Digital Data Preservation and Curation: A Collaboration Among Libraries, Publishers, and the Virtual Observatory

A pilot project aimed at preserving, curating, and enabling access to digital data and associated electronic journals content.

Teresa Ehling, Cornell University

Robert Hanisch, Space Telescope Science Institute

Julie Steffen, University of Chicago Press

Sayeed Choudhury, Tim DiLauro, Alex Szalay, and Ethan Vishniac,
The Johns Hopkins University

Robert Milkey, American Astronomical Society

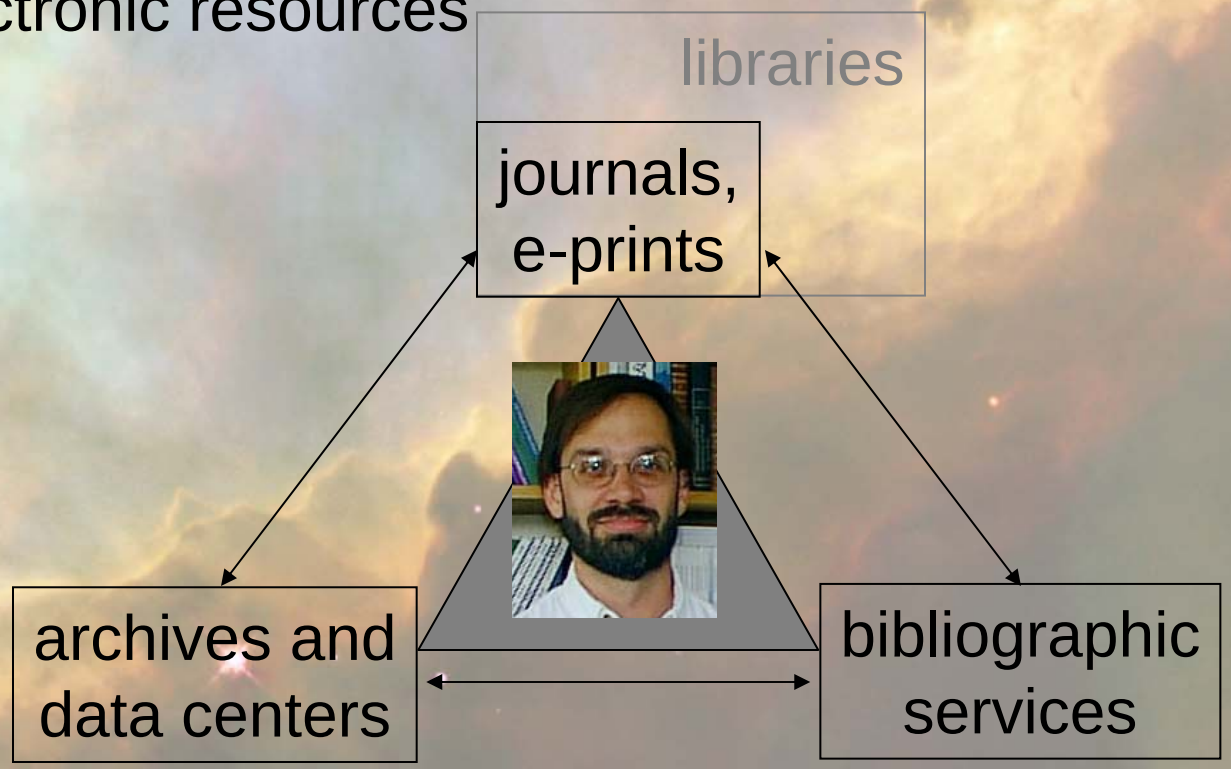
Ray Plante, National Center for Supercomputer Applications

Electronic information in astronomy

- Astronomy was one of the first scientific disciplines to pioneer e-publishing (ApJLett 1995, ApJ and AJ 1996)
- Astronomy has comprehensive e-abstract and bibliographic services
  - Astrophysics Data System, SIMBAD, NED
- Astronomy makes extensive use e-preprints on arXiv.org
- Astronomy data is archived and is generally publicly accessible
  - NASA mission archives
  - ground-based observatories (U.S., Europe, Australia, etc.)
  - data centers (catalogs, tables, value-added services)

Electronic information in astronomy

- E-journals link to underlying data, and data archives link to e-journals, through a system of persistent, unique identifiers
- Astronomers interact with a set of connected electronic resources



The Virtual Observatory

- The *Virtual Observatory* is a framework for providing access to distributed data, distributed services. The VO is about *data discovery, access, and integration*, and combining data with computational services.
- Motivation:
  - The data deluge. Needs tools to locate and sift through immense collections and to correlate data from many resources. ~500 TB of data currently available.
  - Scientific discovery opportunities exist at the intersections of diverse data sets.
- Astronomy, of course! Space science, solar physics, aeronomy, seismology, oceanography, hydrology, biology, genomics, medicine. […]ology and […]onomy.
- Keywords: *Metadata, interoperability*

Data/Information in the VO

- 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
- Name-resolution services
  - “Andromeda Galaxy”, “Messier 31”, “M31”, “NGC 224”, “UGC 454”, etc. ==> ra 00h 42m 44s, dec +41º 16’ 08”
  - Geographic equivalent of “Glenn Dale, MD”, “20769”, “Prince George’s County” ==> 76º 48’ 19 W, 38º 58’ 36” N

not discoverable through text-based search engines

Without VO

n services,
n interfaces

archive 1

archive 2

archive 3

survey 1

survey 2

survey 3

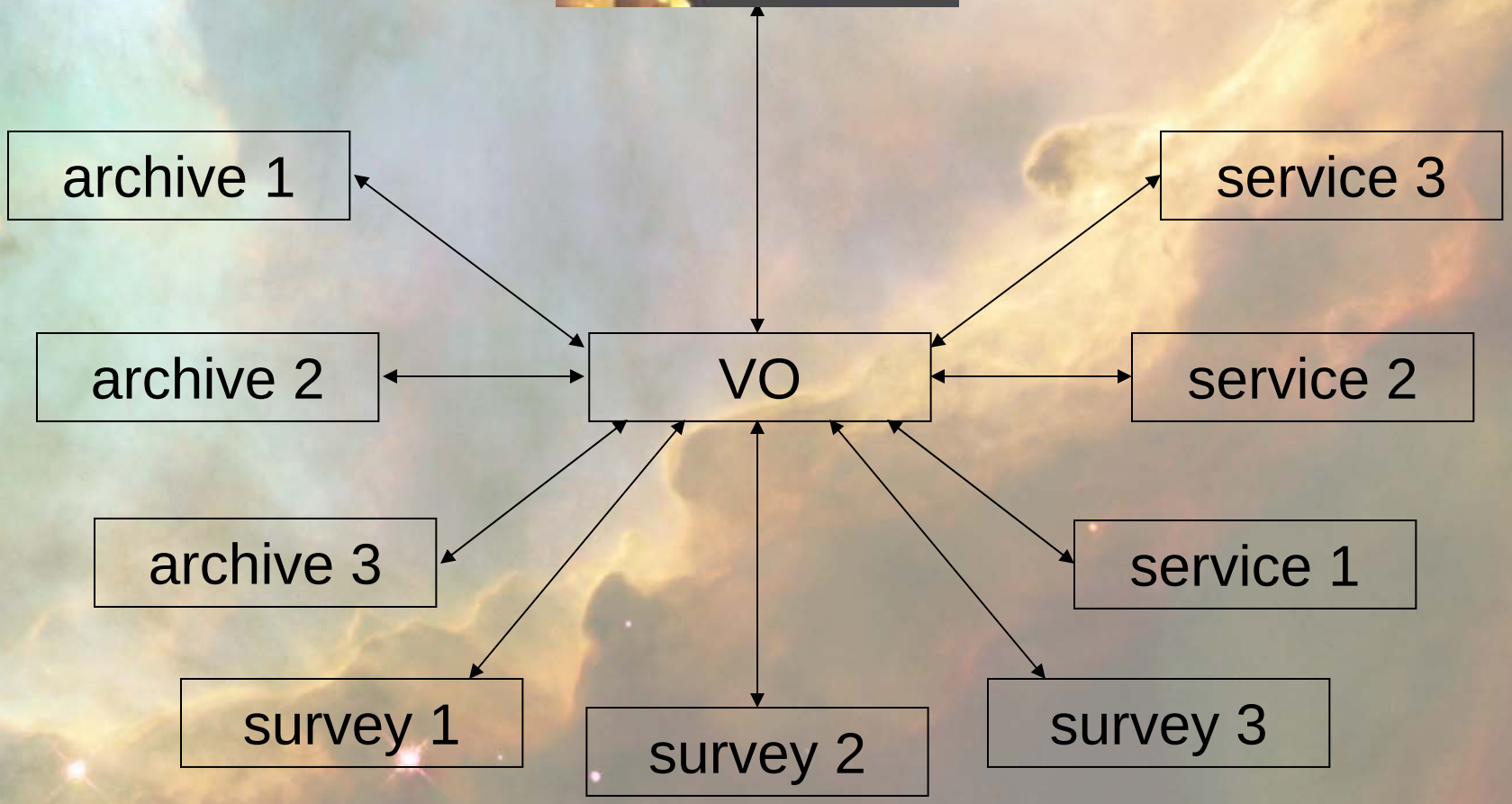
service 1

service 2

service 3

With VO

n services,
“1” interface

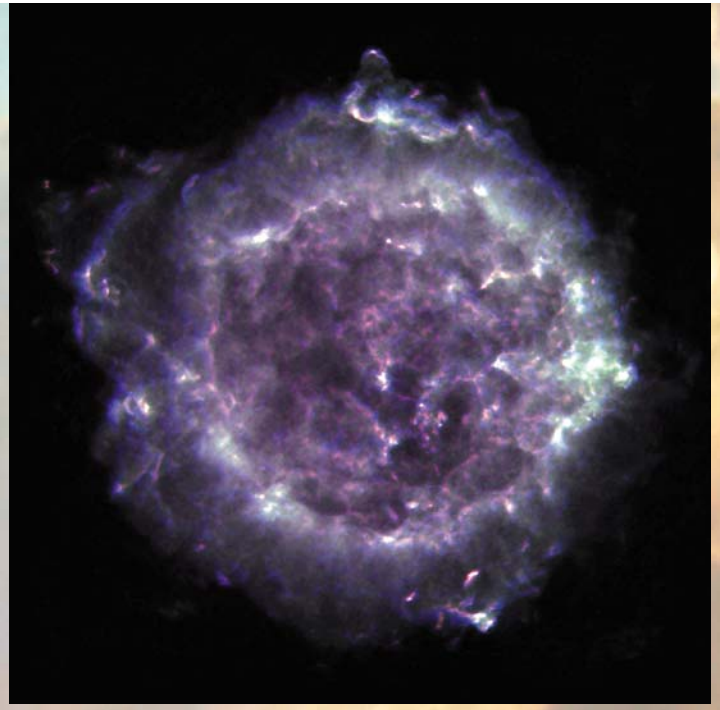


Data integration

Cas A supernova remnant

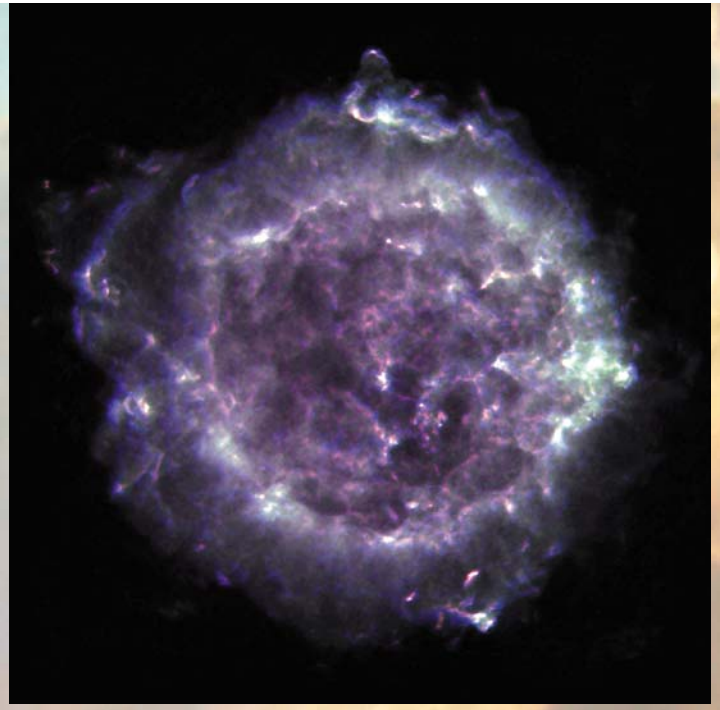
optical (HST)

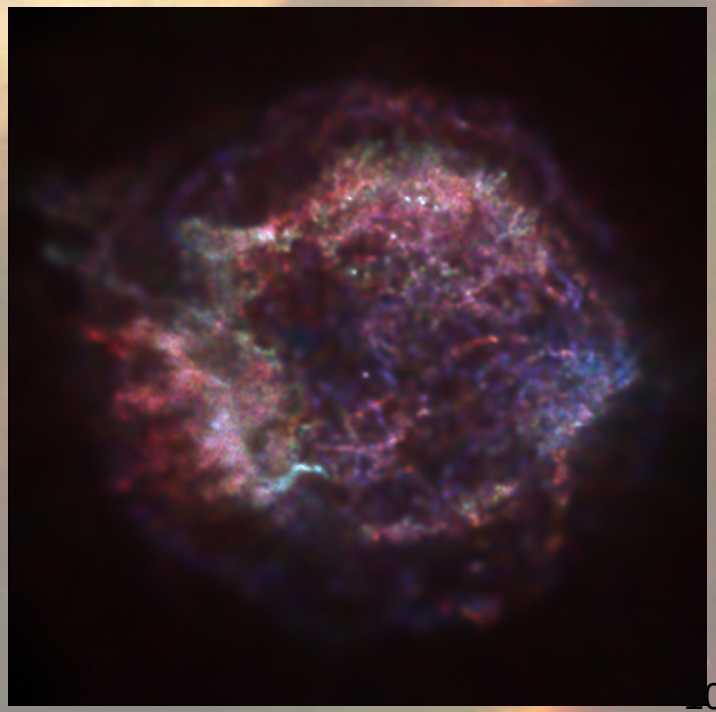
Data integration



radio (VLA)

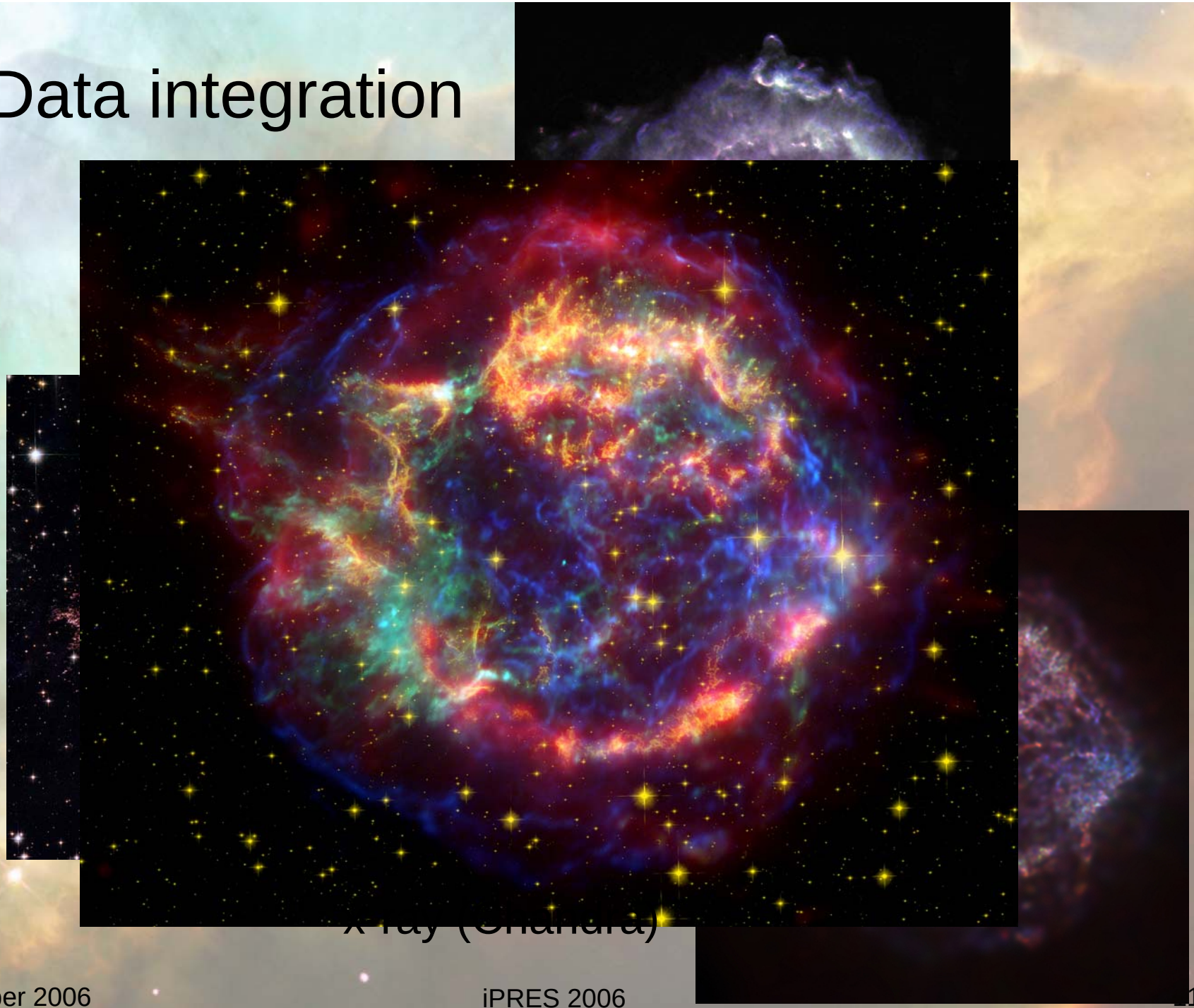
Data integration





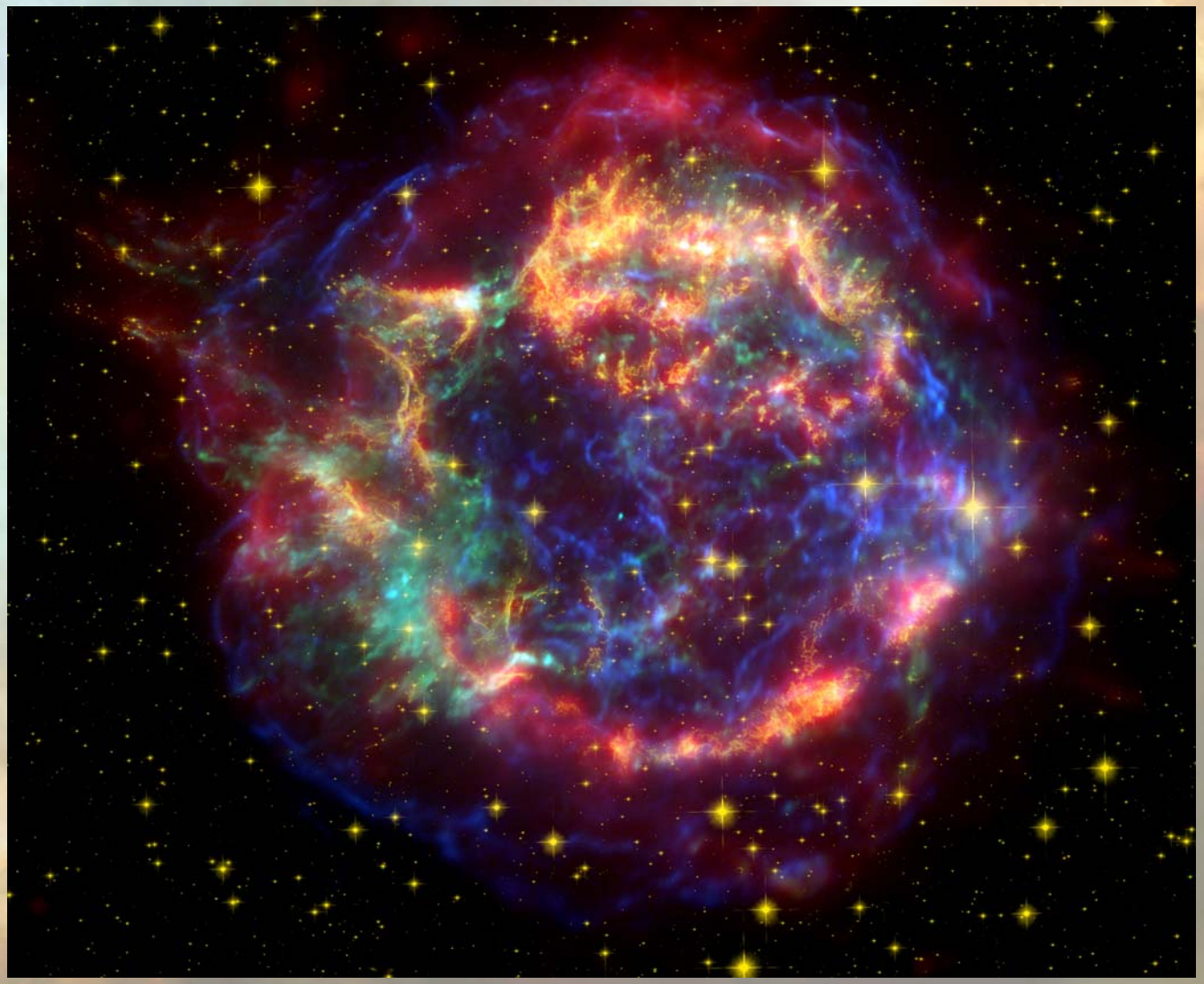
x-ray (Chandra)

Data integration



X-ray (Chandra)

Data integration



The data preservation problem

- Research communities publish peer-reviewed journal papers that describe highly processed data.
- Long-term preservation and curation systems for digital journal content are not currently in place; *only the graphical representations of data are being saved*.
- 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.
- Exploit information management and curation experience in the university libraries and build on long-term institutional commitments to 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!

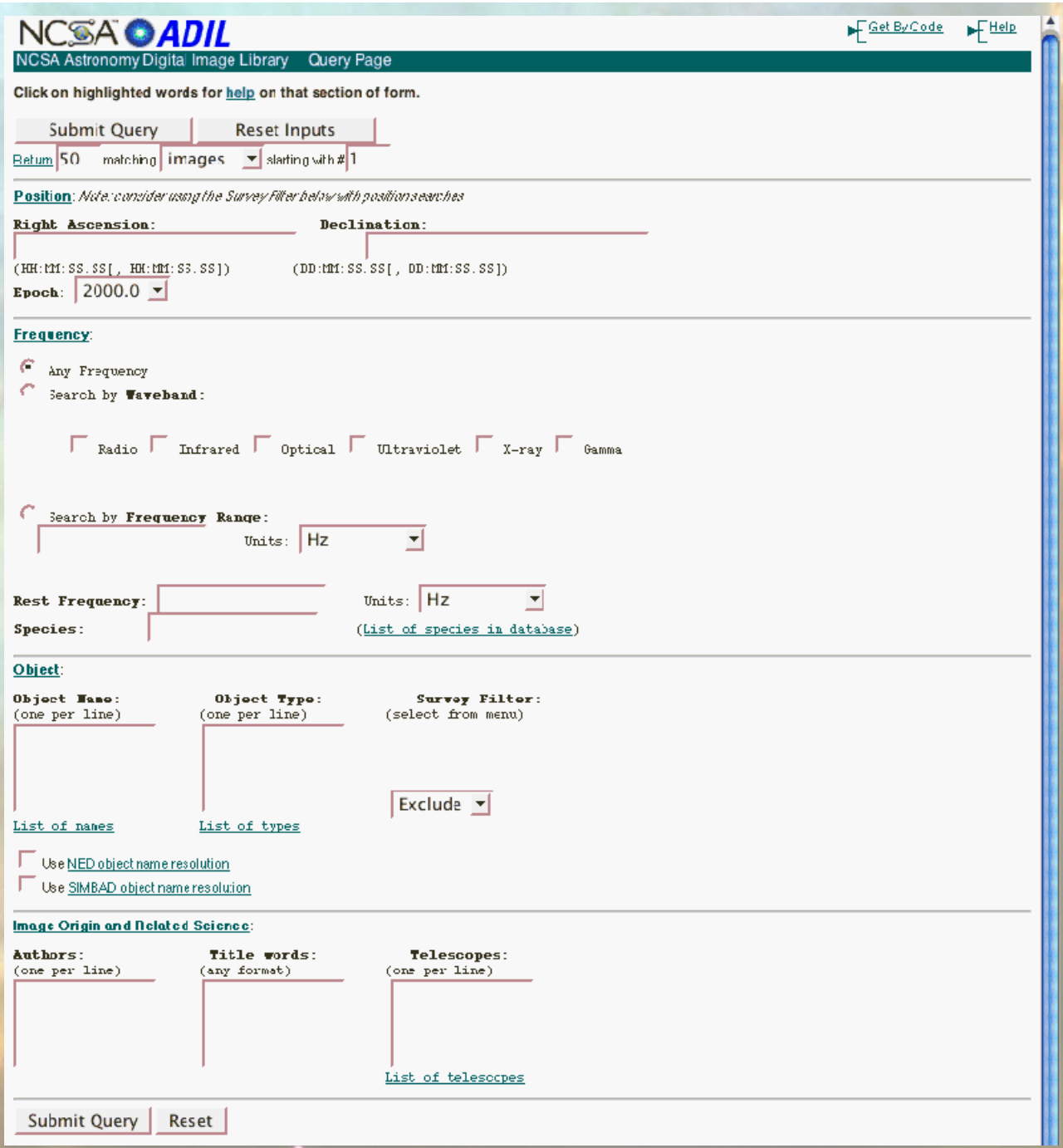
NCSA ADIL

User's Guide | Step 1: Gather Files

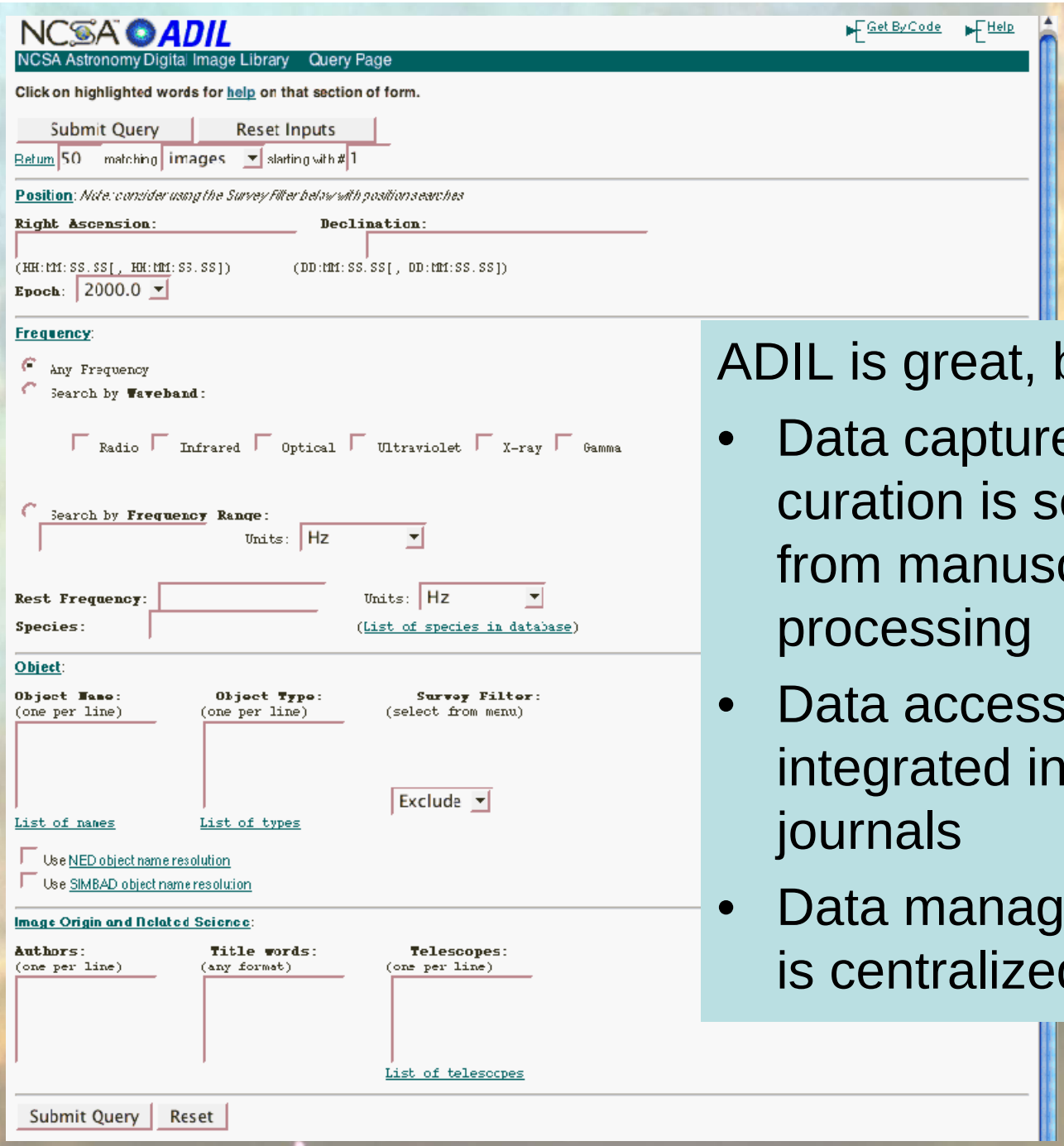
The AstronomyDigital Image Library is a project of Radio Astronomy Imaging Team at the National Center for Supercomputing Applications on the campus of the University of Illinois at Urbana-Champaign

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
- Data management is centralized

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

Stanford et al. 2001). Recently, deep, panoramic multicolor (i', 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.

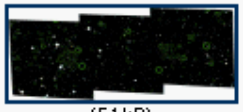

(51 kB)

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.

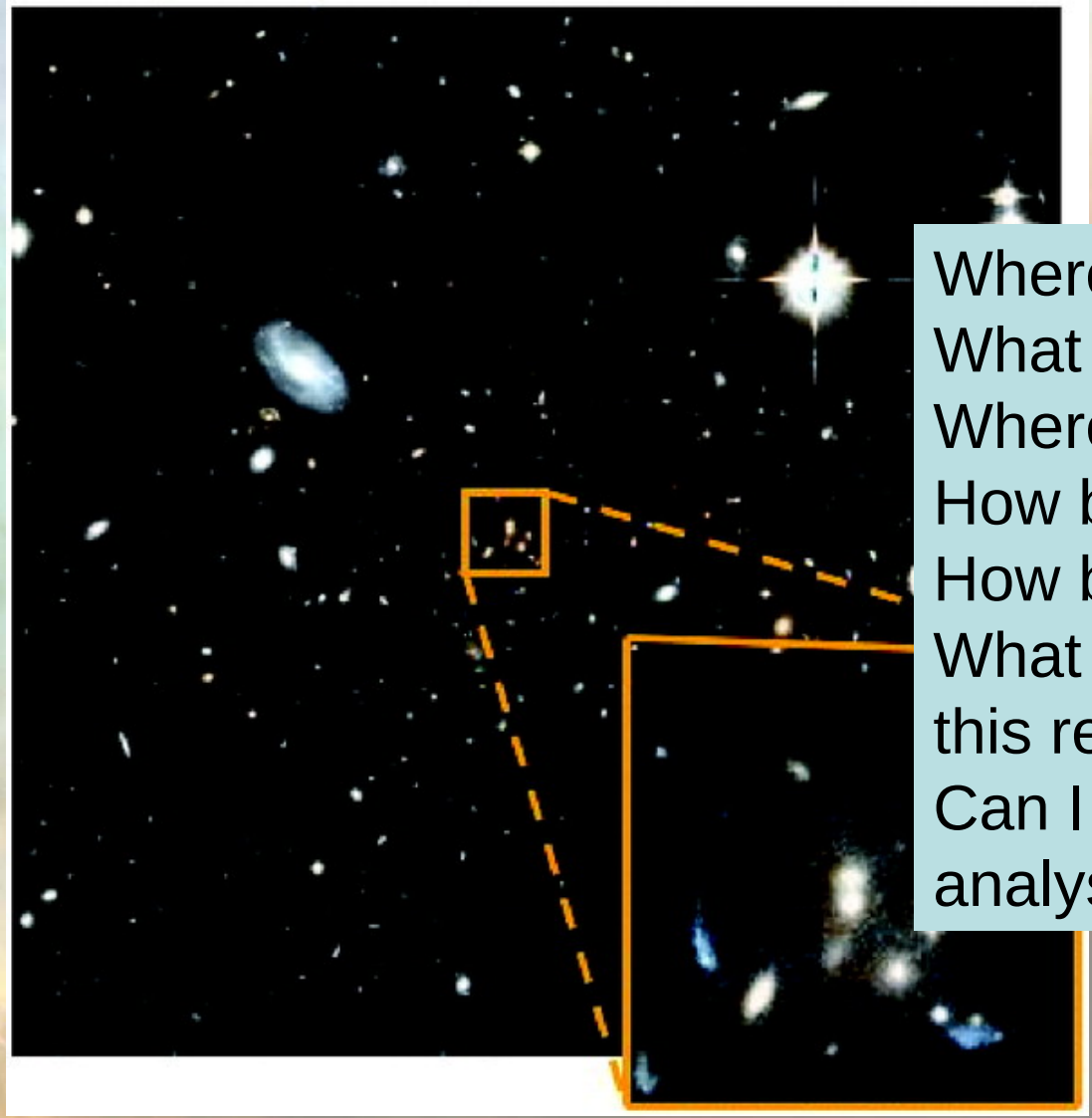

(49 kB)

Fig. 2 Lynx E ACS image (scale is 1′ × 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



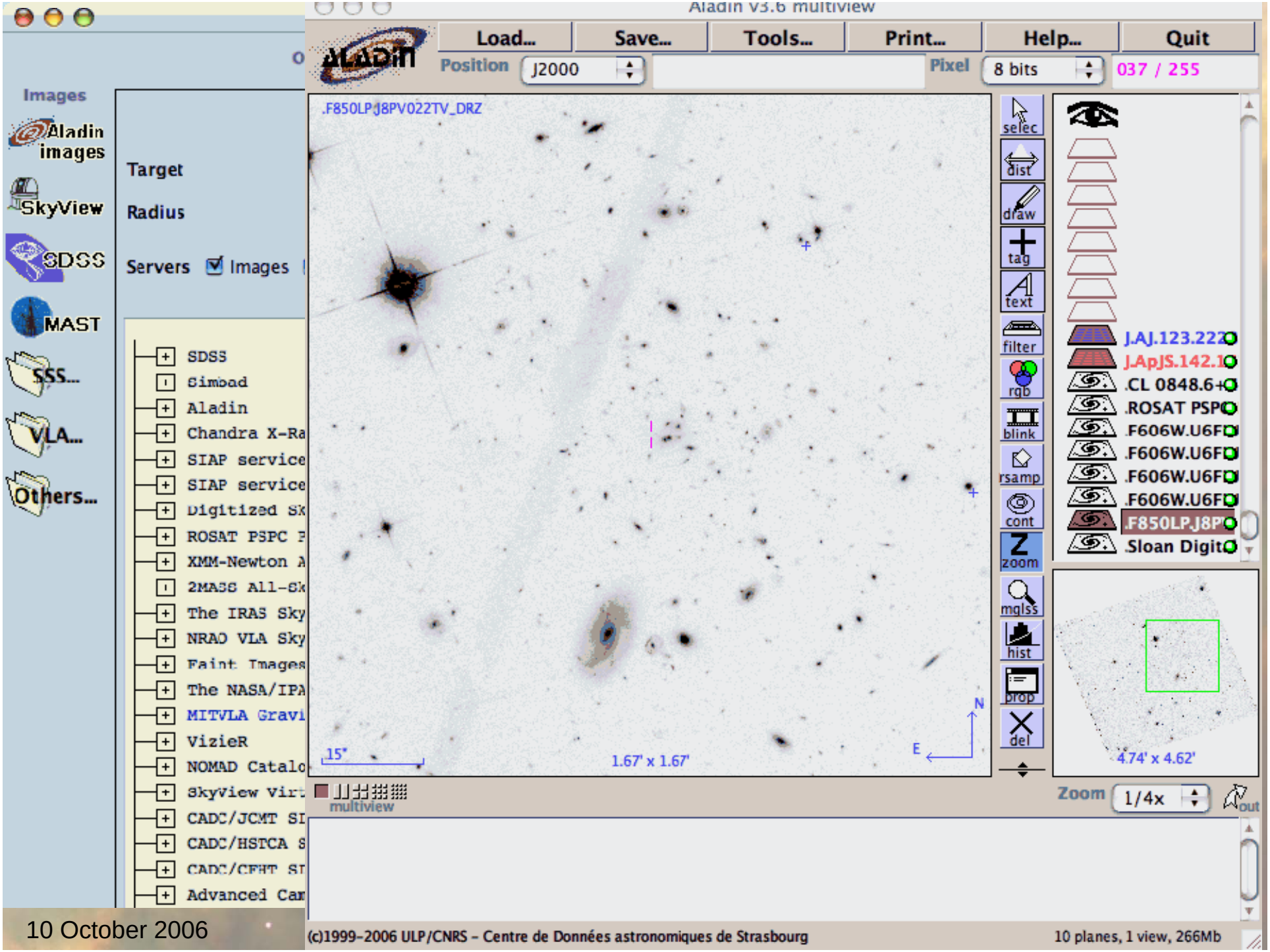
Where is this?
What is the image scale?
Where is north?
How bright is the star?
How bright is the galaxy?
What else is known about this region?
Can I trust the data analysis in this paper?

Storyboard

Save file
Copy to my VOSpace

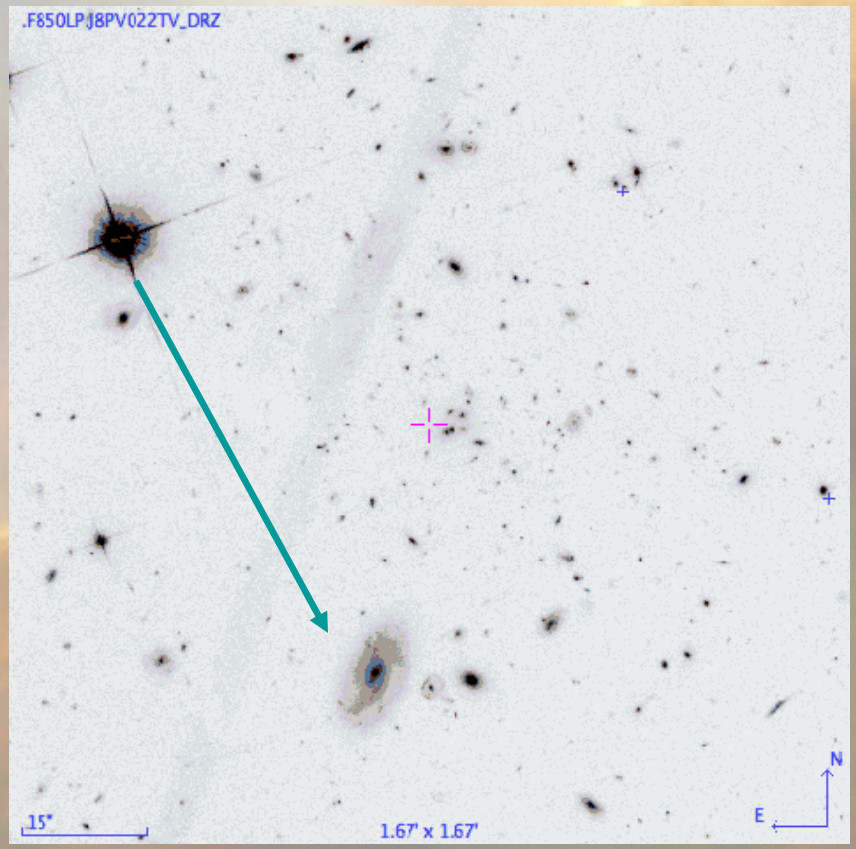
isplay and compare

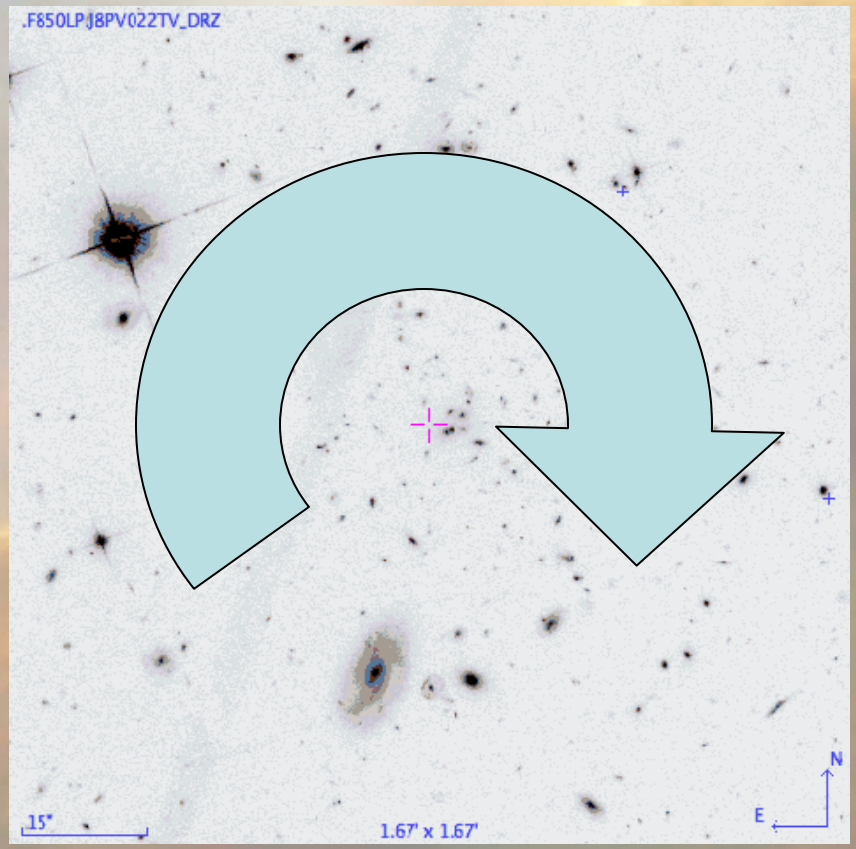

Server selector
Others:
File
all VO
FOV
Images
Aladin images
SkyView
SDSS
MAST
SSS...
VLA...
Others...
VO discovery tool
?
Target
08 49 +44 52
Grab coord
Radius
10.0'
Servers
Images
Catalogs
Spectra
Detailed list...
SDSS
Simbad
Aladin
Chandra X-Ray Observatory Data Archive
SIAP service for the INT wide-field 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 Cataloque
SkyView Virtual Observatory
CADC/JCMT SIA service
CADC/HSTCA SIA service
CADC/CFHT SIA service
Advanced Camera for Surveys
Catalogs
All VizieR
Surveys
Missions
SIMBAD
NED
SkyBot
Others..

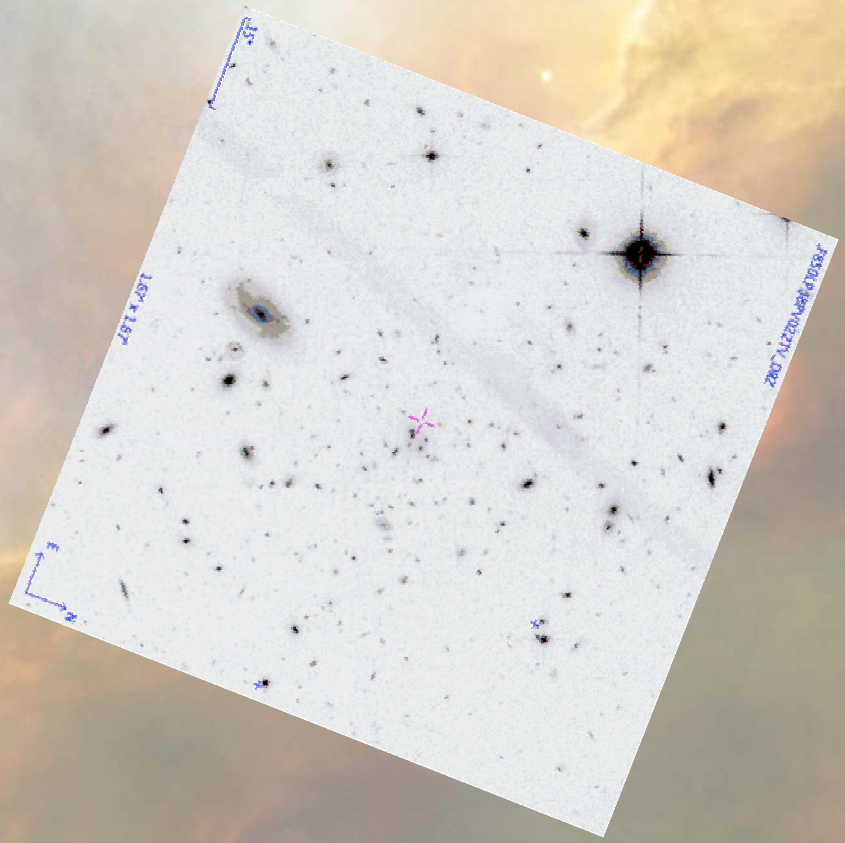

Aladin v3.6 multiview
Load...
Save...
Tools...
Print...
Help...
Quit
Position
J2000
Pixel
8 bits
037 / 255
Images
Aladin images
SkyView
SDSS
MAST
SSS...
VLA...
Others...
Target
Radius
Servers
Images
SDSS
Simbad
Aladin
Chandra X-Ra
SIAP service
SIAP service
Digitized Sk
ROSAT PSPC F
XMM-Newton A
2MASS All-Sk
The IRAS Sky
NRAO VLA Sky
Faint Images
The NASA/IPA
MITVLA Gravi
VizieR
NOMAD Catalo
SkyView Virt
CADC/JCMT SI
CADC/HSTCA S
CADC/CFHT SI
Advanced Cam
.F850LPJ8PV022TV_DRZ
15"
1.67' x 1.67'
N
E
selec
dist
draw
tag
text
filter
rgb
blink
rsamp
cont
zoom
mglss
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

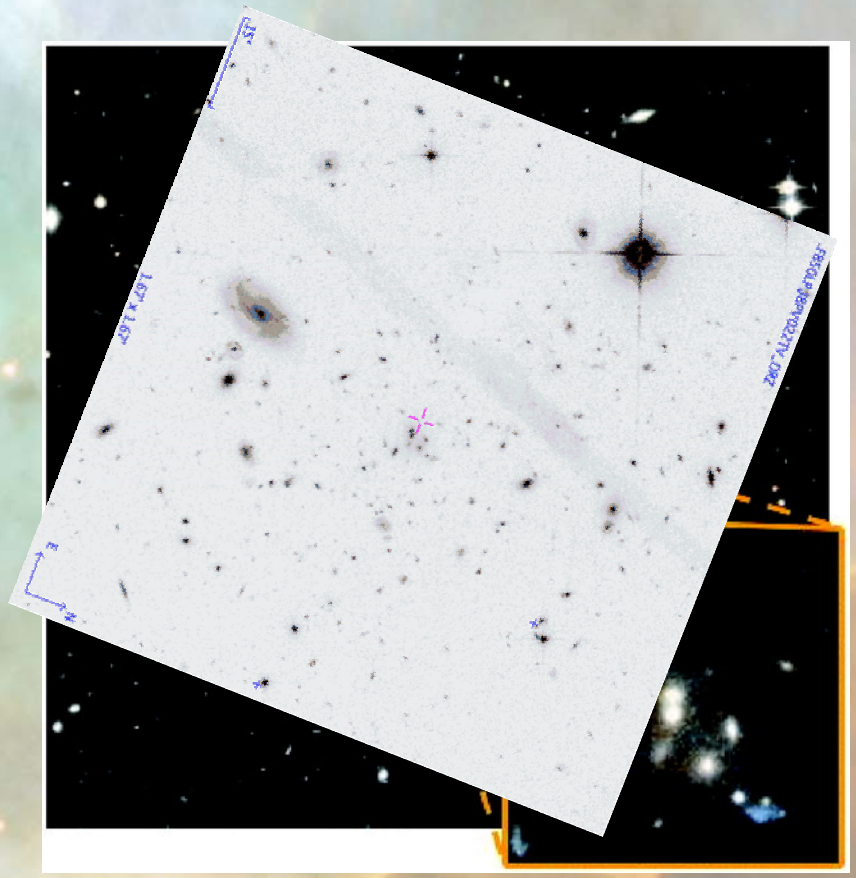
Journal…

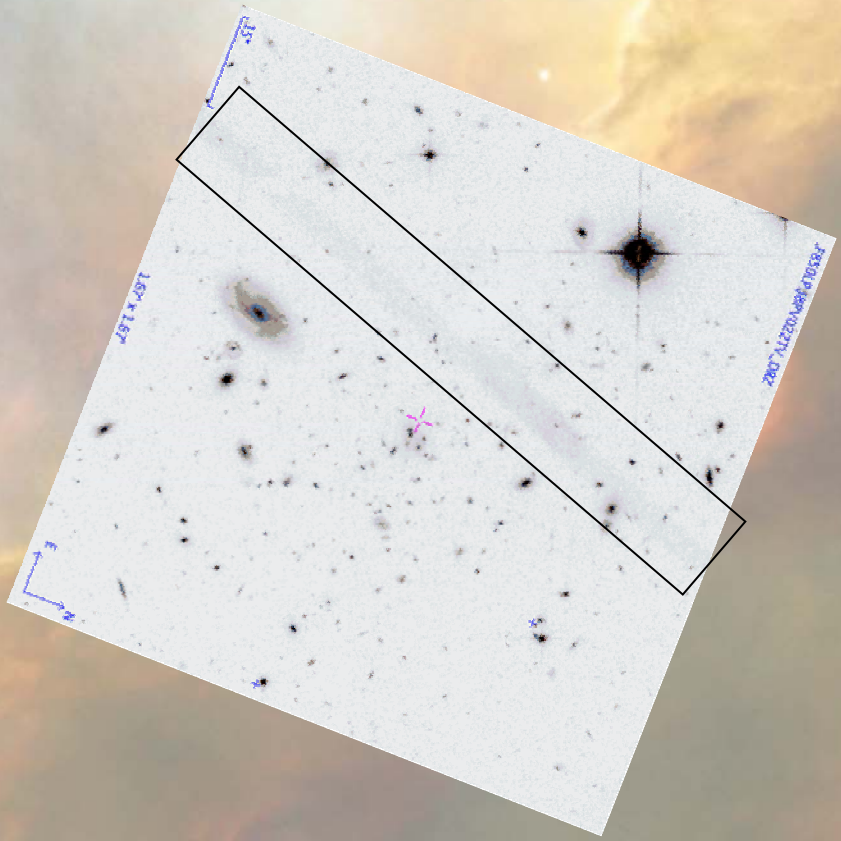
Archive…



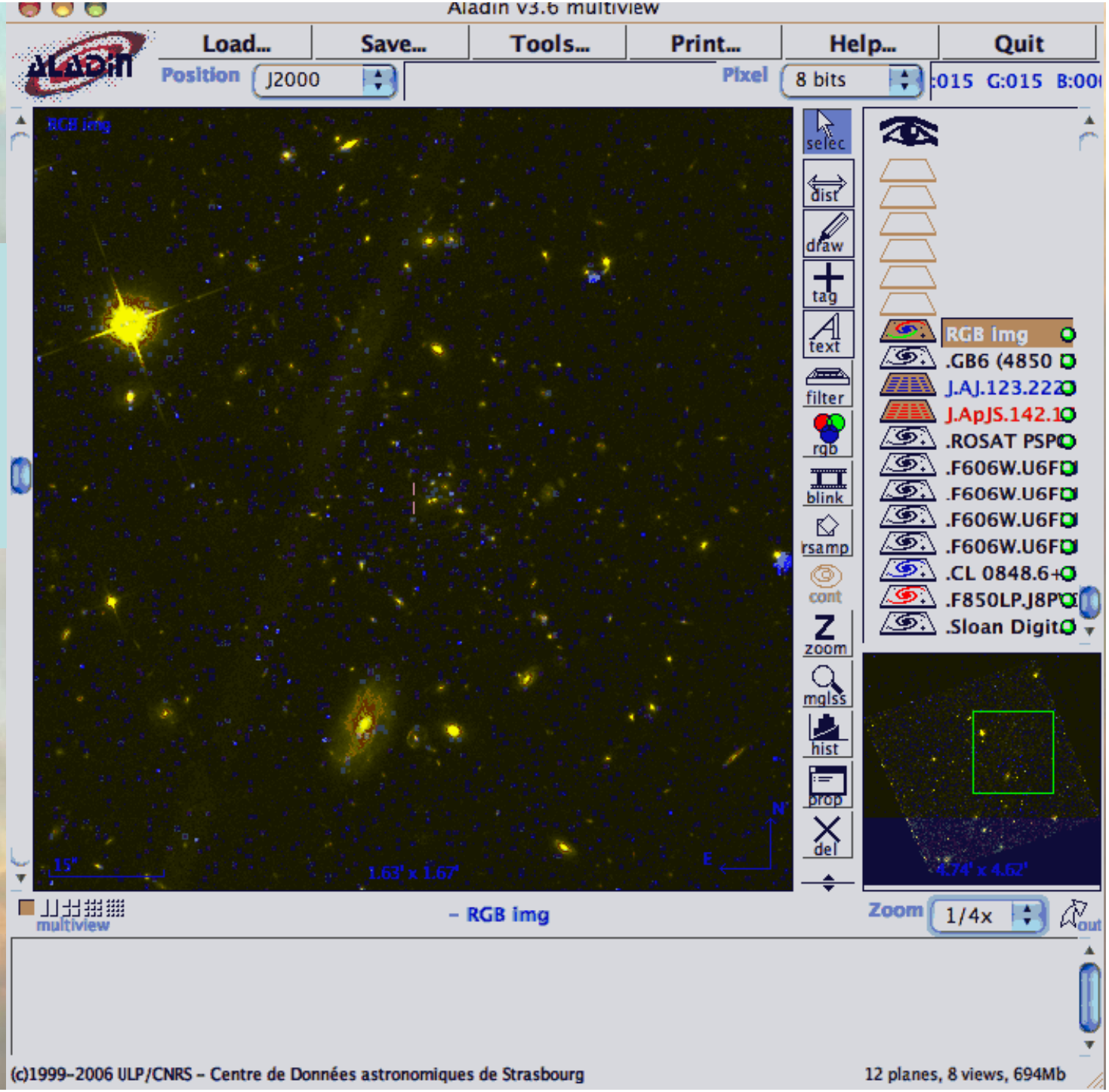

.F850LP.J8PV022TV_DRZ
.15"
1.67' x 1.67'
N
E



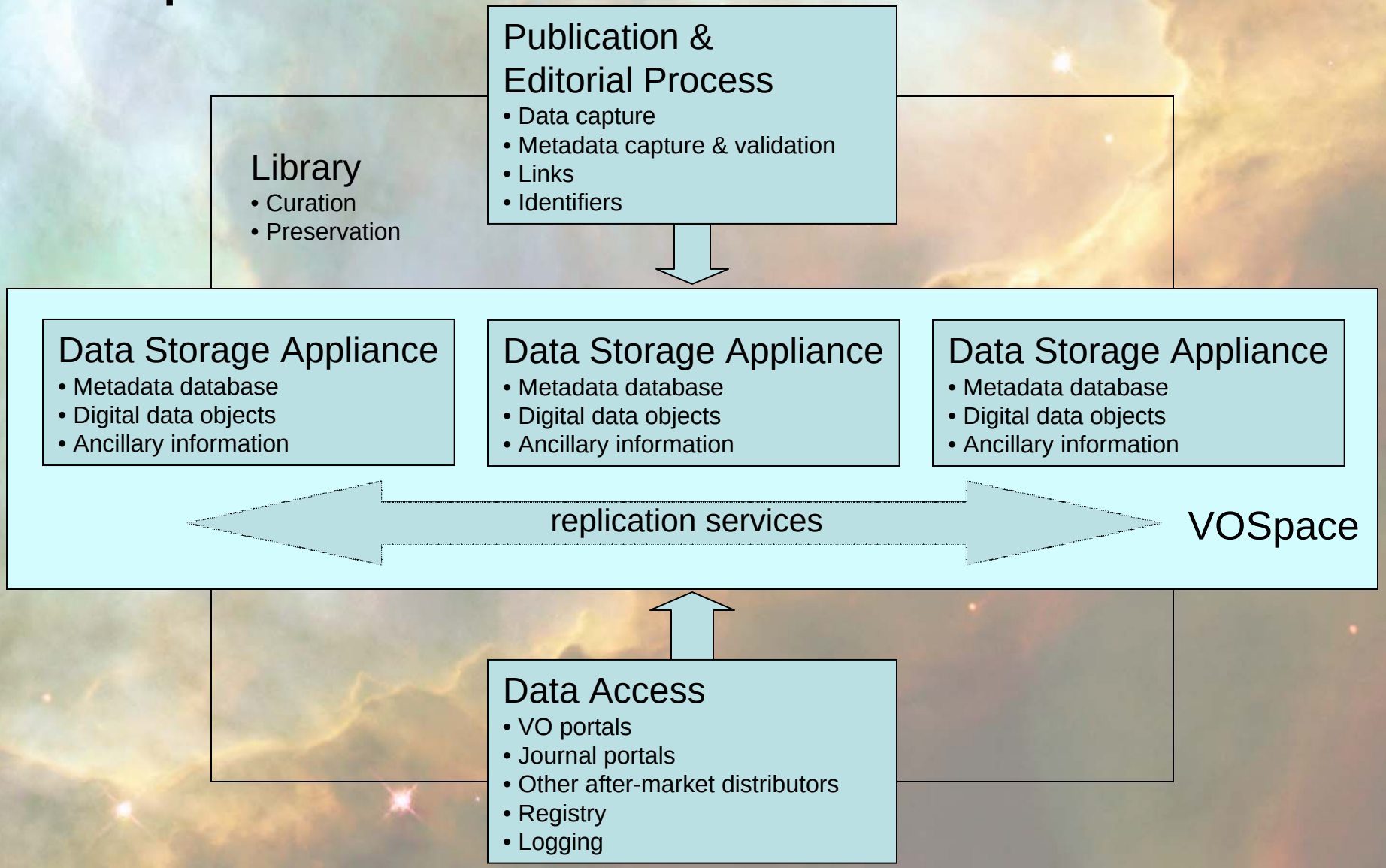

N
E
1.67' x 1.67'
J2000.0 180°00'21.0"
2°11'20.0"



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)

Current collaborators

- American Astronomical Society (journals, editors)
- The University of Chicago Press (publisher for the AAS journals)
- The Johns Hopkins University-Sheridan Library and Cornell University Library (information management, curation & e-pu
- The National Virtual Observatory project (representatives from JHU, Space Telescope Science Institute, and the National Center for Supercomputing Applications)

Status

- Support committed or promised from
  - UK JISC (Joint Information Systems Committee)
  - SPARC (Scholarly Publishing and Academic Resources Coalition)
  - Microsoft
  - TeraGrid
  - NVO
  - IMLS
- Begin development in fall of 2006

Digital data discovery and access is essential for the research community

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