDB Schema Creator
- MDB Dictionary Tool
- MDB Extractor
- MDB Ingestor
- MDB Integrator
- MDB Explorer
- MDB Data Manager

Gaia Transfer System (GTS)

- GTS (Gaia Transfer System is the service that permits data exchange between the Data Processing Centres (DPCs).
- Composed of Aspera (commercial) and the DTSTool (add on).
  - Aspera is a COTS for fast data transfer, it provides the technical platform for the data transfers.
  - The DTSTool is responsible for building an interface between Aspera and the data processing software systems at each DPC.
    - DTSTool developed by ALTEC - *see the booth*.





MDB Requirements

- **14 functional requirements**
- **1 performance requirement**

- MDB Schema Creator: 1 functional requirement
- MDB Dictionary Tool: 24 functional requirements
- MDB Extractor: 12 functional requirements, 1 performance requirement
- MDB Ingestor: 10 functional requirements, 2 performance requirements
- MDB Integrator: 9 functional requirements
- MDB Explorer: 16 functional requirements
- Organically added functionality
  - MDB DataManager: no written requirements

If we take all the functional requirements there are 86.

GTS Requirements

- **13 functional requirements**
- **2 performance requirements**

- 29 other requirements:
  - Portability: 1
  - User Interface: 5
  - Execution and activation: 1
  - File naming convention: 3
  - Interface: 5
  - Safety and Security: 12
  - Performance: 3

Cost effectiveness analysis - factors

Real cost of software goes far beyond the license itself, and includes:

- Implementation, evaluation, integration efforts, influenced by:
  1. Quality of documentation
  2. Responsiveness of support staff
  3. Availability of code examples and other learning aides
- Resources consumed by the use of the software.
- Reliability across all possible use-cases (often the most expensive aspect of cheaper software).
- Flexibility and adaptability relative to the competition.

(See also Poster P2.8)

Cost effectiveness - recommendations

To achieve real cost effectiveness we recommend seeking the following qualities in server software for high-demand workflows:

- Memory-allocation independent (application terminates after execution).
- Fully separable from other functions.
- Availability of well defined and stable API in order to decouple the COTS from SW customisation.
- Proven track-record of reliability in high-volume production environments.

Costing

MDB

- We have a reasonable account of effort booked to the MDB work package 2005-2016:
  - Effort ≈ 18 person years incl. testing and documentation/support.
  - Depending on our cost model that is a cost of €2.8M to €4.5M
  - Must include all consumables, office space, phones, travel etc..
  - lets call it €3M.

GTS

- Aspera licenses and support to date €158K
- DTSTool, customisation, automaton etc. .. effort ≈ 3.5 person years incl. testing documentation.
- Effort cost (similar rate to MDB but actually probably lower) €560K
- Total cost $\approx$ €720K

Cost effectiveness

No good way to do this !!
Constructive Cost Model (COCOMO) good *BEFORE* you build perhaps.
Though similar in requirement the MDB feels much more complex than GTS if we use the functional requirements and say all were met that gives MDB a much higher complexity

- UNIT COST = Cost/Function
- MDB UC= 3M/86 = 34883
- GTS UC= 720K/29 = 24827
- COST EFFECTIVENESS - COTS 1.4 times more cost effective.
- COTS where appropriate .. could not find tool for MDB

Of course you can find many problems with this !!

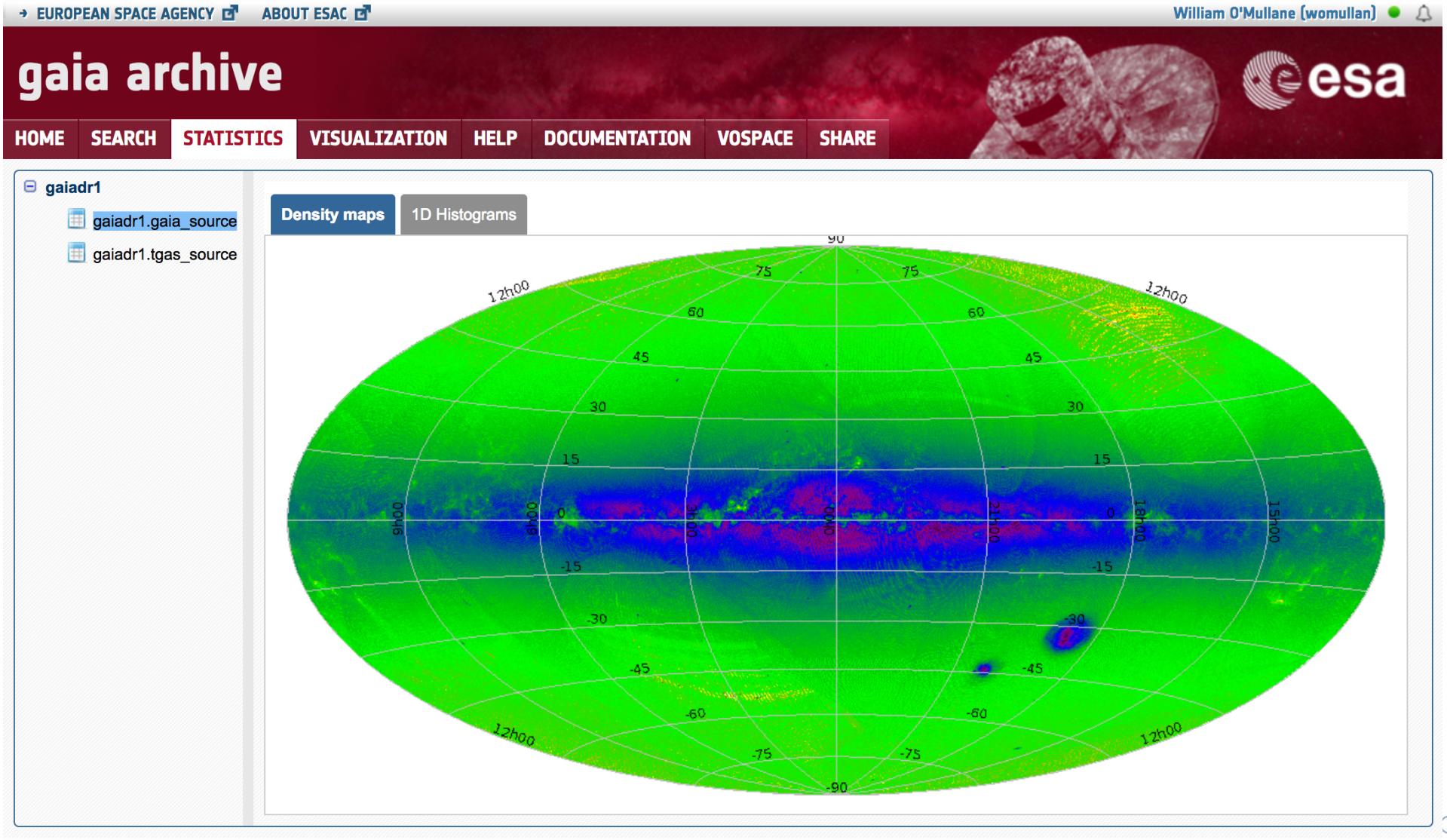
Conclusion

- We may want to take more care about requirements to track this
- COTS items we can use in space science is fairly limited (DBMS, xfer other generic stuff).
  - How can we guess what software *may* be available later
- Choose carefully - that small development can cost a lot cumulatively.
- Open source is ok but can also die - we supported Apache Common Math - now ONLY we support it.
  - Java as a sort of open platform worked out ok
  - software - XML parsing, GUI, plotting libraries.
  - Eclipse IDE , MANTIS (now we use Jira) ..
  - concurrent versioning tools (i.e. SVN, GIT), code quality tools,
  - all good there are many peripheral tools which work and save a lot.
  - ...

Shameless plug and questions

Between COTS and in house the First Gaia Catalogue is out!!

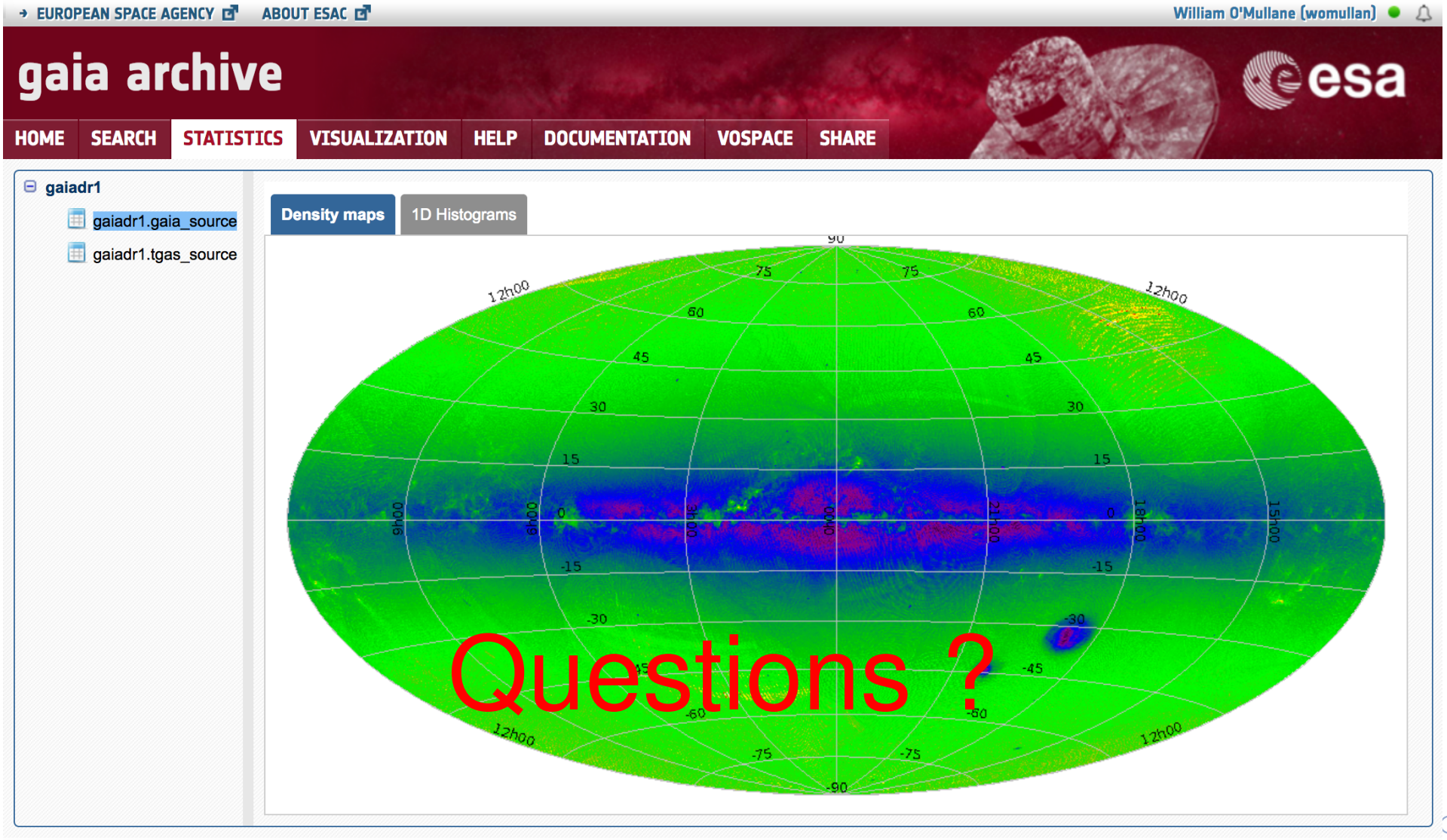
`http://archives.esac.esa.int/gaia/`



Shameless plug and questions

Between COTS and in house the First Gaia Catalogue is out!!

`http://archives.esac.esa.int/gaia/`



Acronyms I

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

Acronym	Description
ADASS	Astronomical Data Analysis Software and Systems
ALTEC	Advanced Logistic Technology Engineering Centre
API	Application Programming Interface
COTS	Commercial-Off-The-Shelf
CU	Coordination Unit (in DPAC)
DBMS	DataBase Management System
DPAC	Data Processing and Analysis Consortium
DPC	Data Processing Centre
DPCE	Data Processing Centre ESAC
ESA	European Space Agency
ESAC	European Space Astronomy Centre (VilSpa)
GB	GigaByte
GTS	Gaia Transfer System
GUI	Graphical User Interface
IDE	Integrated Development Environment
MDB	Main DataBase
MOC	Mission Operations Centre
SOC	Science Operations Centre
SVN	SubVersioN
SW	Software (also denoted S/W)
TOC	Table of Contents
XML	eXtensible Markup Language

References

O'Mullane, W., Lammers, U., Bailer-Jones, C., et al., 2006, ArXiv Astrophysics e-prints, ADS Link

O'Mullane, W., Hernández, J., Hoar, J., Lammers, U., 2009, In: D. A. Bohlender, D. Durand, & P. Dowler (ed.) Astronomical Data Analysis Software and Systems XVIII, vol. 411 of Astronomical Society of the Pacific Conference Series, 470, ADS Link