# Building on Existing Communities: the Virtual Astronomical Observatory (and NIST)

Robert Hanisch
Space Telescope Science Institute
Director, Virtual Astronomical Observatory

# Data in astronomy

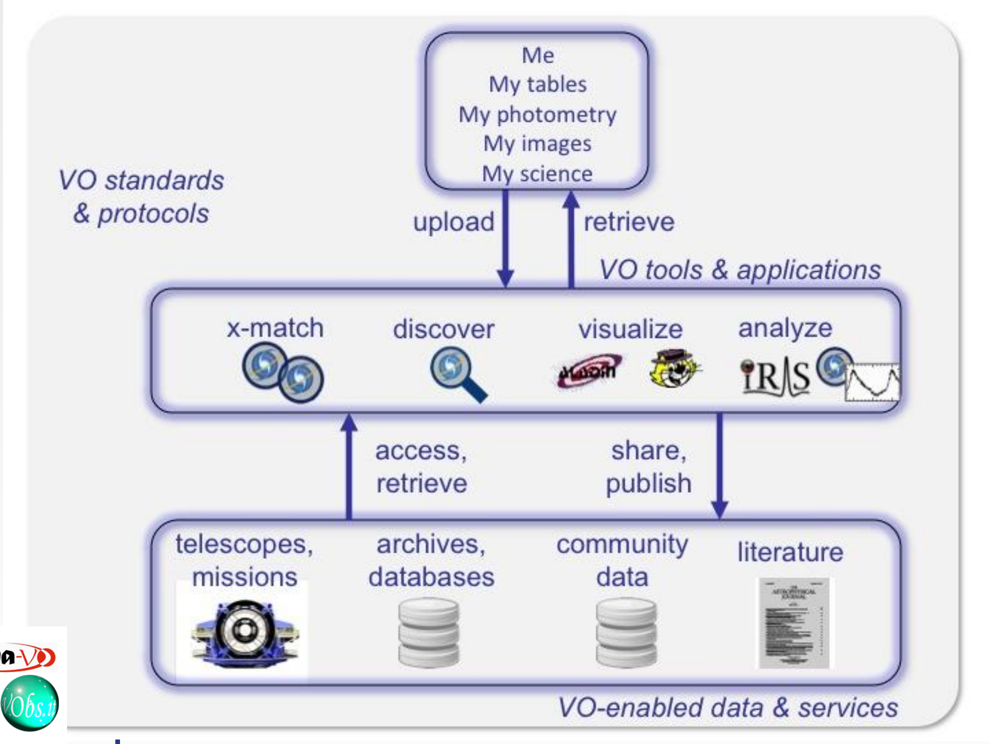
- ~70 major data centers and observatories with substantial on-line data holdings
- ~10,000 data "resources" (catalogs, surveys, archives)
- Data centers host from a few to ~100s TB each, currently at least 2 PB total
- Current growth rate ~0.5 PB/yr, increasing
- Current request rate ~1 PB/yr
- Future surveys will increase data rates to PB/day
  - "For LSST, the telescope is a peripheral to the data system" (T. Tyson)

*How do astronomers navigate through all of this data?*

# The Virtual Observatory

## *The VO is a data discovery, access, and integration facility*

- Images, spectra, time series
- Catalogs, databases
- Transient event notices
- Software and services
- Application inter-communication
- Distributed computing
  - authentication, authorization, process management
- International
  coordination nd
  collaboration
  IVOA
  W3C)



Me
My tables
My photometry
My images
My science

*VO standards & protocols*

upload

retrieve

*VO tools & applications*

x-match

discover

visualize

analyze

access, retrieve

share, publish

telescopes, missions

archives, databases

community data

literature

*VO-enabled data & services*

# Virtual Observatory capabilities

- Data exchange / interoperability / multi-λ (co-observing)

  *Data Access Layer (SIAP, SSAP / time series)*

- Query and cross-match across distributed databases

  *Cone Search, Table Access Protocol*

- Remote (but managed) access to centralized computing and data storage resources

  *VOSpace, Single-Sign-On (OpenID), SciDrive*

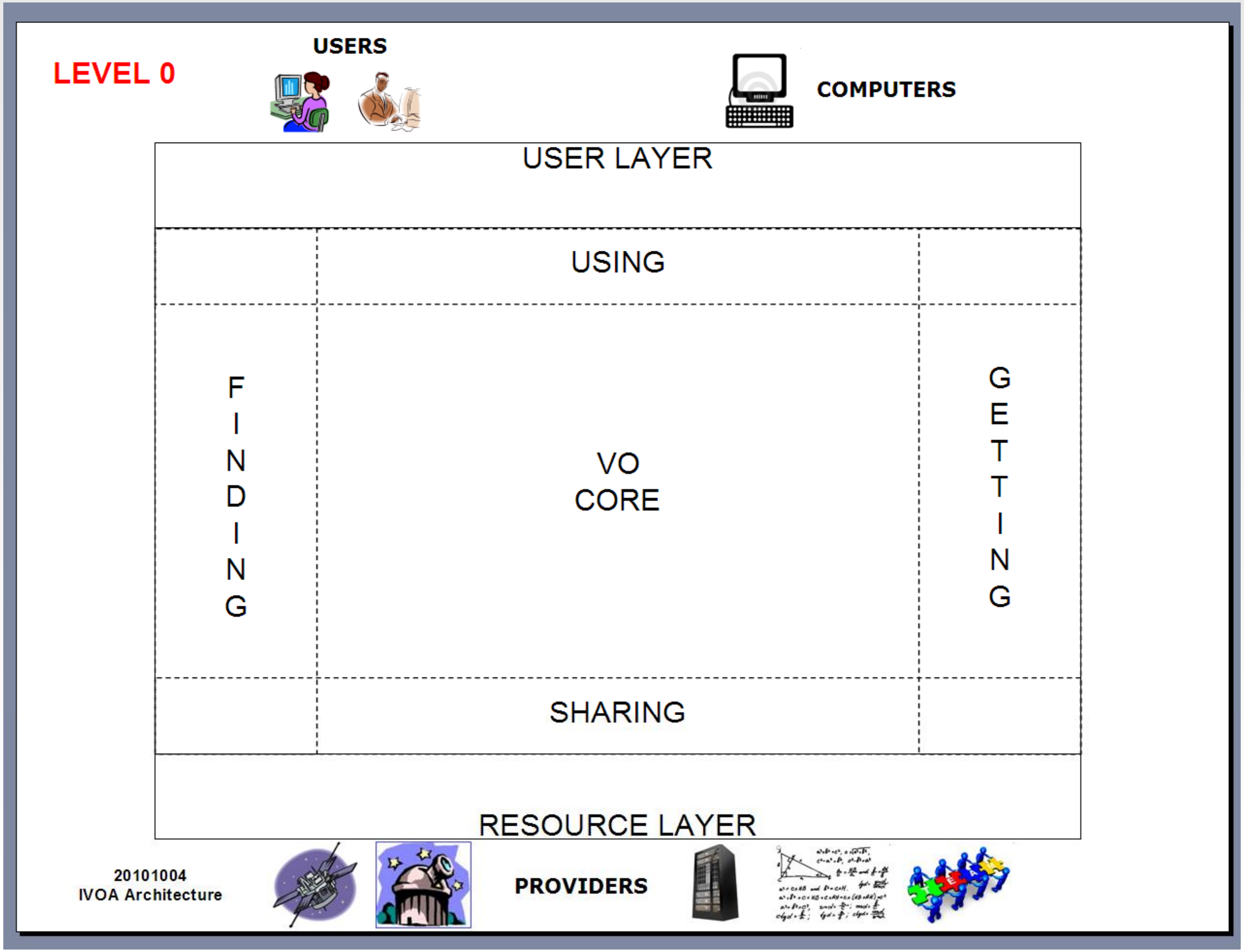
- Transient event notification, scalable to $10^6$ messages/night

  *VOEvent*

- Data mining, characterization, classification, statistical analysis

  *VOStat, Data Mining and Exploration toolkit*

# VO architecture



LEVEL 0

USERS

COMPUTERS

USER LAYER

USING

FINDING

VO CORE

GETTING

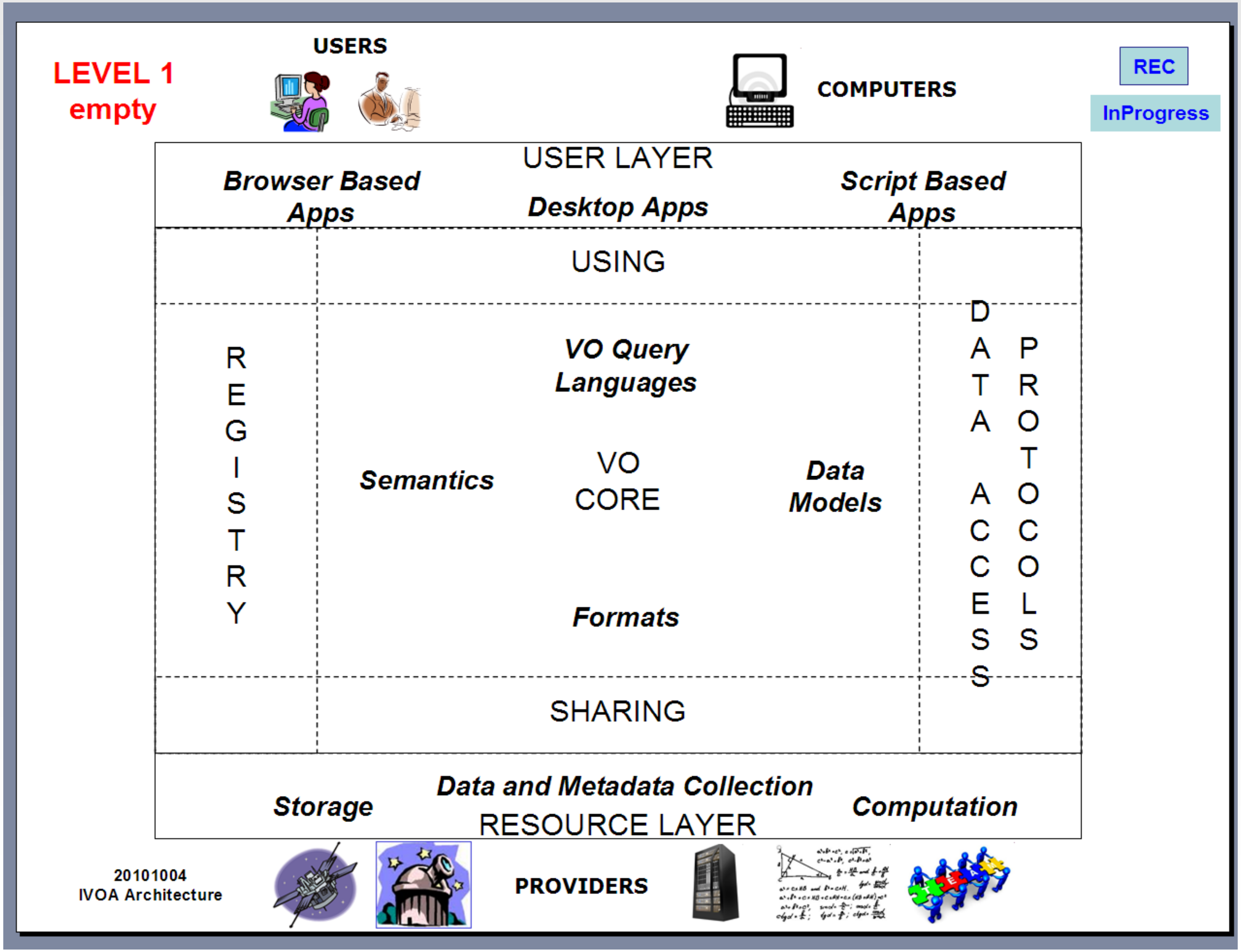
SHARING

RESOURCE LAYER

20101004
IVOA Architecture

PROVIDERS

# VO architecture



LEVEL 1
empty

USERS

COMPUTERS

REC

InProgress

USER LAYER

Browser Based Apps

Desktop Apps

Script Based Apps

USING

REGISTRY

VO Query Languages

Semantics

VO CORE

Data Models

Formats

DATA ACCESS

PROTOCOLS

SHARING

Data and Metadata Collection

Storage

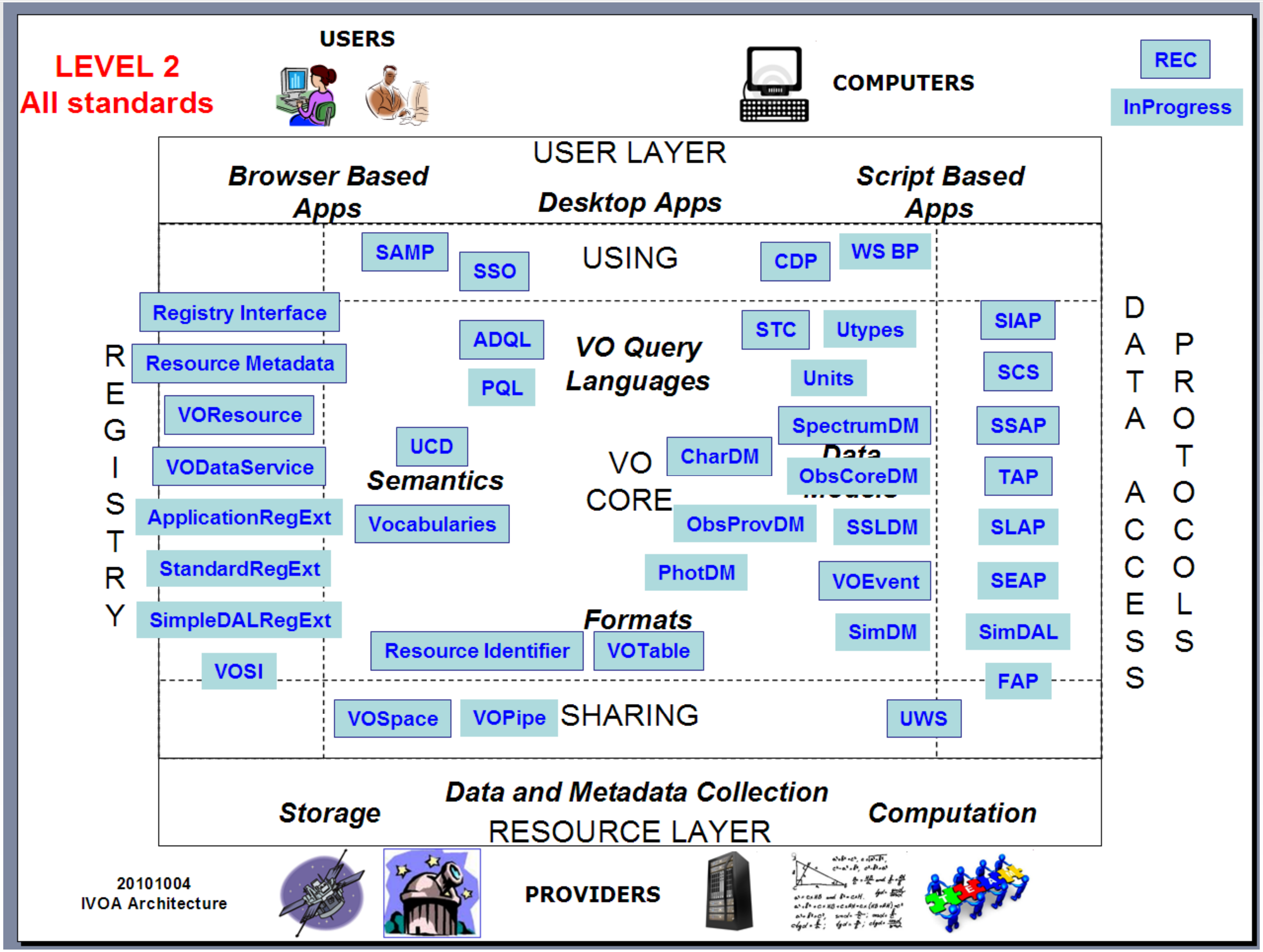
Computation

RESOURCE LAYER

20101004
IVOA Architecture

PROVIDERS

# VO architecture



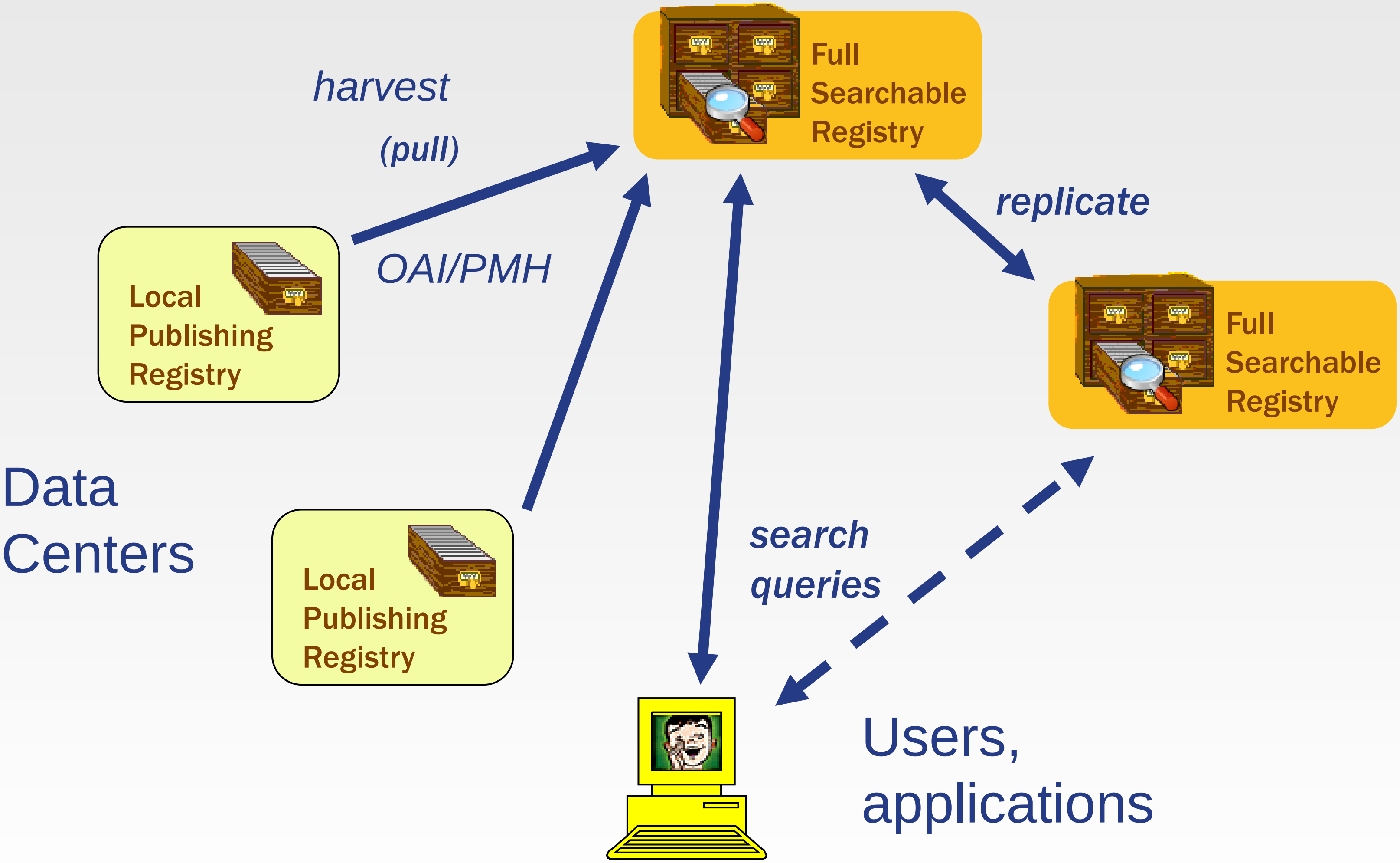
# Key to discovery:  Registry

- Used to discover and locate *resources*—data and services—that can be used in a VO application
- Resource:  anything that is describable and identifiable.
  - Besides data and services:  organizations, projects, software, standards
- Registry: a list of resource descriptions
  - Expressed as structured metadata in XML

    to enable automated processing and searching

    Metadata based on Dublin Core

# Registry framework


harvest

(pull)

OAI/PMH

Full Searchable Registry

replicate

Full Searchable Registry

Local Publishing Registry

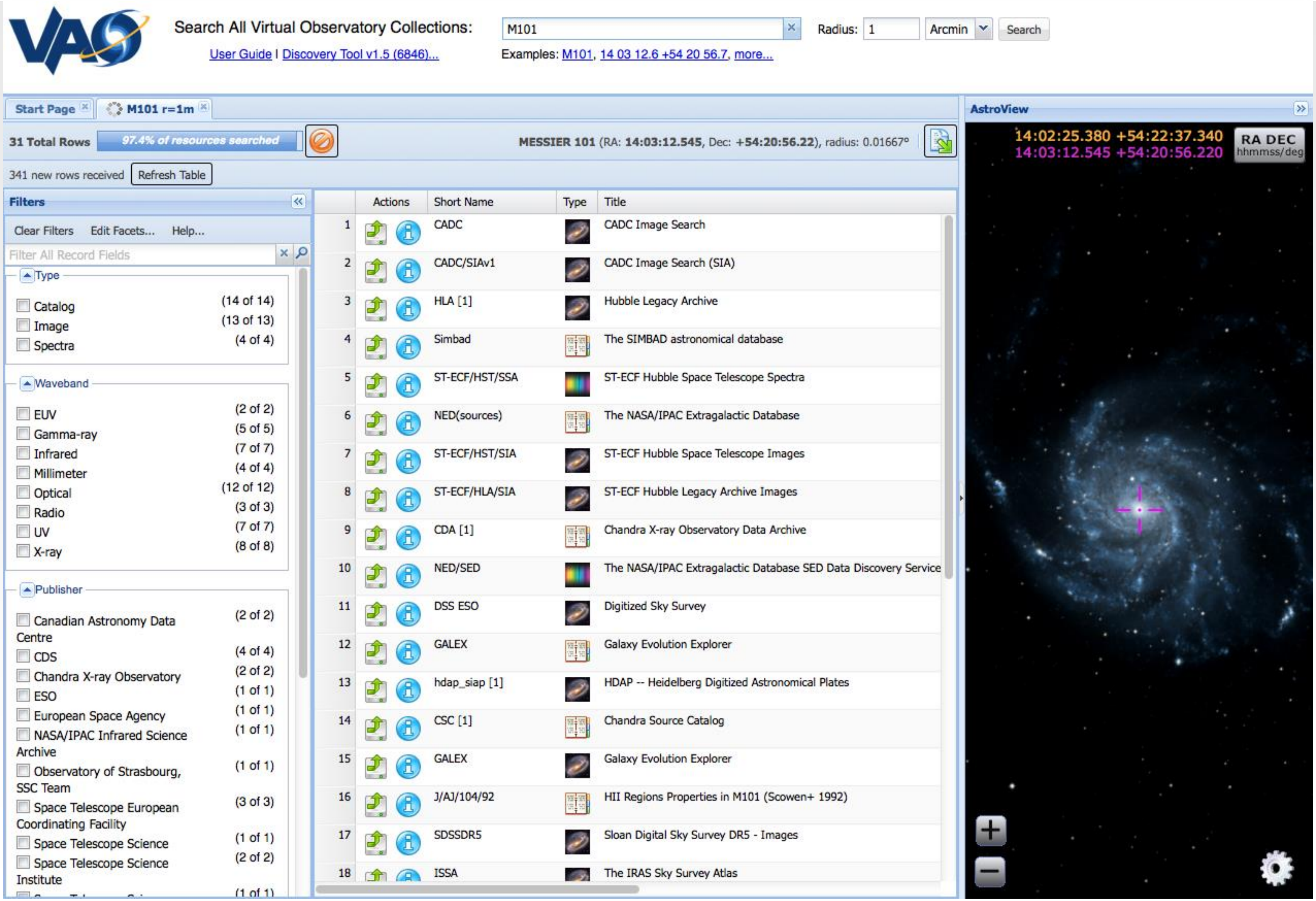
Local Publishing Registry

Data Centers

search queries

Users, applications

# Data discovery



Search All Virtual Observatory Collections: M101 Radius: 1 Arcmin Search

User Guide | Discovery Tool v1.5 (6846)...

Examples: M101, 14 03 12.6 +54 20 56.7, more...

Start Page | M101 r=1m

31 Total Rows 97.4% of resources searched

MESSIER 101 (RA: 14:03:12.545, Dec: +54:20:56.22), radius: 0.01667°

341 new rows received Refresh Table

AstroView

14:02:25.380 +54:22:37.340
14:03:12.545 +54:20:56.220

RA DEC hhmmss/deg

**Filters**

Clear Filters Edit Facets... Help...

Filter All Record Fields

Type

| | |
|---|---|
| □ Catalog | (14 of 14) |
| □ Image | (13 of 13) |
| □ Spectra | (4 of 4) |

Waveband

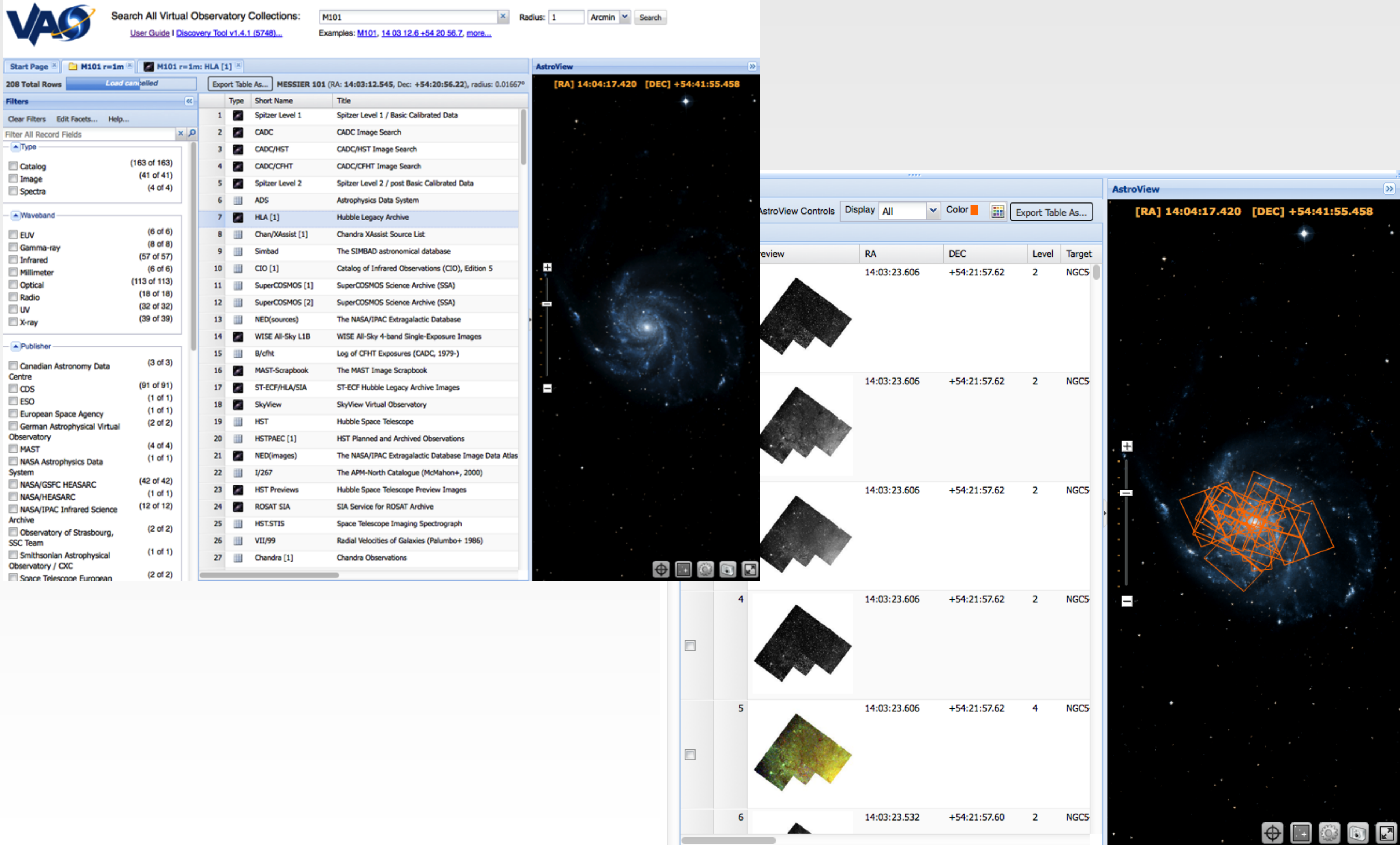
| | |
|---|---|
| □ EUV | (2 of 2) |
| □ Gamma-ray | (5 of 5) |
| □ Infrared | (7 of 7) |
| □ Millimeter | (4 of 4) |
| □ Optical | (12 of 12) |
| □ Radio | (3 of 3) |
| □ UV | (7 of 7) |
| □ X-ray | (8 of 8) |

Publisher

| | |
|---|---|
| □ Canadian Astronomy Data Centre | (2 of 2) |
| □ CDS | (4 of 4) |
| □ Chandra X-ray Observatory | (2 of 2) |
| □ ESO | (1 of 1) |
| □ European Space Agency | (1 of 1) |
| □ NASA/IPAC Infrared Science Archive | (1 of 1) |
| □ Observatory of Strasbourg, SSC Team | (1 of 1) |
| □ Space Telescope European Coordinating Facility | (3 of 3) |
| □ Space Telescope Science | (1 of 1) |
| □ Space Telescope Science Institute | (2 of 2) |
| □ | (1 of 1) |

| | Actions | Short Name | Type | Title |
|---|---|---|---|---|
| 1 | | CADC | | CADC Image Search |
| 2 | | CADC/SIAv1 | | CADC Image Search (SIA) |
| 3 | | HLA [1] | | Hubble Legacy Archive |
| 4 | | Simbad | | The SIMBAD astronomical database |
| 5 | | ST-ECF/HST/SSA | | ST-ECF Hubble Space Telescope Spectra |
| 6 | | NED(sources) | | The NASA/IPAC Extragalactic Database |
| 7 | | ST-ECF/HST/SIA | | ST-ECF Hubble Space Telescope Images |
| 8 | | ST-ECF/HLA/SIA | | ST-ECF Hubble Legacy Archive Images |
| 9 | | CDA [1] | | Chandra X-ray Observatory Data Archive |
| 10 | | NED/SED | | The NASA/IPAC Extragalactic Database SED Data Discovery Service |
| 11 | | DSS ESO | | Digitized Sky Survey |
| 12 | | GALEX | | Galaxy Evolution Explorer |
| 13 | | hdap_siap [1] | | HDAP -- Heidelberg Digitized Astronomical Plates |
| 14 | | CSC [1] | | Chandra Source Catalog |
| 15 | | GALEX | | Galaxy Evolution Explorer |
| 16 | | J/AJ/104/92 | | HII Regions Properties in M101 (Scowen+ 1992) |
| 17 | | SDSSDR5 | | Sloan Digital Sky Survey DR5 - Images |
| 18 | | ISSA | | The IRAS Sky Survey Atlas |

# Data discovery



Search All Virtual Observatory Collections: M101 Radius: 1 Arcmin Search

User Guide | Discovery Tool v1.4.1 (5748)...

Examples: M101, 14 03 12.6 +54 20 56.7, more...

208 Total Rows

MESSIER 101 (RA: 14:03:12.545, Dec: +54:20:56.22), radius: 0.01667°

Filters

Type
- □ Catalog (163 of 163)
- □ Image (41 of 41)
- □ Spectra (4 of 4)

Waveband
- □ EUV (6 of 6)
- □ Gamma-ray (8 of 8)
- □ Infrared (57 of 57)
- □ Millimeter (6 of 6)
- □ Optical (113 of 113)
- □ Radio (18 of 18)
- □ UV (32 of 32)
- □ X-ray (39 of 39)

Publisher
- □ Canadian Astronomy Data Centre (3 of 3)
- □ CDS (91 of 91)
- □ ESO (1 of 1)
- □ European Space Agency (1 of 1)
- □ German Astrophysical Virtual Observatory (2 of 2)
- □ MAST (4 of 4)
- □ NASA Astrophysics Data System (1 of 1)
- □ NASA/GSFC HEASARC (42 of 42)
- □ NASA/HEASARC (1 of 1)
- □ NASA/IPAC Infrared Science Archive (12 of 12)
- □ Observatory of Strasbourg, SSC Team (2 of 2)
- □ Smithsonian Astrophysical Observatory / CXC (1 of 1)
- □ Space Telescope European (2 of 2)

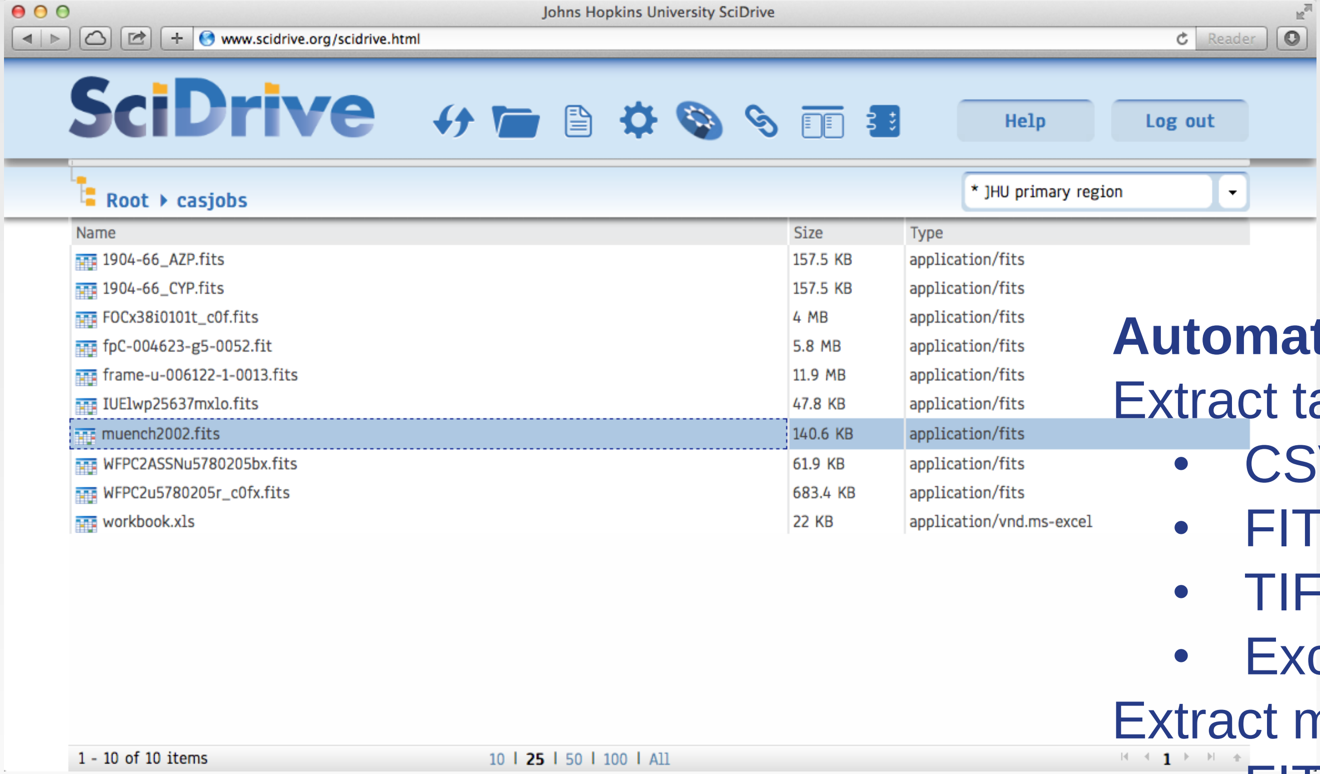
| | Type | Short Name | Title |
|---|---|---|---|
| 1 | ■ | Spitzer Level 1 | Spitzer Level 1 / Basic Calibrated Data |
| 2 | ■ | CADC | CADC Image Search |
| 3 | ■ | CADC/HST | CADC/HST Image Search |
| 4 | ■ | CADC/CFHT | CADC/CFHT Image Search |
| 5 | ■ | Spitzer Level 2 | Spitzer Level 2 / post Basic Calibrated Data |
| 6 | ▦ | ADS | Astrophysics Data System |
| 7 | ■ | HLA [1] | Hubble Legacy Archive |
| 8 | ▦ | Chan/XAssist [1] | Chandra XAssist Source List |
| 9 | ▦ | Simbad | The SIMBAD astronomical database |
| 10 | ▦ | CIO [1] | Catalog of Infrared Observations (CIO), Edition 5 |
| 11 | ▦ | SuperCOSMOS [1] | SuperCOSMOS Science Archive (SSA) |
| 12 | ▦ | SuperCOSMOS [2] | SuperCOSMOS Science Archive (SSA) |
| 13 | ▦ | NED(sources) | The NASA/IPAC Extragalactic Database |
| 14 | ■ | WISE All-Sky L1B | WISE All-Sky 4-band Single-Exposure Images |
| 15 | ▦ | B/cfht | Log of CFHT Exposures (CADC, 1979-) |
| 16 | ■ | MAST-Scrapbook | The MAST Image Scrapbook |
| 17 | ■ | ST-ECF/HLA/SIA | ST-ECF Hubble Legacy Archive Images |
| 18 | ■ | SkyView | SkyView Virtual Observatory |
| 19 | ▦ | HST | Hubble Space Telescope |
| 20 | ▦ | HSTPAEC [1] | HST Planned and Archived Observations |
| 21 | ■ | NED(images) | The NASA/IPAC Extragalactic Database Image Data Atlas |
| 22 | ▦ | I/267 | The APM-North Catalogue (McMahon+, 2000) |
| 23 | ■ | HST Previews | Hubble Space Telescope Preview Images |
| 24 | ■ | ROSAT SIA | SIA Service for ROSAT Archive |
| 25 | ▦ | HST.STIS | Space Telescope Imaging Spectrograph |
| 26 | ▦ | VII/99 | Radial Velocities of Galaxies (Palumbo+ 1986) |
| 27 | ▦ | Chandra [1] | Chandra Observations |

[RA] 14:04:17.420 [DEC] +54:41:55.458

| | eview | RA | DEC | Level | Target |
|---|---|---|---|---|---|
| | | 14:03:23.606 | +54:21:57.62 | 2 | NGC5 |
| | | 14:03:23.606 | +54:21:57.62 | 2 | NGC5 |
| | | 14:03:23.606 | +54:21:57.62 | 2 | NGC5 |
| 4 | | 14:03:23.606 | +54:21:57.62 | 2 | NGC5 |
| 5 | | 14:03:23.606 | +54:21:57.62 | 4 | NGC5 |
| 6 | | 14:03:23.532 | +54:21:57.60 | 2 | NGC5 |

[RA] 14:04:17.420 [DEC] +54:41:55.458

# SciDrive: astro-centric cloud storage


Johns Hopkins University SciDrive
www.scidrive.org/scidrive.html
Reader
SciDrive
Help
Log out

Root ▸ casjobs

* JHU primary region

| Name | Size | Type |
|---|---|---|
| 1904-66_AZP.fits | 157.5 KB | application/fits |
| 1904-66_CYP.fits | 157.5 KB | application/fits |
| FOCx38i0101t_c0f.fits | 4 MB | application/fits |
| fpC-004623-g5-0052.fit | 5.8 MB | application/fits |
| frame-u-006122-1-0013.fits | 11.9 MB | application/fits |
| IUElwp25637mxlo.fits | 47.8 KB | application/fits |
| muench2002.fits | 140.6 KB | application/fits |
| WFPC2ASSNu5780205bx.fits | 61.9 KB | application/fits |
| WFPC2u5780205r_c0fx.fits | 683.4 KB | application/fits |
| workbook.xls | 22 KB | application/vnd.ms-excel |

1 - 10 of 10 items  10 | 25 | 50 | 100 | All  1

**Automatic Metadata Extraction**

Extract tabular data from:

- CSV
- FITS
- TIFF
- Excel

Extract metadata from:

- FITS
- Image files (TIFF, JPG)

Automatically upload tables into relational databases:

- CasJobs/MyDB
- SQLShare

*Controlled data sharing*
*Single sign-on*
*Deployable as virtual machine*

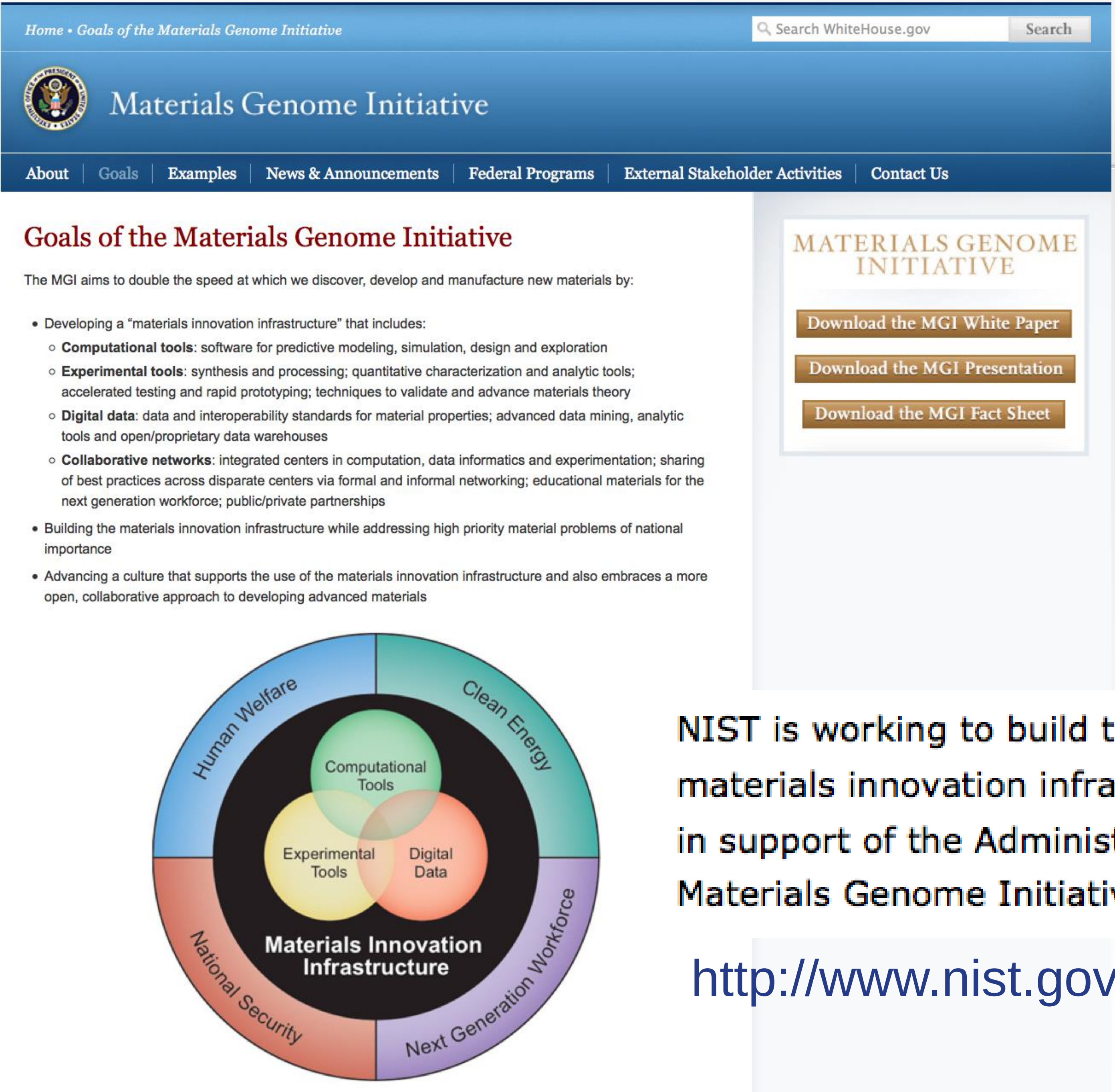
# The VO concept elsewhere

- Space Science
  - Virtual Heliophysics Observatory (HELIO)
  - Virtual Radiation Belt Observatory (ViRBO)
  - Virtual Space Physics Observatory (VSPO)
  - Virtual Magnetospheric Observatory (VMO)
  - Virtual Ionosphere Thermosphere Mesosphere Observatory (VITMO)
  - Virtual Solar-Terrestrial Observatory (VSTO)
  - Virtual Sun/Earth Observatory (VSEO)
- Virtual Solar Observatory
- Planetary Science Virtual Observatory
- Deep Carbon Virtual Observatory
- Virtual Brain Observatory

# Data management at NIST

National Institute of Standards and Technology
U.S. Department of Commerce

- I move to NIST 7/28/2014 as Director, Office of Data and Informatics, Material Measurement Laboratory
  - Materials science, chemistry, biology
  - Materials Genome Initiative
- Foster a culture of data management, curation, re-use in a bench-scientist / PI-dominated organization having a strong record of providing “gold standard” data
- Inward-looking challenges
  - Tools, support, advice, common platforms, solution broker
  - Big data, lots of small/medium data
- Outward-looking challenges
  - Service directory
  - Modern web interfaces, APIs, better service integration
  - Get better sense of what communities want from NIST
- Define standards, standard practices
- Collaboration:  other government agencies, universities, domain repositories



Home • Goals of the Materials Genome Initiative
Search WhiteHouse.gov
Search
Materials Genome Initiative
About
Goals
Examples
News & Announcements
Federal Programs
External Stakeholder Activities
Contact Us

## Goals of the Materials Genome Initiative

The MGI aims to double the speed at which we discover, develop and manufacture new materials by:

- Developing a "materials innovation infrastructure" that includes:
  - **Computational tools**: software for predictive modeling, simulation, design and exploration
  - **Experimental tools**: synthesis and processing; quantitative characterization and analytic tools; accelerated testing and rapid prototyping; techniques to validate and advance materials theory
  - **Digital data**: data and interoperability standards for material properties; advanced data mining, analytic tools and open/proprietary data warehouses
  - **Collaborative networks**: integrated centers in computation, data informatics and experimentation; sharing of best practices across disparate centers via formal and informal networking; educational materials for the next generation workforce; public/private partnerships
- Building the materials innovation infrastructure while addressing high priority material problems of national importance
- Advancing a culture that supports the use of the materials innovation infrastructure and also embraces a more open, collaborative approach to developing advanced materials

MATERIALS GENOME INITIATIVE
Download the MGI White Paper
Download the MGI Presentation
Download the MGI Fact Sheet
Human Welfare
Clean Energy
National Security
Next Generation Workforce
Computational Tools
Experimental Tools
Digital Data
Materials Innovation Infrastructure

NIST is working to build the materials innovation infrastructure in support of the Administration's Materials Genome Initiative.

http://www.nist.gov/mgi

# NDS and domain repositories

- Domain repositories are discipline-specific
- Various business models in use; long-term sustainability is a major challenge*
- Potential NDS roles
  - Customizable data management and curation tools built on a common substrate
  - Access to cloud-like storage but at non-commercial rates
  - A directory of ontology-building and metadata management tools
  - A directory of domain repositories
  - Accreditation services
  - Advice, referral services, “genius bar”

* “Sustaining Domain Repositories for Digital Data: A White Paper,” C. Ember & R. Hanisch, eds. http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf

# Technologies/standards to build on

- Just use the VO standards!
  - OK, seriously… NIH syndrome
  - Much could be re-used in terms of architecture
  - Generic, collection-level metadata
- Cross-talk with Research Data Alliance (ANDS, EUDAT)
  - Data Citation WG
  - Data Description Registry Interoperability WG
  - Data Type Registries WG
  - Domain Repositories IG
  - Long Tail of Research Data IG
  - Metadata IG
  - Metadata Standards Directory WG
  - Preservation e-Infrastructure WG
  - and others…

*Dataverse, Dryad, iRODS, DSpace, etc.*

# Lessons learned re/ federation

- It takes more time than you think
  - Community consensus requires buy-in early and throughout
- Top-down imposition of standards likely to fail
- Balance requirements coming from a research-oriented community with innovation in IT
- Marketing is very important
  - Managing expectations
  - Build it, and they might come
- Coordination at the international level is essential
  - But takes time and effort
- Data models – sometimes seem obvious, more often not
- Metadata collection and curation are eternal but essential tasks

# Lessons learned re/ federation

- For example, the Cancer Biomedical Informatics Grid (caBIG) [$350M]
  - "…goal was to provide shared computing for biomedical research and to develop software tools and standard formats for information exchange."
  - "The program grew too rapidly without careful prioritization or a cost-effective business model."
  - "…software is overdesigned, difficult to use, or lacking support and documentation."
  - "***The failure to link the mission objectives to the technology shows how important user acceptance and buy-in can be.***" (M. Biddick, Fusion PPT)

J. Foley, *InformationWeek*, 4/8/2011

http://www.informationweek.com/architecture/report-blasts-problem-plagued-cancer-research-grid/d/d-id/1097068?