*Astroinformatics*
*Proceedings IAU Symposium No. 325, 2016*
*M. Brescia, S.G. Djorgovski, E. Feigelson,*
*G. Longo & S. Cavuoti, eds.*


doi:10.1017/S1743921316013132

# Evolution of the NASA/IPAC Extragalactic Database (NED) into a Data Mining Discovery Engine

**Joseph M. Mazzarella and the NED Team**[1]

[1]California Institute of Technology,
IPAC, MS 100-22, Pasadena, CA 91125
email: mazz@ipac.caltech.edu

**Abstract.** We review recent advances and ongoing work in evolving the NASA/IPAC Extragalactic Database (NED) beyond an object reference database into a data mining discovery engine. Updates to the infrastructure and data integration techniques are enabling more than a 10-fold expansion; NED will soon contain over a billion objects with their fundamental attributes fused across the spectrum via cross-identifications among the largest sky surveys (e.g., GALEX, SDSS, 2MASS, AllWISE, EMU), and over 100,000 smaller but scientifically important catalogs and journal articles. The recent discovery of super-luminous spiral galaxies exemplifies the opportunities for data mining and science discovery directly from NED's rich data synthesis. Enhancements to the user interface, including new APIs, VO protocols, and queries involving derived physical quantities, are opening new pathways for panchromatic studies of large galaxy samples. Examples are shown of graphics characterizing the content of NED, as well as initial steps in exploring the database via interactive statistical visualizations.

**Keywords.** galaxies: general, galaxies: statistics, astronomical data bases, surveys, catalogs

## 1. Introduction

Research topics in many areas of extragalactic astrophysics and cosmology require a reliable fusion of observational data across the EM spectrum. For example: What is accelerating the expansion of the universe (dark energy)? What constitutes about 83% of all matter but is invisible (dark matter)? How do the cosmic web and environments of galaxies influence their evolution? What is the history of galaxy assembly and star formation over cosmic time? How do properties of galaxies and AGN (fueling of super-massive black holes) co-evolve? At the same time, thousands of journal articles are being published per year containing data sets with increasing size and complexity. Further, space missions and ground-based telescopes are generating immense data archives spanning from gamma rays through radio frequencies. Therefore, research teams and planners of future missions are faced with an ever increasing challenge of staying current and leveraging this wealth of data to investigate these and other questions.

The NASA/IPAC Extragalactic Database (NED†) is an information system provided for the astronomical community that facilitates and accelerates multi-wavelength research on objects beyond our Milky Way galaxy that would otherwise be impossible or impractical to accomplish. The NED team is continuously integrating data from the literature, space mission archives, and large sky surveys to produce and serve a comprehensive census of the observed universe. Currently the database consists of information synthesized from NASA missions such as Spitzer and HST, large surveys such as GALEX and SDSS, and

† http://ned.ipac.caltech.edu

information gleaned from more than 103,000 journal articles, catalogs, and astronomical telegrams. Content includes object names, coordinates, redshifts, redshift-independent distances, fluxes, sizes, classifications and attributes, along with derived quantities such as cross-identifications, redshift-based distances, metric sizes, spectral energy distributions (SEDs), foreground Galactic extinction estimates, luminosities, velocity corrections, and cosmological corrections. A repository for images and spectra contributed by authors of journal articles ("data behind the plots") is also supported to simplify reproducibility of published results and, joined with other data in NED, to aid in making new discoveries.

Catalogs are not duplicated or served in their original form. Rather, selected data are fused in a unified data model to enable panchromatic queries. NED simplifies and accelerates extragalactic research by distilling and synthesizing data across the spectrum and providing value-added derived quantities. This enables astronomers to find answers to queries such as: What objects have z > 2.0 and an available flux in the GALEX NUV band? What is the most precise z-independent distance measurement to M82? What is the SED for quasar 3C 279? Which spiral galaxies have stellar bars and Type 2 AGNs?

Objects can be queried by name, near name or near position (cone search), by reference, and by author. Galaxy samples can be constructed by parameter constraints on redshift, sky area, object types, survey names, flux density (mag), or by filtering galaxy classifications and attributes. In addition, the LEVEL 5 Knowledgebase augments review articles in extragalactic astrophysics and cosmology with object names and graphical content within the articles linked directly to relevant database queries. These and other NED capabilities have been described in more detail elsewhere (e.g., Helou *et al.* 1990, Mazzarella *et al.* 2007, 2014). This work summarizes recent advances and ongoing work in evolving NED beyond an object reference database into a data mining discovery engine.

## 2. Recent advances

### 2.1. *Cross-matching*

A process intended to capture two decades of experience and expertise in cross-matching sources from astronomical catalogs in the context of integrating selected data into NED has recently been codified and extended to include a probabilistic algorithm in a computer program called *MatchEx*. The local density of objects is used to estimate the background contamination rate, and Poisson statistics are used to balance completeness versus reliability of matches. Scientific vetting is applied to the results of trial runs in order to tune parameters and refine the algorithms. Further information is presented by Ogle *et al.* (2015). Ongoing and future enhancements will go beyond proximity, adding comparisons of redshifts, object classifications, sizes, and fluxes to the matching decisions.

A recent re-implementation of *MatchEx* with parallel processing is enabling effective and efficient cross-matching of surveys with $10^7 - 10^9$ sources, in addition to maintaining applicability to smaller data sets extracted from the literature. Most recently, *MatchEx* was used to process 42 million sources from the Spitzer Enhanced Imaging Products (SEIP) Source List (Teplitz *et al.* 2012), yielding 37 million new objects and 5 million cross-identifications to prior NED objects. Over 360 million photometric measurements in four IRAC bands (3.6, 4.5, 5.8, 8.0 $\mu$m) and the MIPS 24 $\mu$m band have been integrated into the SEDs for NED objects; Figure 1 shows an example.

### 2.2. *Meeting Big Data Challenges*

Many challenges confronted by the NED team pertain to the four "V's of Big Data".

Volume. Over the next few years, NED will be growing by more than 10-fold to support research with data joined for $\sim 2$ billion objects and $\gtrsim 20$ billion attributes (Fig. 2). This

https://doi.org/10.1017/S1743921316013132 Published online by Cambridge University Press

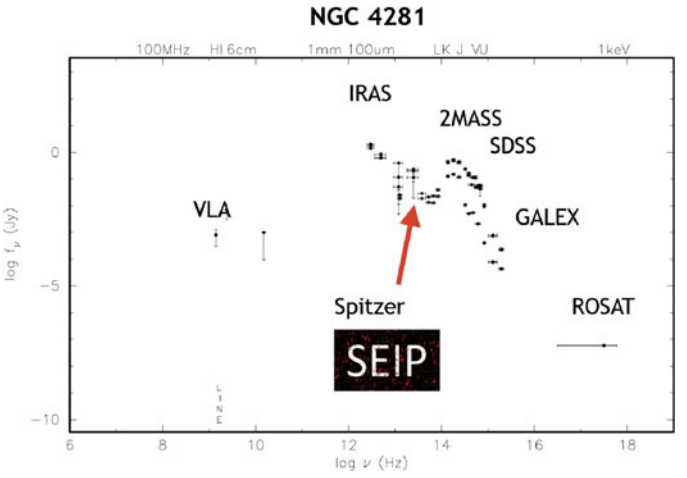


**Figure 1.** SED of NGC 4281 exported directly from the NED interface, with annotations added to illustrate the recent integration of photometry from the Spitzer Enhanced Imaging Products (SEIP) Source List with other major surveys in NED.

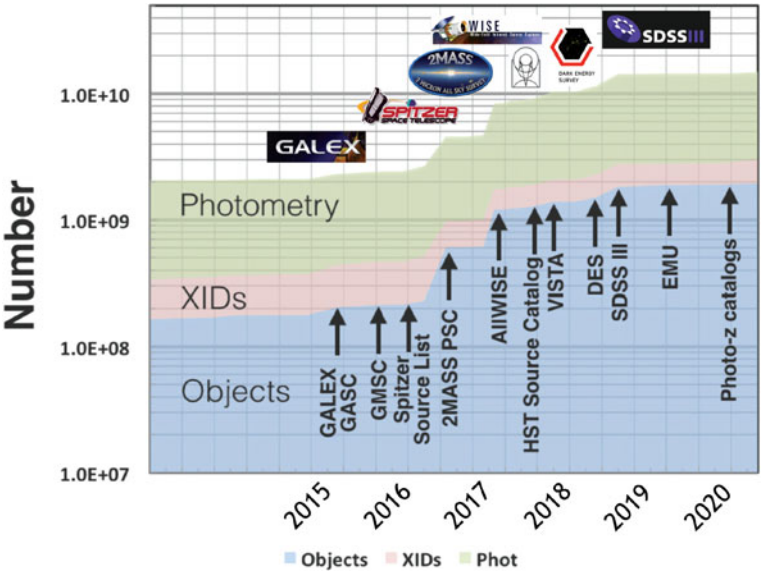


**Figure 2.** Recent and projected growth of NED is dominated by very large sky surveys.

is being accomplished by upgrading storage and servers; refactoring the database schema to improve scalability, extensibility, and performance; and supporting queries on complex regions via spatial indexing using PostgreSQL extensions (Q3C, PgSphere).

*Variety.* NED spans gamma rays through radio frequencies, including data from over 22 NASA missions and 100,000 articles and catalogs, each presenting their data differently. These challenges are being met by continuously updating data capture and fusion techniques, and serving (meta)data using VO standards for interoperability with analysis tools. A summary of featured NED holdings at the time of writing is given in Table 1.

*Velocity.* Data are being published in sky surveys and the literature at an increasing rate. This challenge is being met by accelerating cross-matching and data integration via parallel processing, replacing legacy data processing tools with a new pipeline

https://doi.org/10.1017/S1743921316013132 Published online by Cambridge University Press

**Table 1.** Featured holdings as of 2016 October.

| Data type | Number |
|---|---|
| Photometric measurements | $2.33 \times 10^9$ |
| Diameters | $6.09 \times 10^8$ |
| Multi-wavelength source XIDs | $2.98 \times 10^8$ |
| Distinct astrophysical objects | $2.52 \times 10^8$ |
| Object links to references | $3.70 \times 10^7$ |
| Redshift measurements | $7.15 \times 10^6$ |
| Objects with $\geqslant 1$ redshift | $5.27 \times 10^6$ |
| Images | $2.53 \times 10^6$ |
| Spectra | $5.73 \times 10^5$ |
| Object classifications | $5.02 \times 10^5$ |
| Redshift-independent distances | $1.06 \times 10^5$ |
| References | $1.03 \times 10^5$ |

developed in Python, Perl and C, as well as a small pilot project to apply machine learning techniques to classification of data types in the literature to streamline ingestion.

*Veracity*. Data in the literature are not refereed to the same degree as scientific results; thus there are often issues that impede efficient processing such as incomplete (meta)data, missing uncertainties or observation time-stamps, or ambiguous object identifiers. To help mitigate these issues, the NED team has published *Best Practices for Data Publication to Facilitate Integration into NED* (Schmitz *et al.* 2014)† as a reference guide for authors and referees. The publishers of MNRAS have included a link to this document in their Instructions to Authors, and the AAS journals (ApJ, AJ) intend to do so soon.

### 2.3. *Functionality*

Recent improvements to the user interface include support for asynchronous queries with long run times, connectivity to the IRSA Finder Chart service from NED image query reports, and access to the image archive through the VO Simple Image Access (SIA) protocol. Implementation of a VO Table Access Protocol (TAP) service is in progress, which will enable 'power users' to run ADQL queries against the NED object directory. A tool to explore the environments of galaxies with available redshifts in NED is also experiencing increasing usage and is cited in recent research (e.g., Hagen *et al.* 2016).

## 3. Discovery Using NED

*Super-luminous spiral galaxies*. The recent discovery of a new class of super-luminous spiral galaxies based on data synthesized within NED (Ogle *et al.* 2016) demonstrates its power as a discovery engine. At optical wavelengths, super spirals are as luminous as the brightest elliptical galaxies in clusters. Some basic properties include: $L_r > 8 \times L^*, D = 60 - 130$ kpc, $M_* = 30 - 300 \times 10^9 M_\odot$, and SFR $= 5 - 65$ $M_\odot$/yr. This previously unknown class challenges theories of galaxy formation and evolution. Data joined within NED that enabled this work includes redshifts, object types, diameters, SEDs and derived quantities: luminosity (SDSS), stellar mass (2MASS), and SFR (GALEX, Herschel).

*Science with z-independent distances*. Distance measurements independent of redshift are fundamental to astrophysics. In 2006, NED began providing a comprehensive compilation of redshift-independent extragalactic distance estimates, referred to as NED-D. A decade later, this compendium contains > 100,000 estimates for > 28,000 galaxies

† http://ned.ipac.caltech.edu/docs/BPDP/NED_BPDP.pdf

https://doi.org/10.1017/S1743921316013132 Published online by Cambridge University Press

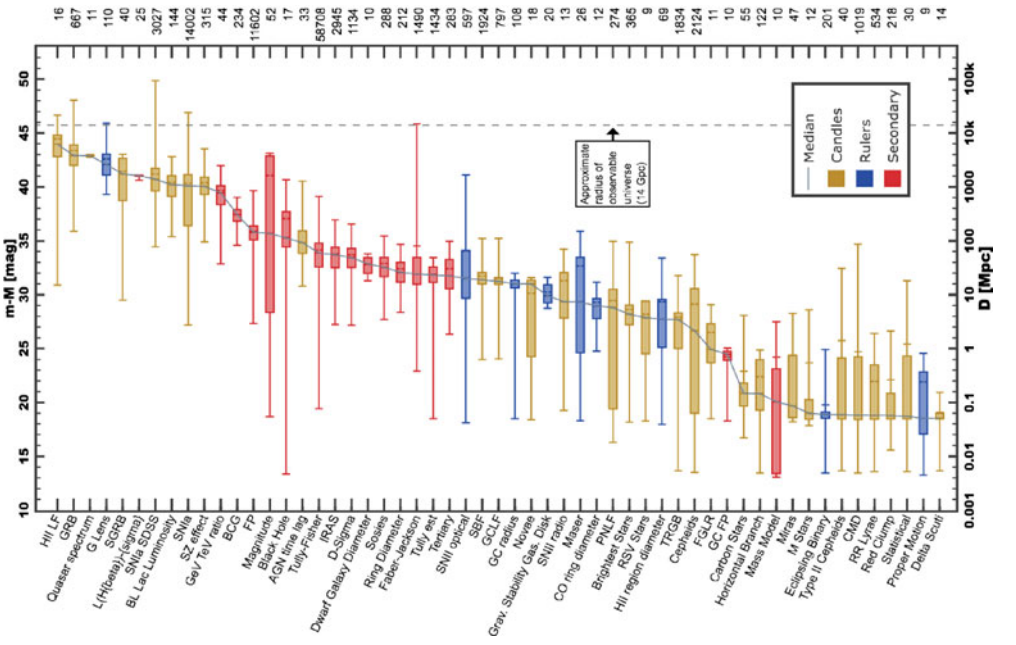


**Figure 3.** Comparison of redshift-independent distance indicators, shown in order of decreasing median distance. For indicators with $\geqslant$ 9 estimates, a 'box plot' represents the distribution of the distances: the left (lower) and right (upper) sides of each box represent the 25th and 75th percentiles, and lines extend to the minimum and maximum values. From Steer *et al.* (2017).

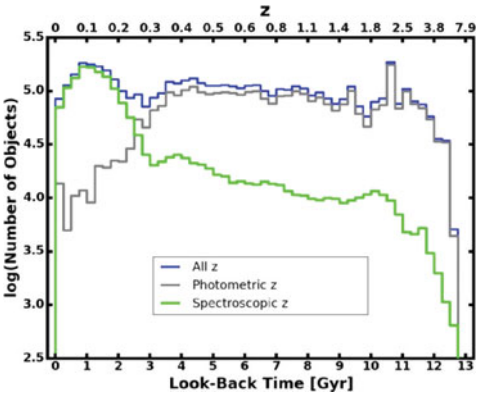


**Figure 4.** The number of objects in NED (Nov 2016) versus redshift and cosmological look-back time, with spectroscopic redshifts in green, photometric redshifts in gray, and their sum in blue.

collated from > 2,000 references. A recent article by Steer *et al.* (2017) describes the methodology, content, and uses of NED-D. Currently 75 different distance indicators (methods) are in use in the astrophysics literature. For 55 methods with more than 8 distance estimates available, Fig. 3 presents a visualization that compares the indicators in terms of their ranges of applicability. This and other figures in the article are static versions of interactive graphics that will be updated on the website as the available data grow, to enhance exploration and understanding of the data.

*Visualizations of content and completeness.* The NED team is developing visualizations of the content of NED, and completeness with respect to published models, in

https://doi.org/10.1017/S1743921316013132 Published online by Cambridge University Press

order to facilitate query formulation and statistical studies. One example is shown in Figure 4, and others are being added and will be kept up-to-date on the web interface.

*Next few years and beyond*. Over the next few years, the number of NED objects with photometry in 3-6 spectral regions will be growing ∼10 fold, enabling SED analysis with broad coverage for millions of galaxies. Other new and improved services that are in development and planned include: completion of the transition to the new user interface; information about galaxy memberships in published pairs, groups, and clusters (hierarchy); information about survey coverage (footprints); K-corrections between observed and rest-frame photometric measurements; source list uploads and improvements to customized tabular output, and support for ADQL queries using the IVOA Table Access Protocol (TAP). We are improving capabilities of science queries on observations joined across missions, along with derived physical quantities, while also streamlining NED access from popular analytics environments such as Python, R and visualization tools.

## 4. Summary

Use of NED is no longer dominated by scientists interactively looking up a few facts about their favorite galaxies one at a time; it is dominated by programmed queries. The discovery of a new class of super-luminous spiral galaxies using NED's unique data synthesis demonstrates its power as a discovery engine. The team is continuously evolving the system with advances at the nexus of astronomy and informatics. The system is growing to serve data fused across the spectrum for billions of galaxies, and delivering new capabilities to exploit this unique resource for scientific discovery.

NED is operated by the California Institute of Technology, under contract with NASA. Current NED team members are Kay Baker, Ben Chan, Tracy Chen, Rick Ebert, Cren Frayer, George Helou, Jeff Jacobson, Tak Lo, Barry Madore, Joseph Mazzarella, Patrick Ogle, Olga Pevunova, Ian Steer, Marion Schmitz, and Scott Terek.

## References

Helou, G., Madore, B. F., Bicay, M. D., *et al.* 1990, in: G. Fabbiano, J. S. Gallagher, & A. Renzini (eds.), *Windows on Galaxies*, Astrophysics & Space Science Library, 160, 109

Hagen, L. M. Z., *et al.* 2016, *ApJ*, 826, 210

Mazzarella, J. M., & NED Team, 2007 in: R. Shaw, F. Hill & D. Bell (eds.), *ADASS XVI*, ASP Conference Series, 376, 153

Mazzarella, J. M., *et al.* 2014, *AAS Meeting* #223, #302.04

Ogle, P. M., Mazzarella, J., & Ebert, *et al.* 2015, in: A. R. Taylor & E. Rosolowsky (eds.), *ADASS XXIV*, ASP Conference Series, 495, 25

Ogle, P. M., Lanz, L., Nader, C., & Helou, G. 2016, *ApJ*, 817, 109

Schmitz, M., Mazzarella, J. M., Madore, B. F., *et al.* 2014, *AAS Meeting* #223, #302.05

Steer, I., Madore, B., Mazzarella, J. M., *et al.* 2017, *AJ*, in press

Teplitz, H. I., Capak, P., Hanish, D., *et al.* 2012, *AAS Meeting* #219, #428.06