From ISO to Gaia : a 20 year journey through ESA Science Archives

*Christophe **Arviset**, Deborah **Baines**, Isa **Barbarisi**, Sebastien **Besse**, Guido **de Marchi**, Beatriz **Martinez**, Arnaud **Masson**, Bruno **Merin**, Jesus **Salgado***

ESAC Science Data Centre
ADASS 2016, Trieste, 17/10/2016

Issue/Revision: 1.0
Reference: A 20yr journey through ESA Archives
Status: Draft

ESAC Archives : an ever growing family

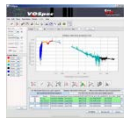
iso

xmm-newton

giotto

cassini-huygens

1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007



vospec

mars express

planck

herschel

hubble space telescope

double star

exomars

2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017

exosat

rosetta

proba-2

cluster

esasky

gaia

soho

smart-1

ulysses

hipparcos

venus express

ISS-SolACES

lisa pathfinder

Archives in different phases

Legacy Archives

- ISO Data Archive
- Exosat Science Archive
- Ulysses Final Archive

Active Archives

- Herschel Science Archive
- XMM-Newton Science Archive
- European HST Archive
- Planck Legacy Archive
- Gaia Archive
- Lisa PathFinder Archive

Planetary Science Archive

- Rosetta
- Mars Express
- Exomars 2016
- Venus Express
- Huygens
- SMART-1
- Giotto

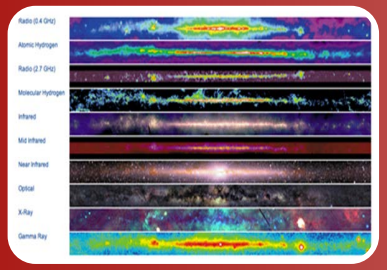
- SOHO Science Archive
- Proba-2 Archive
- Cluster Science Archive
- Double Star Science Archive
- ISS-SOLACES Archive

Archives in development

- Euclid Archive Systems
- Solar Orbiter Archive
- BepiColombo Archive Core Sys.
- JWST

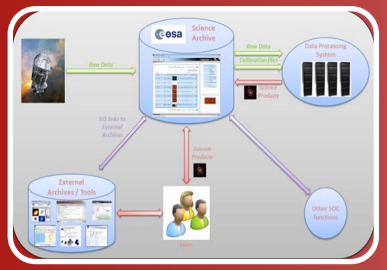
http://archives.esac.esa.int/

ESAC Science Archives Strategy



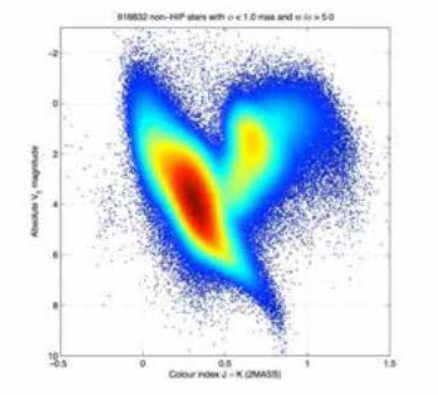
Enable maximum ***scientific exploitation*** of data sets

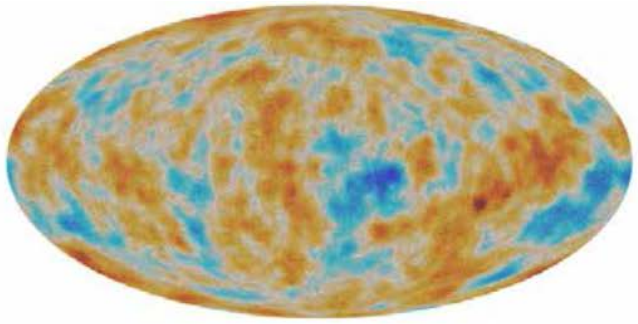
Enable efficient ***long-term preservation*** of data, software and knowledge, using modern technology

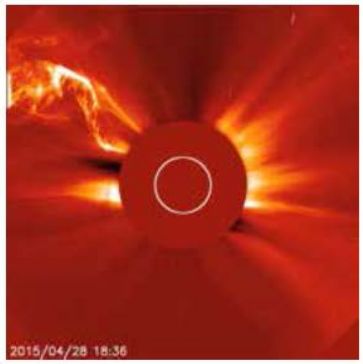


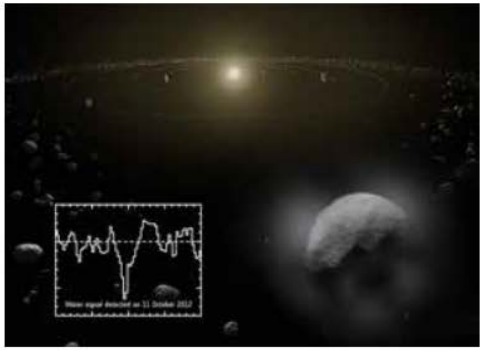
Enable cost-effective archive production by ***integration in, and across, projects***

Enable Maximum Science Exploitation









Scientists and Engineers working very closely together

Science driven Archives

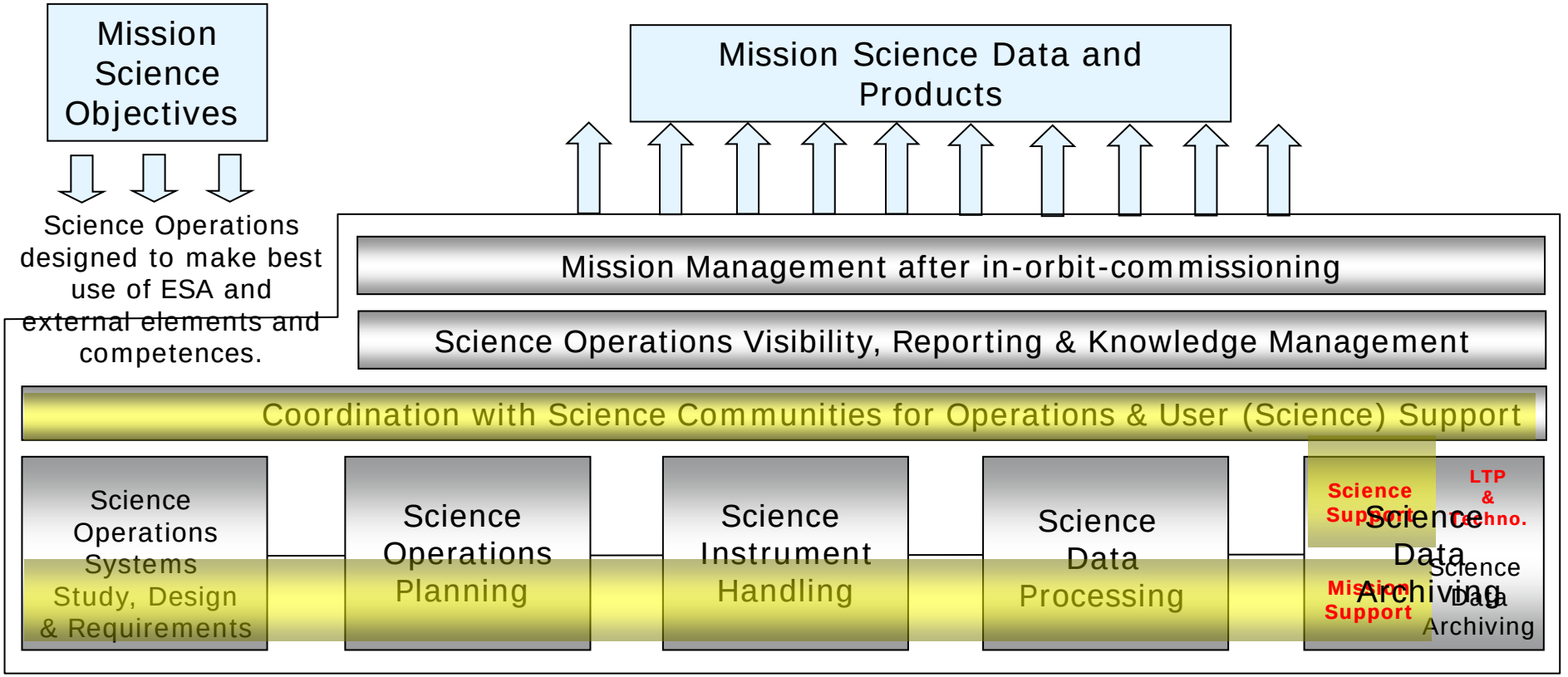
Long Term Preservation Strategy

1. Consolidation of all ESA Space Science Archives at ESAC, with strong re-use across projects, ensuring easier and cheaper long term data preservation
   - Hardware infrastructure
   - Software architecture and code, including technology migration
   - Human technical and scientific expertise
   - Multi mission, multi instruments science exploitation

2. Long term preservation of data processing capabilities (on-going)
   - Preserve software coming from various places
   - Provide data processing capabilities as a “service”
   - Bring the “user software to the data” instead of the “data to the user”

3. Sharing and preservation of knowledge, including international cooperation
   - IVOA, IPDA

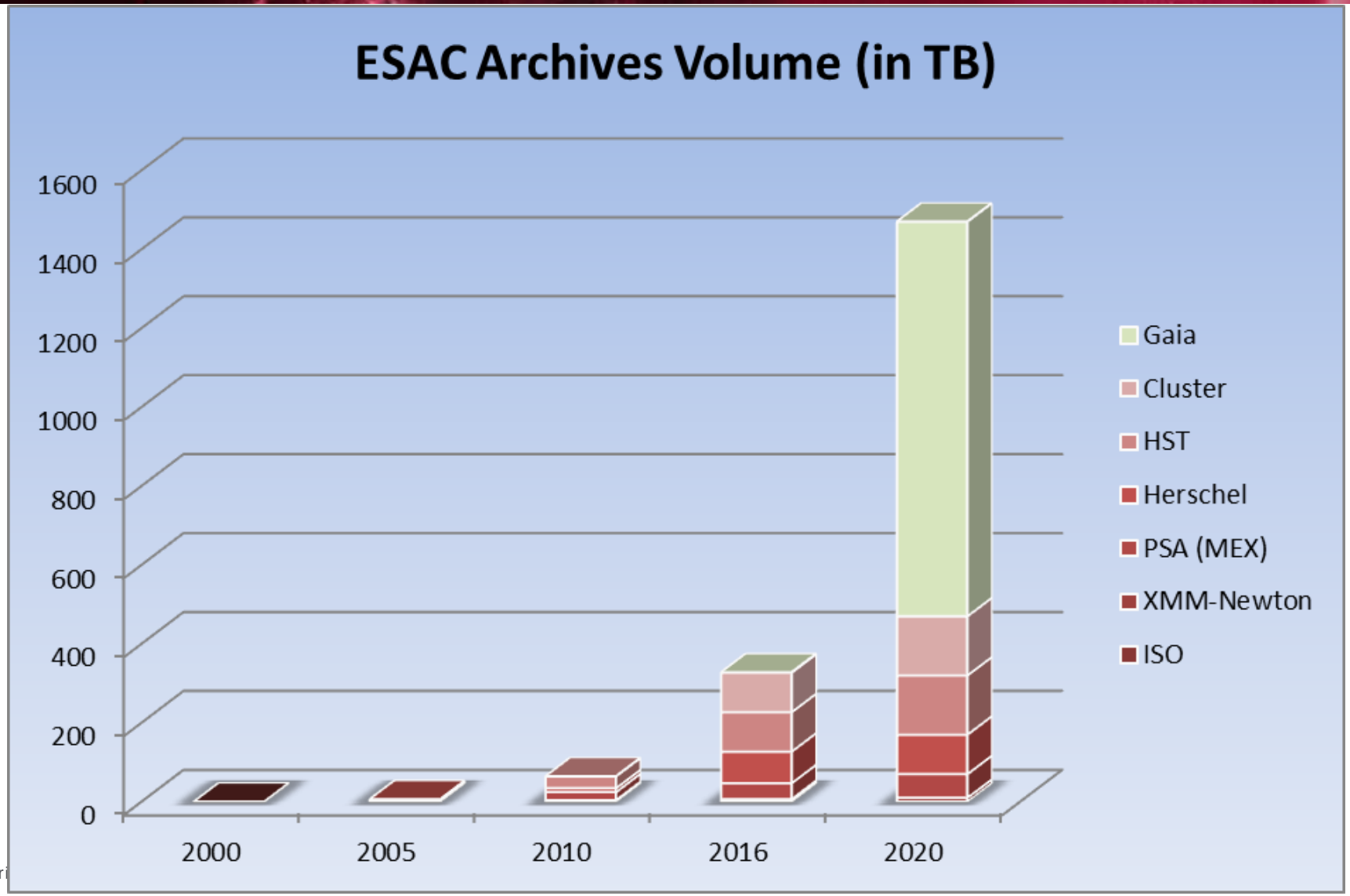
Archives as a Core Component of ESA Science Operations



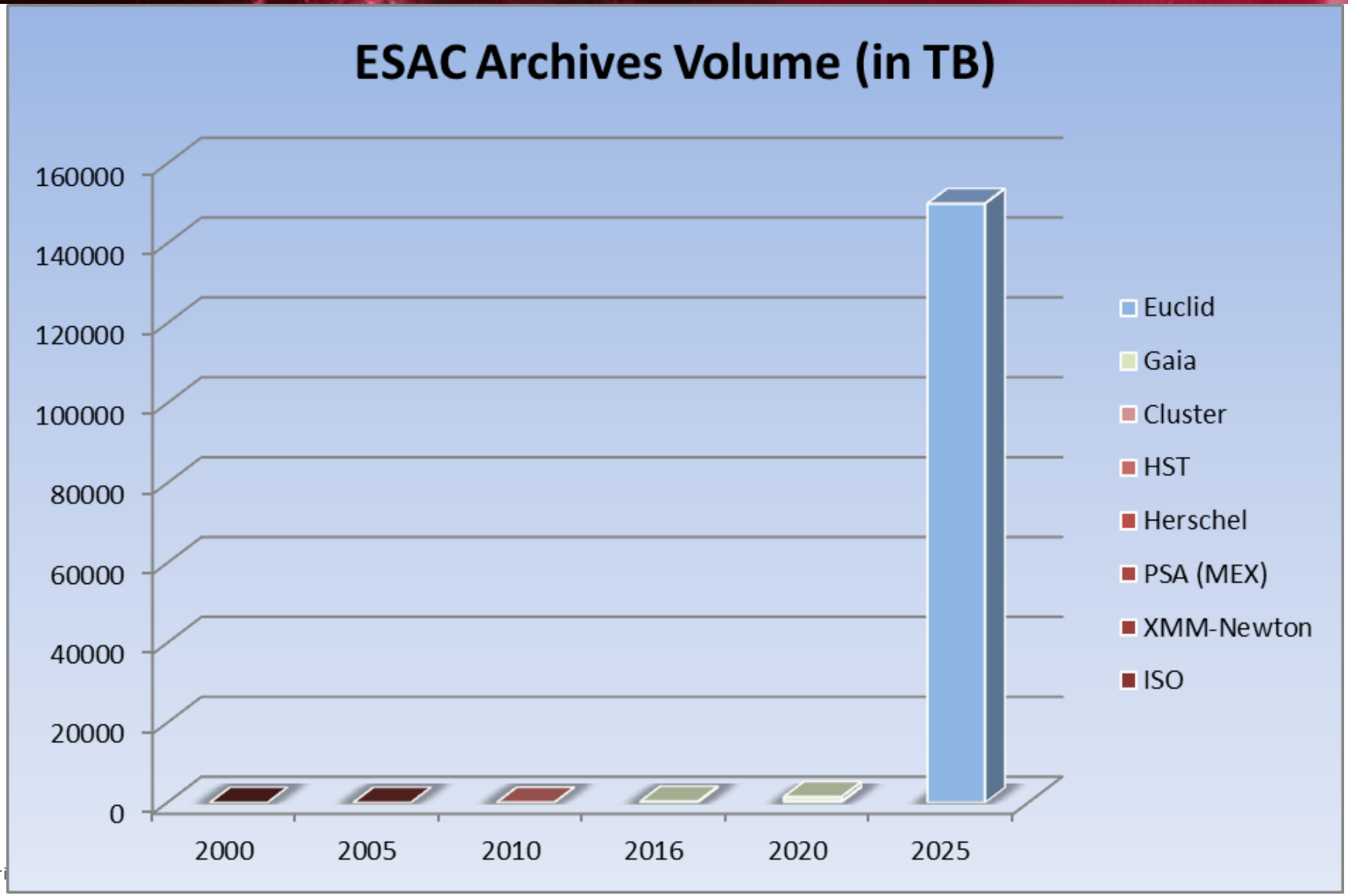
Archive is NOT an afterthought at the end of the in-orbit operations !

Archive development starts in the early phase of the project !

Archives Volume evolution (2000-2020)



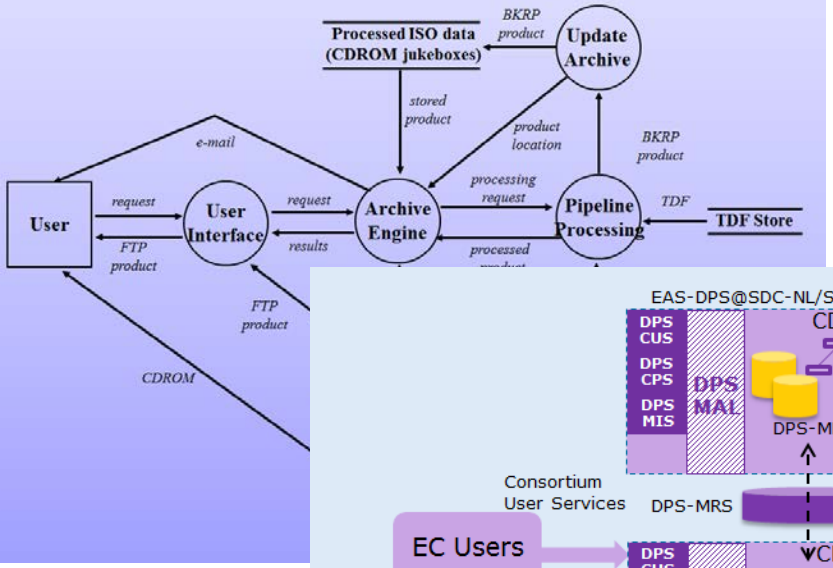
Archives Volume evolution (2000-2025)



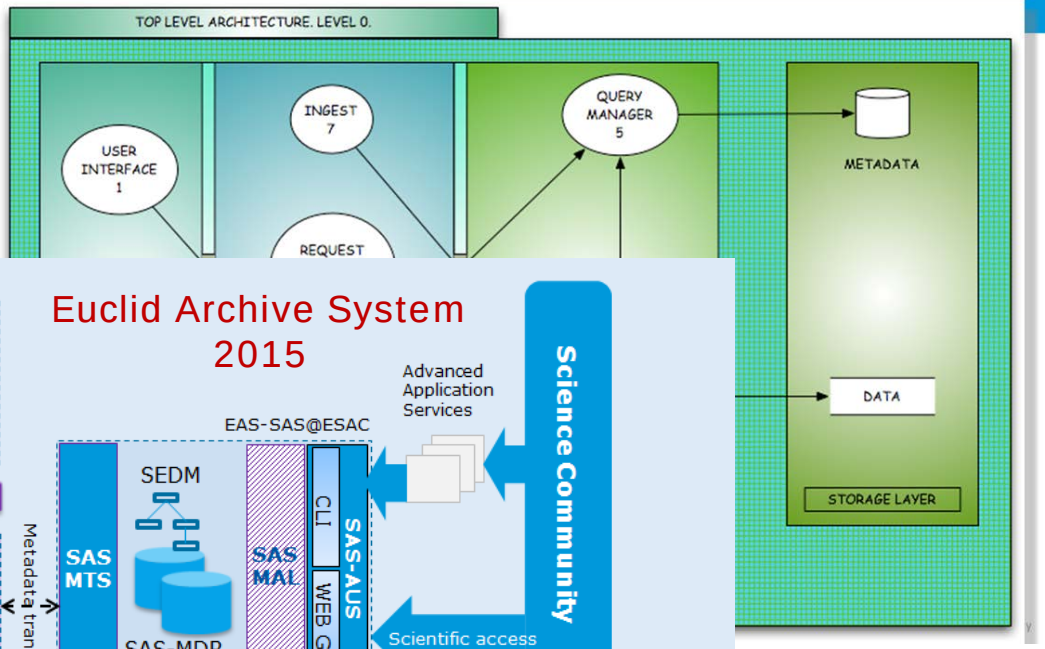
Chri

Software Design DOES Matter !

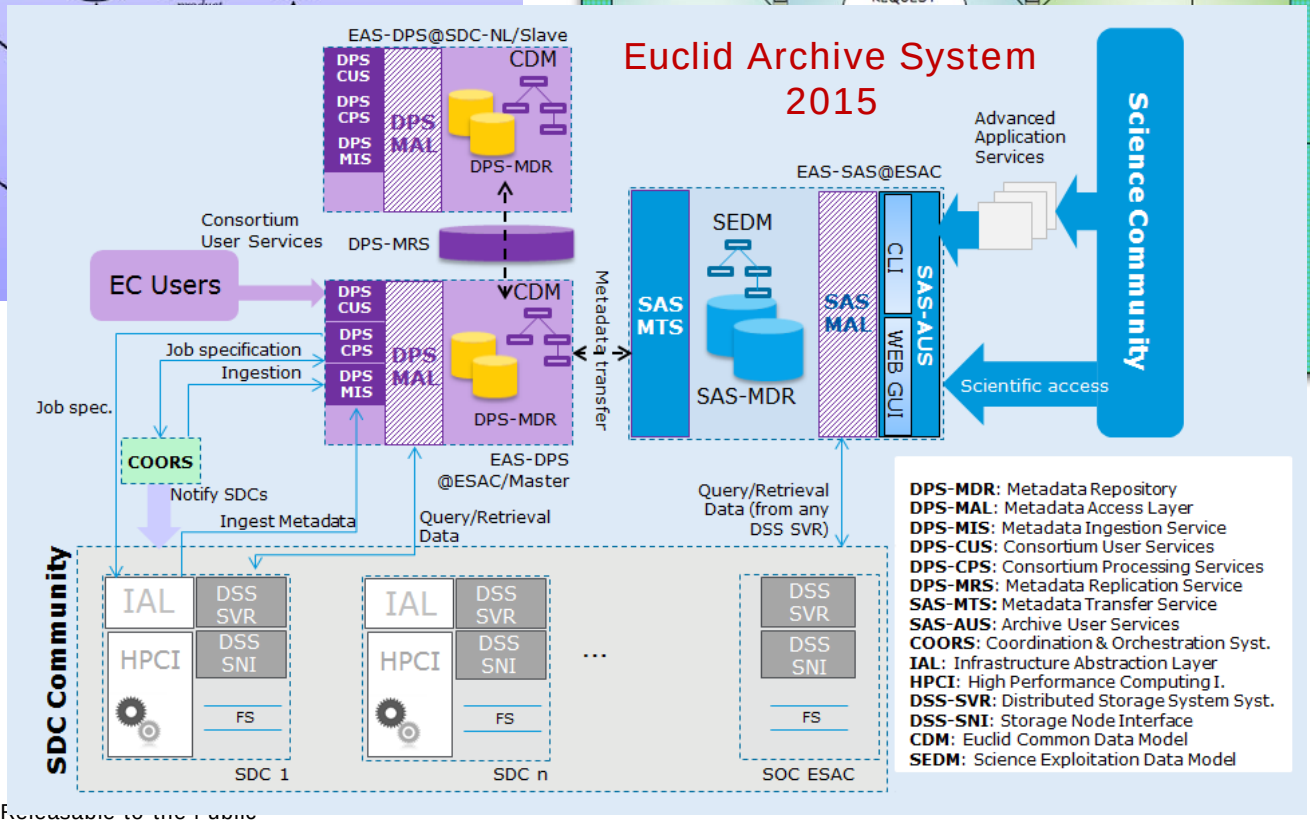
ISO Data Archive Top Level System Design - 1998



ESAC Archives Common Architecture - 2009



Euclid Archive System 2015



Archive Technology Evolution

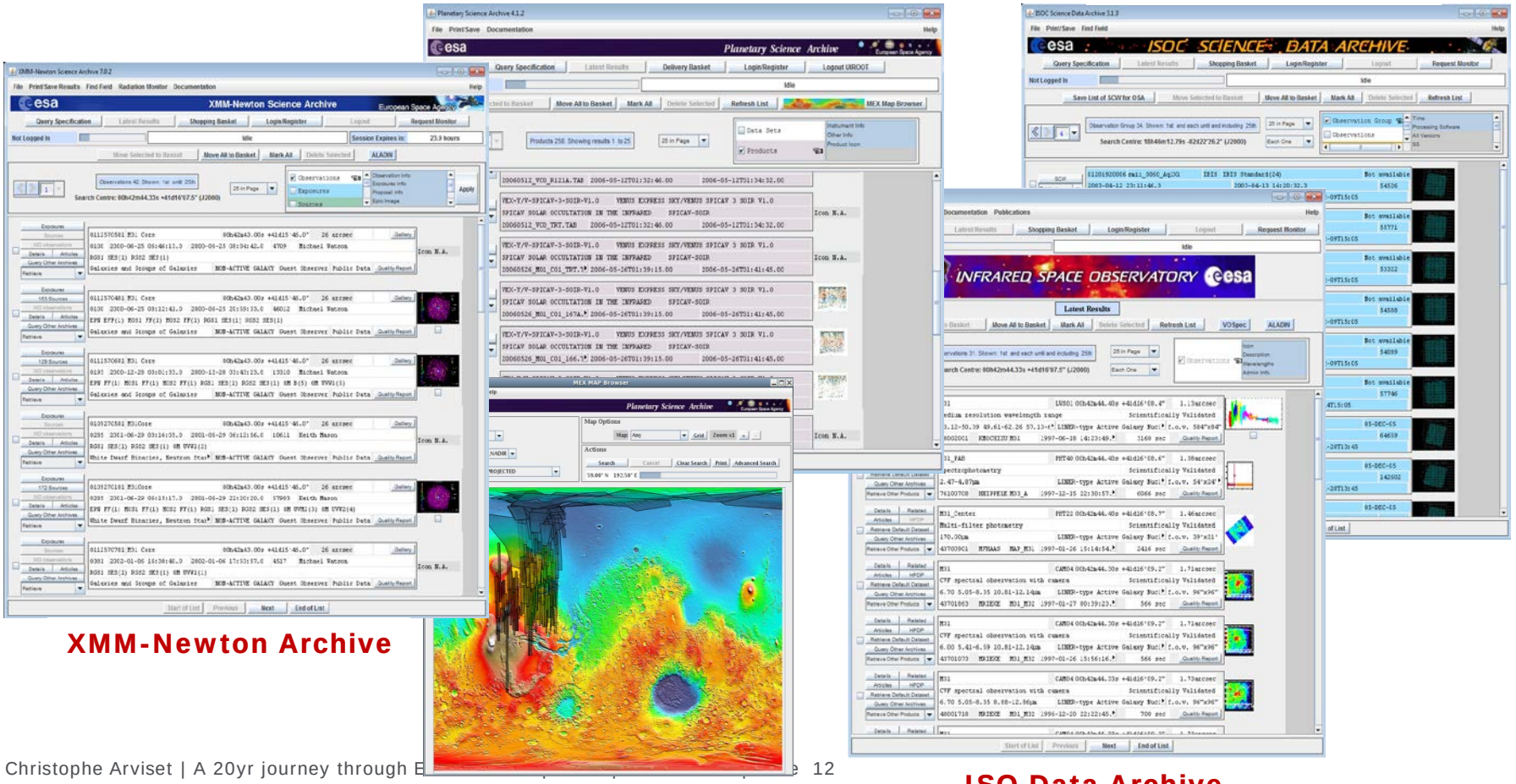
1997	2008	2014	2018+
1st Generation	2nd Generation	3rd Generation	4th Generation
Applet	Webstart	Web Thin Client	Big Data
Java 1.1	Java 1.5	GWT vaadin HTML CSS	docker Bring code to data
in house RPC client server connection	Spring Java Application Framework	https	Lucene Solr hadoop Data mining
FTP	VO layer	VO inside	High availability
ORACLE DATABASE SYBASE COTS DB	PostgreSQL PgSphere Open Source DB	PostgreSQL PgSphere	....
		Viz. tools ALADIN	

1st generation of Archives (1998…)

Java Applets, html could not support all requirements at the time

Planetary Science Archive

Integral SOC Archive

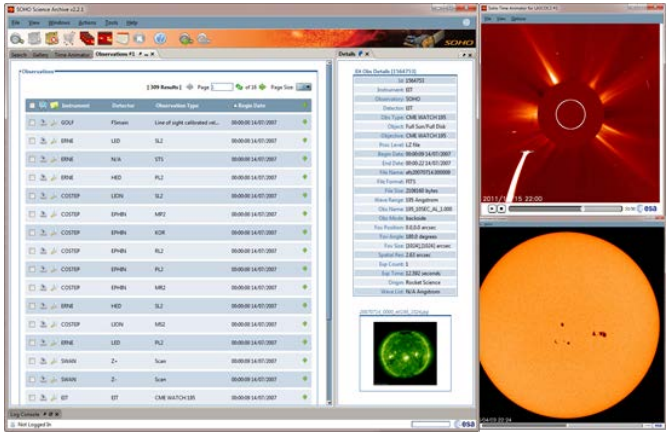


XMM-Newton Archive

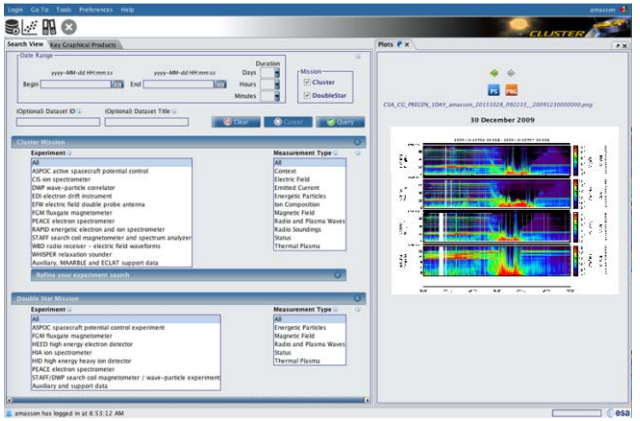
ISO Data Archive

2nd generation of Archives (~2005…)

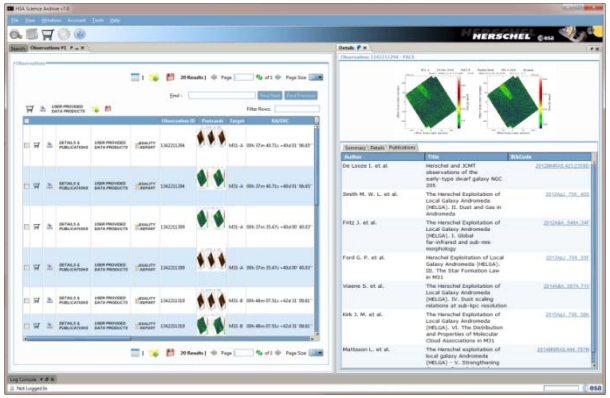
Java web start, more user friendly GUI



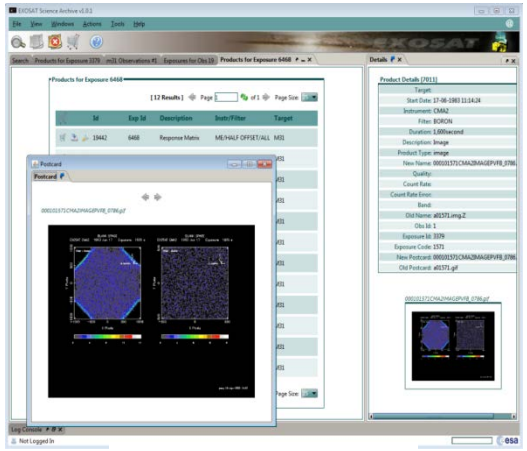
Soho Science Archive
Including Proba-2 data



Cluster Science Archive,
Including Double Star data



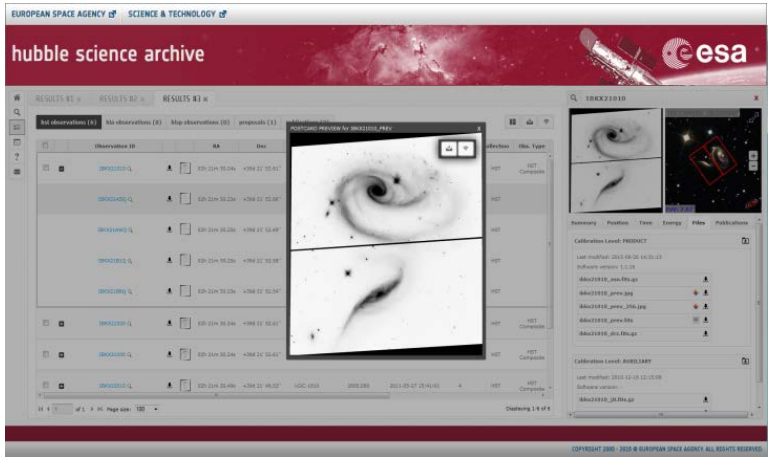
Herschel Science Archive



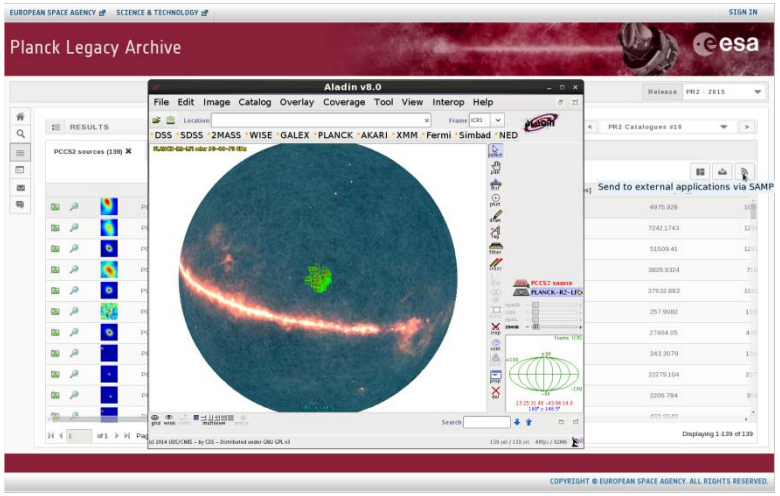
Exosat Archive

3rd generation of Archives (~2013-...)

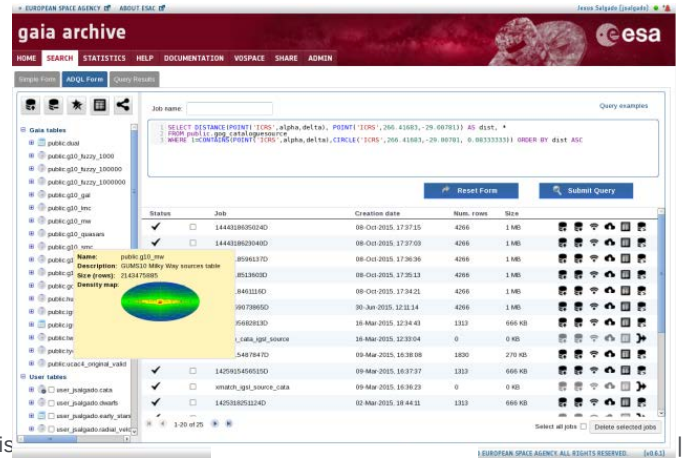
Thin layer web clients



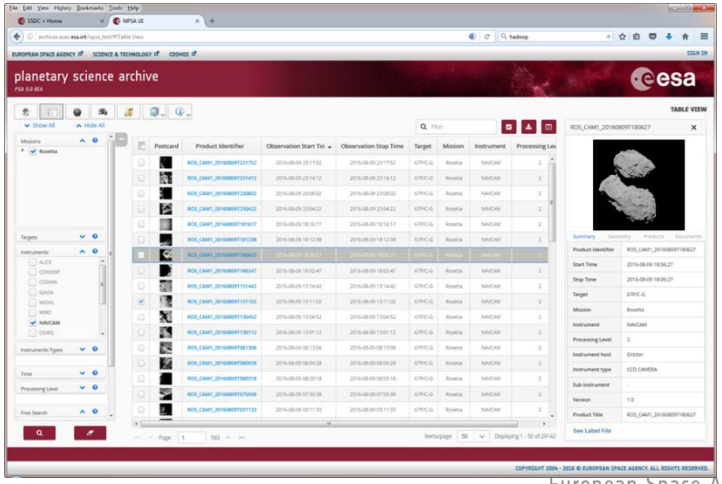
European HST Archive



Planck Legacy Archive

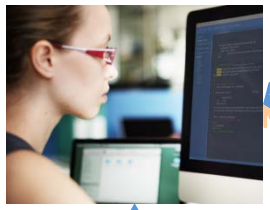


Gaia Archive



Planetary Science Archive

Standard Archives Architecture



Command line

http

xmm-newton science archive

Browser GUI

http

Database

SAMP

SIAP, SSAP, TAP

VO services

Data Repository

VO Apps

ftp

User Disk

Need for new paradigm

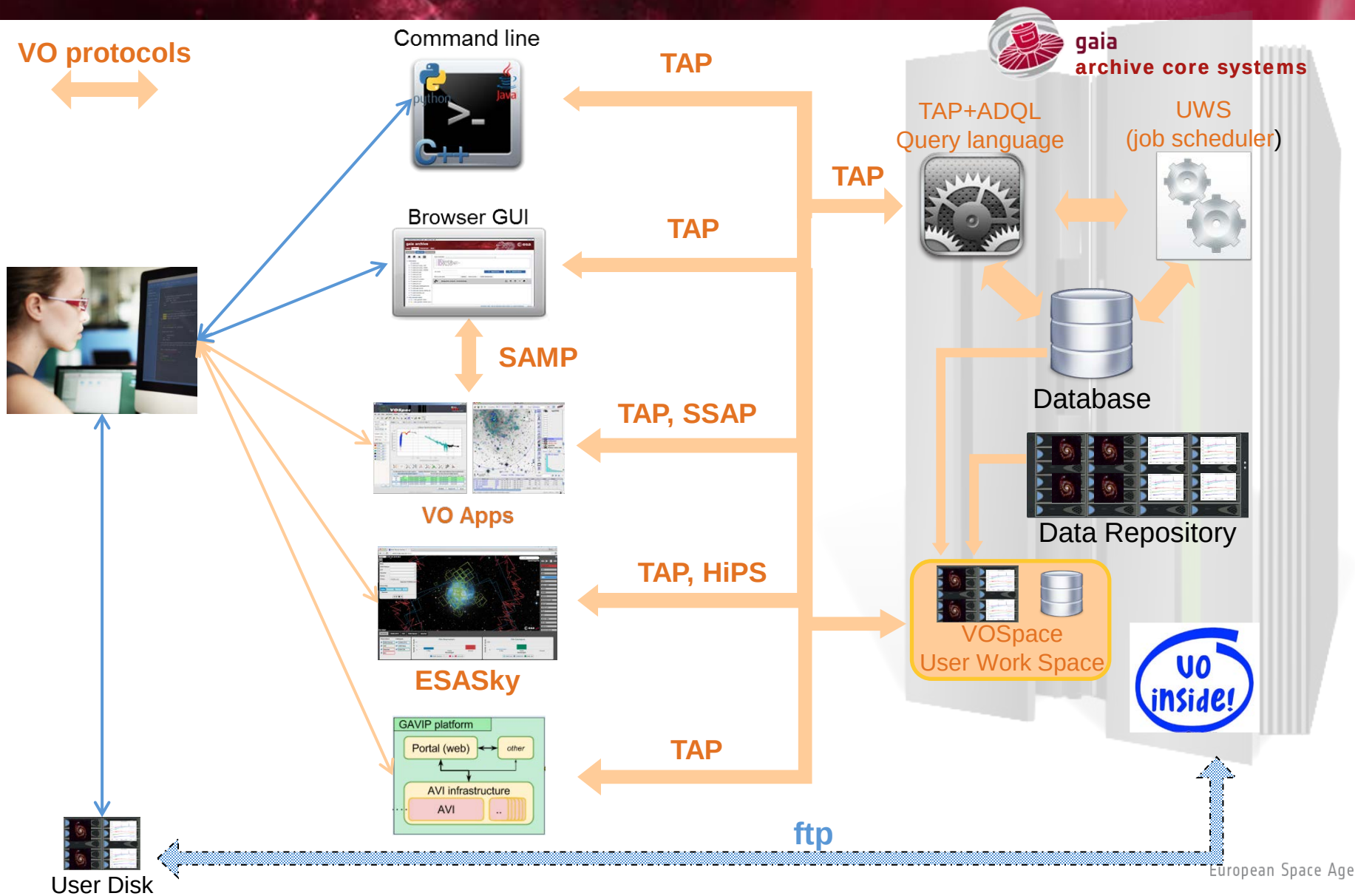
1. New ways required to access the Gaia catalogue and associated data
   - Powerful query mechanism, asynchronicity of results
   - One “query interface” for all archive services and VO services

2. User can not download all catalogue and all data
   - Need to have user workspaces *IN* the Archive
     - User database space, user disk space
   - User workspace shareable amongst various users

3. Bring user code to the data
   - Part of the user workspace in the archive
   - Share code with other users

The user works with the data WHERE the data is : Archive 2.0 concept

Gaia Archive Architecture



Towards multi wavelengths Science data exploitation

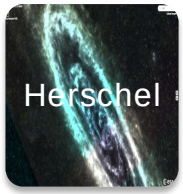
- **Goal**: to facilitate data discovery and archival science for ALL users
  - Multi-wavelength
  - Project agnostic
  - Exploration

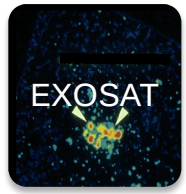
- Interface “on top of” all ESA astronomy archives

ESASky

VO inside!

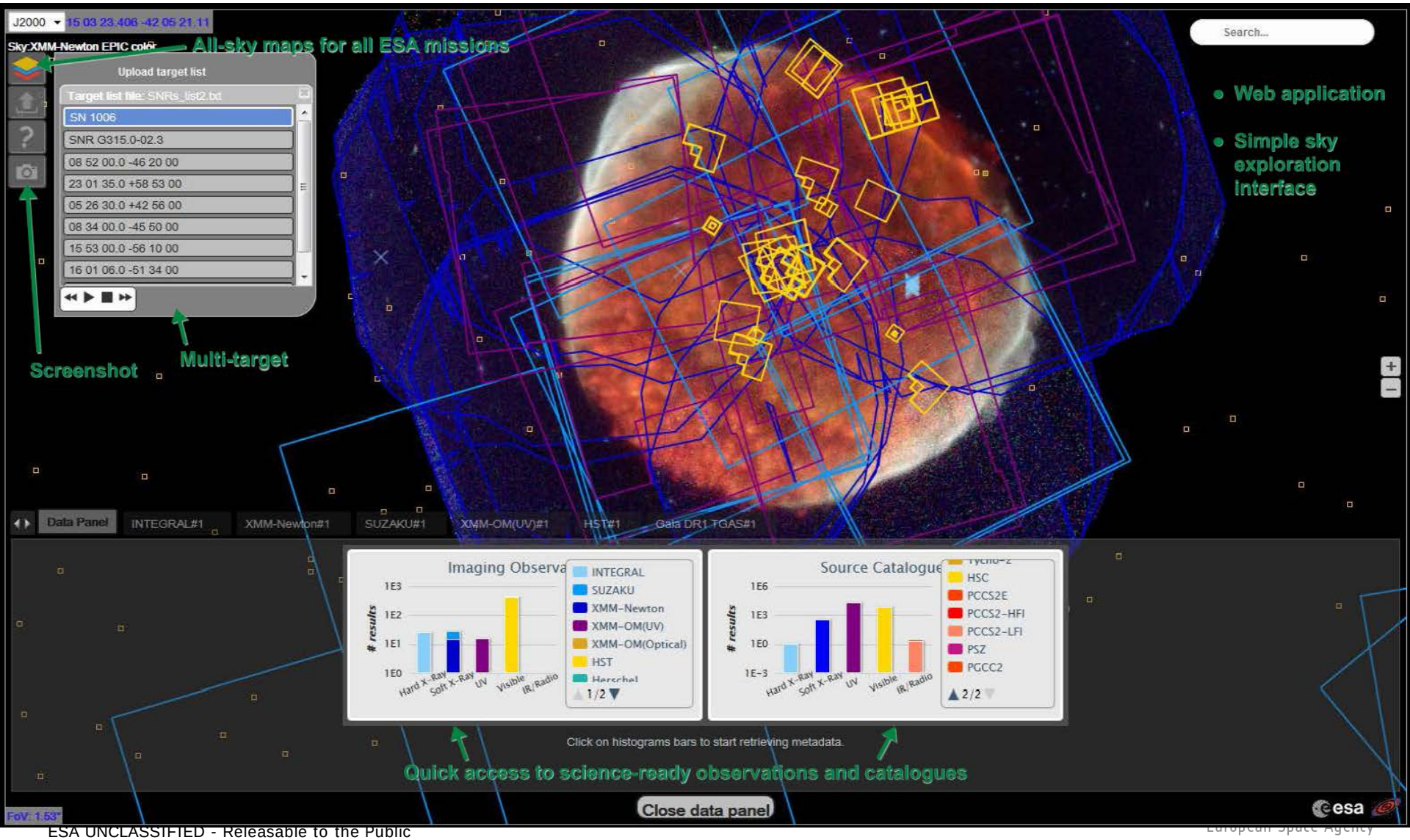
Legacy: IUE, ..





Future: Gaia, JWST, Euclid, ..

sky.esa.int


J2000 15 03 23.406 -42 05 21.11
Sky:XMM-Newton EPIC color
All-sky maps for all ESA missions
Upload target list
Target list file: SNRs_list2.txt
SN 1006
SNR G315.0-02.3
08 52 00.0 -46 20 00
23 01 35.0 +58 53 00
05 26 30.0 +42 56 00
08 34 00.0 -45 50 00
15 53 00.0 -56 10 00
16 01 06.0 -51 34 00
Screenshot
Multi-target
Search...
Web application
Simple sky exploration interface
Data Panel
INTEGRAL#1
XMM-Newton#1
SUZAKU#1
XMM-OM(UV)#1
HST#1
Gaia DR1 TGAS#1
Imaging Observa
results
1E3
1E2
1E1
1E0
Hard X-Ray
Soft X-Ray
UV
Visible
IR/Radio
INTEGRAL
SUZAKU
XMM-Newton
XMM-OM(UV)
XMM-OM(Optical)
HST
1/2
Source Catalogue
1E6
1E3
1E0
1E-3
HSC
PCCS2E
PCCS2-HFI
PCCS2-LFI
PSZ
PGCC2
2/2
Click on histograms bars to start retrieving metadata.
Quick access to science-ready observations and catalogues
Close data panel
FoV: 1.53°
esa

Login or not Login : that is the question

Semi Open access (Free access to query, need login to retrieve data)

- "Burden" for the user, especially for public data
- Required to access proprietary data
- Enables asynchronous services
  - (long queries, big volume download, user spaces)
- Easier usage stats, better knowledge of the users

Full Open access (no login)

- Instant access for the users
- No reason to login for public data access
- Usage stats based on IP addresses more complex and not as accurate

Conclusion : combine both !
Open access is default, offer login for extra services

Mission and Archives metrics ?

1. Mission success usually measured by number of scientific papers

ESA-led Missions	Launch	<2000	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Total
COS-B	1975	173	1																174
IUE	1978	3,217	147	130	76	70	89	31	21	30	25	26	39	30	15	19	15	14	3,994
Exosat	1983	706	8	3	4			2	1	2	1	2			1	2	3		735
Hipparcos	1989	729	194	154	147	133	107	100	120	92	89	89	85	67	89	71	59	42	2,367
Ulysses	1990	898	83	138	39	89	57	56	36	42	52	38	59	34	43	34	34	19	1,751
ISO	1995	490	180	138	146	129	123	107	67	35	33	33	28	10	5	5	3	2	1,534
SOHO	1995	588	292	203	283	301	324	321	271	368	332	323	274	298	243	189	211	122	4,943
Huygens	1997		7	8	15	7	12	24	21	26	26	16	15	8	11	7	5		208
XMM-Newton	1999		23	91	100	231	329	298	327	334	278	314	332	352	298	330	306	220	4,163
Cluster	2000	54	1	38	22	65	127	168	145	123	183	176	192	232	187	171	155	124	2,163
INTEGRAL	2002					79	27	67	99	84	84	96	80	60	67	45	47	41	876
SMART-1	2003						2	8	10	5	10	12	6	5	7	7	7	1	80
Mars Express	2003					4	21	49	87	75	83	89	109	96	82	64	63	19	841
Rosetta	2004						22	16	36	58	29	17	39	25	54	29	41	54	420
Venus Express	2005								21	44	79	48	41	54	75	60	40	52	514
Herschel	2009												228	109	255	323	336	241	1,492
Planck	2009													48	69	118	328	225	788
PROBA-2	2009													7	7	22	13	6	55
Gaia	2013																32	25	57
Non ESA-led																			
HST	1990	2,329	537	558	602	607	611	681	714	728	705	678	734	788	844	784	825	510	13,235
Cassini	1997	22	1	16	18	12	49	108	150	157	173	199	234	184	177	135	127	38	1,800
Double Star	2003						1	27	7	12	40	11	15	11	6	4	4	2	140
Suzaku	2005								20	50	78	138	92	109	101	94	78	2	762
AKARI	2006								2	29	22	21	47	50	61	37	29	3	301
Hinode	2006									62	105	135	133	134	129	100	97	63	958
IRIS	2013															3	22	33	58

Miss

1. Mis
2. Hov

ESA Science Archives Usage Statistics

Mission	In-orbit operations	Period	Active users / month	Volume downloaded / month (TB)	% of total volume downloaded / month	Archive total size (TB)	Notes
Active Archives							
Cluster	2000-	2015-H1	171	3.43	3%	108.80	CSA @ESAC data distribution since Oct 2014 Since 2015-H2, archive size reflects only Science Data
		2015-H2	194	2.57	3%	90.00	
		2016-H1	171	2.34	2%	95.98	
		2016-H2	124	2.85	3%	97.46	
Herschel	2009-2013	2015-H1	901	10.99	21%	52.67	Archive used as prime mean for data distribution
		2015-H2	855	7.89	20%	38.57	
		2016-H1	944	14.74	23%	65.29	
		2016-H2	797	15.09	19%	78.35	
HST	1990-	2015-H1	204	1.01	1%	95.23	European HST Archive now at ESAC since June 2012
		2015-H2	181	0.19	0%	99.06	
		2016-H1	188	0.42	0%	100.51	
		2016-H2	231	0.13	0%	103.63	
Planck	2009-2013	2015-H1	758	14.01	36%	39.40	Public opening of the Planck Archive on 21st March 2013
		2015-H2	615	12.16	23%	52.00	
		2016-H1	609	1.23	2%	52.00	
		2016-H2	554	1.74	3%	52.00	
PSA (MEX, VEX, Rosetta, Exomars, Smart-1, Huygens and Giotto)		2015-H1	395	5.74	20%	28.40	MEX, Rosetta and Huygens data also available via NASA PDS
		2015-H2	444	9.14	27%	34.20	
		2016-H1	403	5.52	15%	37.40	
		2016-H2	143	2.63	6%	41.60	
Soho @ ESAC	1995-	2015-H1	168	0.34	5%	7.14	Main Soho Archive located @ GSFC, USA, maintained by ESA
		2015-H2	152	0.04	1%	7.42	
		2016-H1	165	0.01	0%	8.00	
		2016-H2	152	0.01	0%	8.09	
XMM-Newton	1999-	2015-H1	717	2.92	45%	6.46	Archive used as prime mean for data distribution
		2015-H2	788	2.01	30%	6.79	
		2016-H1	930	2.07	30%	6.90	
		2016-H2	941	2.37	32%	7.39	
Gaia	2013-	2016-H2	1,153	73.00	5214.29%	1.40	Gaia Archive at ESAC, released on 14/09/2016
ISS-SolACES	2008-	2016-H2	21	0.11	8%	1.30	ISS-SolACES at ESAC released 31/08/2016 (values in GB)
Legacy Archives							
Exosat	1983-1985	2015-H1	0	1.50	2%	83.90	Legacy Archive (values in GB)
		2015-H2	1	0.93	1%	83.90	
		2016-H1	0	4.24	5%	83.90	
		2016-H2	1	0.51	1%	83.90	
ISO	1995-1998	2015-H1	88	0.49	0%	171.00	Legacy Archive since 2002 (values in GB)
		2015-H2	82	0.65	0%	171.00	
		2016-H1	91	0.96	1%	171.00	
		2016-H2	78	1.48	1%	171.00	
Ulysses	1990-2009	2015-H1	27	26.19	13%	204.78	Legacy Archive Released in February 2013 (values in GB)
		2015-H2	18	23.38	11%	204.78	
		2016-H1	33	18.88	9%	204.78	
		2016-H2	24	10.33	5%	204.78	
active users	means	user logged in and downloading data at least once a month					
		IP address downloading data at least once a month					

Mission and Archives metrics ?

1. Mission success usually measured by number of scientific papers

2. How to measure success of the archives ?
   - Number of active users / IP addresses ?
   - Volume of data being downloaded vs size of the archive ?
   - Ratio of data-out vs data-in ?

Mission and Archives metrics ?

1. Miss
2. How

Mission	Data-in	Data-out	Ratio
Cluster	30	29	1.0
SOHO	4	5	1.2
PSA	19	262	13.8
Herschel	53	206	3.9
HST	67	28	0.4
Planck	42	168	4.0
XMM	3	98	32.7
Total (TB)	**67**	**173**	**2.6**

Mission and Archives metrics ?

1. Mission success usually measured by number of scientific papers

2. How to measure success of the archives ?
   - Number of active users / IP addresses
   - Volume of data being downloaded vs size of the archive
   - Ratio of data-out vs data-in ?

3. Still not perfectly unified metrics, still big differences
   - between disciplines,
   - between missions,
   - between missions phases

4. Would be useful to hear how others archive centre do ?

Looking back and looking forward

Archives serves various functions

- Maximum science exploitation of data
- Ensure long term preservation of data, software and knowledge
- Support Science Ground Segments from early phases

Scientists and Engineers working together

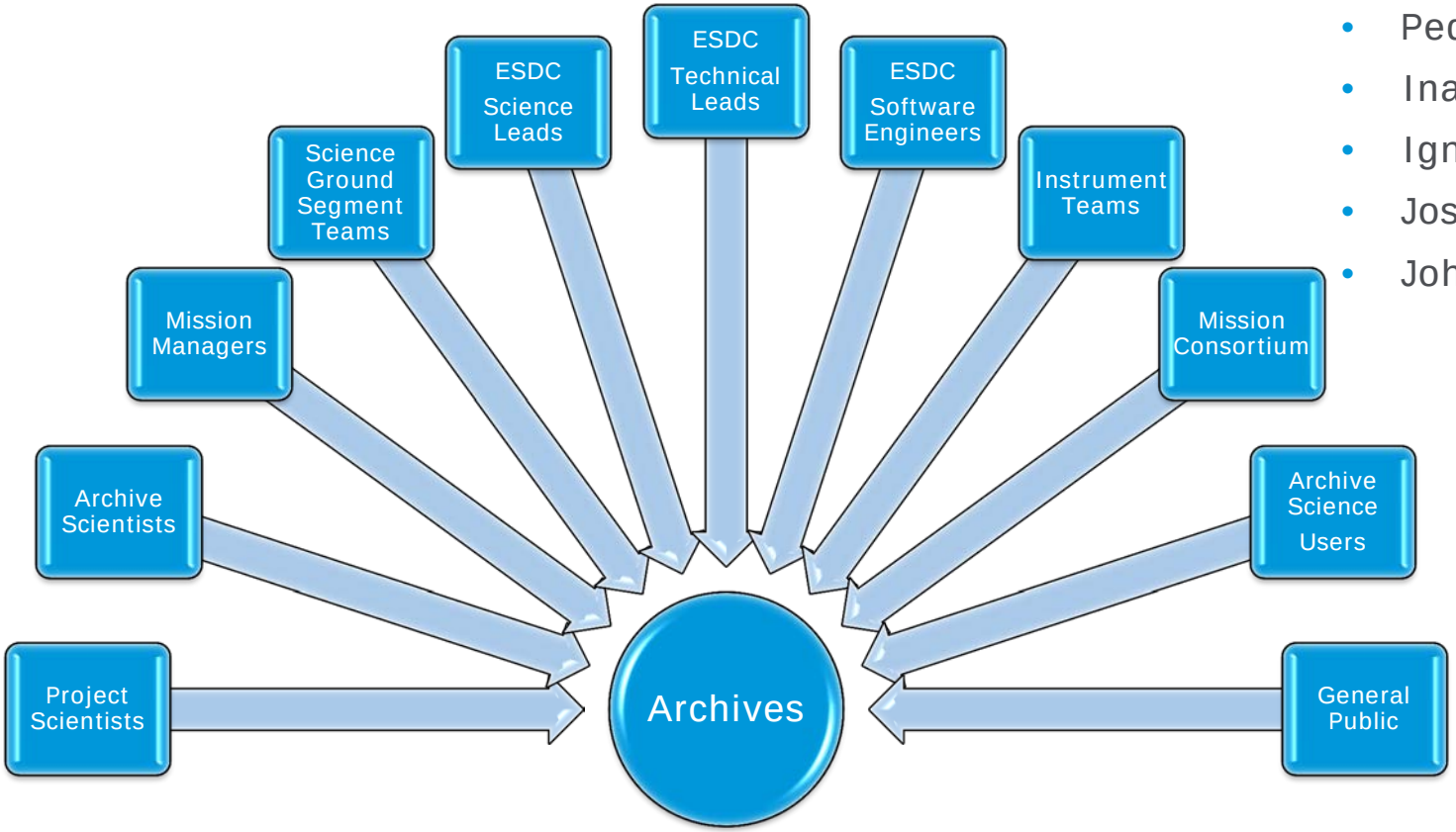
- Archives must be science driven
- Archives require strong IT expertise
- Regular technology evolution through time

Towards Archives 2.0

- Open data, open source, open archives
- Paradigm shift towards “bring your code to the data”

Acknowledgements to all Archive stakeholders

Many thanks to all the **ESAC Science Data Centre**



And to people who played a special role in the ESAC Archives lifetime

- Pedro Osuna
- Inaki Ortiz
- Ignacio Leon
- Jose Hernandez
- John Dowson