EPJ Web of Conferences **186**, 03001 (2018) https://doi.org/10.1051/epjconf/201818603001
LISA VIII

Evolution of the NASA/IPAC Extragalactic Database (NED) into a Data Mining Discovery Engine. II. Current Contents and Future Plans

Marion Schmitz[1,2,*]

[1] *California Institute of Technology*
[2] *IPAC, MS 100-22, Pasadena, CA 91125*

Abstract. Recent advances have been made in evolving the NASA/IPAC Extragalactic Database (NED) into a data mining discovery engine. Infrastructure changes and data integration techniques are enabling more than a 10-fold expansion. NED will soon contain over a billion objects with their fundamental attributes (such as names, positions, redshifts, fluxes, and diameters) determined via cross-identifications among the largest sky surveys and over 100,000 smaller but scientifically important catalogs and journal articles. In addition, enhancements to the user interface, including new APIs, VO protocols, and queries involving derived physical quantities, will provide new pathways for multi-wavelength studies of large extragalactic samples.

1 Introduction

The NASA/IPAC Extragalactic Database (NED[1]) is an information system provided for the astronomical community that facilitates and accelerates multi-wavelength research on objects beyond our Milky Way that would otherwise be impossible or impractical to accomplish. The NED team is continuously integrating data from the literature, NASA 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 over 22 NASA missions, large surveys such as GALEX and SDSS, and information gleaned from over 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 (XIDs), 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 making new discoveries in combination with other data in NED.

*e-mail: mschmitz@ipac.caltech.edu ORCID: 0000-0002-2055-7549
[1] http://ned.ipac.caltech.edu/

2 NED Basics

At the time of this writing (July 2017) the NED contents included 253 million objects (with names and positions) obtained from 108 thousand distinct references. Data for these objects includes photometry (2.3 billion data points), redshifts (over 11 million), redshift-independent distances for 27,000 objects, 2.5 million images, and 600,000 spectra. The Level5 Knowledgebase contains 893 articles of interest for extragalactic astronomy and cosmology.

The NED interface allows the database to be queried by name, near name or position (cone search), by reference, and by author. Galaxy samples can be constructed by parameter constraints on redshift, sky area, object types, survey names, and flux density (magnitudes), or by filtering galaxy classifications and attributes. In addition, the LEVEL5 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). Refer to Paper I (Mazzarella et al. 2017) for more information about the evolution of NED.

3 NED and Big Data

3.1 The four "V"'s

(The following is extracted from Paper I.)

Many challenges confronted by the NED team pertain to the four "V's of Big Data" - Volume, Variety, Velocity, and Veracity.

Volume: Over the next few years, NED will be growing by more than 10-fold to support research with data joined for about 2 billion objects and more than 20 billion attributes. This 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 the entire spectral domain from gamma rays through radio frequencies. 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.

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 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]) [2] 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] http://ned.ipac.caltech.edu/docs/BPDP/NED_BPDP.pdf

3.2 Software improvements

Since 1990 the NED team have been gaining experience in cross-matching sources in astronomical catalogs in the context of integrating selected data. This expertise has recently been codified and extended to include a probabilistic algorithm in an in-house 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 implementation of MatchEx with parallel processing enables effective and efficient cross-matching of surveys with tens of millions of sources, in addition to maintaining applicability to small data sets extracted from the literature. Most recently, MatchEx was used to process 42 million sources from the Spitzer Enhanced Imaging Products Source List, (SEIP [3]) yielding 37 million new objects and 5 million cross-ids to prior NED objects. Over 360 million photometric measurements in four IRAC bands and the MIPS 24micron band have been integrated into the SEDs for NED objects.

3.3 Database Content additions

In addition to the basic information that NED has traditionally contained (see Section 2), the next few years will see new data and meta-data added in order to enhance data discovery. Galaxy memberships in published pairs, groups, and clusters will be systematically added. Details about the observations (such as the telescope, instrument, and filters used) will be searchable. Survey coverage (footprints) of major surveys will be included.

3.4 Improvements of User Access to the Database

Recent improvements to the user interface include support for asynchronous queries with long run times, connectivity to the IRSA Finder Chart [4] 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.

Source list uploads and improvements to user-specified tabular output will be implemented. 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 World-wide Coordination

NED is a resource which relies on coordinated efforts with other resources worldwide.

The authors who provide the published and peer-reviewed data are essential to NED. Additional involvement will include requests that the authors provide NED with data ready for ingestion and to help more in the validation of published data.

[3] http://irsa.ipac.caltech.edu/data/SPITZER/Enhanced/SEIP/docs/seip_explanatory_supplement_v3.pdf
[4] http://irsa.ipac.caltech.edu/applications/FinderChart

NED and astronomy librarians have always had a close working relationship. Plans are underway to incorporate telescope bibliographies [5] and the Unified Astronomical Thesaurus (UAT [6]) which have been coordinated by the librarians.

Journal publishers have been and continue to be helpful in allowing NED access to the original journal articles as well as the data tables and supplementary materials contained within them. Mutual linking between the publishers and NED makes access to each more efficient for the researcher.

Three decades of close coordination of NED with CDS and ADS has proven beneficial to all three projects. NED and CDS share common challenges and have often worked together to determine appropriate solutions. These collaborative exchanges will continue to be helpful in the future.

NED's involvement with the NASA Astronomical Virtual Observatories (NAVO [7]) and the International Virtual Observatory Alliance (IVOA [8]) via contributions and implementations will continue.

5 Summary

NED usage had long been dominated by scientists interactively looking up a few facts about their favorite galaxies one at a time. Although this type of access will continue, current usage is dominated by programmed queries. The NED team is continuously evolving the hardware, software, and data content in response to advances in 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, Olga Pevunova, Ian Steer, Marion Schmitz, and Scott Terek.

References

[1] G. Helou, B.F. Madore, M.D. Bicay et al. , *Windows on Galaxies* edited by G. Fabbiano, J. S. Gallagher, A. Renzini (1990), Vol. 160 of Astrophysics & Space Science Library, p. 109

[2] J.M. Mazzarella, and NED Team, *NED for a New Era* in *Astronomical Data Analysis Software and Systems XVI* edited by R.A. Shaw, F. Hill, D.J. Bell (2007), Vol. 376 of Astronomical Society of the Pacific Conference Series, p. 153–162

[3] J.M. Mazzarella, P.M. Ogle, D. Fadda et al., *xplosive Growth and Advancement of the NASA/IPAC Extragalactic Database (NED)* in *American Astronomical Society Meeting Abstracts #223* (2014), Vol. 223, #302.04

[4] J.M. Mazzarella, and NED Team, *Evolution of the NASA/IPAC Extragalactic Database (NED) into a Data Mining Discovery Engine* in *Astroinformatics* edited by M. Brescia, S.G. Djorgovski, E.D. Feigelson (2017), Vol. 325 of IAU Symposium, p. 379–384 (Paper I)

[5] P.M. Ogle, J. Mazzarella, R. Ebert et al., *Rule-based Cross-matching of Very Large Catalogs* in *Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV)* edited by A.R. Taylor, E. Rosolowsky (2015), Vol. 495 of Astronomical Society of the Pacific Conference Series, p. 25–35

[6] M. Schmitz, J.M.. Mazzarella, B.F. Madore et al., *Best Practices for Data Publication to Facilitate Integration into NED: A Reference Guide for Authors* in *American Astronomical Society Meeting Abstracts #223* (2014), Vol. 223, #302.05

[5] http://aspbooks.org/custom/publications/paper/492-0099.html
[6] http://astrothesaurus.org/
[7] https://heasarc.gsfc.nasa.gov/vo/summary/
[8] http://www.ivoa.net/