Next Steps for Enhancing Scientific Discovery from the Astrophysics Literature

Joseph M. Mazzarella
Project Scientist & Manager
NASA/IPAC Extragalactic Database
Caltech, IPAC

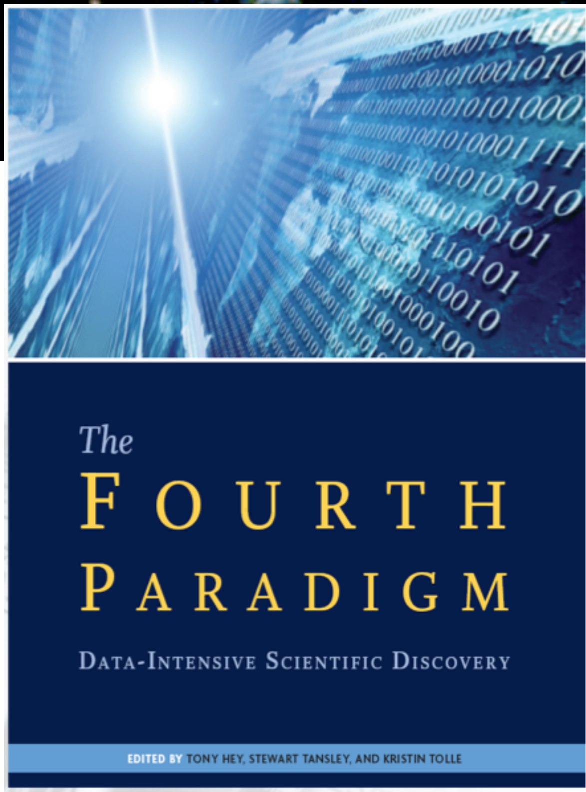
Outline

- *e-Science/X-info Context*
- *Data Capture and Multi-λ Fusion by NED*
- *Lessons Learned: Practical Obstacles to Automation*
- *Next Steps: Fulfilling the e-Science Vision*

AstroInformatics2010

Pasadena, June 16 - 19, 2010

AI2010 Context



As discussed in Tony Hey's talk …

Jim Gray's Call to Action

Listed 7 key areas for action by Funding Agencies:

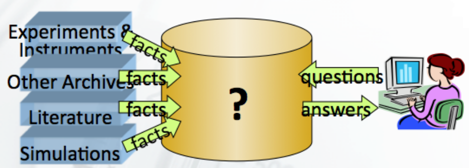
1. Fund both development and support of software tools
2. Invest at all levels of the finding 'pyramid'
3. Fund development of 'generic' Laboratory Information Management Systems
4. Fund research into scientific data management, data analysis, data visualization, new algorithms and tools

Remaining three key areas for action relate to the future of Scholarly Communication and Libraries:

5. Establish Digital Libraries that support the other sciences like the NLM does for Medicine
6. Fund development of new authoring tools and publication models
7. Explore development of digital data libraries that contain scientific data (not just the metadata) and support integration with published literature

X-Info

- The evolution of X-Info and Comp-X for each discipline X
- How to codify and represent our knowledge

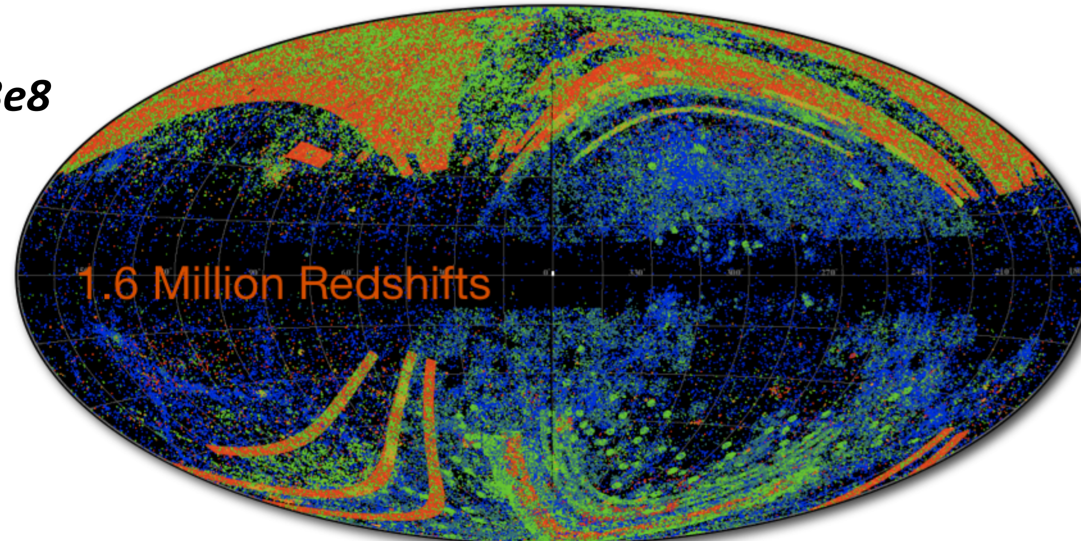


The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *reorganize* it
- How to share with others
- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation

NED: fusion of multi-λ data for extragalactic objects

New data are continuously integrated from the astrophysics *literature, missions, and sky surveys*



- ✧***Unique Sources (w/ Cross-IDs): 1.63e8***
- ✧***Multi-λ Source Cross-IDs: 1.77e8***
- ✧*Multi-λ Source Associations: 6.40e5*
- ✧***Galaxies with ≥ 1 Redshift: 1.61e6***
- ✧*Diameters: 6.09e8*
- ✧*Images: 2.25e6*
- ✧*Spectra: 2.91e5*
- ✧*Detailed Object Notes: 6.01e4*
- ✧*Redshift-Independent Distances: 3.55e4*
- ✧*Searchable Classifications & Derived Attributes: 1.67e5*
- ✧***5.61e6 Objects linked to 7.54e4 Journal Articles***
- ✧***1.72e9 Photometric measurements: Radio to Gamma Rays, SEDs, spectral lines***
- ✧*LEVEL5 Knowledgebase: Review articles linked to database queries ("Live Docs")*

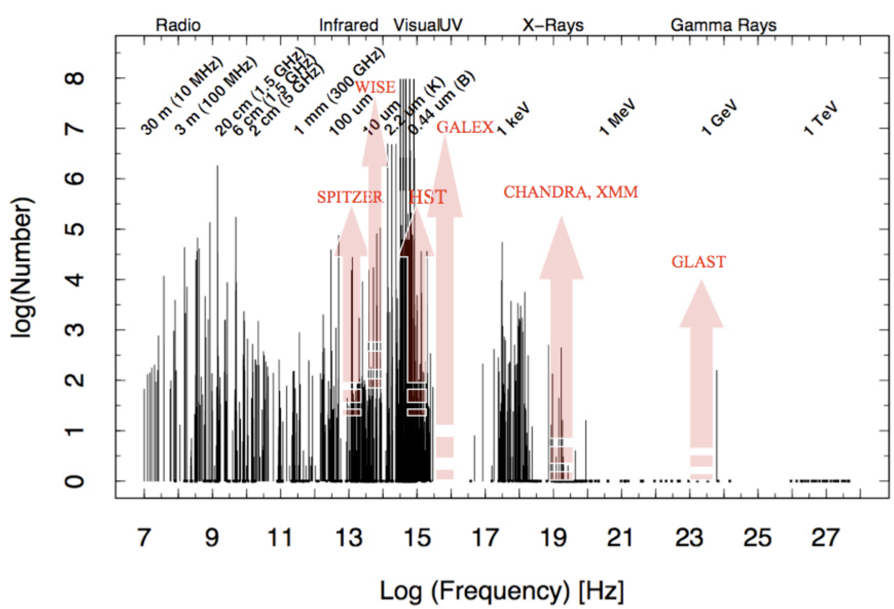
NED: Derived Quantities & Services

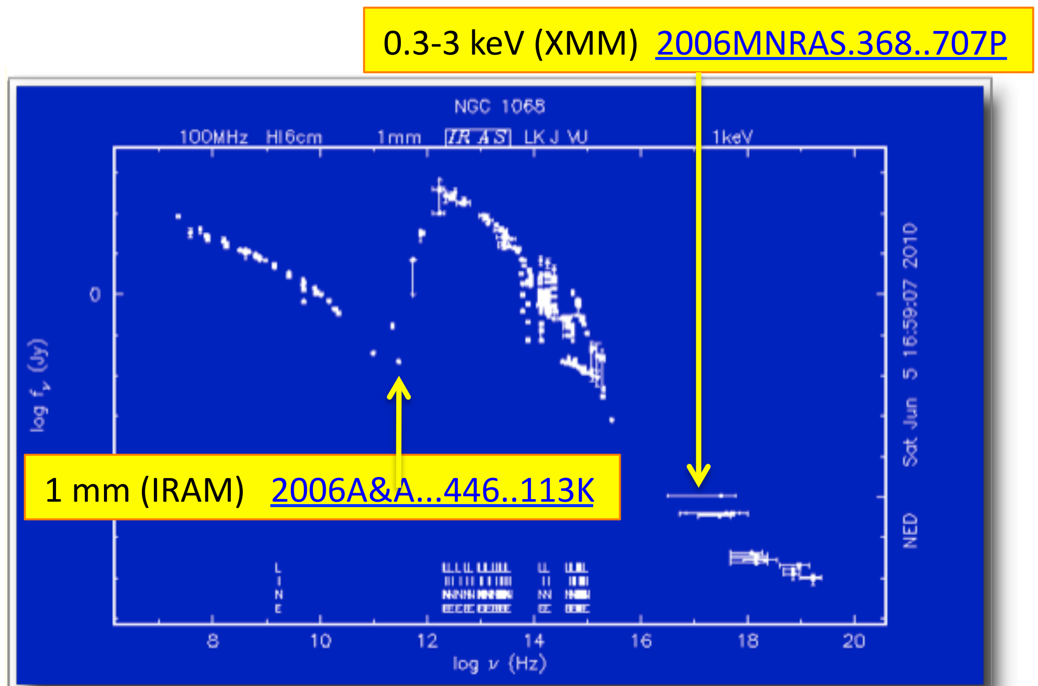
Core services

- Cross-correlation of extragalactic observations across the EM spectrum
- Fusion of information into a unified database
- Continuous refinements of data relationships as new information emerges
- Query by: Source Name, Position, Cone, Parameters, Attributes, Reference, etc
- *"Name Resolver" for best available source coordinates (utilizing X-IDs)*
- *Pointers to objects in journal articles shared with ADS*
- *Connectivity with distributed archives and tools*
- Value-added Derived Quantities
    - *Cosmological Quantities: Luminosity Distance, Scale, Surface Brightness Dimming*
    - *Radial velocity corrections to GSR, Local Group, CMB, Infall (Attractors)*
    - *Search in volumes around objects*
    - *NEW in 2010: Searchable attributes derived from galaxy Classifications*
        - *Ben Chan's talk on Saturday – Astrosemantics*

Importance of Information from the Literature

- ◆Knowledgebase upon which new observations are planned and analyzed
- ◆Links with parameters in large sky surveys are crucial for multi-λ research
- ◆Unique data in wavebands unavailable in large-scale surveys
  - ❖Large sky surveys (i.e., 2MASS, SDSS) provide many objects but limited λ coverage
  - ❖The widest λ coverage in SEDs comes from the literature





✧***1.72e9 photometric measurements, standardized flux units, and Cross-IDs ➔ SEDs***

NASA/IPAC EXTRAGALACTIC DATABASE

Search the NED spectral database

Help | Comment | NED Home

Object name: e.g.: 2dFGRS S899Z002, NGC 4321, NGC 4666, UGC 08335, 2QZ J011010.0-320009

Extended name search? Only if input name has 'multiple-components' Never Always

Result type: Simplified (with links to details) Detailed only

Show slit and spectrum previews? Yes No

Number of spectra per page: 20

Submit Query Reset

Optional Search Filters:

2.91e5 Spectra

Reference:

ANY
1992ApJS...80..109B -- Optical spectra of 87 Low-redshift Quasi-Stellar objects(QSO); Boroson and Green. (87 spectra)
1993ApJS...86....5K -- UV spectra of star-forming galaxies; Kinney et al. (143 spectra)
1995ApJS...98..129K -- Optical spectra of luminous infrared galaxies; Kim et al. (414 spectra)
1995ApJS...98..477H -- Optical spectra of nearby galaxy nuclei; Ho et al. (966 spectra)
1998A&AS..130..333T -- H I spectra of 905 Galaxies from LEDA; Theureau et al. (1806 spectra)
2000ApJ...532L..21H -- The Mid-Infrared Spectra of Normal Galaxies; Helou et al. (7 spectra)
2003A&A...402..849O -- Testing the Seyfert unification theory: Chandra HETGS observations of NGC 1068; Ogle et al. (2 spectra)
2003A&A...412...57P -- H I data for 848 galaxies from HYPERLEDA; Paturel et al. (848 spectra)
2003A&A...412..651C -- Optical Spectroscopy of BL Lac objects: New redshift and mis-identified sources; Carangelo et al. (2 spectra)
20032dF...C...0000C -- Optical spectra for 2dF Galaxy Redshift Survey; Colless et all. (233,221 spectra)
2004AJ....128...16K -- H I spectra of 1000 Brightest HIPASS Galaxies; Koribalski et al. (1,000 spectra)
2004MNRAS.349.1397C -- The 2dF QSO Redshift Survey - XII; Croom et al.(40,220 spectra)
2005AJ....129..559S -- ESO VLT Optical Spectroscopy of BL Lac objects: I.New Redshifts; Sbarufatti et al. (16 spectra)
2005ApJS..160..149S -- H I spectra of 8848 Galaxies; Springob et al.. (8,488 spectra)
2005A&A...430..373T -- 21-cm line measurements of 758 galaxies; Theureau et al. (758 spectra)
2006A&A...457...35S -- Optical Spectroscopy of BL Lac objects: Broad lines, companion galaxies, and redshift lower limits; Sbarufatti et al. (14 spectra)
2006AJ....132....1S -- ESO VLT Optical Spectroscopy of BL Lac objects: II. New Redshifts, featureless objects, and classification assessments; Sbarufatti et al. (38 spectra)
2007SINGS.5......0: -- SINGS spectroscopic survey of 75 nearby galaxies; Kennicutt et al. (756 spectra)
2007A&A...465...71T -- 21-cm line measurements of 550 galaxies; Theureau et al.
2009AJ....137..337S -- ESO VLT Optical Spectroscopy of BL Lac objects: III. An extension of the sample; Sbarufatti et al. (14 spectra)

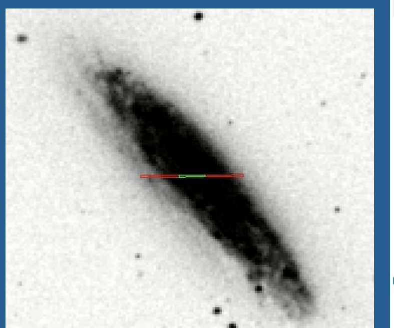
Bandpass:
ANY
Xray
Ultraviolet
Optical
Infrared
mm
Radio

Line:
ANY
HI
CO

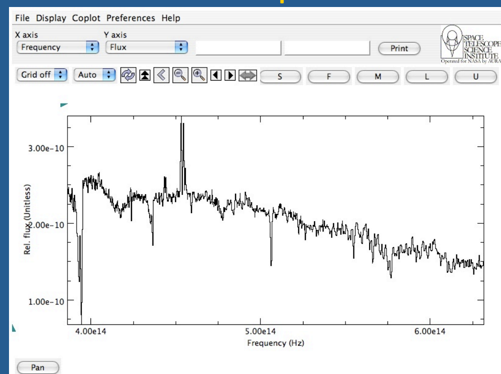
Submit Query Reset

NED Spectral Archive - Query Response 1

Aperture (slit) on DSS

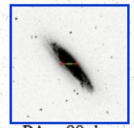


Launch Specview



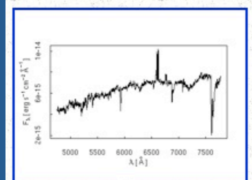
Extensive metadata

Slit Orientation



PA = 90 deg

Spectrum Previews



Launch Specview

Retrieve Data

FITS 4.7kb
Author-ASCII 25.7kb
NED-ASCII 131.6kb
VOTable 122.7kb

OBJECT INFORMATION

	Published Values		NED Cross-Identified Values	
Name:	NGC 4666		NGC 4666	
Position (J2000):	12h45m08.43s	-00d27m43.96s	12h45m08.68s	-00d27m42.90s
Origin: Published-Value System: Equatorial Type: sexagesimal Equinox: B1950	12h42m34.6s	-00d11m20s		

OBSERVATIONAL DETAILS

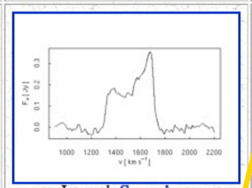
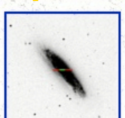
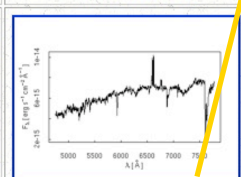
Observed Region:	Nucleus	Telescope Name:	MKO 2.2m
Band:	Optical	Instrument Name:	Faint Object Spectrograph
Start Channel:	4747.7Å	Disperser Type:	Blazed-Grating
End Channel:	7770.0Å	Detector Type:	CCD
Channel Step:	2.9Å	Observation Mode:	Long-Slit
Spectral Resolution:	7.0Å	Aperture/Beam Size:	1.5 arcsec × 74.0 arcsec
Mean Resolving Power:	894.1	Extracted Region Size:	1.5 arcsec × 18.85 arcsec
Seeing:	1.00 arcsec	PA of Slit:	90 deg
Airmass:	1.90	Dimension of Spectrum:	1 D
Integration Time:	400 sec	Number of Pixels:	1025
Flux Units:	$ergs^{-1}cm^{-2}Å^{-1}$	Reference Frame:	Observed
Flux Calibration:	Yes		
Absolute Calibration:	Yes	Reference:	1995ApJS...98..129K
Sky Subtraction:	Yes		
Notes:	N/A		

Basic metadata:

NASA/IPAC EXTRAGALACTIC DATABASE

Help | Comment | NED Home

Spectral data in NED archive for object NGC 4666

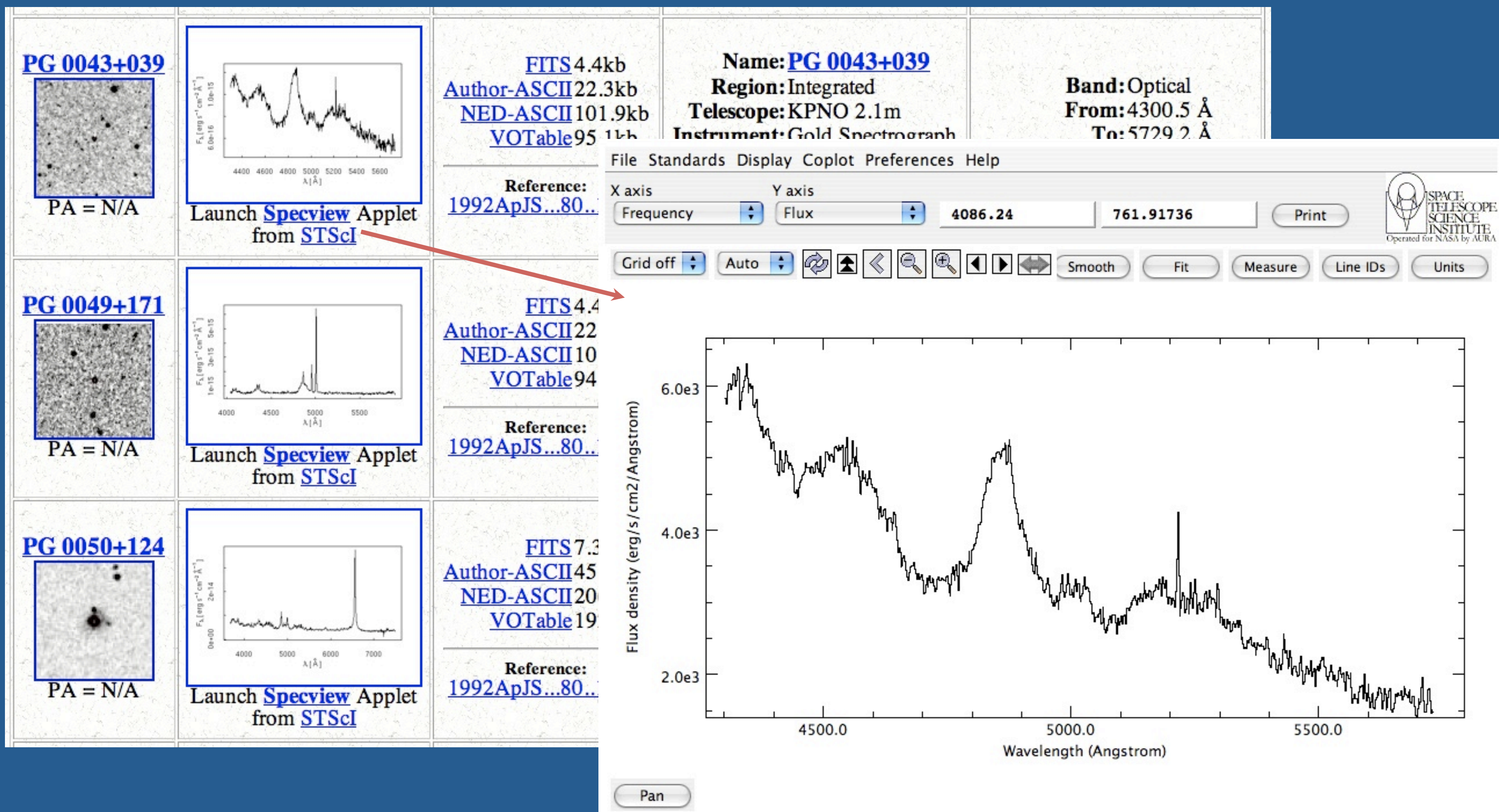
Slit Orientation	Spectrum Previews	Retrieve Data	Observational Information	Spectral Coverage & Resolution
 PA = N/A	 Launch Specview	FITS N/A Author-ASCII 4.4kb NED-ASCII 12.9kb VOTable 12.5kb External Resource Reference: 2004AJ....128...16K	Region: Integrated Telescope: Parkes Instrument: 21 cm Multibeam Receiver Abs-Cal: Yes Ref-Frame: Observed Full description	Line: HI From: $906.2 kms^{-1}$ To: $2199.1 kms^{-1}$ Step: $13.2 kms^{-1}$ Resol^{n}: $18.0 kms^{-1}$
 PA = 90 deg	 Launch Specview	FITS 4.7kb Author-ASCII 25.7kb NED-ASCII 131.6kb VOTable 122.7kb Reference: 1995ApJS...98..129K	Region: Nucleus Telescope: MKO 2.2m Instrument: Faint Object Spectrograph Abs-Cal: Yes Ref-Frame: Observed Full description	Band: Optical From: 4747.7Å To: 7770.0Å Step: 2.9Å Resol^{n}: 7.0Å

Remote resource:
HI Parkes All-Sky Survey (HIPASS)

NED Spectral Archive - Query Response 2

Optical spectra of 87 Low-Redshift QSOs

Boroson & Green 1992, ApJS, 80, 109



Challenges with (Meta)Data in the Literature - 1

Lessons learned "in the trenches" by the NED Team while capturing and fusing data from >75,000 journal articles and catalogs

- Substantial problems with data published in the literature
  - *Create impediments to automating extraction and analysis*
  - *Limit efficiency of data integration*
  - *Produce unnecessary obstacles to our complete e-Science vision*
- Meta-Problem: Data are not refereed to same extent as science results
  - *"There are no rules." (Well, OK, some minimal rules)*
  - *Every paper presents data details differently*
  - *A multi-faceted problem as summarized next…*

Challenges with (Meta)Data in the Literature - 2

- Problem (1): Data are not published following standards
  - *README files (created from LaTeX tables) are very useful, but limited*
  - *Uniform column descriptors lacking*
    - *Ex: multiple ways to define photometric filters*
    - Ex: CIT vs. Johnson H mag impacts SED
  - *Crucial (meta)data is scattered throughout the text (unstructured)*
  - *Traditional keywords are very limited*
- Solution (1): Coordinate with journal publishers and authors to make application of VO standards user-friendly
  - *Hide complexities of data models "under the hood" of web-based UIs*
  - *Enforce uniformity without authors realizing it's happening*
  - *Q&A interview type approach for beginners; scripted approach for experts*
  - *Apply a rich set of standard key phrases (tags with controlled vocabulary)*

Challenges with (Meta)Data in the Literature - 3

- Problem (2): Data often "published" on personal URLs (in footnotes)
  - *Rarely refereed to meet standards of formatting, column descriptions, etc.*
  - *Often data move (w/ data providers) or disappear completely => unstable*
  - *The paradigm of self-publishing data is flawed*
- Solution (2): Require data to be deposited in established data center repositories accessible with VO standards
- Problem (3): IAU rules for source nomenclature often not followed
  - *Ambiguous object names (i.e., SN named "Bob"!?)*
  - *Truncation of coordinate-based source IDs; uncertainties in coords omitted*
  - *Makes establishment of cross-IDs/associations very difficult (if not impossible)*
- *Solution (3): With journal article submission*
  - *Include list of valid astronomical object IDs with articles*
  - *Validate using name resolvers (NED/NStED/SIMBAD) <u>before publication</u>*

Challenges with (Meta)Data in the Literature - 4

- Problem (4): Data “behind the plots” not part of traditional publications
  - *Greatly reduces reproducibility and ability to build upon published results*
  - *Valuable highly-processed images and spectra are*
  - *NED captures what it can after publication; but only a fraction*
- Solution (4): Repositories for highly-processed images and spectra with metadata required for utilization in science queries
  - *Expand upon NED’s approach for author-contributed images and spectra*
  - *Capture data (and metadata) at the time of publication*
    - *Get authors involved before they lose the data and/or lose interest (move on)*
    - *Capture many more images and spectra than currently possible*
    - *Enforce uniform VO standards (under the hood) via web-based tools*
  - Broaden to all astrophysics disciplines with specialized repositories
    - Exoplanet systems
    - Theory data sets
    - Galactic targets: moving objects, nebulae, etc.
    - Any astrophysics data lacking a long-term “home” for curation and integration
    - Enable multi-disciplinary collaborations via coordination with DataNet
  - Replication

Challenges with (Meta)Data in the Literature - 5

- Problem (5): e-journals keep changing formatting/markup
  - Structures adequate for human readers are not suitable for automated data analysis software
  - Programs that follow links to retrieve e-tables, seek metadata, keywords, etc. need constant maintenance to keep up with changes at journal cites
  - Even within a single journal (i.e., ApJ) there are multiple digital formats depending on the year of publication
  - Past and current data sets require human expertise to pick up where automation fails: locate, re-format and apply uniform metadata before data can be integrated into DBMS and served with VO standards
- Solution (5): Convince publishers to provide stability and uniformity
  - See following meta-problem and proposed solution

Challenges with (Meta)Data in the Literature - 6

- Meta-Problem: Limitations in past and current data sets require human expertise to pick up where automation fails
  - Locate data and metadata (often scattered throughout paper)
  - Re-format and apply uniform metadata so that data can be cross-matched, integrated into DBMS and served with VO standards

Challenges with (Meta)Data in the Literature - 7

- Meta-Solution: Coordination between VAO, journal publishers and authors to apply standards and capture (meta)data more efficiently
  - Coordination of AI/VAO community with AAS Pub Board, IOP, authors
  - What about other journals? Another challenge for IVOA:
    - Can this be done in lock-step with Euro journals: A&A, MNRAS?
    - Can this be done in lock-step with Asian journals? Others?
  - On what time-scale will traditional journals evolve with AI community?
  - If collaboration with traditional journal publishers fails (or moves too slowly): collaborate with arxiv.org and/or PLoS to develop new, refereed e-journal(s) linked intimately with VO repositories and query services?
  - Will the community buy in?

Maximizing Science Impact - 1

- Build upon what's working and expand on it
  - NED image archive took ~10 years to reach current take-up
  - NED spectral archive now at about same stage as images at ~2-year mark
  - Metadata are as important as the data
  - Only scratches the surface of what's possible (for reasons outlined here…)
- Keep the barriers low: ***Very user-friendly tools are needed***
  - Or take-up by the community will be minimal
  - Domain knowledge (astronomers in the loop) is essential
- Automated tools are needed to "referee" (meta)data
  - Replace links to personal web sites with established data repositories
  - Validate adherence to VO standards

Maximizing Science Impact - 2

- The benefits of a small amount of extra work for authors of journal articles will enable
  - More rapid and error-free integration of information into services that astronomers use heavily today (NED, NStED, ADS, SIMBAD…)
  - The entire universe of emerging and future VO/AI services
  - Long-term preservation of unique, highly-processed data
  - Enhanced discovery capabilities
  - More reuse of data, leading to
  - More discoveries, publications, and citations to publications
  - Happy funding agencies

Maximizing Science Impact - 3

- The potential for the VAO to support our dream of advanced capabilities in Astroinformatics (highly automated, seamless knowledge discovery) depends on rapid solutions to these fundamental limitations of the way data and metadata are traditionally published
- We can do better, but it will take a major coordination effort and the cooperation of authors, e-journal publishers (AAS Pub Board, IOP), VAO, and its partners
- Our goal is a major transformation of the methodology and culture of publishing data and metadata with the derived knowledge to greatly enhance science capabilities

Summary of Specific Next Steps

- Goal: capture all tables, images and spectra "behind the plots" in journal articles into VO-enabled repositories
- Build on experience of NED, ADS, NVO, IVOA, journal publishers, and emerging DataNet project, coordinated by VAO
- Engage authors during the publication process
- Apply VO standards to data and metadata
- Provide easy-to-use tools to hide underlying complexity of DMs, etc.
- Enhance linkages between literature, objects and data
- Enable increasingly sophisticated Astroinformatics applications to boost science discovery capabilities
- ***This work has high priority in the 2010 VAO Science Council report, and the Program Council is working on an implementation plan***