PARSEC:

Propelling Astrophysics Research through Semantically Enriched Catalogues

Giorgos Flouris [1*], Andreas Zezas [2,3*#], Vasilis Efthymiou [1,4], Iosif Oikonomakis [1], Katerina Papantoniou [1], Ariadne Pratikaki [1], Elias Kyritsis [2,3,5], Ioanna Leonidaki [2,3], Michael Mavrommatis [2,3], Dimitrios Rompogiannakis [2,3]

(1) Institute of Computer Science, FORTH; (2) Institute of Astrophysics, FORTH; (3) Department of Physics, University of Crete; (4) Department of Informatics and Telematics Harokopio University of Athens; (5) Max Planck Institute for Extraterrestrial Physics

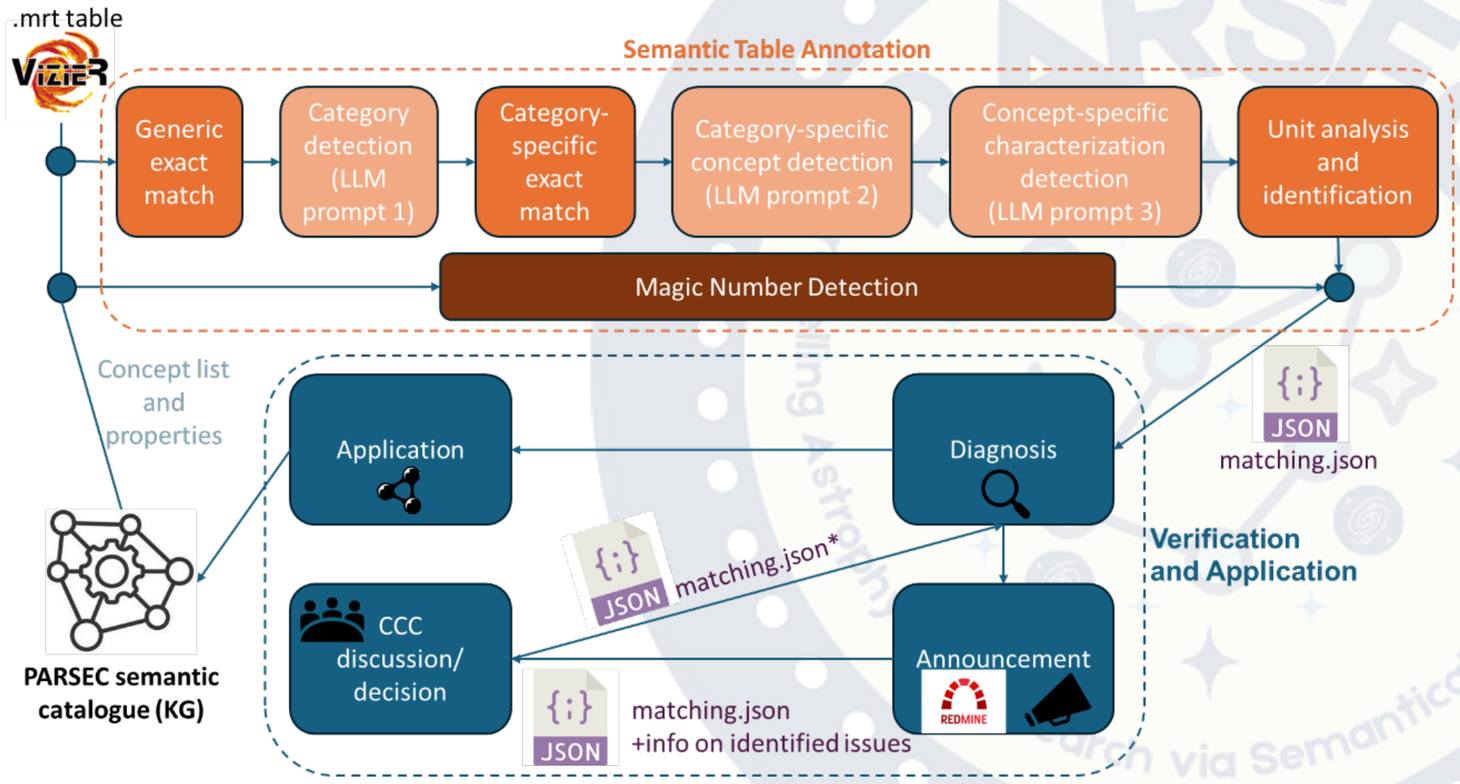
The availability of major wide-area time-domain surveys across the electromagnetic spectrum produces a flood of data and data products at a rate that require automated processes for their ingestion in the existing body of astrophysical data. This need becomes even more critical at the era of multi-messenger Astrophysics that require rapid access to multiple types of data for the characterization of non-electromagnetic events (e.g. gravitational-wave signals, neutrino or cosmic-ray events etc). For this reason we have embarked in a major project aiming in the development of a platform that enables the automated mining of data from publications and their ingestion in a semantically annotated database. In this process we include a self-check mechanism for the correct characterization of the data and resolution of problems, automatic recognition and homogenization of data units, detection of magic numbers, and a user-friendly query environment.

The PARSEC approach

In the first phase of this program we focus on tabulated data for two different types of astronomical objects: Galaxies and Supernova Remnants.

Our pipeline consists of the following steps:

Step 1: Mining of catalogue data (currently mrt tables).
Step 2: Characterization of tabular data into concepts.
Step 3: Unit recognition and homogenization.
Step 4. Magic number detection.
Step 5. Diagnosis of problems in the data semantic annotation.
Step 6. Resolution of problems by the Catalogue Curation Committee (CCC).
Step 7. Ingestion of the data in the PARSEC Semantic catalogue.
Step 8. Query of the catalogue via a user-friendly environment.



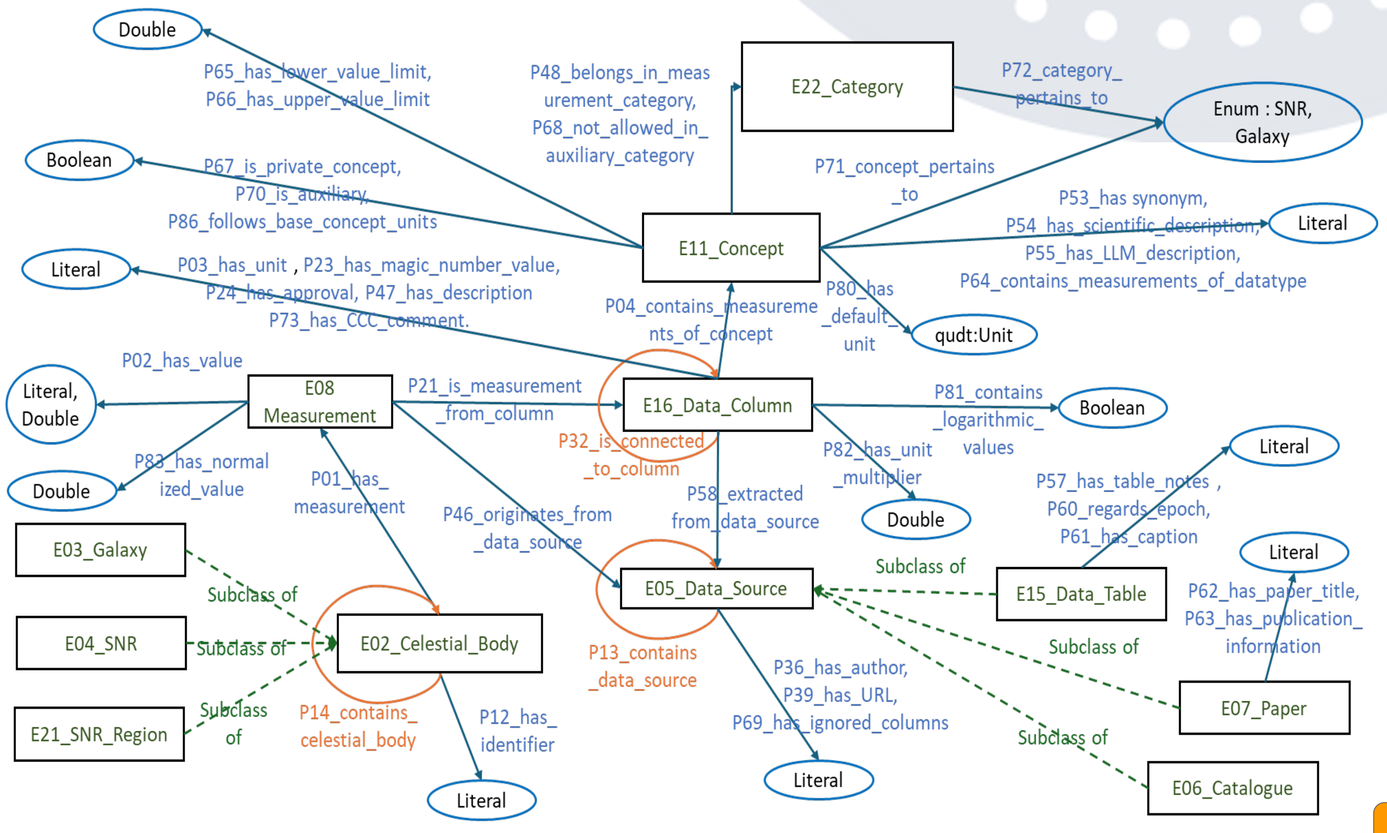


Figure 2. An outline of the PARSEC ontology

Each table column is characterized into astrophysical quantities (concepts) which are organized into categories.
For each measurement we record the celestial object it belongs to (allowing for multiple regions), uncertainties (if available), units, origin (if provided), associated information.

Contacts: azezas@physics.uoc.gr, fgeo@ics.forth.gr

The PARSEC novelties

The PARSEC pipeline includes several state-of-the art aspects:

- A hierarchical concept organization into:
  Categories – Concepts – Accompanying concepts (e.g. errors, comments, additional data, method characterization etc).
- A hierarchical celestial object framework allowing for the ingestion of data for spatially extended objects.
- Automatic recognition and homogenization of data units.
- Automatic recognition of problems in the data annotation.
- Ability to extend and modify the concepts list to accommodate the growing variety of data.

Figure 1. A flowchart of the PARSEC pipeline

Top: The steps leading to the semantic annotation of the tabular data. This is done in four phases: (a) identification of the category in which a column belongs to, (b) characterization of the concept describing its contents, (c) identification of any accompanying concepts, and (d) identification of the data units. These steps are performed through a combination of exact matches and LLM queries. In parallel we identify any magic numbers in the data.
Bottom: Rectification of uncharacterized or inconsistently characterized table columns. This involves diagnosis of the problem, suggestion of solutions, intervention of the Catalogue Curation Committee (CCC), and finally ingestion in the database.
The communication between the different steps is done with .json files.

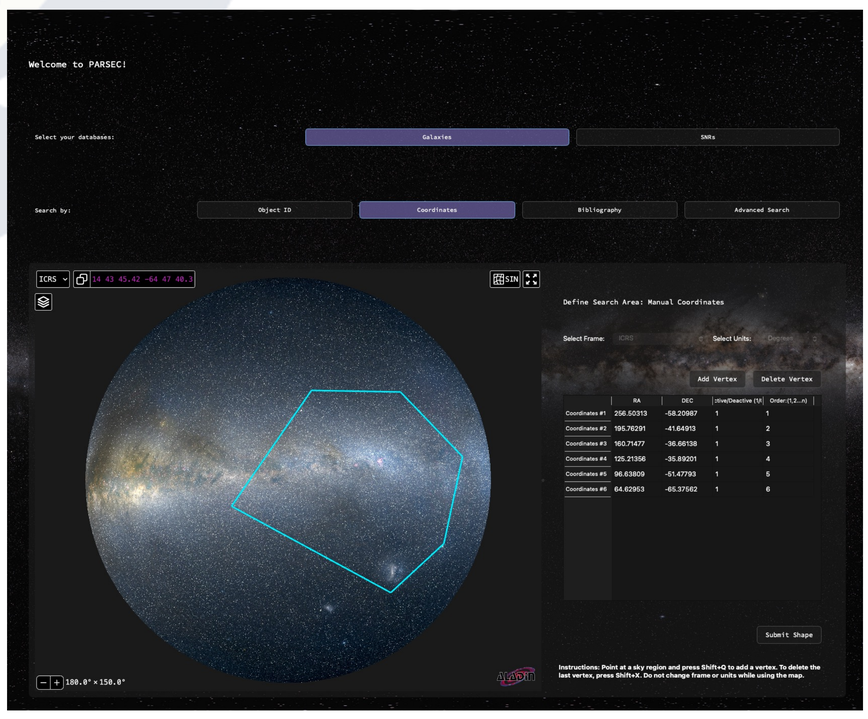


Figure 3. A user friendly graphical query interface

The graphical interface allows queries on object name, coordinates, sky regions, or object properties. In the latter the user can perform complex queries by combining multiple concepts from the database. The user can view the data or download them in tabular form for off-line analysis.

Future steps

The PARSEC project places the foundation for a more complete tool including further innovations such as:

- Mining of data from the textual body of the publications.
- Homogenization of data in the semantic catalogue.
- Tools for on-line analysis in the query interface.
- Extension to other types of astronomical objects including transients (e.g. supernovae, γ-ray bursts, gravitational-wave events) and variable sources (e.g. Tidal Disruption Events, X-ray binaries).