Digital Data in Astronomy

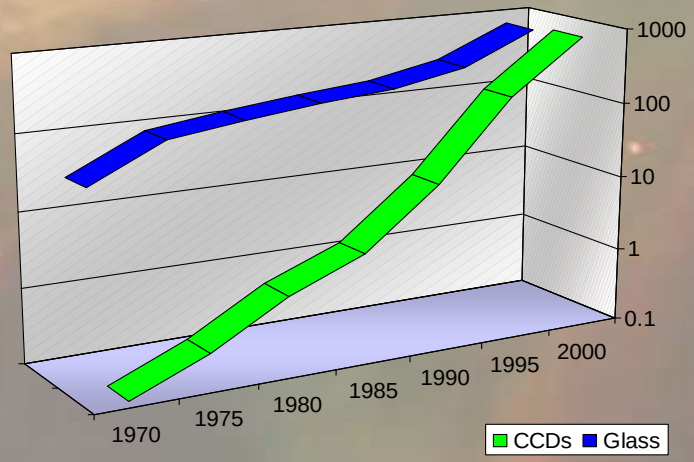
Robert Hanisch
US National Virtual Observatory
Space Telescope Science Institute
Baltimore, MD

Sayeed Choudhury, Alex Szalay, Tim DiLauro (JHU)
Carl Lagoze (Cornell)
Ethan Vishniac (McMaster)
Ray Plante (UIUC)

Data/Information in Astronomy

- Basic data
  - digital images, spectra, time series, catalogs, tables
- Simulations
  - models (results, computer codes, computational services)
  - virtual observations
- Analysis and interpretation
  - journals, e-preprints
  - reprocessed and enhanced data
- ~several PB worldwide, with 1-2 year doubling rate

not discoverable through text-based search engines



Astronomical Data and the Virtual Observatory

- The VO provides standard protocols for obtaining data from *distributed collections.*
  - Metadata about collections, stored in *registries*, based on Dublin Core, shared/coordinated through OAI-PMH
  - Metadata about data objects (catalogs, images, spectra, theoretical models and simulations)
  - Metadata about applications, libraries, computational resources
- The VO is NOT a centralized repository.
- The Virtual Observatory enables research by greatly enhancing access to data and computing resources. The VO makes it easy to locate, retrieve, and analyze data from archives and catalogs worldwide.

Without VO

n services,
n interfaces

archive 1

archive 2

archive 3

survey 1

survey 2

survey 3

service 3

service 2

service 1

With VO

n services,
“1” interface

archive 1

archive 2

archive 3

survey 1

survey 2

survey 3

VO

service 3

service 2

service 1

The Key to the VO: Interoperability

- Metadata standards
- Data discovery
- Data requests
- Data delivery
- Database queries
- Distributed applications; web services
- Distributed storage; replication
- Authentication and authorization

Data from NASA missions and most major observatories is archived and accessible, but we have a major gap…

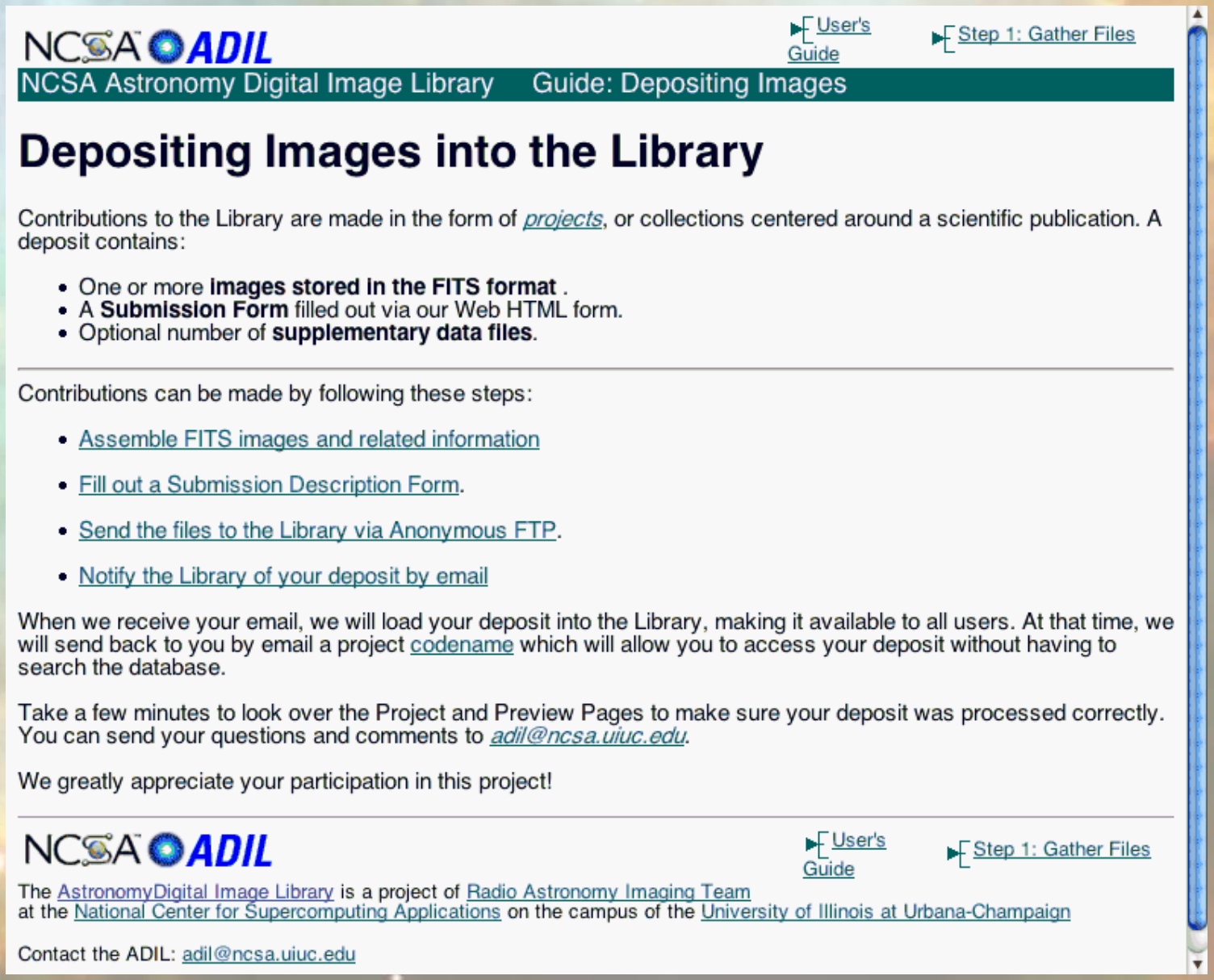
The Data Preservation Problem

- Research communities publish peer-reviewed journal papers that describe highly processed data.
- Text and graphics are now being preserved, but the digital data behind the graphics are “homeless”
- The research cannot be verified and the results cannot be easily compared to other data in order to broaden impact.
- Public funds invested in scientific research do not have maximum return on investment. Essential legacy datasets are being lost.

Approach

- Integrate digital data management into the publication process (data capture, review, metadata tagging and validation, storage).
- Exploit emerging information technology standards for managing distributed data collections, including digital journals.
- Provide multiple access methods to digital data to maximize visibility and re-use.
- Build on information management and curation experience in the university libraries and their institutional commitments for long-term preservation.

Astronomy Digital Image Library



NCSA ADIL

User's Guide | Step 1: Gather Files

NCSA Astronomy Digital Image Library Guide: Depositing Images

Depositing Images into the Library

Contributions to the Library are made in the form of *projects*, or collections centered around a scientific publication. A deposit contains:

- One or more **Images stored in the FITS format** .
- A **Submission Form** filled out via our Web HTML form.
- Optional number of **supplementary data files**.

Contributions can be made by following these steps:

- Assemble FITS images and related information
- Fill out a Submission Description Form.
- Send the files to the Library via Anonymous FTP.
- Notify the Library of your deposit by email

When we receive your email, we will load your deposit into the Library, making it available to all users. At that time, we will send back to you by email a project codename which will allow you to access your deposit without having to search the database.

Take a few minutes to look over the Project and Preview Pages to make sure your deposit was processed correctly. You can send your questions and comments to *adil@ncsa.uiuc.edu*.

We greatly appreciate your participation in this project!

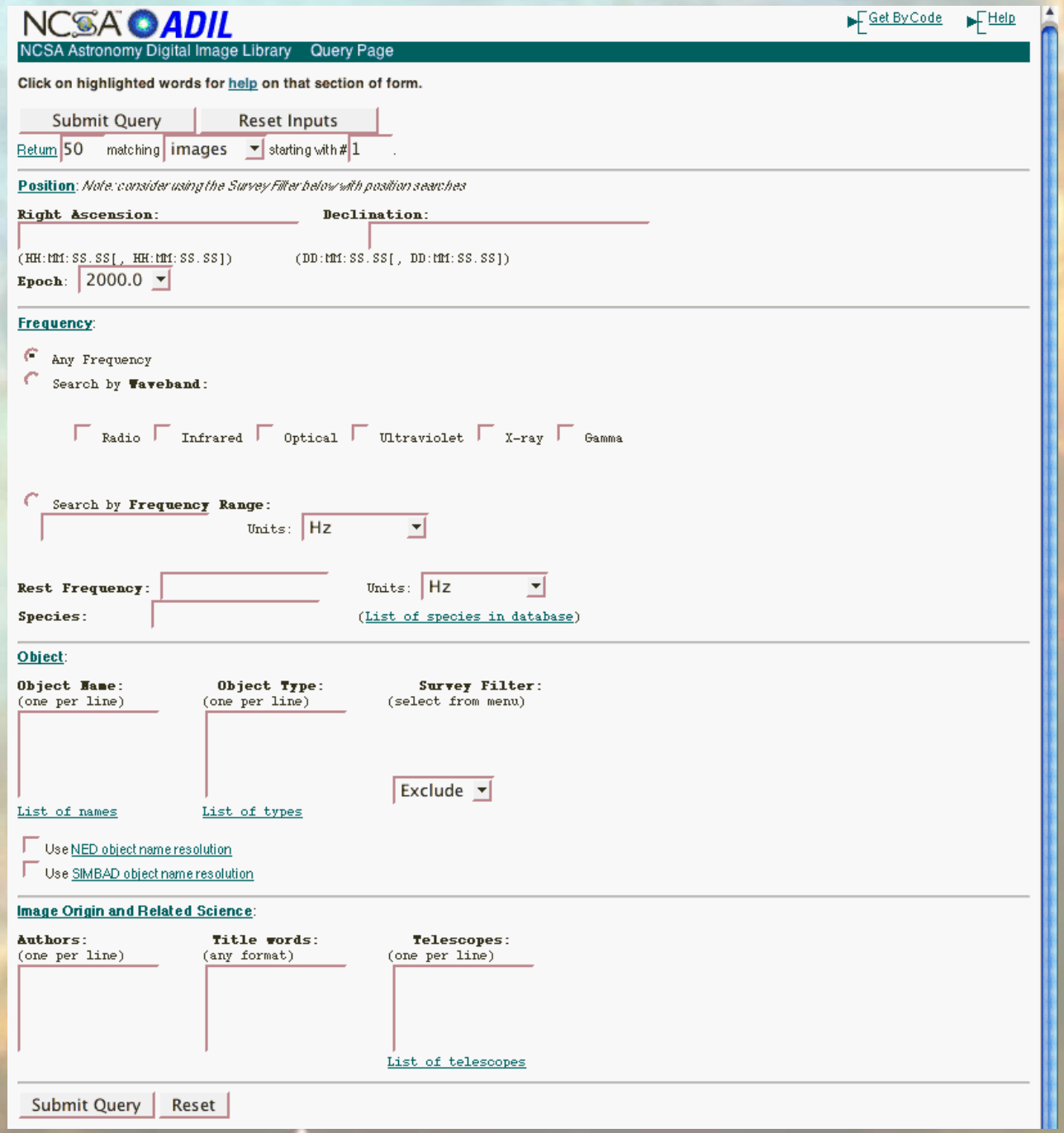
NCSA ADIL

User's Guide | Step 1: Gather Files

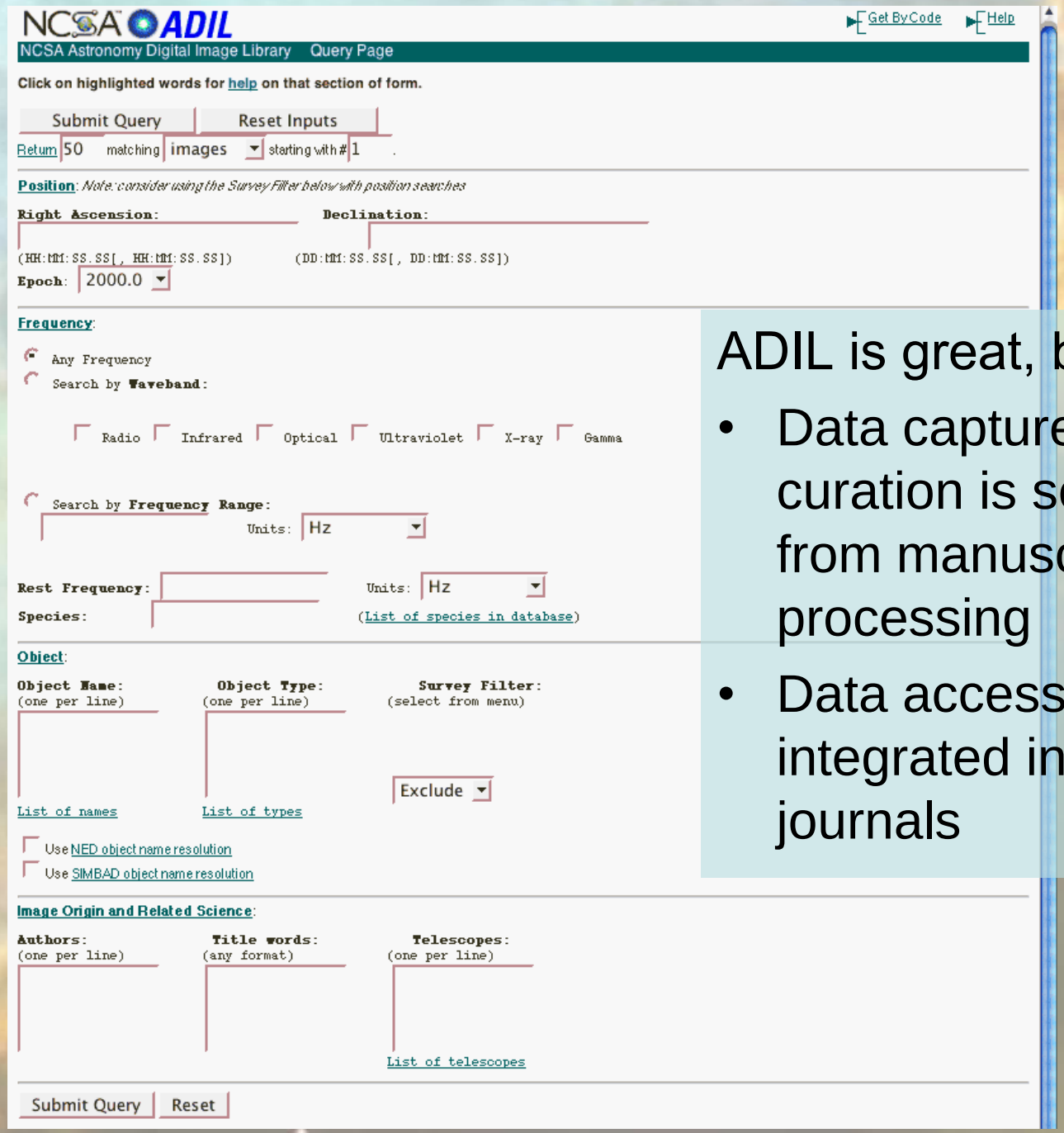
The AstronomyDigital Image Library is a project of Radio Astronomy Imaging Team
at the National Center for Supercomputing Applications on the campus of the University of Illinois at Urbana-Champaign

Contact the ADIL: adil@ncsa.uiuc.edu

ADIL query


NCSA ADIL
Get By Code
Help
NCSA Astronomy Digital Image Library Query Page
Click on highlighted words for help on that section of form.
Submit Query
Reset Inputs
Return 50 matching images starting with # 1 .
Position: Note: consider using the Survey Filter below with position searches
Right Ascension:
Declination:
(HH:MM:SS.SS[, HH:MM:SS.SS])
(DD:MM:SS.SS[, DD:MM:SS.SS])
Epoch: 2000.0
Frequency:
Any Frequency
Search by Waveband:
Radio
Infrared
Optical
Ultraviolet
X-ray
Gamma
Search by Frequency Range:
Units: Hz
Rest Frequency:
Units: Hz
Species:
(List of species in database)
Object:
Object Name:
(one per line)
Object Type:
(one per line)
Survey Filter:
(select from menu)
Exclude
List of names
List of types
Use NED object name resolution
Use SIMBAD object name resolution
Image Origin and Related Science:
Authors:
(one per line)
Title words:
(any format)
Telescopes:
(one per line)
List of telescopes
Submit Query
Reset

ADIL query



ADIL is great, but…

- Data capture and curation is separate from manuscript processing
- Data access is not integrated into the journals

Storyboard

The Astrophysical Journal, 644:759–768, 2006 June 20

Evolution of the Color-Magnitude Relation in High-Redshift Clusters: Early-Type Galaxies in the Lynx Supercluster at $z \sim 1.26$

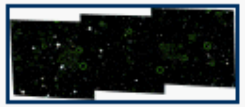
Simona Mei,[1] Brad P. Holden,[2] John P. Blakeslee,[1,3] Piero Rosati,[4] Marc Postman,[1,5] Myungkook J. Jee,[1] Alessandro Rettura,[4,6] Marco Sirianni,[5] Ricardo Demarco,[1] Holland C. Ford,[1] Marijn Franx,[7] Nicole Homeier,[1] and Garth D. Illingworth[2]

Received 2005 October 10; accepted 2006 February 24

ABSTRACT

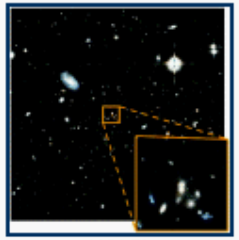
Color-magnitude relations (CMRs) have been derived in two high-redshift clusters, RX J0849+4452 and RX J0848+4453 (with redshifts of $z = 1.26$ and 1.27, respectively), that lie in the highest redshift cluster superstructure known today, the Lynx Supercluster. The CMR was determined from ACS imaging in the WFC F775W (i_{775}) and F850LP (z_{850}) filters combined with ground-based spectroscopy. Early-type cluster candidates have been identified according to the Postman et al.

Stanford et al. 2001). Recently, deep, panoramic multicolor (i' and z' bands) imaging around these two central clusters identified seven galaxy groups (Nakata et al. 2005) with photometric redshift $z_{phot} \sim 1.26$. This makes the Lynx region a unique laboratory, being the only supercluster observed at such a high redshift today, and for this reason, one of the best regions at $z > 1$ in which we can study properties of evolving galaxies within a structure that is still assembling, and in different environments.



(51 kB)

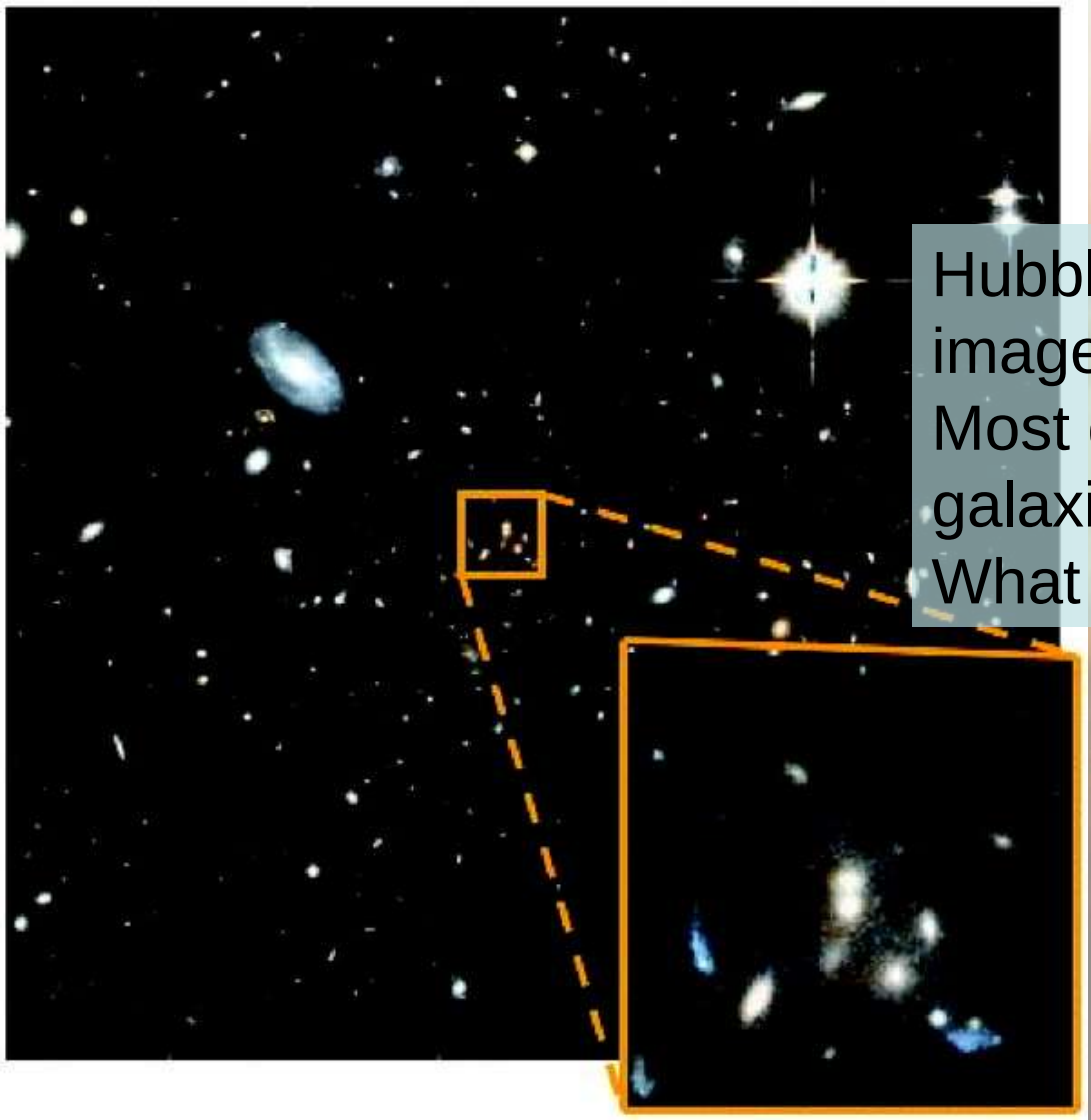
Fig. 1 *Chandra* X-ray contours overlaid on the ACS color composite image for Lynx E (*on the left*) and Lynx W (*on the right*). The contours are adaptively smoothed with a minimum significance of 3 σ. We refined the alignment of the *Chandra* image with respect to the ACS using the X-ray point sources.



(49 kB)

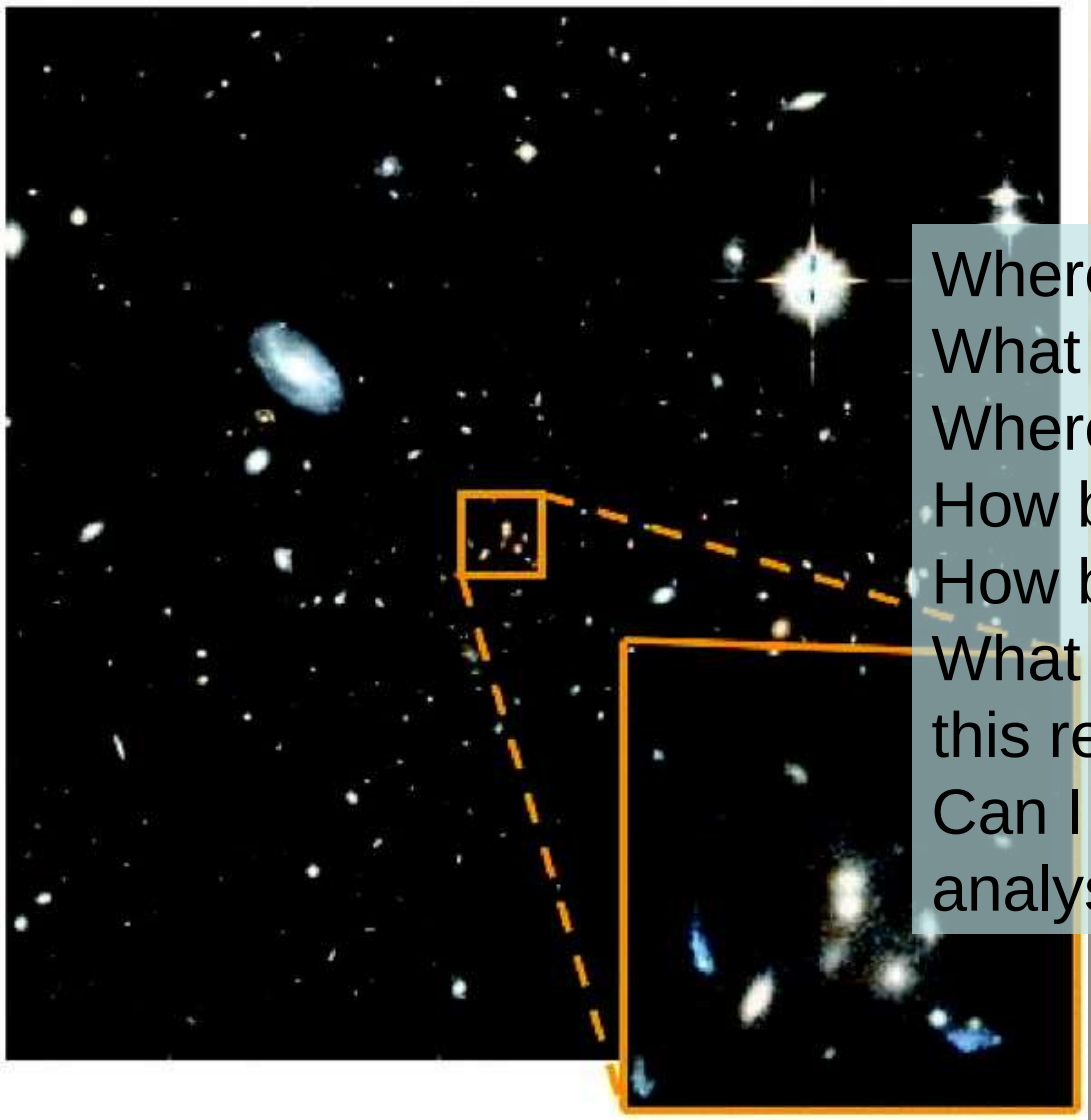
Fig. 2 Lynx E ACS image (scale is $1' \times 1'$). The central ongoing merger is magnified to also show a gravitational arc and its likely counterimage.

Storyboard



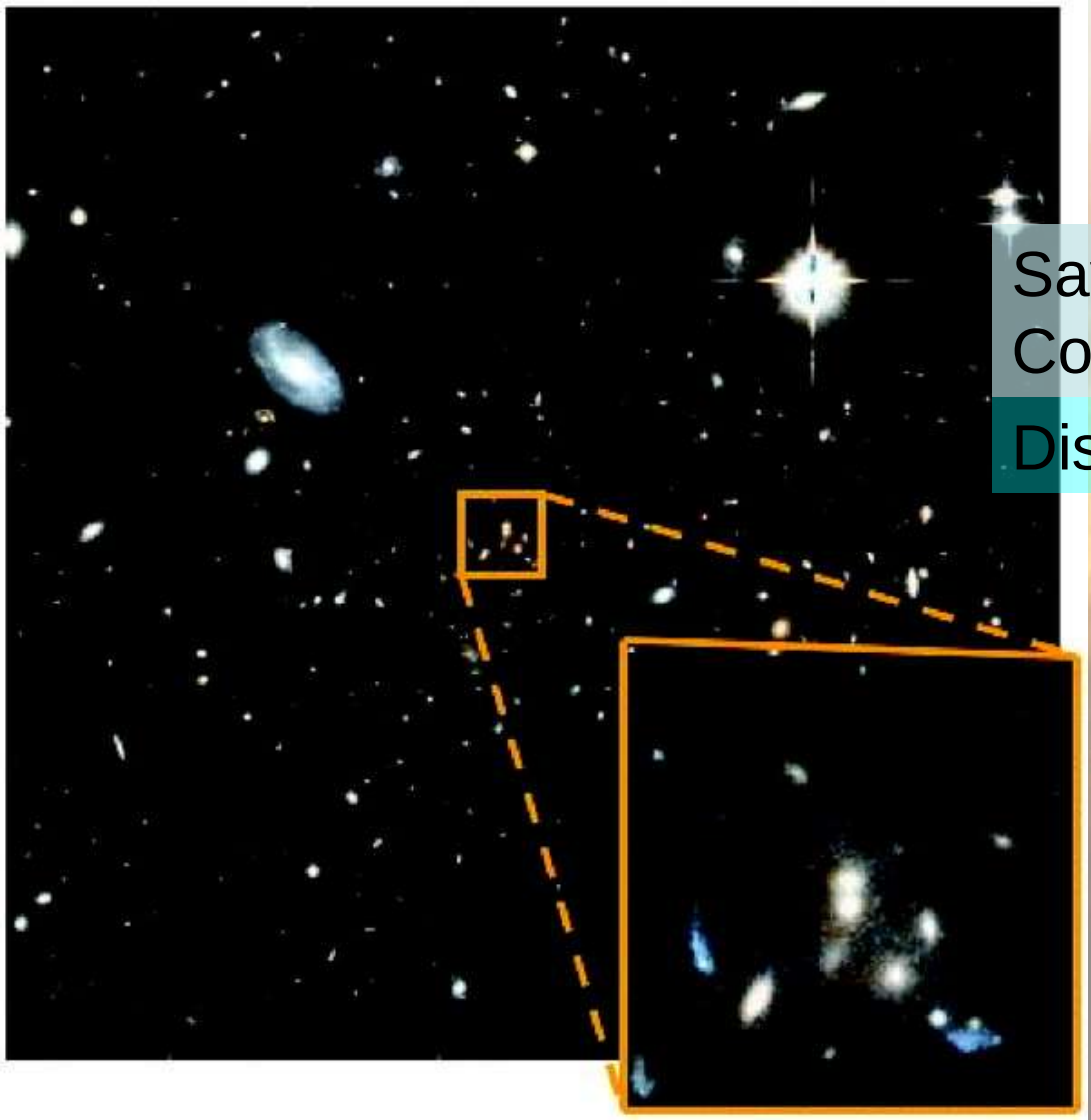
Hubble Space Telescope image.
Most distant cluster of galaxies known.
What more can I find out?

Storyboard



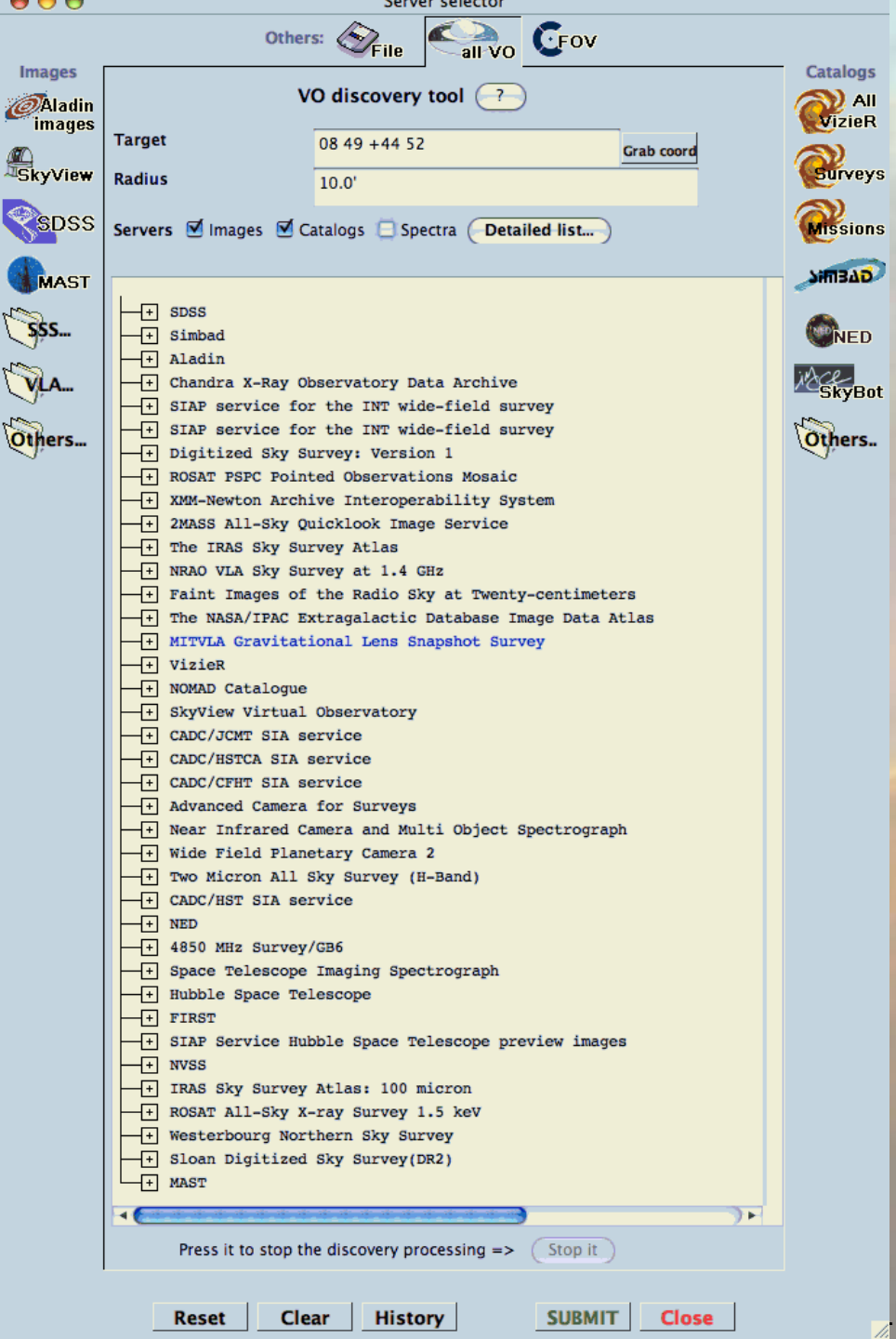
Where is this?
What is the image scale?
Where is north?
How bright is the star?
How bright is the galaxy?
What else is known about this region?
Can I trust the data analysis in this paper?

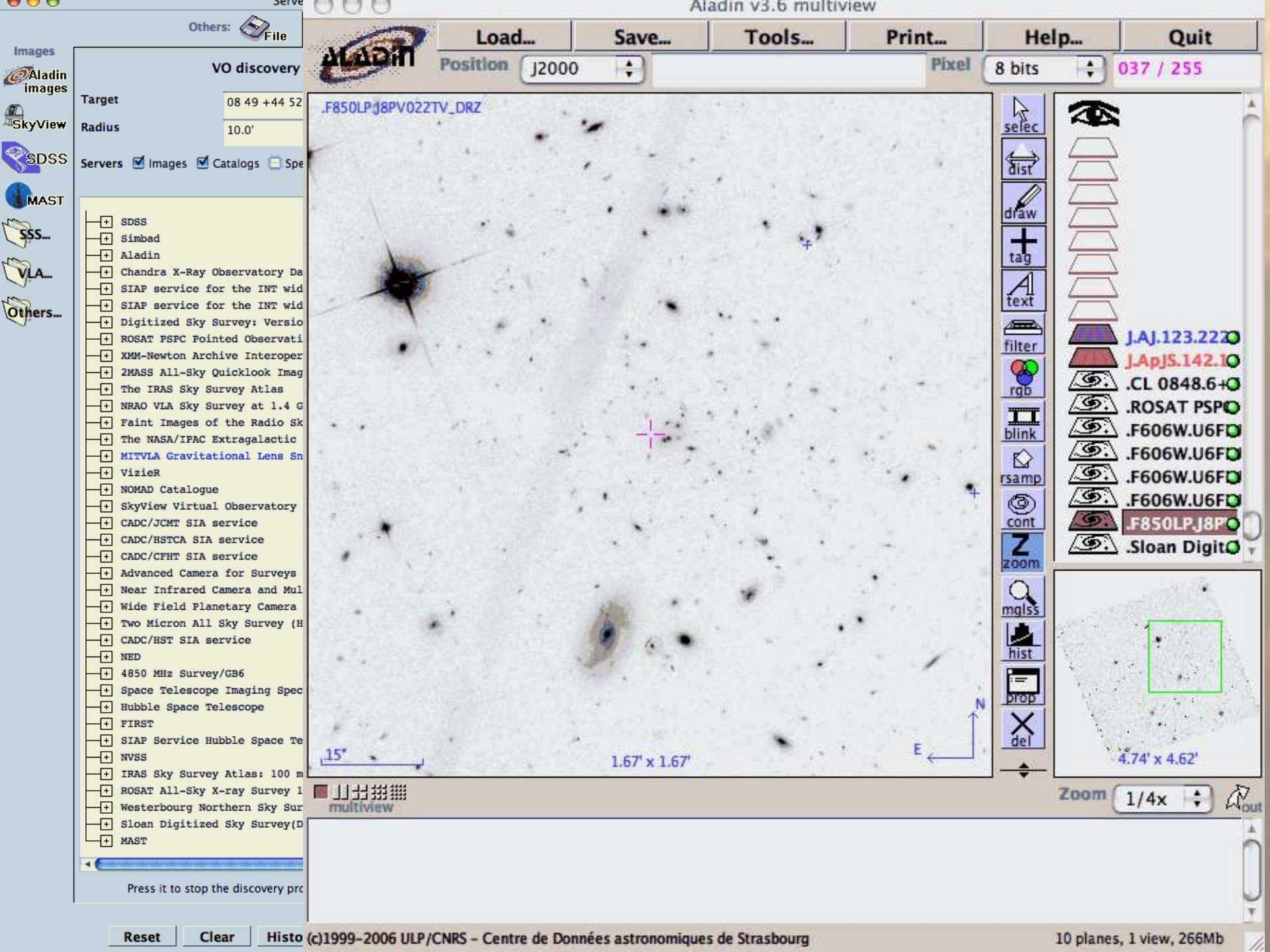
Storyboard



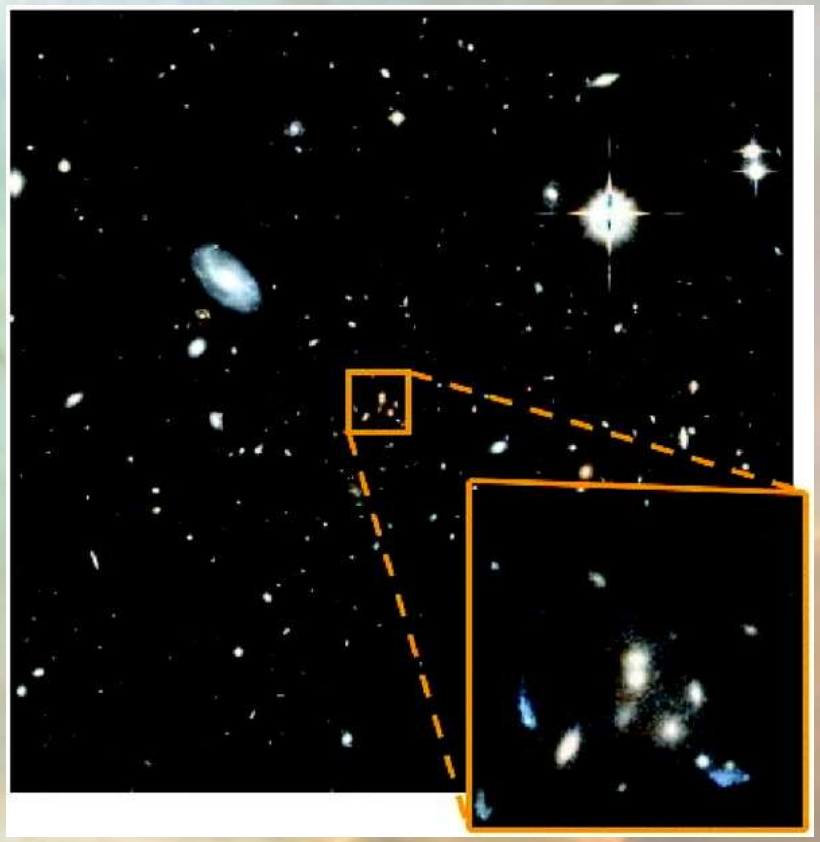
Save file
Copy to my VOSpace

Display and compare

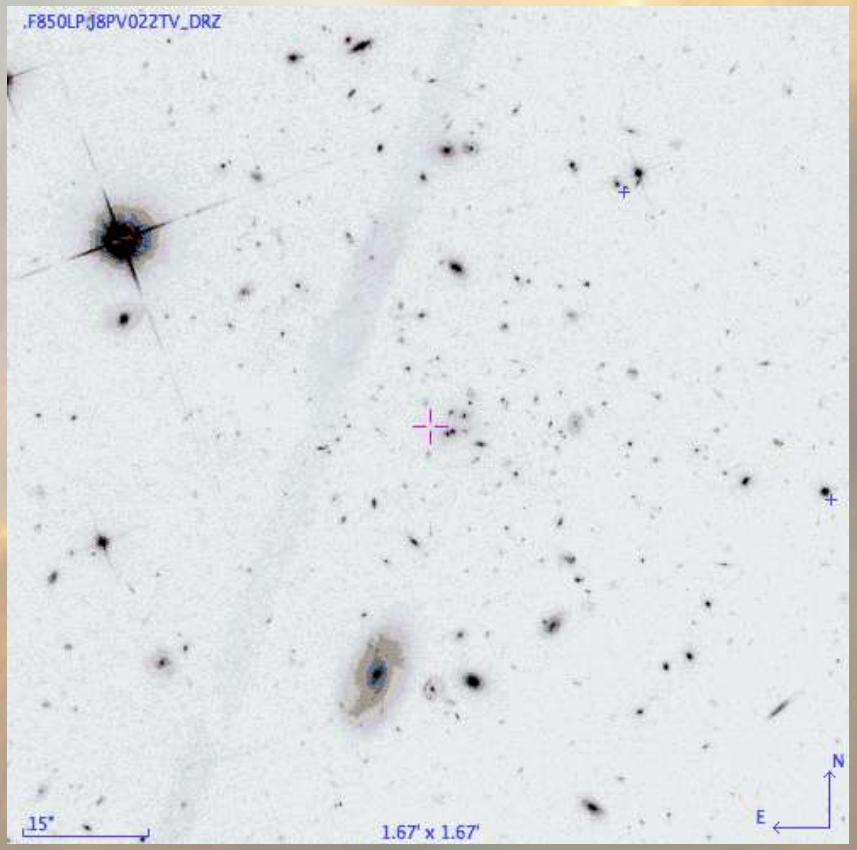

Server selector
Others:
File
all VO
FOV
Images
Aladin images
SkyView
SDSS
MAST
SSS...
VLA...
Others...
Catalogs
All VizieR
Surveys
Missions
SIMBAD
NED
SkyBot
Others...
VO discovery tool
?
Target
08 49 +44 52
Grab coord
Radius
10.0'
Servers
Images
Catalogs
Spectra
Detailed list...
SDSS
Simbad
Aladin
Chandra X-Ray Observatory Data Archive
SIAP service for the INT wide-field survey
SIAP service for the INT wide-field survey
Digitized Sky Survey: Version 1
ROSAT PSPC Pointed Observations Mosaic
XMM-Newton Archive Interoperability System
2MASS All-Sky Quicklook Image Service
The IRAS Sky Survey Atlas
NRAO VLA Sky Survey at 1.4 GHz
Faint Images of the Radio Sky at Twenty-centimeters
The NASA/IPAC Extragalactic Database Image Data Atlas
MITVLA Gravitational Lens Snapshot Survey
VizieR
NOMAD Catalogue
SkyView Virtual Observatory
CADC/JCMT SIA service
CADC/HSTCA SIA service
CADC/CFHT SIA service
Advanced Camera for Surveys
Near Infrared Camera and Multi Object Spectrograph
Wide Field Planetary Camera 2
Two Micron All Sky Survey (H-Band)
CADC/HST SIA service
NED
4850 MHz Survey/GB6
Space Telescope Imaging Spectrograph
Hubble Space Telescope
FIRST
SIAP Service Hubble Space Telescope preview images
NVSS
IRAS Sky Survey Atlas: 100 micron
ROSAT All-Sky X-ray Survey 1.5 keV
Westerbourg Northern Sky Survey
Sloan Digitized Sky Survey(DR2)
MAST
Press it to stop the discovery processing =>
Stop it
Reset
Clear
History
SUBMIT
Close

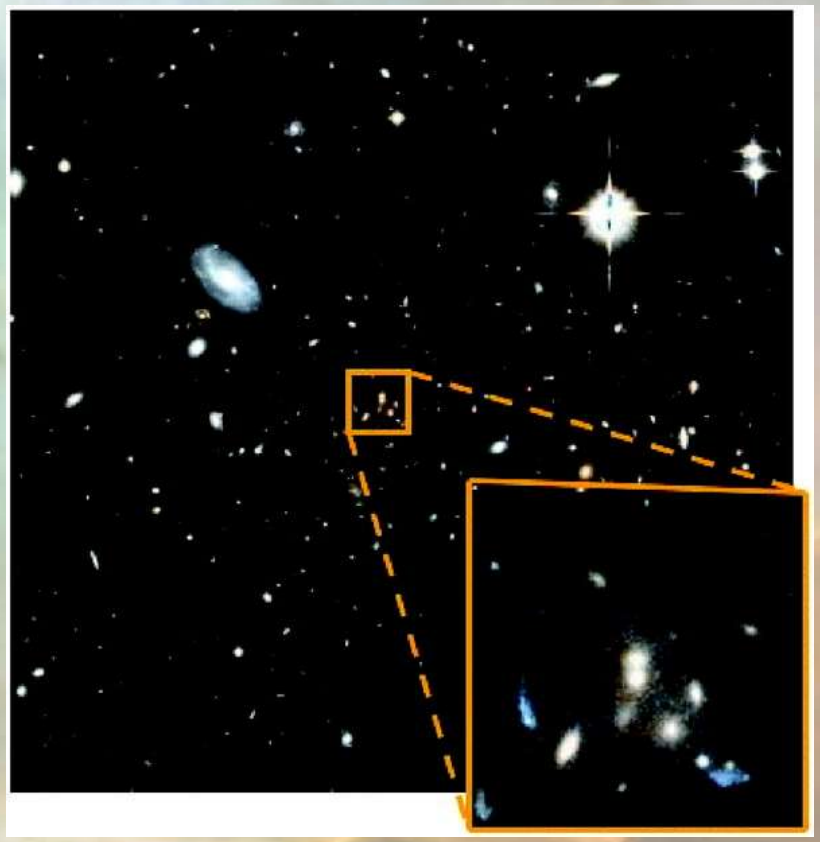

Aladin v3.6 multiview
Load...
Save...
Tools...
Print...
Help...
Quit
Position
J2000
Pixel
8 bits
037 / 255
Others:
File
Images
Aladin images
SkyView
SDSS
MAST
SSS...
VLA...
Others...
VO discovery
Target
08 49 +44 52
Radius
10.0'
Servers
Images
Catalogs
SDSS
Simbad
Aladin
Chandra X-Ray Observatory Da
SIAP service for the INT wid
SIAP service for the INT wid
Digitized Sky Survey: Versio
ROSAT PSPC Pointed Observati
XMM-Newton Archive Interoper
2MASS All-Sky Quicklook Imag
The IRAS Sky Survey Atlas
NRAO VLA Sky Survey at 1.4 G
Faint Images of the Radio Sk
The NASA/IPAC Extragalactic
MITVLA Gravitational Lens Sn
VizieR
NOMAD Catalogue
SkyView Virtual Observatory
CADC/JCMT SIA service
CADC/HSTCA SIA service
CADC/CFHT SIA service
Advanced Camera for Surveys
Near Infrared Camera and Mul
Wide Field Planetary Camera
Two Micron All Sky Survey (H
CADC/HST SIA service
NED
4850 MHz Survey/GB6
Space Telescope Imaging Spec
Hubble Space Telescope
FIRST
SIAP Service Hubble Space Te
NVSS
IRAS Sky Survey Atlas: 100 m
ROSAT All-Sky X-ray Survey 1
Westerbourg Northern Sky Sur
Sloan Digitized Sky Survey(D
MAST
Press it to stop the discovery prc
Reset
Clear
Histo
.F850LPJ8PV022TV_DRZ
15"
1.67' x 1.67'
N
E
selec
dist
draw
tag
text
filter
rgb
blink
rsamp
cont
zoom
mgISS
hist
prop
del
J.AJ.123.222
J.ApJS.142.1
.CL 0848.6+
.ROSAT PSP
.F606W.U6F
.F606W.U6F
.F606W.U6F
.F606W.U6F
.F850LPJ8P
.Sloan Digit
4.74' x 4.62'
multiview
Zoom
1/4x
out
(c)1999-2006 ULP/CNRS - Centre de Données astronomiques de Strasbourg
10 planes, 1 view, 266Mb

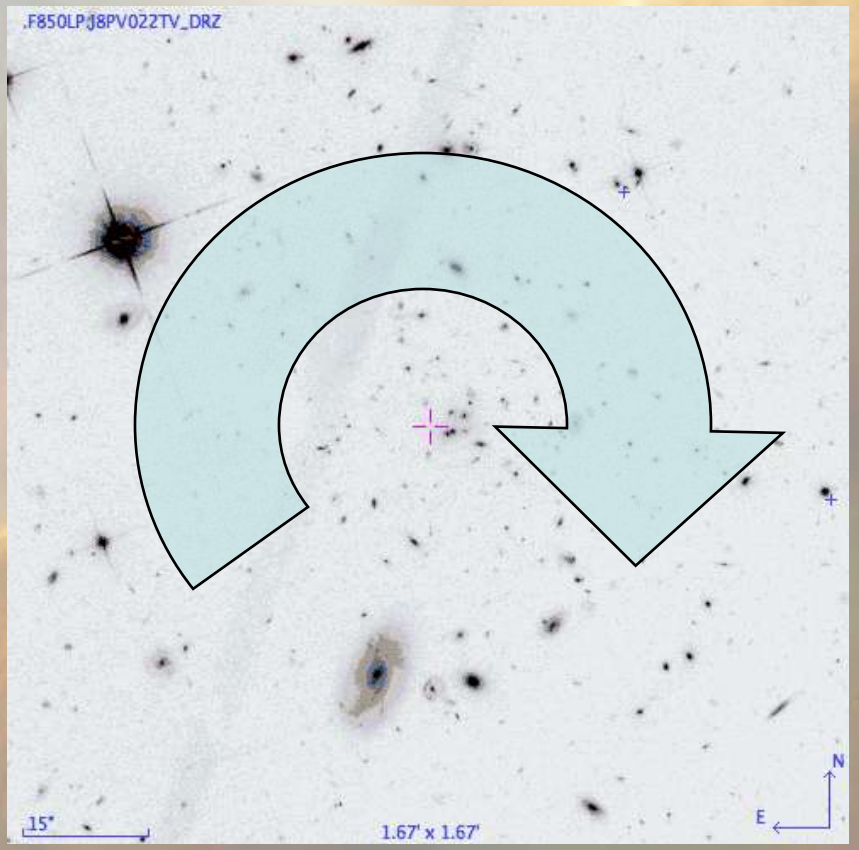
Journal…

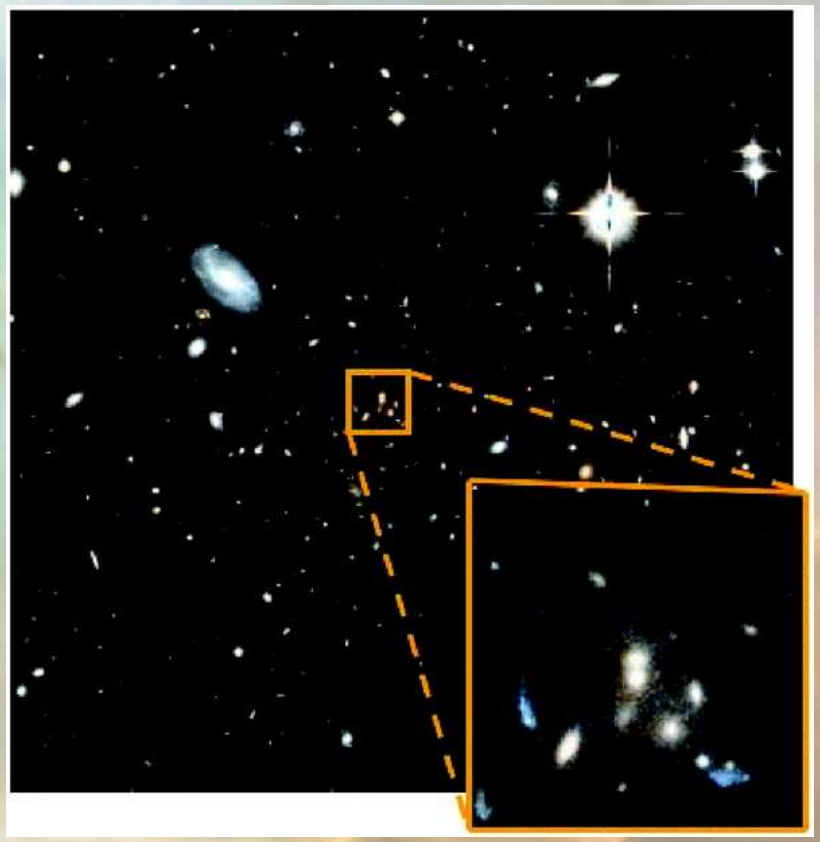


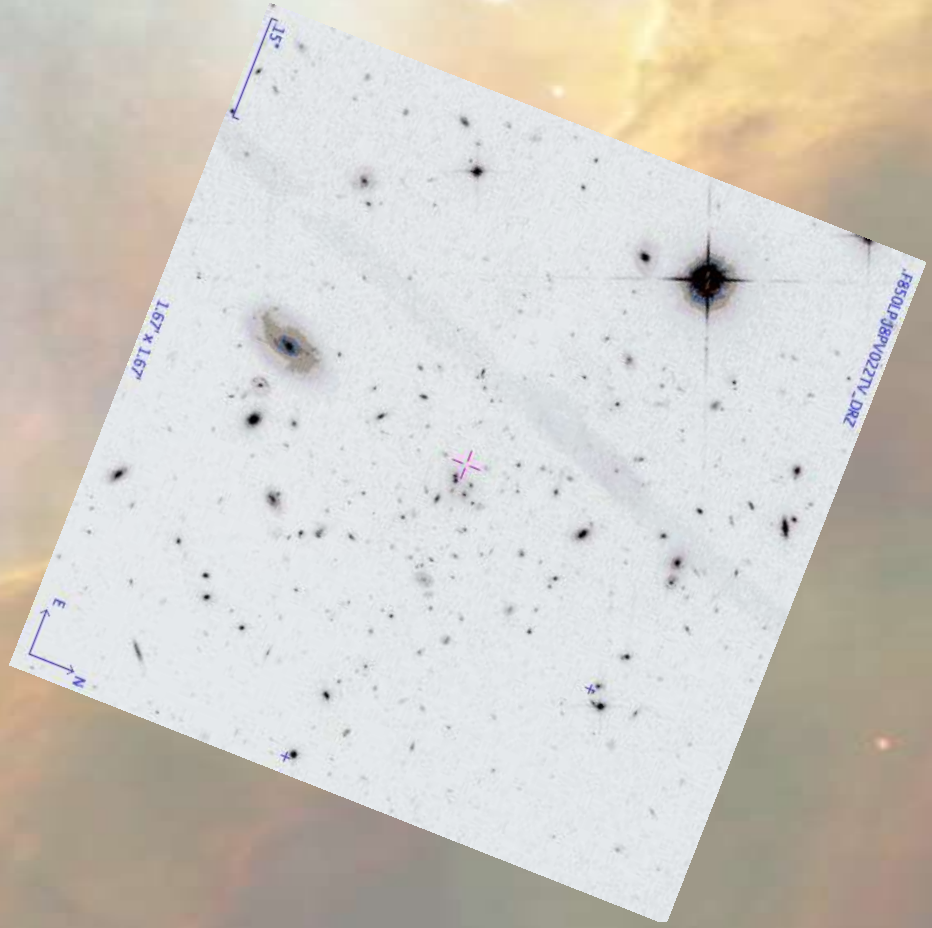
Archive…

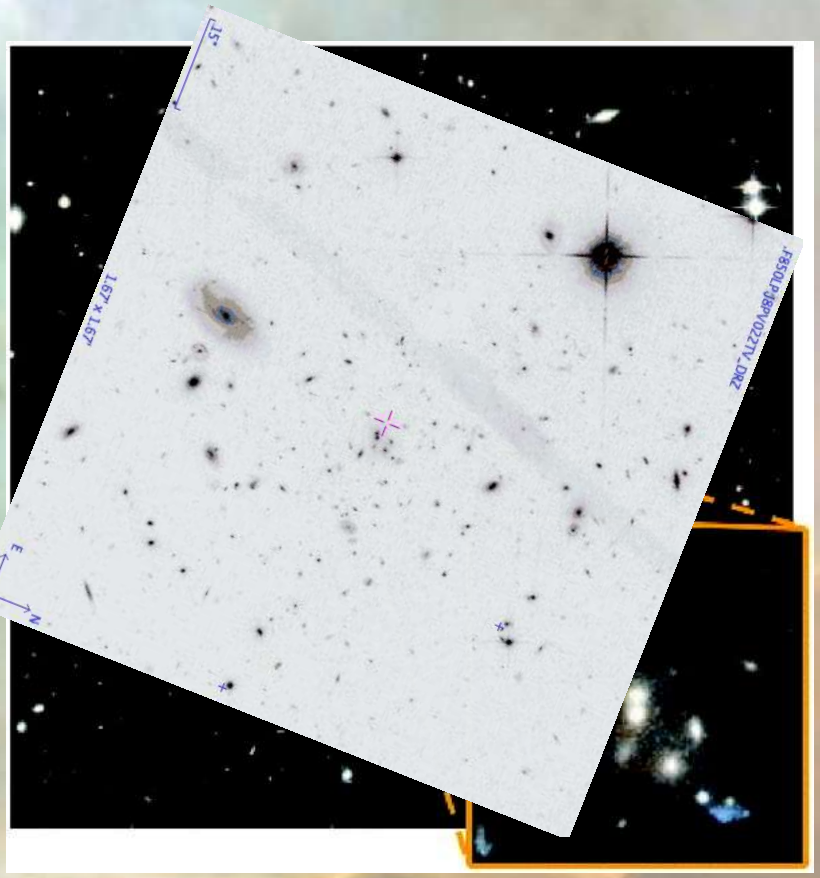




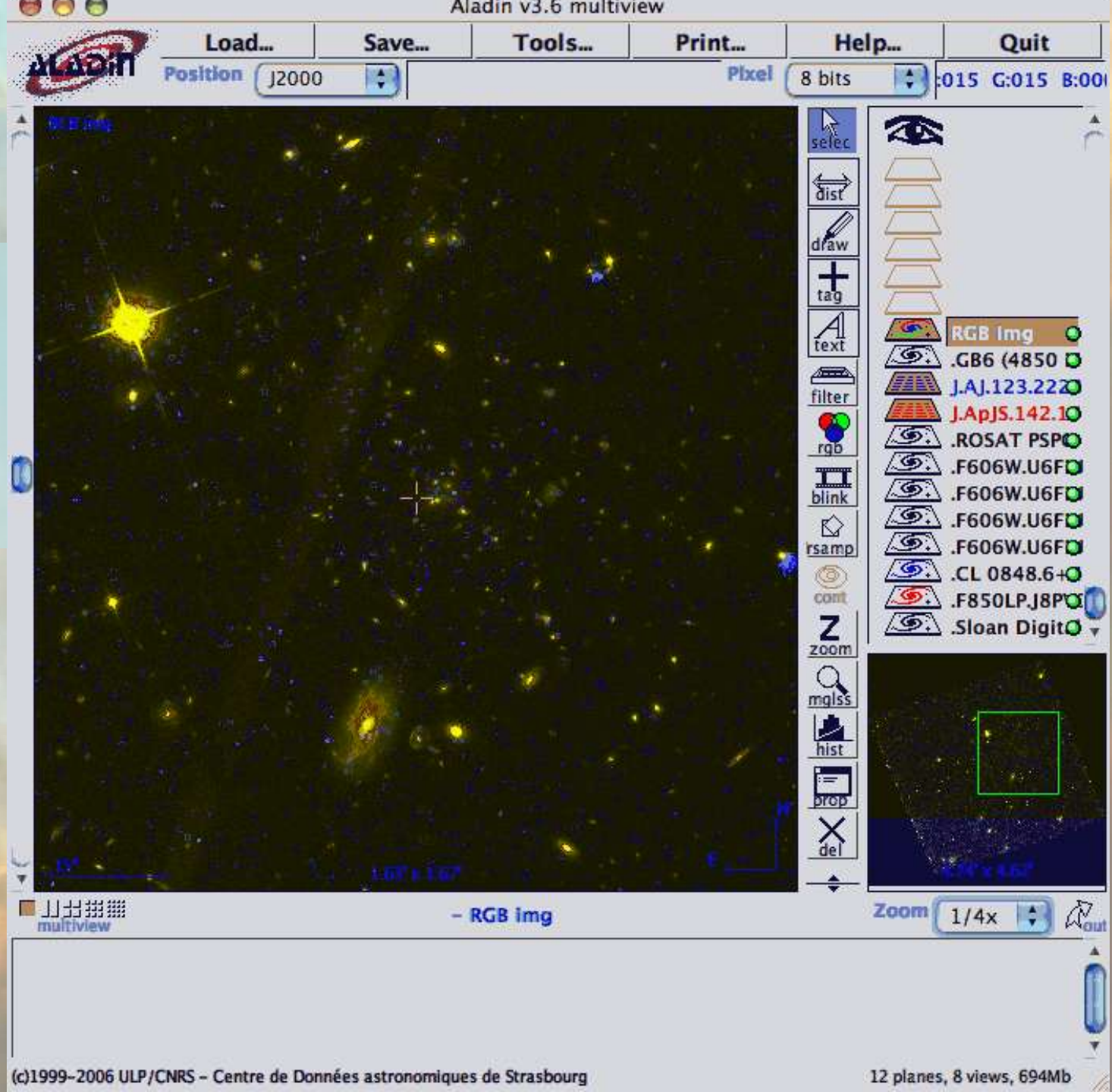

F850LP_J8PV022TV_DRZ
15"
1.67' x 1.67'
N
E



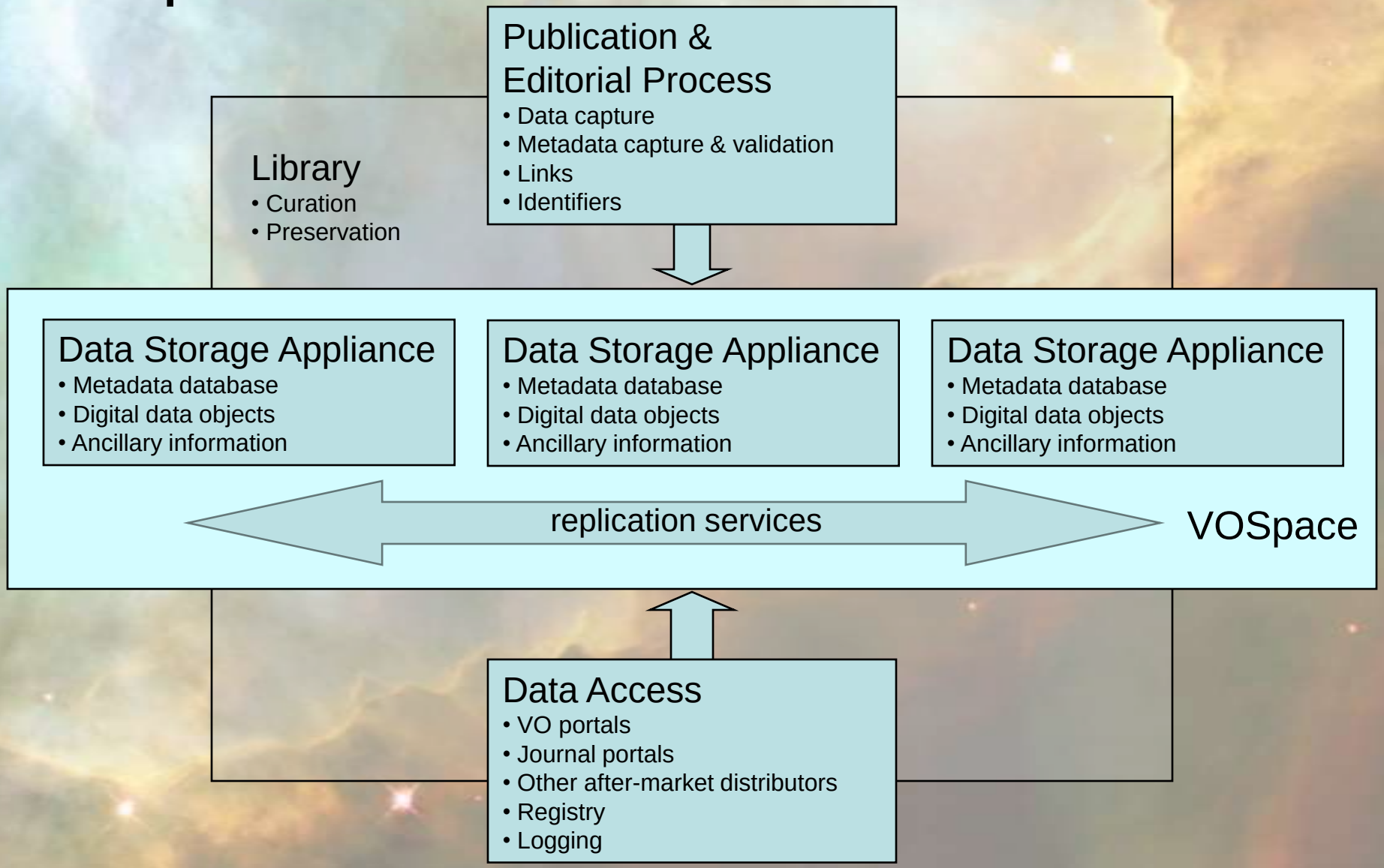

15'
1.67' x 1.67'
N
E



Is there any X-ray emission from this cluster of galaxies?



Components


Publication &
Editorial Process
• Data capture
• Metadata capture & validation
• Links
• Identifiers
Library
• Curation
• Preservation
Data Storage Appliance
• Metadata database
• Digital data objects
• Ancillary information
Data Storage Appliance
• Metadata database
• Digital data objects
• Ancillary information
Data Storage Appliance
• Metadata database
• Digital data objects
• Ancillary information
replication services
VOSpace
Data Access
• VO portals
• Journal portals
• Other after-market distributors
• Registry
• Logging

A prototype project

- Implement end-to-end prototype using astronomy scholarly publications as a test-bed.
- Understand operational costs and develop long-term business plan for preservation of peer-reviewed journal content and associated supporting data.
- Develop associated policies affecting data accessibility (e.g., move toward requiring digital data availability as requirement for publication).
- Utilize commodity open-source technologies and partner with Virtual Observatory to maximize return on investment, flexibility, adaptability.
- Long-term: evaluate impact on citations and productivity resulting from having ready access to digital data.

A prototype project

- Tasks (partners)
  - Metadata definition (VO, library)
  - Content management tool evaluation/selection (Fedora) (VO, library)
  - Physical storage and replication (VO, library, publisher)
  - Publication process revisions and testing (publisher, editorial staff)
  - Policy development (editorial staff, professional society)
  - Business model development (publisher, professional society)

A prototype project

- Initial work underway
  - IMLS grant
  - Microsoft grant
  - NVO collaboration
- Hoping to do future work in context of NSF Datanet program
  - Pending support for Data Conservancy project led by S. Choudhury (JHU)

Impact of digital data preservation

- Data re-use
- Increasing the discovery space
- Optimization of public investment in science
- Creation of a research legacy
- Integrity in scientific publication

Good metadata

Title?	Galaxy Evolution Explorer		
Harvested From			
Shortname?	GALEX	Identifier?	ivo://archive.stsci.edu/siap/galex
ContactName?	Alberto Conti	ContactEmail?	archive@stsci.edu
Creator?	Chris Martin	Publisher?	Space Telescope Science Institute
Contributor?	California Institute of Technology	Subject?	Cosmology, Galaxies, Star Formation
ResourceType	SIAP		
Description?	The Galaxy Evolution Explorer (GALEX), a NASA Small Explorer mission, is performing the first all-sky, deep imaging and spectroscopic ultraviolet surveys in space. The prime goal of GALEX is to study star formation in galaxies and its evolution with time.		
Related Resources			
Type?	Archive	Instrument?	Far and Near Ultraviolet microchannel plates, grism spectrograph
Date?	11/22/2005 12:22:51 PM	Version?	V1.0
ReferenceURL?	http://galex.stsci.edu	ServiceURL?	http://galex.stsci.edu/gxWS/SIAP/gxSIAP.aspx?
CoverageSpatial?	circle (FK5, 0.0, 0.0, 180.0)	CoverageTemporal?	2003-04 to present
RegionOfRegard?	0	CoverageSpectral?	Ultraviolet
ContentLevel?	Research	Facility?	GALEX
ModificationDate	3/11/2006 6:43:25 PM	ValidationLevel	2
Interfaces			
Footprint	circle J2000 0.0 0.0 180.0		

Simple Image Access

Format					
VOTableColumns					
ImageServiceType		MaxqueryRegionSizeLat	0	MaxqueryRegionSizeLong	0
MaxRecords	0	MaxImageExtentLat	0	MaxImageExtentLong	0
MaxFileSize	0	MaxImageSizeLat	0	MaxImageSizeLong	0

Pretty good metadata

Title?	Markarian Galaxies		
Harvested From	http://heasarc.gsfc.nasa.gov/cgi-bin/OAI/XMLFile/nvo/oai.pl?verb=ListRecords&metadataPrefix=ivo_vor&from=2000-01-01T		
Shortname?	Markarian	Identifier?	ivo://nasa.heasarc/markarian
ContactName?	Michael Preciado	ContactEmail?	preciado@milkyway.gsfc.n
Creator?	Markarian, Lipovetskii, Stepanian	Publisher?	NASA/GSFC HEASARC
Contributor?	NOT PROVIDED	Subject?	Galaxy
ResourceType	CONE		
Description?	This catalog contains machine-readable version of the lists "Galaxies with Ultraviolet Continuum" (MKN; Markarian 1967, 1969, 1969, Markarian and Lipovetskii 1971, 1972, 1973, 1974, 1976a, 1976b, Markarian et al., 1977a, 1977b, 1979a, 1979b, 1979c, 1982).		
Related Resources	service-for NASA/GSFC Exploration of the Universe Division ivo://nasa.heasarc/eud		
Type?	Catalog	Instrument?	NOT PROVIDED
Date?	5/25/2007 12:00:00 AM	Version?	NOT PROVIDED
ReferenceURL?	http://adsabs.harvard.edu/cgi-bin/bib_query?1967Afz.....3...55M	ServiceURL?	http://heasarc.gsfc.nasa.gov/cgi-bin/vo/cone/coneGet.pl?table=markarian&
CoverageSpatial?	All-sky<regionOfRegard xmlns="http://www.ivoa.net/xml/VODataSer	CoverageTemporal?	<startTime xmlns="http://www.ivoa.net/xml/VODataSer
RegionOfRegard?	0.25	CoverageSpectral?	Optical
ContentLevel?	Research	Facility?	NOT PROVIDED
ModificationDate	5/25/2007 3:00:20 AM	ValidationLevel	2

Interfaces	Number	Type	QType	ResultType	AccessURL
	Params	0	PARAMHTTP	GET	text/xml	http://heasarc.gsfc.nasa.gov/cgi-bin/vo/cone/coneGet.pl?table=markarian&

Footprint			
ConeSearch			
VOTableColumns			
MaxSR	180	MaxRecords	99999

Poor metadata

Title?	SIAP service for the INT wide-field survey		
Harvested From	http://hydra.star.le.ac.uk:8080/astrogrid-registry/OAIHandlerv0_10?verb=ListRecords&metadataPrefix=ivo_vor&from=2005-0		
Shortname?	NOT PROVIDED	Identifier?	ivo://org.astrogrid/INT-WFS.SIAP
ContactName?	Guy Rixon	ContactEmail?	gtr@ast.cam.ac.uk
Creator?	Guy Rixon	Publisher?	Cambridge Astronomical Survey Unit
Contributor?	NOT PROVIDED	Subject?	SIAP
ResourceType	SIAP		
Description?	A Simple Image Access Protocol service for the reduced images of the Isaac Newton Telescope wide-field survey.		
Related Resources			
Type?	Archive	Instrument?	NOT PROVIDED
Date?	4/8/2005 9:35:19 AM	Version?	NOT PROVIDED
ReferenceURL?	http://wiki.astrogrid.org/bin/view/Astrogrid/IntWfsDataModel	ServiceURL?	http://archive.ast.cam.ac.uk/cgi-bin/wfs-siap-atlas/queryImage
CoverageSpatial?	NOT PROVIDED Could not find a part of the path "c:\Inetpub\nvo\voregistry\region.xsl".	CoverageTemporal?	NOT PROVIDED
RegionOfRegard?	0	CoverageSpectral?	
ContentLevel?		Facility?	NOT PROVIDED
ModificationDate	4/8/2005 10:35:48 AM	ValidationLevel	2

Interfaces	Number	Type	QType	ResultType	AccessURL
	Params	0	WEBBROWSER		http://archive.ast.cam.ac.uk/cgi-bin/wfs-siap-atlas/queryImage

Footprint	

Simple Image Access

Format					
VOTableColumns					
ImageServiceType		MaxqueryRegionSizeLat	0	MaxqueryRegionSizeLong	0
MaxRecords	5000	MaxImageExtentLat	0	MaxImageExtentLong	0
MaxFileSize	0	MaxImageSizeLat	0	MaxImageSizeLong	0

Poor metadata

Title?	SIAP service for the INT wide-field survey		
Harvested From	http://hydra.star.le.ac.uk:8080/astrogrid-registry/OAIHandlerv0_10?verb=ListRecords&metadataPrefix=ivo_vor&from=2005-0		
Shortname?	NOT PROVIDED	Identifier?	ivo://org.astrogrid/INT-WFS.SIAP
ContactName?	Guy Rixon	ContactEmail?	gtr@ast.cam.ac.uk
Creator?	Guy Rixon	Publisher?	Cambridge Astronomical Survey Unit
Contributor?	NOT PROVIDED	Subject?	SIAP
ResourceType	SIAP		
Description?	A Simple I Telescope		
Related Resources			
Type?	Archive		
Date?	4/8/2005		
ReferenceURL?	http://wiki.a		/queryImage
CoverageSpatial?	NOT PROVID the path "c:\Inetpu		
RegionOfRegard?	0		
ContentLevel?			
ModificationDate	4/8/2005 10:35:48 AM	ValidationLevel	2

Currently have more than 10,000 resources like this needing metadata curation and service validation

Interfaces	Number	Type	QType	ResultType	AccessURL
	Params	0	WEBBROWSER		http://archive.ast.cam.ac.uk/cgi-bin/wfs-siap-atlas/queryImage

Footprint	

Simple Image Access

Format					
VOTableColumns					
ImageServiceType		MaxqueryRegionSizeLat	0	MaxqueryRegionSizeLong	0
MaxRecords	5000	MaxImageExtentLat	0	MaxImageExtentLong	0
MaxFileSize	0	MaxImageSizeLat	0	MaxImageSizeLong	0

Metadata granularity

- Desire to capture “fine-grained” metadata
  - Rich data discovery
  - Automated workflow
  - Automated serialization, class construction
- But, is this feasible?
  - Astronomical databases can contain dozens of tables and thousands of attributes
  - They *should* be documented, but often are not
  - Dynamic access vs. cache

Implications, challenges

- Fielded metadata is important for data discovery and access in disciplines whose content is not text-based
- Curation effort is substantial, requires domain expertise, and is ongoing
- Level of detail required is not yet clear
- Are ontologies necessary?
- New role: *data scientist*
  - IS + domain expertise
  - Supports data center and library