UniversAI: Exploring the Universe with Artificial Intelligence
Proceedings IAU Symposium No. ,
I. Liodakis, V. Efthymiou, eds.
doi:10.1017/xxxxx

Automated development of an astrophysical semantic catalogue in PARSEC

Iosif Oikonomakis[1], Vasilis Efthymiou[1], Giorgos Flouris[1], Elias Kiritsis[2], Ioanna Leonidaki[2], Michael Mavrommatis[2], Dimitrios Rompogiannakis[2], Andreas Zezas[2]

[1]FORTH - Institute of Computer Science
{sifisoik,vefthym,fgeo}@ics.forth.gr
[2]FORTH - Institute of Astrophysics
ekyritsis@physics.uoc.gr, {ioanna, azezas}@ia.forth.gr, {ph5604, ph4502}@edu.physics.uoc.gr

Abstract. This short paper presents the activities of PARSEC, a research project that aims to develop and apply a machine-assisted methodology for constructing and curating semantically enriched astrophysical catalogues in the form of knowledge graphs (KGs). The ultimate objective of this methodology is to automate the process of constructing astrophysical catalogues and to simplify their analysis and integration, allowing astrophysicists to more easily generate new knowledge and scientific insights.

Keywords. semantic annotation, Knowledge Graphs, PARSEC project, HECATE, SNR, AI in Astrophysics

1. Introduction

Telescopes, artificial satellites and other astrophysical instruments constantly probe the sky, gathering astronomical data and resulting to large-scale multi-wavelength surveys. The resulting data are pre-processed and organised in *astrophysical catalogues* (Makarov et al. 2014; NASA/IPAC 2025; Wenger et al. 2000), which contain information about celestial sources such as galaxies, stars and transient events.

Astrophysical research crucially relies on the availability of such catalogues to enable cross-domain research tasks (Budavári and Szalay 2008). However, the exponentially growing volume of astronomical data highlights the need for a new approach in astronomical catalogue curation, as these surveys (and the resulting catalogues) constitute enormous datasets that contain valuable information, but often lack the semantic structure needed to fully support reuse, integration, and automated exploration appropriate for downstream tasks as described by Accomazzi et al. (2000).

In practice, catalogue curation remains a manual and time consuming task. As the volume of incoming data increases, along with the different data formats, the task of catalogue curation poses major challenges. Catalogues are currently stored as static tables in CSV or other textual formats, or, in the best scenario, as a single-table relational database (Hanisch 2006). As a result, these resources risk becoming outdated, difficult to maintain, and hard to exploit in new research contexts (Richards et al. 2011).

Semantic technologies and Knowledge Graphs (KGs) could provide a solution to this problem, by representing the data as entities and relationships, aligned with a domain-specific ontology that defines concepts and their properties (Antoniou et al. 2012). This model could be used to perform data querying in more expressive and semantically meaningful ways, and can allow efficient data searching, querying, reasoning, data analysis, pattern identification, provenance, trust and credibility analyses, catalogue maintenance and enrichment, data reuse, catalogue interlinking, and others. Nowadays, the astrophysical community have introduced ontologies for astronomical objects and metadata (Karray et al. 2016; IVOA 2025), but large-scale catalogue conversion into KGs remains rare due to the scale and complexity of the

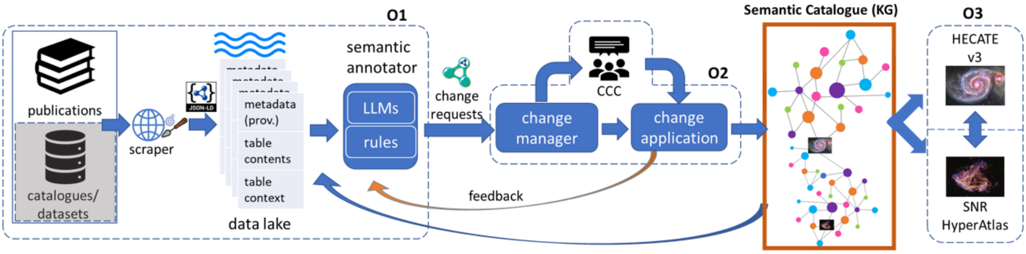

Figure 1. Overview of the workflow of the PARSEC project

existing datasets. Achieving the representation of astrophysical knowledge into a KG will simplify several downstream tasks that will eventually lead to new scientific discoveries.

The development and application of a machine-assisted methodology for constructing and curating semantically enriched astrophysical catalogues in the form of KGs is the main objective of the PARSEC project†, whose workflow is visualized in Figure 1. More specifically, the project's aim is to develop a KG (that we call *semantic catalogue*), supported by a carefully designed *ontology* (see Section 2), which will model astrophysical knowledge found in scientific publications. To populate this KG, we will mine information from tables found in publications, and employ AI techniques from the field of semantic annotation to appropriately interpret and store the information found in those tables in the semantic catalogue (objective O1 in Figure 1, and Section 3). The process will be overviewed by the Catalogue Curation Committee (CCC), a group of experts who will curate and manage the semantic catalogue (objective O2 in Figure 1, and Section 4); note however that we aim to keep human involvement to a minimum. We consider two use cases, related to galaxies and supernova remnants (SNRs) respectively, each leading to a different (but interrelated) KG (objective O3 in Figure 1, and Section 5).

2. Designing the semantic catalogue

Knowledge Graphs are often accompanied with ontologies, whose aim is to provide semantics to the terms used in the KG. In PARSEC, we developed our own domain-specific ontology, *DOCBO* (*Domain Ontology for Celestial Bodies and Observations* – see Figure 2), which provides a lightweight but expressive model to describe astrophysical records expressed in tabular formats (the most commonly used format for astrophysical data). DOCBO contains 15 classes, 18 object properties and 15 data properties, which capture the desired semantics, while allowing alignment with other relevant ontologies (planned for future work). A more detailed documentation of these classes and properties can be found in Zenodo (https://zenodo.org/records/16281112).

3. The knowledge acquisition process

The knowledge acquisition process is the first and most important stage of the workflow, aiming to extract structured information (tables along with their textual context) from raw astrophysics data (html, pdf, mrt, csv, png) and to semantically annotate it in order to be incorporated into the semantic catalogues. A neuro-symbolic semantic annotation process, combining logical rules and LLMs, will be employed.

Specifically, the semantic annotation process will receive tabular data with astrophysical measurements as input, and will identify their overlap, both in terms of meta-data (aka schema) and instance-level data, to our semantic catalogue. This process will help us understand the

† Propelling Astrophysics Research via Semantically Enriched Catalogues - https://isl.ics.forth.gr/PARSEC/.

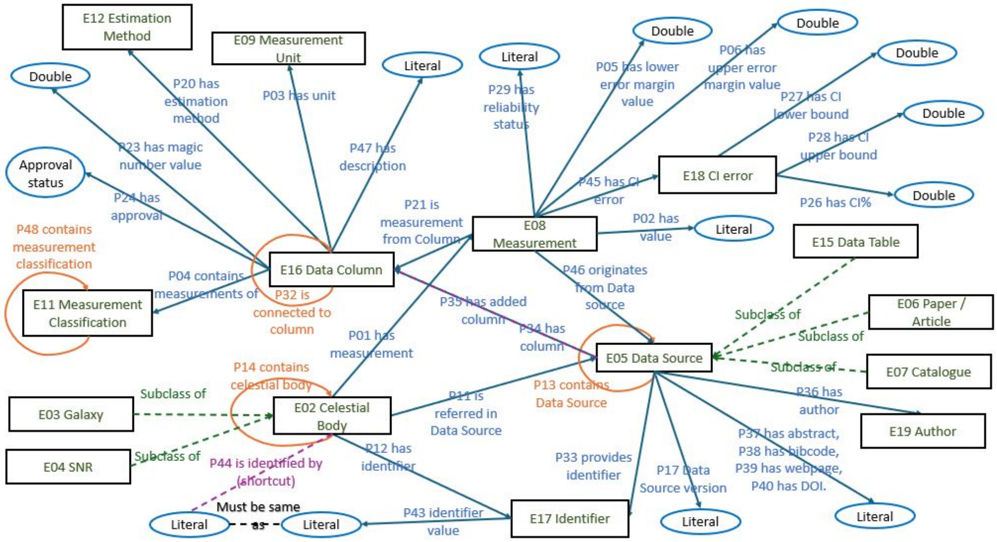

Figure 2. The DOCBO ontology

data contained in scientific publications, but also enrich our catalogue in terms of completeness (new data) and correctness (more accurate/homogenised values for existing objects). The annotation process will be largely built on top of LLMs, but, in order to avoid hallucinations, as well as to reduce the search space of potential matches to our catalogues, we will also employ logical rules (e.g., restricting the options based on the domain or the data type of each column).

4. Change management and curation of the semantic catalogue

The information acquired by the semantic annotation process will be issued in the form of a *change request* for the semantic catalogues, and will be applied to the KG using the change manager. Our aim is that, in most cases, the process will be automatic and no human intervention will be necessary. However, when the new information is found to be inconsistent with the existing knowledge, the CCC will be employed in order to decide how to incorporate the new knowledge. All changes will be recorded using a versioning mechanism.

5. The envisioned semantic catalogues

We plan to develop two interrelated, but independent, astrophysical catalogues, namely *HECATE v3* and *SNR HyperAtlas*.

HECATE v3 will be an extended version of the *HECATE* catalogue (Kovlakas et al. 2021) to include all known galaxies within a sphere of 500 Mpc, matching the volume accessible by the LIGO O5 run (expected to start in 2026). It will provide robust and homogenized information, enabling the rapid characterization of the hosts of gravitational-wave sources and other transient events such as those identified in time-domain experiments (e.g., eROSITA, LSST). HECATE is currently used routinely for the characterization of sources in X-ray surveys (Tranin et al. 2024; Ponti et al. 2023; Salvato et al. 2022; Dage et al. 2021) and transient and gravitational-wave Astrophysics (Hosseinzadeh et al. 2024; Förster et al. 2022).

SNR HyperAtlas will develop a KG including spatially resolved information for extended objects. SNRs provide important information on stellar evolution. Key information for different regions of these objects (e.g., temperature, density, velocity) is currently scattered in a large variety of publications; PARSEC will systematically gather and homogenize this information, enabling us to study the physical properties of SNRs as a function of their age and environment and to perform systematic comparisons with SNR evolution models. The developed methodology can be also applied to the curation of catalogues of spatially resolved information for other extended objects such as galaxies, interstellar nebulae, star clusters and stellar associations. Our goal is to provide a semantically enriched astrophysical catalogue, using collections of publications (Green 2025), observations (Safi-Harb 2025), or multi-wavelength surveys. This catalogue will not extend or replace the existing ones, but will aggregate astrophysical data that is not available in these catalogs in a systematic way ready/valid for statistical use.

6. Discussion

This short paper described PARSEC, an interdisciplinary research project aiming to develop and apply a machine-assisted methodology for constructing and curating semantically enriched astrophysical catalogues in the form of knowledge graphs, leading to two interrelated semantic catalogues, namely HECATE v3 and SNR HyperAtlas. The main challenges associated with this process are related to the richness and diversity of knowledge found in astrophysics papers, the sheer size of the data, and the characteristics of astrophysical data pertaining to the need for homogenization, to the uncertainty and errors in measurements, to missing data and magic numbers, etc.

References

A. Accomazzi, G. Eichhorn, M.J. Kurtz, C.S. Grant, and S.S. Murray. The NASA Astrophysics Data System: Architecture. *A&A Supplement Series*, 143(1):85–109, April 2000.

Grigoris Antoniou, Paul Groth, Frank Van Harmelen, and Rinke Hoekstra. *A Semantic Web Primer*. MIT Press, 3rd edition, 2012.

Tamás Budavári and Alexander S. Szalay. Probabilistic cross-identification of astronomical sources. *ApJ*, 679(1):301–309, May 2008. ISSN 1538-4357.

Kristen C. Dage, Noah Vowell, Erica Thygesen, et al. Ultraluminous x-ray sources in seven edge-on spiral galaxies. *MNRAS*, 508:4008–4016, 2021.

F. Förster, Muñoz A.A.M., I. Reyes-Jainaga, et al. DELIGHT: deep learning identification of galaxy hosts of transients using multiresolution images. *AJ*, 164(5):195, oct 2022.

D.A. Green. SNR catalogue (online resource). https://www.mrao.cam.ac.uk/surveys/snrs/snrs.paper.html, 2025. Accessed: August 21, 2025.

R.J. Hanisch. Data standards for the international virtual observatory. *Data Science Journal*, 5, 2006.

G. Hosseinzadeh, K. Paterson, J.C. Rastinejad, et al. SAGUARO: time-domain infrastructure for the fourth gravitational-wave observing run and beyond. *ApJ*, 964(1):35, mar 2024.

IVOA. IVOA vocabulary. https://www.ivoa.net/rdf/object-type/2020-10-06/object-type.html, 2025. Accessed: August 21, 2025.

Faten Karray, Mohamed Ben Ahmed, and Ahmed Hadj Kacem. ASON: An OWL-S based ontology for astrophysical services. *Computers & Geosciences*, 94:71–80, 2016.

K. Kovlakas, A. Zezas, JJ. Andrews, et al. The Heraklion Extragalactic Catalogue (HECATE): a value-added galaxy catalogue for multimessenger astrophysics. *MNRAS*, 506:1896–1915, 2021.

D. Makarov, P. Prugniel, N. Terekhova, H. Courtois, and I. Vauglin. HyperLEDA. III. the catalogue of extragalactic distances. *A&A*, 570:A13, oct 2014.

NASA/IPAC. NASA/IPAC extragalactic database (NED). http://ned.ipac.caltech.edu, 2025. Accessed: August 21, 2025.

G. Ponti, J. S. Sanders, N. Locatelli, et al. Characterizing the patchy appearance of the circumgalactic medium and the influence of foreground absorption. *A&A*, 670:A99, 2023.

J.W. Richards, D.L. Starr, N.R. Butler, et al. On machine-learned classification of variable stars with sparse and noisy time-series data. *ApJ*, 733(1), apr 2011. ISSN 1538-4357.

S. Safi-Harb. SNRcat - high energy observations of galactic supernova remnants (online resource). http://snrcat.physics.umanitoba.ca/, 2025. Accessed: August 21, 2025.

M. Salvato, J. Wolf, T. Dwelly, et al. The eROSITA final equatorial-depth survey (eFEDS): Identification and characterization of the counterparts to point-like sources. *A&A*, 661:A3, 2022.

Hugo Tranin, Natalie Webb, Olivier Godet, and Erwan Quintin. Statistical study of a large and cleaned sample of ultraluminous and hyperluminous X-ray sources. *A&A*, 681:A16, 2024.

M. Wenger, F. Ochsenbein, D. Egret, et al. The SIMBAD astronomical database. the CDS reference database for astronomical objects. *A&A Suppl. Ser.*, 143:9–22, April 2000.