ESA Science Archives Architecture Evolution

Iñaki Ortiz de Landaluce
Science Archives Team
13th Sept 2013

Outline

- Introduction: ESA Science Archives

- Archives Architecture Evolution
  - User Interfaces and the Web 2.0
  - Application Frameworks
  - Databases, Spatial Indexing and Big Data

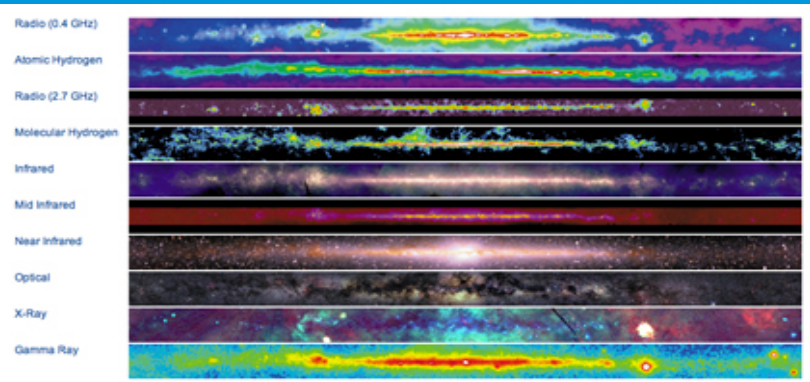
Outline

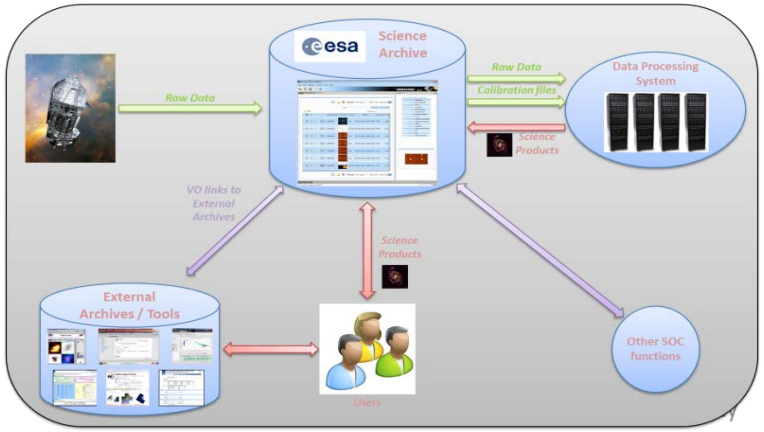
- Introduction: ESA Science Archives

- Archives Architecture Evolution
  - User Interfaces and the Web 2.0
  - Application Frameworks
  - Databases, Spatial Indexing and Big Data

ESAC Science Archives Strategy

- Large set of science archives co-located at ESAC are a major research asset for community.
- Need to be kept readily available for future users and novel uses.
- Thus, must plan now for next 5–20+ years.
- Planning based around 3 major goals:
  - 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.

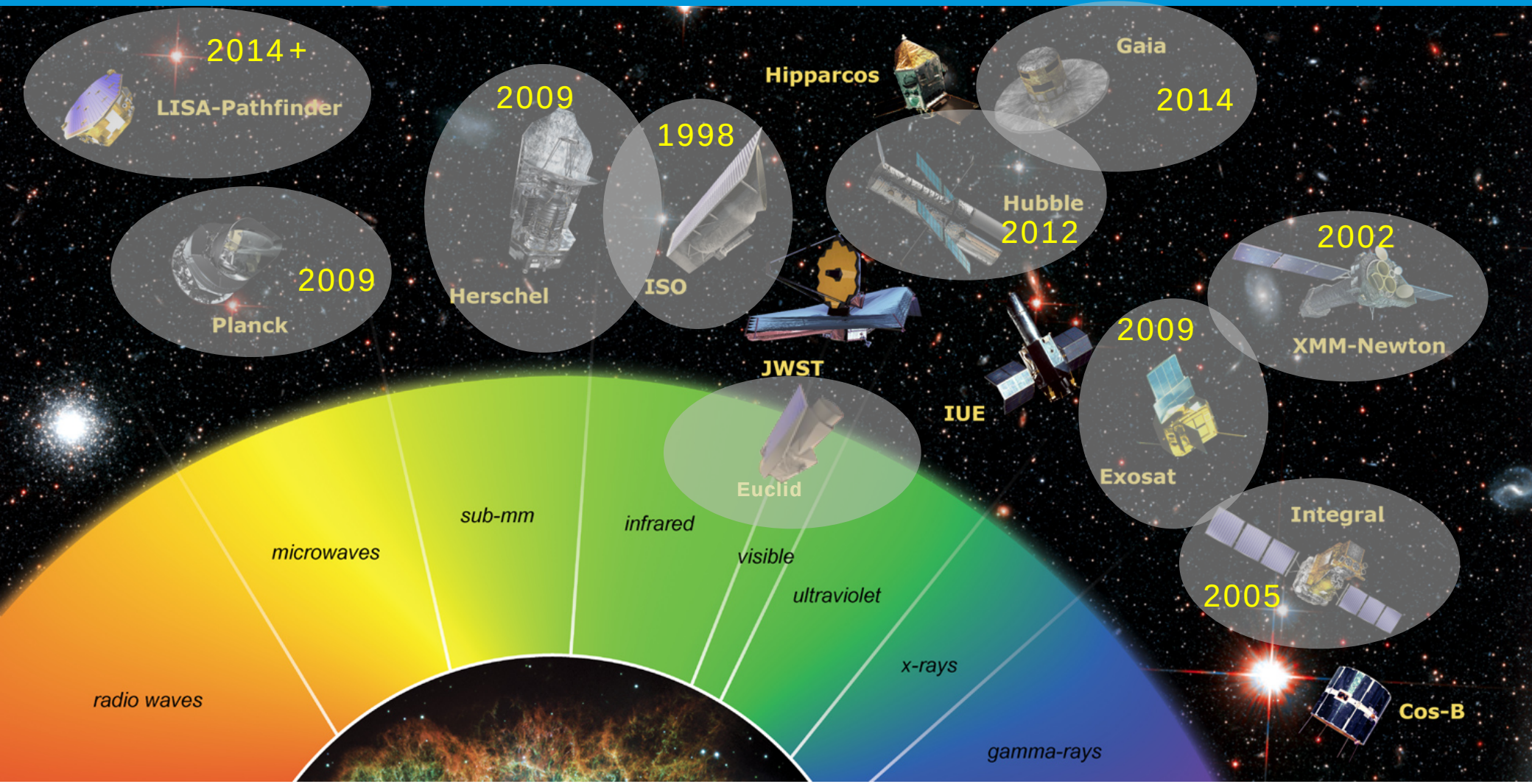




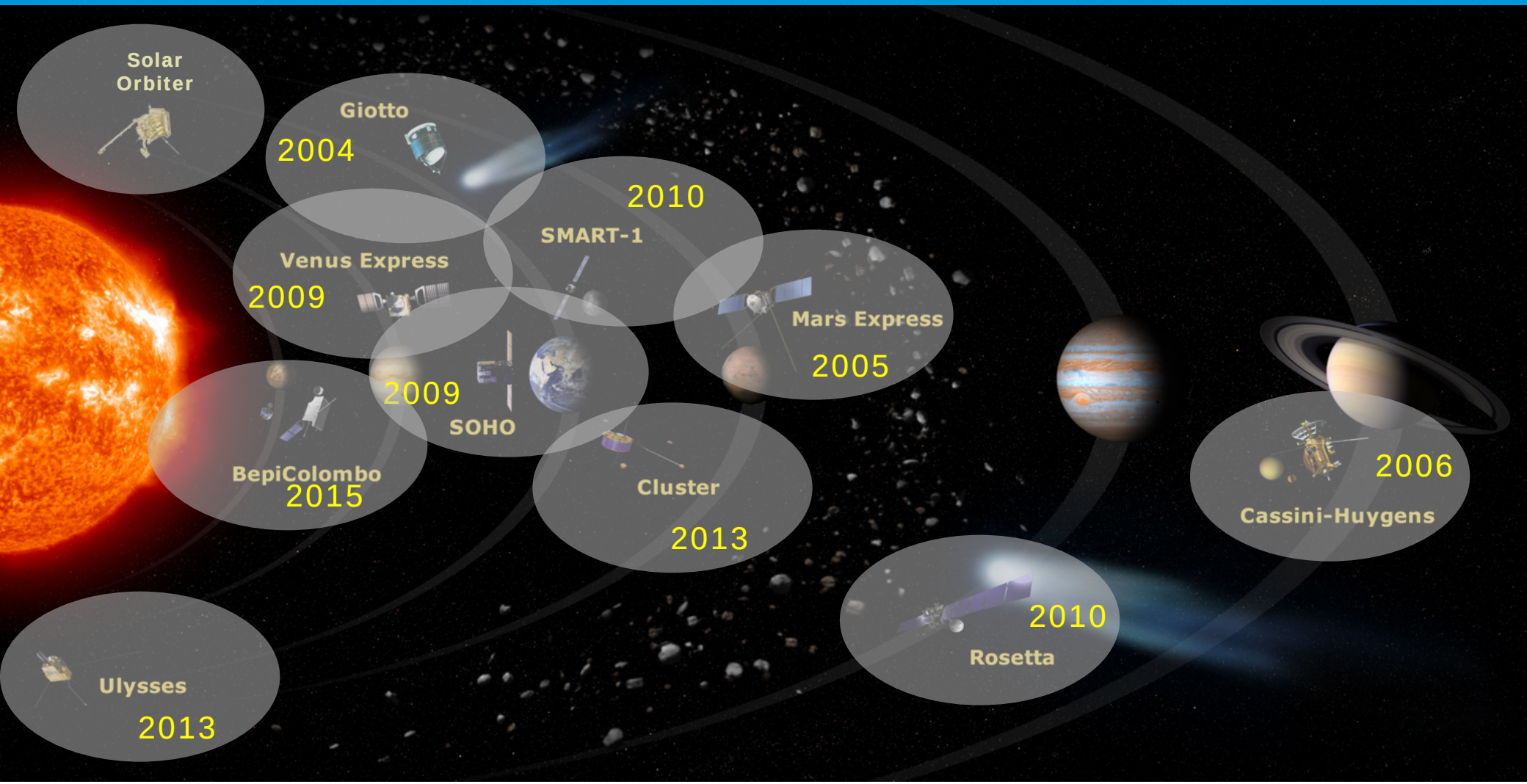
Introduction: ESA Science Archives

- Different types of
  - Missions: Astronomy, Planetary, Solar System, ...
  - Data: Raw data, calibrated processed data, high level data products, ...
  - Users:
    - Scientific Community (public access)
    - PI team and observers (controlled access)
    - Science Operations Team (privileged access)
- Common Architecture and Look and Feel
  - Better corporate image for ESA

ESA Science Archives - Astronomy

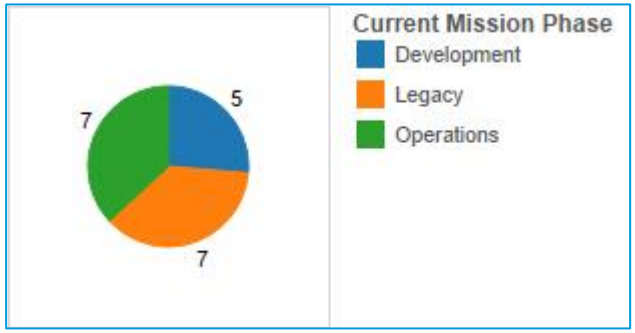


ESA Science Archives – Planetary & Solar System



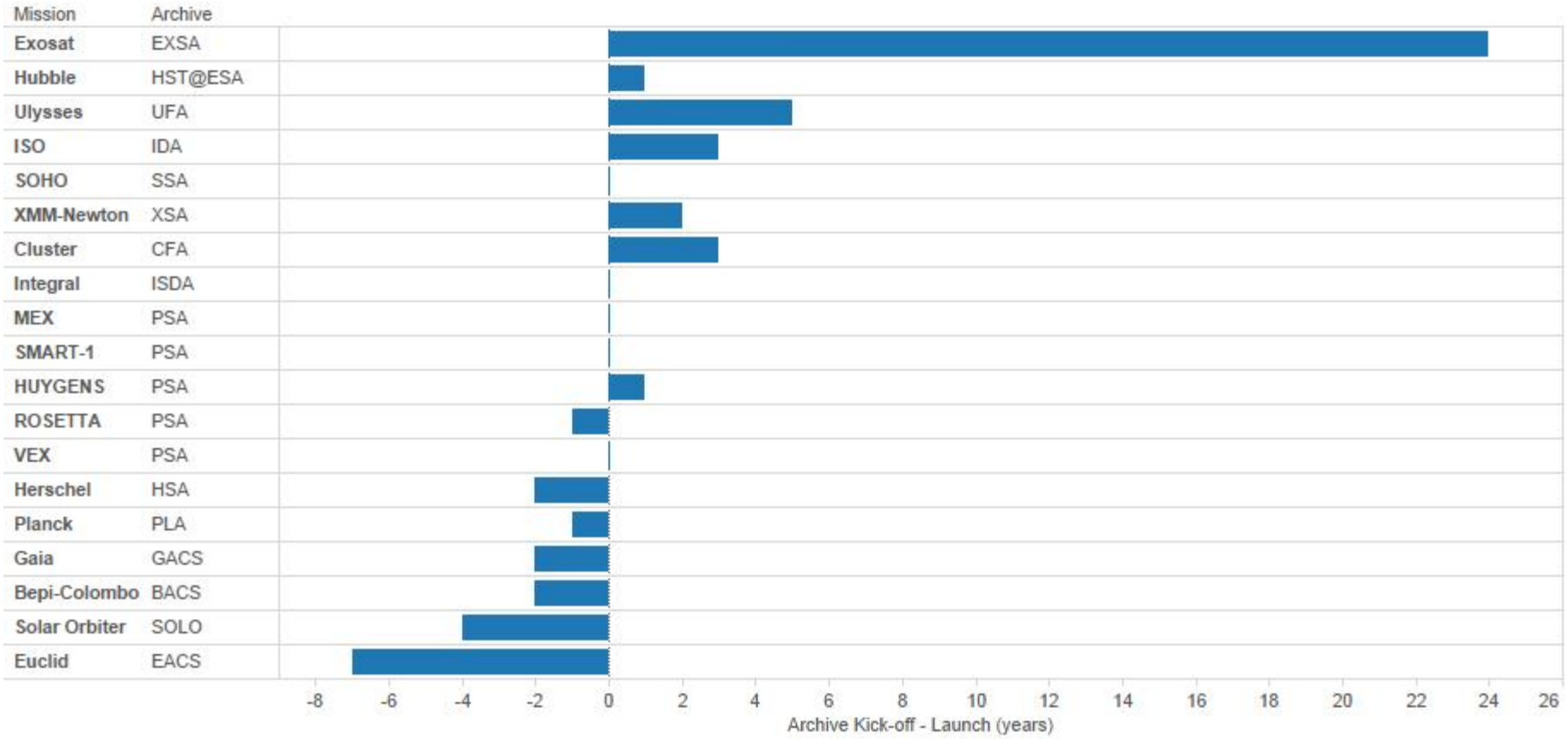
ESA Science Archives - Mission Phases Support

- ESA Science Archives support missions in different phases
  - Development
  - Operations (EOP, CP, PVP, SDP, RP...)
  - Post-operations and Legacy Archive



- Early start of archiving activities within the mission phases

Archive Kick-off vs Launch



Outline

- Introduction: ESA Science Archives

- Archives Architecture Evolution
  - User interfaces and the Web 2.0
  - Application Frameworks
  - Databases, Spatial Indexing and Big Data

ESA Science Archives Architecture Evolution

- Technology has evolved enormously since 1998.
- Also, new and heterogeneous requirements had to be addressed:
  - Different mission types: multi-wavelength astronomy, solar and planetary
  - Increasing number of archives
  - New network and security policies
- That resulted into 3 different generation of archives:

• **ISO** Data Archive
• **INTEGRAL** Science Data Archive
• Planetary Science Archive
  • **MEX**
  • **VEX**
  • **ROSETTA**
  • **HUYGENS**
  • and others
	• **SOHO** Science Archive
	• **EXOSAT** Science Archive
	• **Planck** Legacy Archive
	• **Herschel** Science Archive
	• **Cluster**
		• **Ulysses** Final Archive
		• **XMM-Newton** Science Archive
		• ESA **Hubble** Science Archive
		• **Euclid**
		• **GAIA**
		• **Solar Orbiter**
		• **Bepi Colombo**

Outline

- Introduction: ESA Science Archives

- Archives Architecture Evolution
  - User Interfaces and the Web 2.0
  - Application Frameworks
  - Databases, Spatial Indexing and Big Data

User Interfaces and Web 2.0

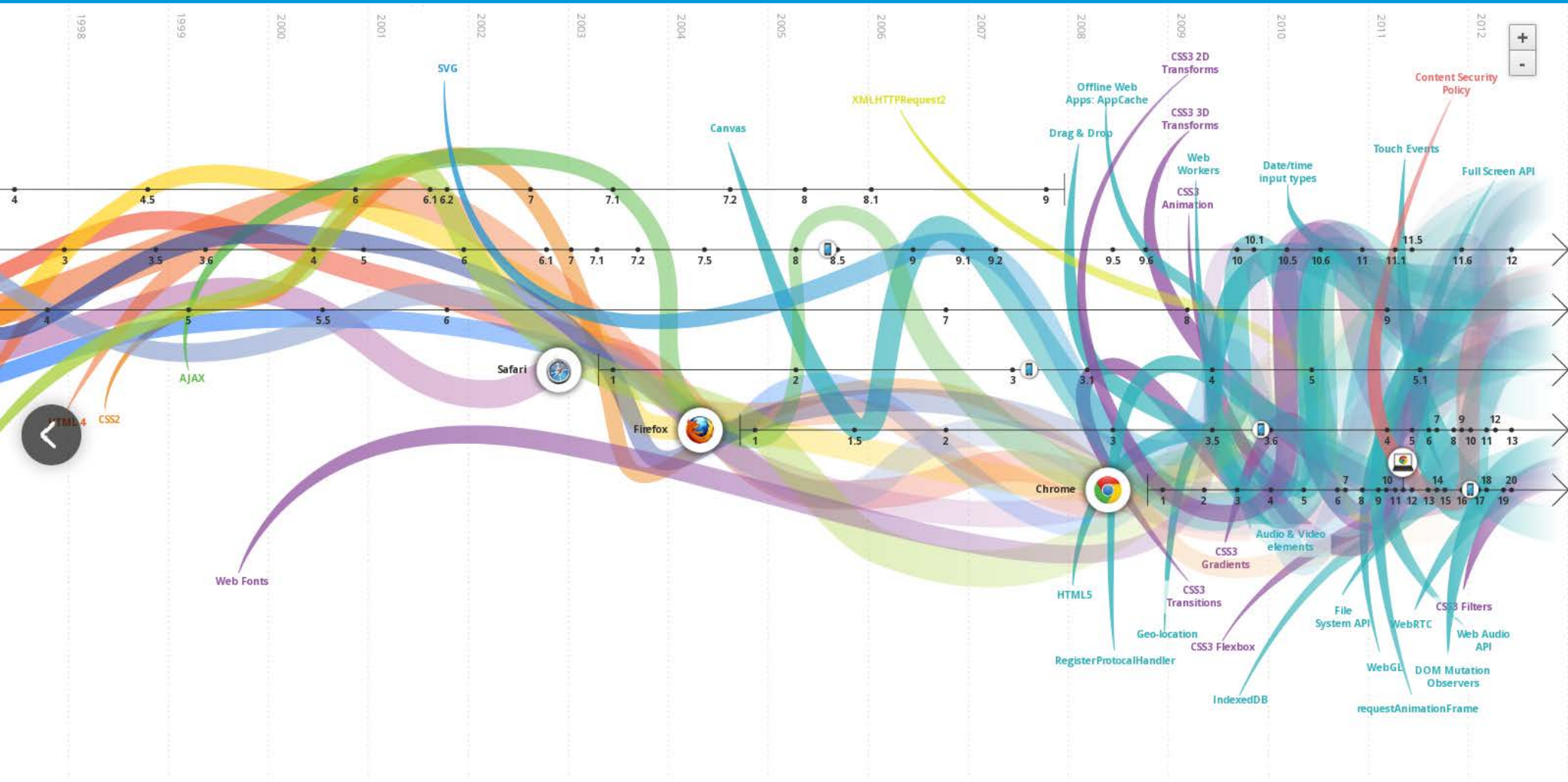
- **Year 1998**: First generation of ESA Science Archives (ISO)
  - HTML 3, Netscape and IE
  - Web could not satisfy user requirements
    - visualization
    - dynamic content
  - Small and dynamic applications could run from a browser using plug-ins
    - Flash, Java Applets

- **2006**: Second generation of ESA Science Archives (SOHO, EXOSAT,..)
  - Java Applets replaced by Java Web Start Technology (JNLP)
  - Desktop application
  - Java Version updating and Pack200 compression
  - Security enhancements

User Interfaces and Web 2.0

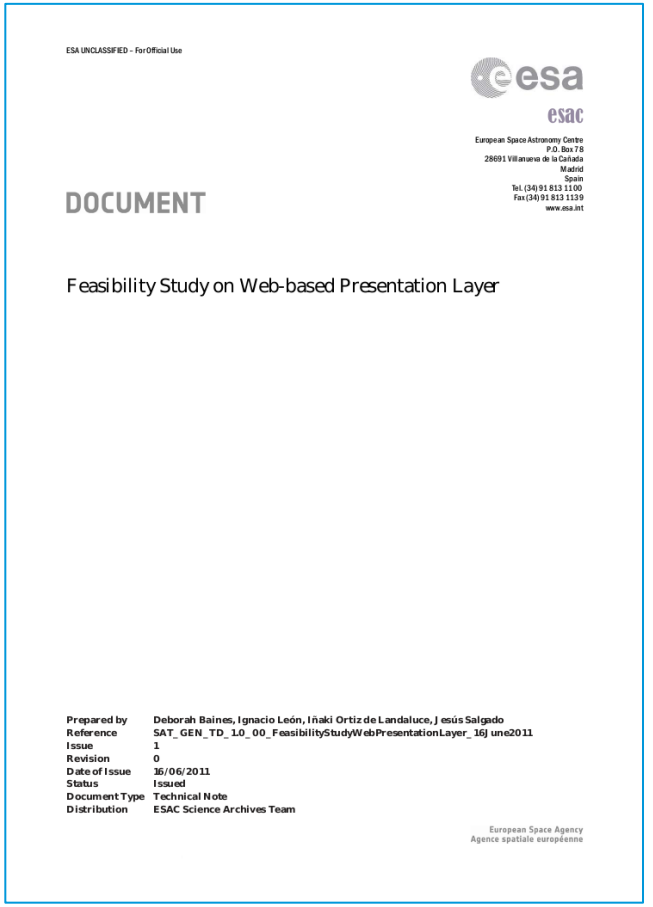
- **The web revolution**: From Dot-com to Web 2.0 and beyond
  - Fully-featured dynamic contents (HTML5, CSS3, AJAX)
  - Video and Audio support
  - No plug-ins required
  - 2D/3D Graphics Rendering (WebGL)
  - Ubiquity, The Web becomes social

User Interfaces and Web 2.0



User Interfaces and Web 2.0

- **2011:** Third generation of archives kick-off
  - Feasibility study to implement Web-based Archive User Interfaces
    - Web-based technologies assessment
    - Impact on current ESA Archives Architecture
    - SWOT Analysis
    - Analysis of existing web-based archives: HEASARC, CDS ASDC, ISDC, Spitzer, Hubble, CADC, ESO, PDS, SDSS
  - GWT is the selected technology
    - Small learning curve for Java developers
    - Long Term Support expected (it's Google)
    - Wide community
- **2013:** Release of first web-based ESA Science Archives
  - Ulysses Final Archive
  - XMM-Newton Science Archive



ESA UNCLASSIFIED - For Official Use

esa

esac

European Space Astronomy Centre
P.O. Box 78
28691 Villanueva de la Cañada
Madrid
Spain
Tel. (34) 91 813 1100
Fax (34) 91 813 1139
www.esa.int

DOCUMENT

Feasibility Study on Web-based Presentation Layer

Prepared by	Deborah Baines, Ignacio León, Iñaki Ortiz de Landaluce, Jesús Salgado
Reference	SAT_GEN_TD_1.0_00_FeasibilityStudyWebPresentationLayer_16June2011
Issue	1
Revision	0
Date of Issue	16/06/2011
Status	Issued
Document Type	Technical Note
Distribution	ESAC Science Archives Team

European Space Agency
Agence spatiale européenne

User Interfaces Evolution

1st Generation

- Applet
- Java 1.1+ Abstract Window Toolkit
- In-house RPC connection through TCP/IP

2nd Generation

- Web Start
- Java 1.5+ Java Swing, JGoodies
- spring
- Inter-operability

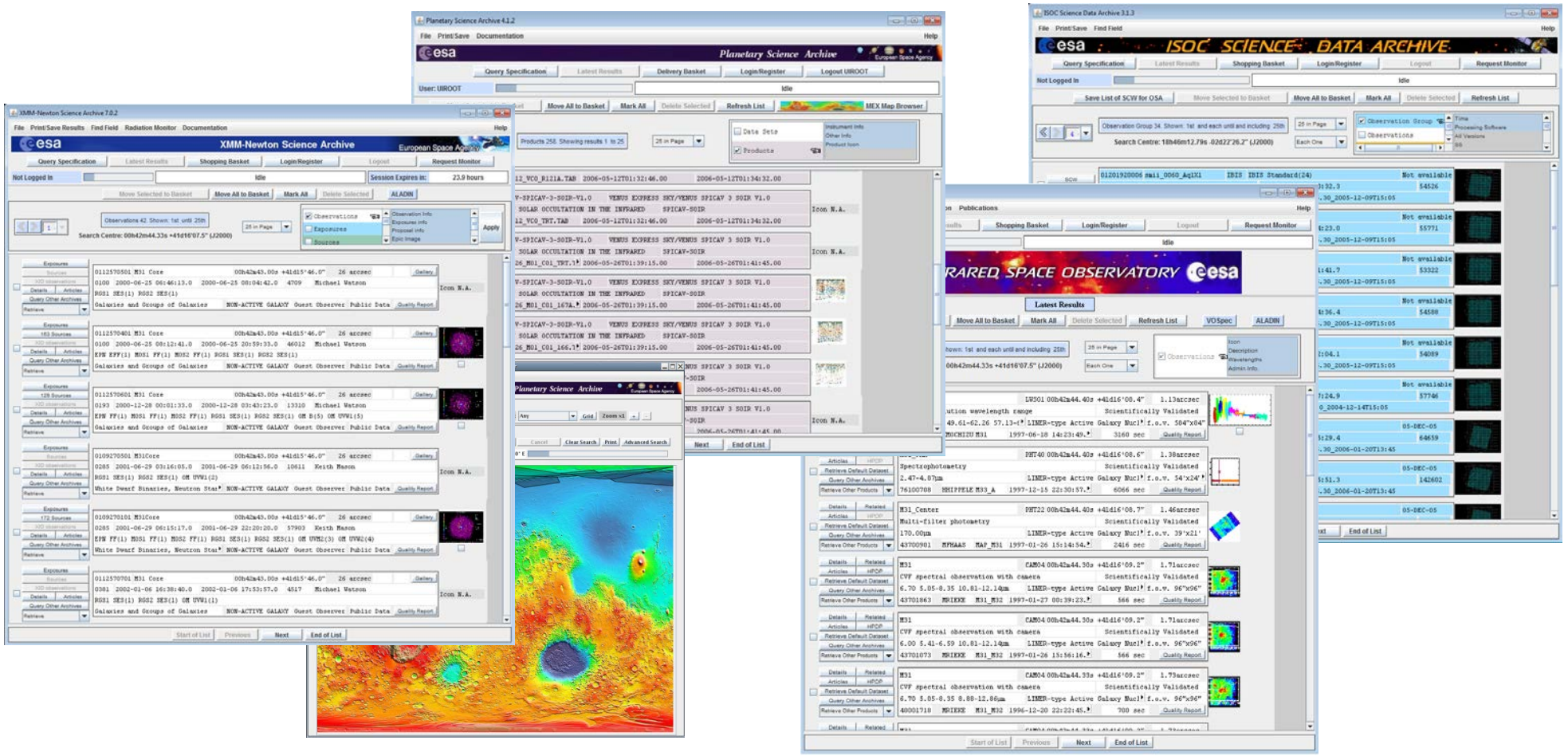
3rd Generation

- Web based
- G
- HTTP
- Inter-operability (Web Profile)

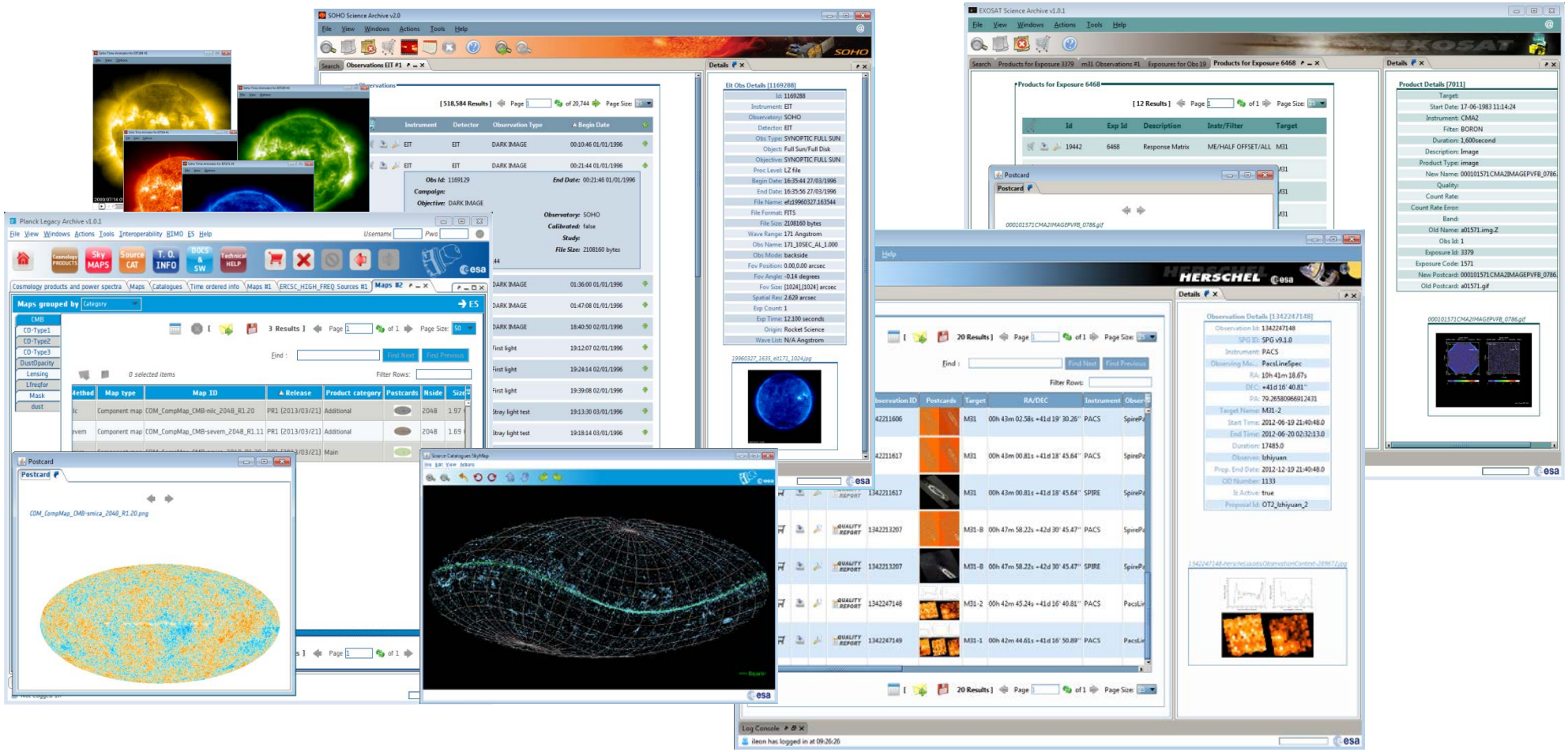
In Progress

- WebGL
- Enhanced visualization tools

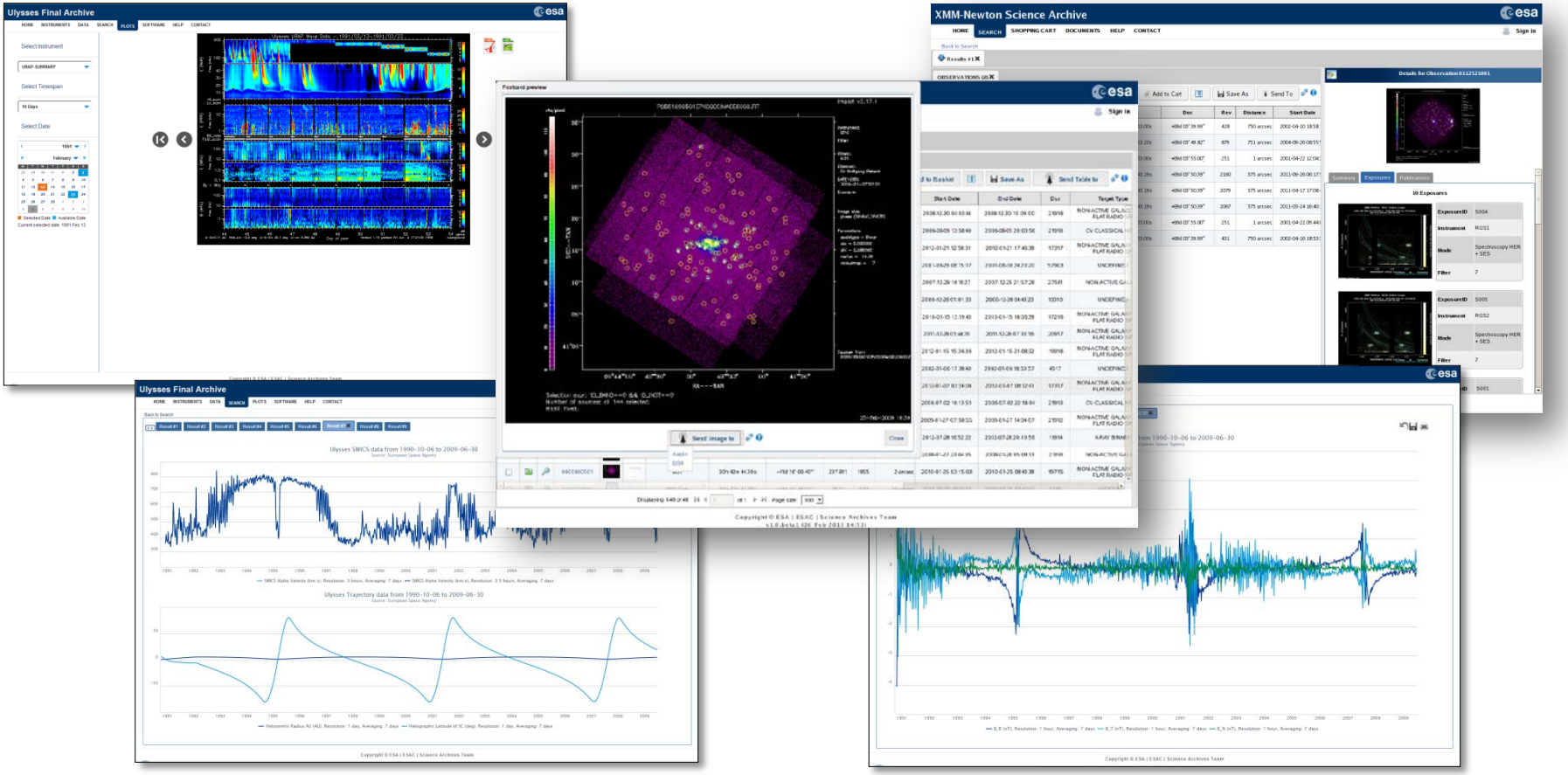
User Interfaces - 1st Generation of ESA Science Archives – Applets



User Interfaces – 2nd Generation of ESA Science Archives – Java Web Start



User Interfaces – 3rd Generation of ESA Science Archives – Web applications



Outline

- Introduction: ESA Science Archives

- Archives Architecture Evolution
  - User Interfaces and the Web 2.0
  - Application Frameworks
  - Databases, Spatial Indexing and Big Data

Application Frameworks

1st Generation

- In-house developed java server in non standard ports
- In-house RPC connection through TCP/IP
- In-house developed load balancing
- File retrieval through port 21
- VO data access protocols (SIAP, SSAP, SLAP)

2nd Generation

- Apache
- spring
- Use of standard ports and protocols
- File retrieval through port 80

3rd Generation

- Single software in servlet container
- HTTP

In Progress

- High availability
- Load balancing
- + VO compatible interfaces (TAP, VOSpace)

Outline

- Introduction: ESA Science Archives

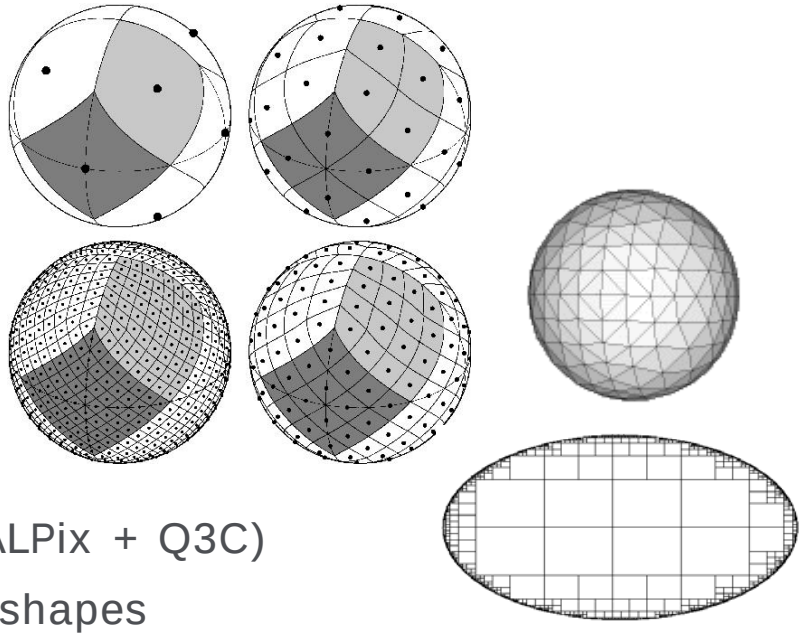
- Archives Architecture Evolution
  - User Interfaces and the Web 2.0
  - Application Frameworks
  - Databases, Spatial Indexing and Big Data

Databases

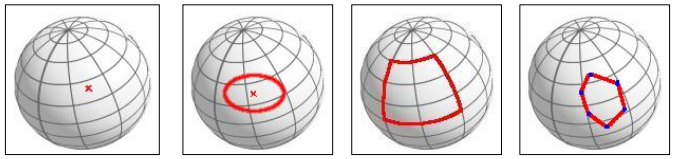
1st Generation	2nd Generation	3rd Generation (In Progress)
SYBASE, ORACLE	PostgreSQL	Postgres-XC, Greenplum
JDBC Access	HIBERNATE	hadoop, HIVE
In-house developed connection pooling	Standard connection pooling (c3p0)	High Availability
Non standard, ad-hoc pagination	Standard Pagination (Hibernate + Postgresql)	
	EHCACHE	
	Spatial indexing (q3c, h3c, pgSphere)	

Spatial Indexing

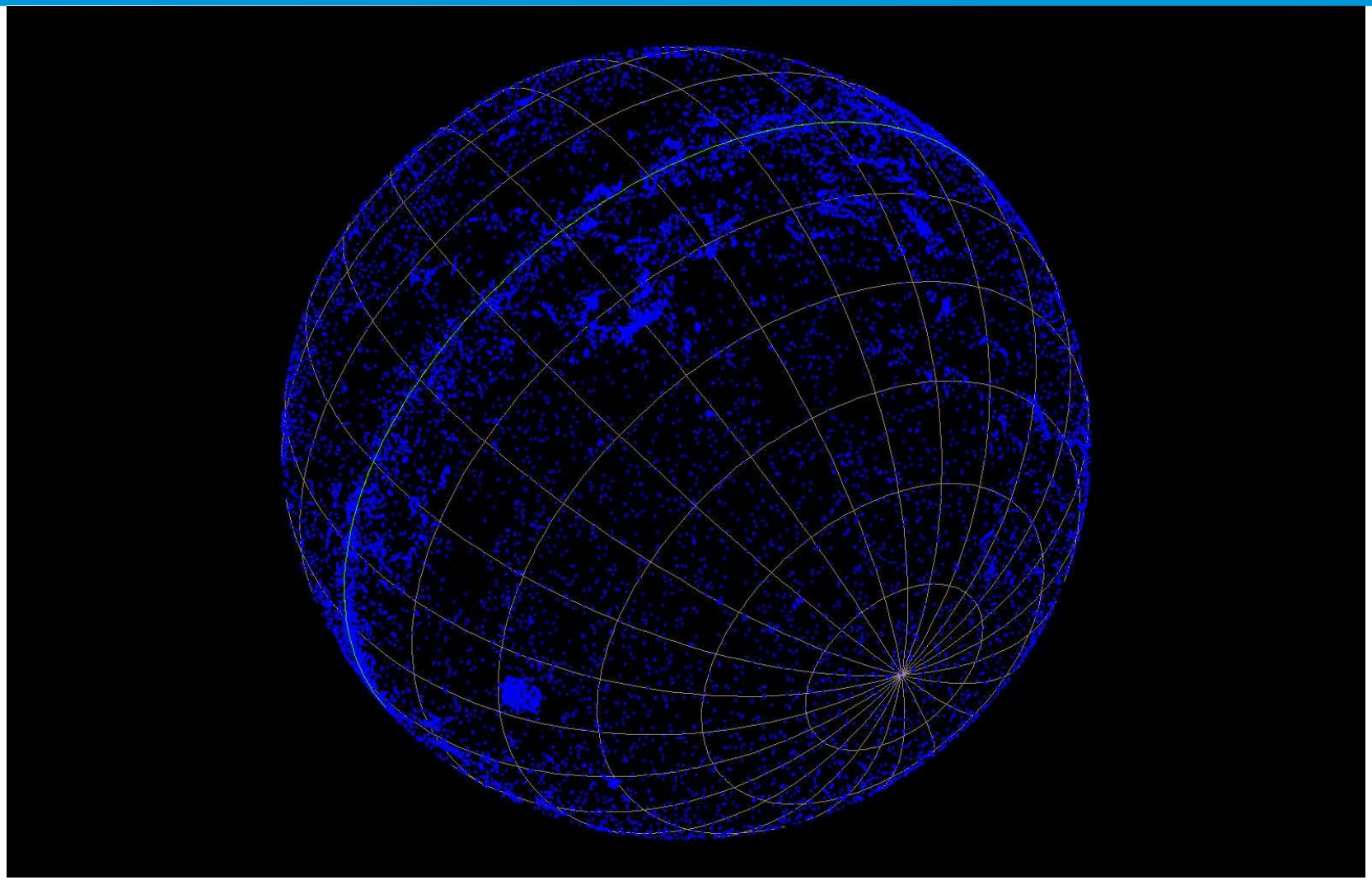
- Allows better performance on complex geometrical queries
  - Cone-Search, X-Match
  - Complex FOV overlap operations
- Avoids squaring ROI and post-processing overhead
- Some Sky-Pixelation schemas
  - HEALPix: Hierarchical Equal Area Iso Latitude
  - HTM: Hierarchical Triangular Mesh
  - Q3C: Quad Tree Cube



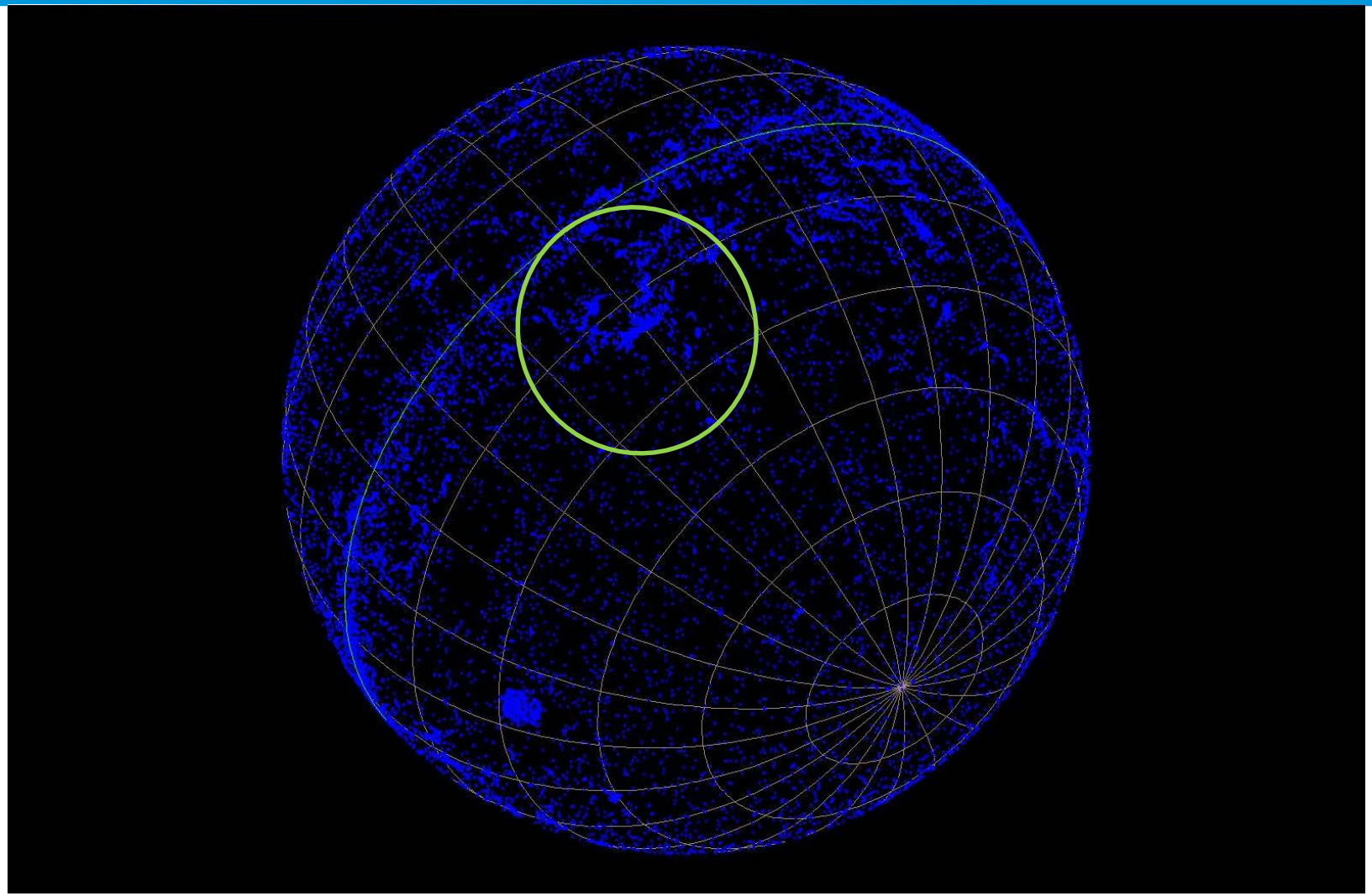
- PostgreSQL Plug-ins: PgSphere, Q3C and H3C (HEALPix + Q3C)
  - Database manages Geometrical operators and shapes
  - FOVs of stored observations are pre-computed



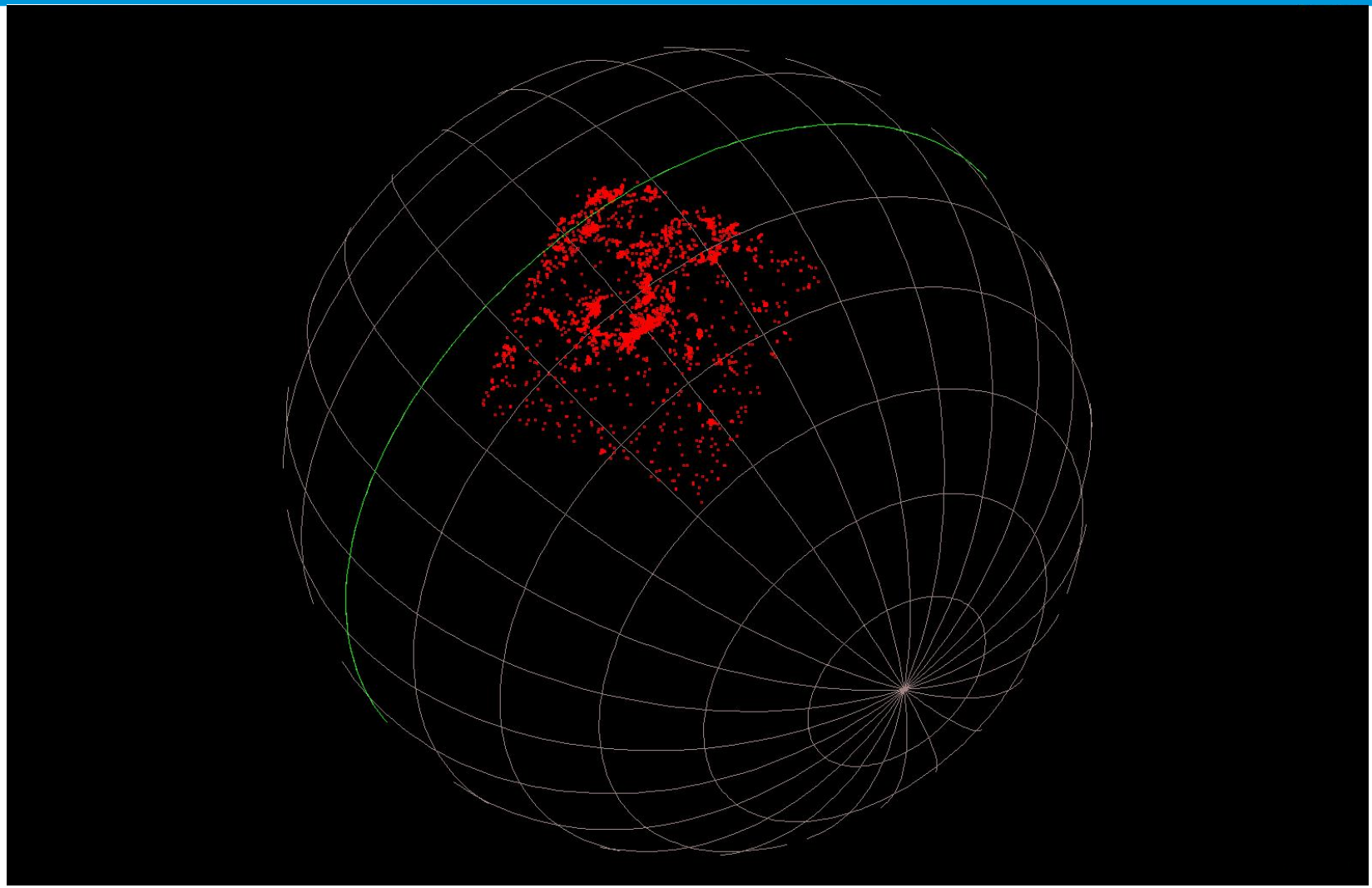
Spatial Indexing



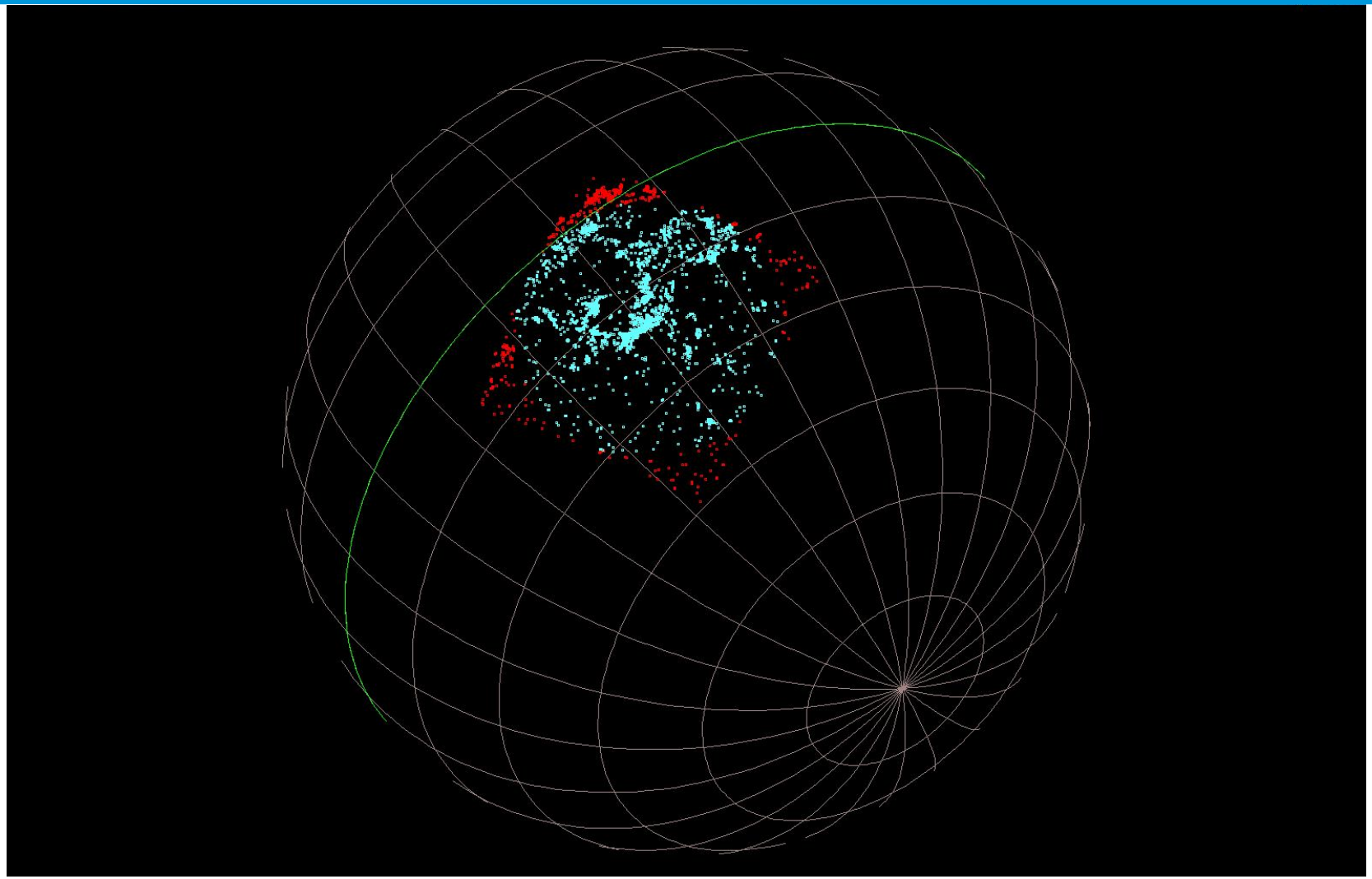
Spatial Indexing



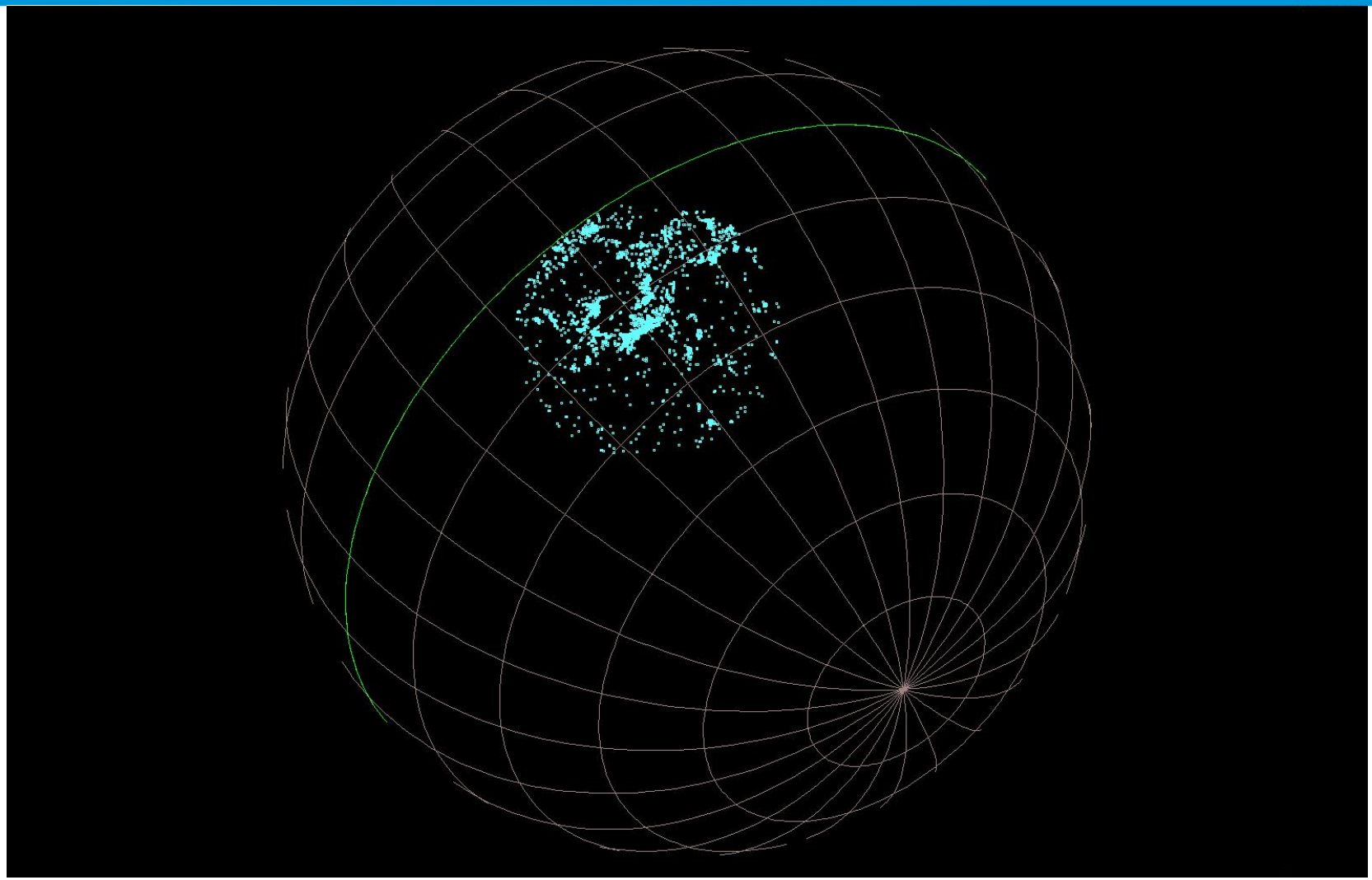
Spatial Indexing



Spatial Indexing

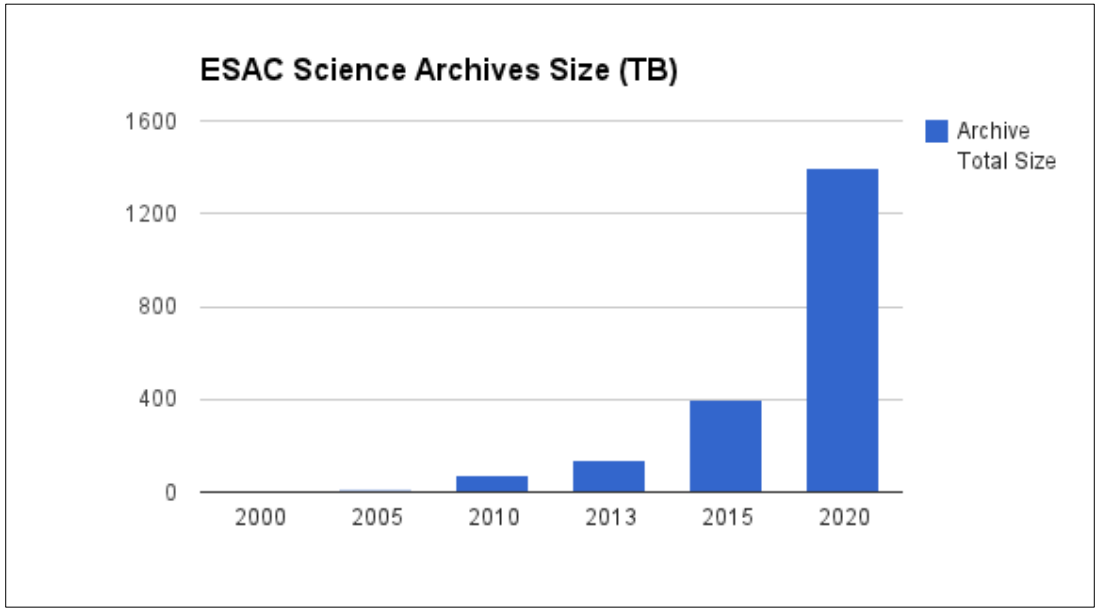


Spatial Indexing



Big Data

- ESA Science Archive data volume increasing exponentially



Big Data

- ESA Science Archive data volume increasing exponentially
- GAIA:
  - 1 billion sources
  - GUMS catalogue (synthetic) ~2 billion sources
  - Map-Reduce paradigm applied to PostgreSQL RDBMS
  - Hadoop cluster for advanced applications
  - Some numbers
    - Positional + Magnitude X-Match
      - GUMS Stellar Sources Catalogue ~2 billion sources
      - Fuzzy Synthetic Catalogue ~100 million sources
      - **1 degree radius, 9 seconds**

Query:

```
SELECT g.*, m.*, f.* FROM g10_ss_noclust AS g, test_xmatch_table AS m,
g10_fuzzy_100000000 AS f
WHERE g.source_id=m.g10_id AND m.id=f.source_id
AND g.source_id IN (
     SELECT g10_id FROM test_xmatch_table WHERE id = f.source_id
     ORDER BY (dist+mag_diff) ASC LIMIT 1
)
AND g.pos @ scircle'<(266.41683d, -29.00781d), Rd>'
```

Service times – Interactive Query

	GACSDB01 - fresh	GACSDB01 - cached	Output size
R=0.1d	3.7 s	96 ms	100 rows
R=0.5d	3.4 s	84 ms	100 rows
R=1d	4.4 s	148 ms	100 rows

Service times – Full output

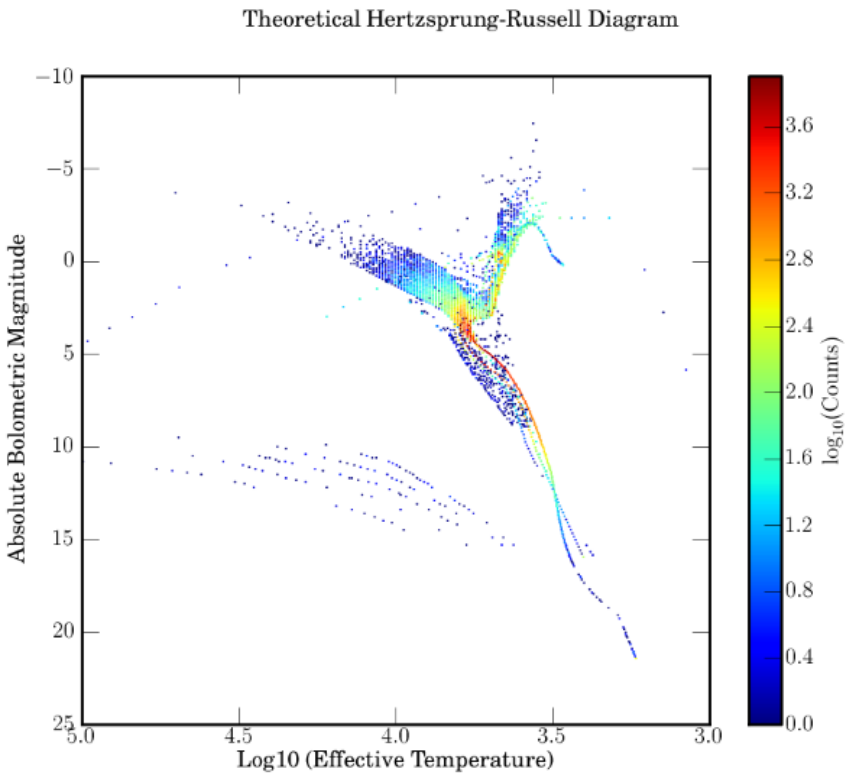
	GACSDB01 - fresh	GACSDB01 - cached	Output size
R=0.1d	7 s	198 ms	372 rows, 428 KB
R=0.5d	150 s	2.2 s	9628 rows, 11 MB
R=1d	678 s	8.6 s	44210 rows, 50 MB

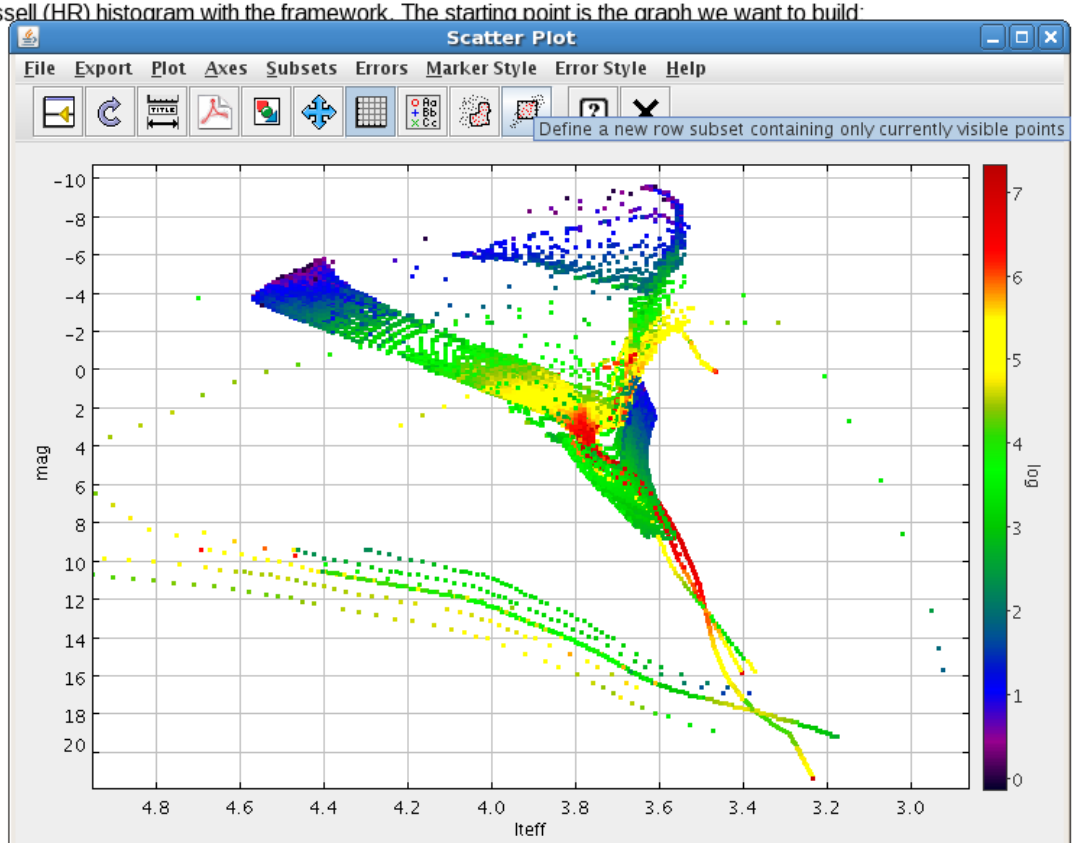
Big Data

- ESA Science Archive data volume increasing exponentially
- GAIA:
  - 1 billion sources
  - GUMS catalogue (synthetic) ~2 billion sources
  - Map-Reduce paradigm applied to PostgreSQL RDBMS
  - Hadoop cluster for advanced applications
  - Some numbers
    - Positional + Magnitude X-Match
      - GUMS Stellar Sources Catalogue ~2 billion sources
      - Fuzzy Synthetic Catalogue ~100 million sources
      - **1 degree radius, 9 seconds**
    - H-R Diagram, Full GUMS catalogue
      - **30 minutes**

Theoretical Hertzsprung-Russell histogram example

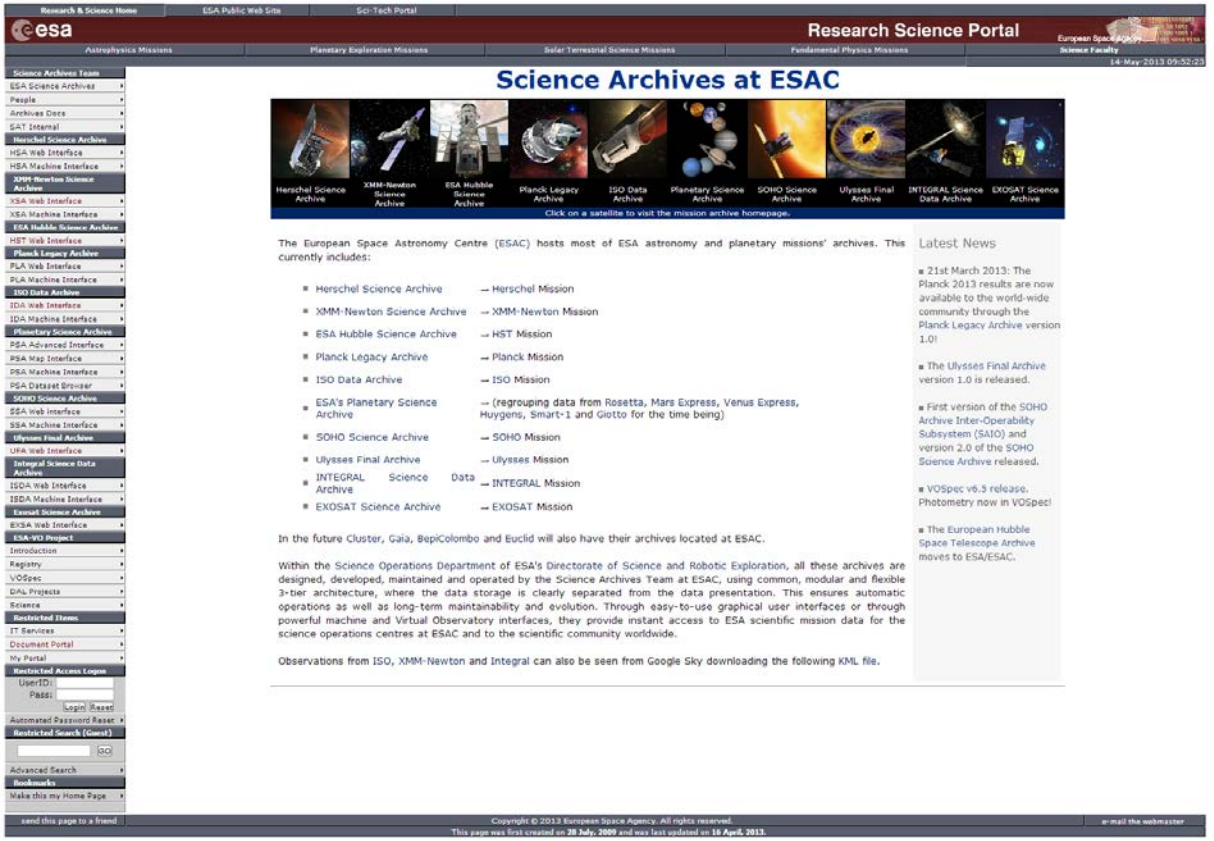
In this section, we will go through the steps needed for computing a theoretical Hertzsprung-Russell (HR) histogram with the framework. The starting point is the graph we want to build:





Come and visit us...

http://archives.esac.esa.int



THANK YOU

Science Archives Architecture Evolution

inaki.ortiz@sciops.esa.int