

Galaxy Data Mines

Comparing Astronomical Object Classifications in NED and SIMBAD

Final Progress Report

Matthew Shubat
CS4490Z Thesis
Department of Computer Science
Western University
April 8, 2019

Supervisor: Dr. Pauline Barnby
Course Instructor: Prof. Nazim Madhavji

Structured Abstract

Context and Motivation

Astronomers inherently face a staggering amount of data. To make sense of this data, large aggregate “data mines” have been developed to collect and organize it. Two significant instances of these are NED and SIMBAD.

Question/Problem

For researchers working with both NED and SIMBAD it is clear that each of these databases share many common objects, however these objects are often given different classifications by NED’s linear classification scheme and SIMBAD’s hierarchical one. This produces confusion as to what the objects class truly is as well as the degree to which these two systems agree with one another.

Principle ideas/results

In this project, we automate some of this inter-database comparison process. Through this means, we found a classification similarity of over 90% between the two systems and established an increase in similarity at larger distances and for larger objects.

Contribution

By systematically comparing classifications between NED and SIMBAD, insight is given into the optimal use of each system. A tool called gdmimes was created to obtain this insight and is openly available for others to use.

1. Introduction

Due to the fundamental nature of their work, which spans the scope and scale of the universe, astronomers are faced with a vast set of data. Organizing and searching through this information is a substantial undertaking. There are many repositories of astronomical literature and data available online, each with their own methodology of organization and access. Although fruitful, this landscape can be somewhat dispersed and segmented.

To help mitigate this challenge, several large aggregate data repositories have been developed. Two substantial instances of these are NED and SIMBAD, which pull data from broad areas of astronomical literature. They have been immensely helpful in addressing this information organization and access challenge and have helped birth a new era of astronomical discovery [1].

Proper classification of astronomical objects is one challenge in the field of astronomy. There are billions of objects to be classified; furthermore, there are inherent difficulties in classifying and matching astronomical objects in general [2]. Given these realities, and since both NED and SIMBAD pull data from various sources online, each has taken its own unique approach to organizing the objects it does track. NED uses a linear, 1-dimensional classification scheme, whereas SIMBAD uses a hierarchical one. Objects added to each system are then given the most accurate classification available in the respective scheme.

This system works well when using NED or SIMBAD in isolation. However, when a researcher begins to work between the two systems a problem emerges: overlapping objects—objects present in both databases—may be given different classifications. If the classifications given are very similar or at the same level of a classification hierarchy, then it can be simple to see whether or not the two systems agree. However, often these differences are less clear. One system may provide a higher-level classification in the hierarchy, or perhaps over-generalize. This situation produces confusion as to what an objects class truly is, as well as the degree to which these two systems agree. As a result, a researcher is made less confident and is tasked with further investigation.

In this paper we focus on answering the question: how related are overlapping objects' classifications between NED and SIMBAD? Furthermore, we try to quantify these systems' relatedness. To achieve this insight, we developed the `gdm` (galaxy data mines) command line tool to help automate this process and answer these questions. The tool is openly available for others to use [3].

We applied the `gdm` tool to the famous Messier object set [4] containing a total of 110 objects, which range from star clusters to galaxies. Choosing this set of objects was important, as queries centered on the coordinates of a messier object will return thousands of results to be compared. This allowed us to get a better overall picture of how each system's classifications compared. This is the first time the contents of NED and SIMBAD have been compared in a systematic way, and our results can provide insight into how researchers can best utilize these resources.

In the upcoming sections we walk through some relevant background information, list our objectives and their significance, go through our methodology, and finally give results of the system development and the Messier object study. We also give an interpretation of these results and suggest future work.

2. Background and Related Work

2.1. Overview of Current Landscape

- There are many astronomical knowledge and data resources available online to both researchers and the general public [5].
- Some of these resources house journal articles, others large compilations of astronomical datasets, which can be individually searched for an object or area of interest.
- Although the information is available, its dispersed placement and differing formats has led to many disjoint compilations of research: each with their own method of organization.
- Several aggregate databases have been created in an effort to organize and unify this information.
- There are inherent difficulties in the classification and matching of objects, and each aggregate database must source its data from a variety of sources with varying degrees of uncertainty.
- Each aggregate system chooses its own philosophy and organization approach when assigning object classifications in its database.
- These compilations are extremely beneficial and allow a researcher to query one location or object and get results sourced from a vast array of resources.
- Two substantial instances of these are NED and SIMBAD.

2.2. NED

- NED stands for “NASA/IPAC Extragalactic Database”.
- It is operated by the Jet Propulsion Laboratory (JPL), California Institute of Technology (Caltech), under contract with the National Aeronautics and Space Administration (NASA) [6].
- NED is primarily focused on collecting information on extragalactic objects.
- NED is broader in scope and contains few classification categories.
- NED uses one-dimensional classification system with a total of 63 classes. See [[7], Fig. 1] for a portion of this system.
- As of today, NED contains 667,024,593 distinct objects [6].

2.3. SIMBAD

- SIMBAD stands for “Set of Identifications, Measurements, and Bibliography for Astronomical Data”.
- It is managed by the Centre de Données astronomiques de Strasbourg (CDS) [8].
- SIMBAD is focused on extrasolar objects, objects outside our solar system, primarily within our own galaxy, the Milky Way Galaxy.
- SIMBAD provides greater detail in its classification of objects.
- SIMBAD uses a hierarchical classification system with 238 possible classifications in 17 different “groups”. See [9, Fig. 2] for a subsection of this system.
- As of today, SIMBAD contains 10,285,241 objects [8].

2.4. Analysis and Research Gap

These two independent systems have greatly expedited the research process, enabled new kinds of discoveries to be made, and helped usher in a new era of discovery in astronomy [1].

With a general look at NED and SIMBAD it is clear that they both have their own areas of focus. NED contains many more objects than SIMBAD [6], [8] which mirrors its larger, more broad extragalactic scope. However, SIMBAD provides more object classifications and sub-types than NED, as reflected in its hierarchical classification scheme [9].

Although different in several ways, NED and SIMBAD share many common, overlapping objects. These are objects which have been determined to share the same position in the astronomical coordinate system, with some degree of uncertainty, and are listed in both NED and SIMBAD. Although sharing the same position—and otherwise appearing to be the same object—a researcher working between these two systems may notice that, due to the different classification systems used, overlapping objects in each system are often given different classifications.

Naturally some questions arise from this situation: “Which classification is correct?”, “Are the classifications incompatible?”, “To what degree do the classifications given by NED and SIMBAD to overlapping objects agree?”, “Which source am I better off using and for what objects?”.

Given how important these databases are to the future of the field [1], a better understanding of how such systems relate to one another could prove to be immensely helpful.

*	Star or Point Source
**	Double star
*Ass	Stellar association
*Cl	Star cluster
AbLS	Absorption line system
Blue*	Blue star
C*	Carbon star
EmLS	Emission line source
EmObj	Emission object
exG*	Extragalactic star (not a mer
Flare*	Flare star
G	Galaxy
GammaS	Gamma ray source
GClstr	Cluster of galaxies
GGroup	Group of galaxies
GPair	Galaxy pair
GTrpl	Galaxy triple
G_Lens	Lensed image of a galaxy
HII	HII region
TrS	Infrared source

Fig. 1 Subsection of NED Linear Classification Scheme.

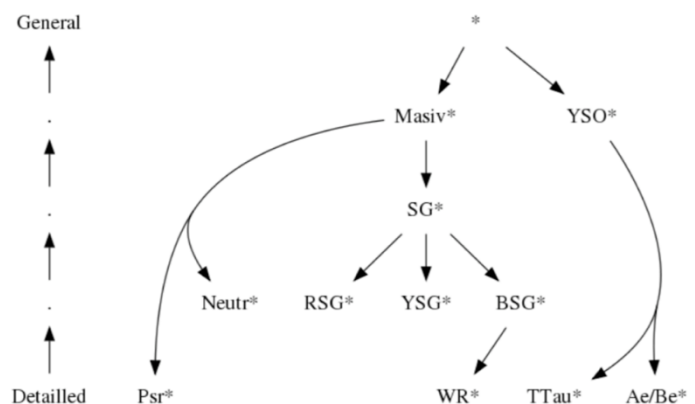


Fig. 2 Subsection of SIMBAD Hierarchical Classification Scheme.

3. Research Objectives

G – Goal, O – Objective

- **G1:** Create a tool to automate the inter-database object classification comparison process.
 - **O1-1:** Relate hierarchical classification scheme of SIMBAD to the linear one of NED, define match types to model the nuanced relationships between the classification systems, and create the corresponding algorithms for each.
 - **O1-2:** Expand data sources from local test data to each system’s online data.
 - **O1-3:** Produce meaningful output of comparisons in a variety of formats.
- **G2:** Quantify and infer overall similarity of object classifications between NED and SIMBAD.
 - **O2-1:** Obtain regions/names of all Messier objects and sort by distance from earth.
 - **O2-2:** Query each object in the set from closest to farthest and save results of queries: plots, statistics, tables, logs.
 - **O2-3:** Review the results and look for any obvious patterns. Further patterns should be found via match type statistics, plots and further analysis.
 - **O2-4:** If an overall trend emerges, develop a fitting hypothesis and try to extrapolate these findings to the entire NED and SIMBAD system.
- **Significance:**
 - NED and SIMBAD are widely used and valuable tools; increasing our knowledge of the classification methods used by these systems, and how they relate to one another, would provide useful insight to researchers and the field.
 - With this information, researchers would be able to make more informed decisions about where to best gather data, and for which objects. The gdmimes tool can also help researchers gain a clearer picture of an object’s classification.

4. Methodology

Galaxy Data Mines Tool

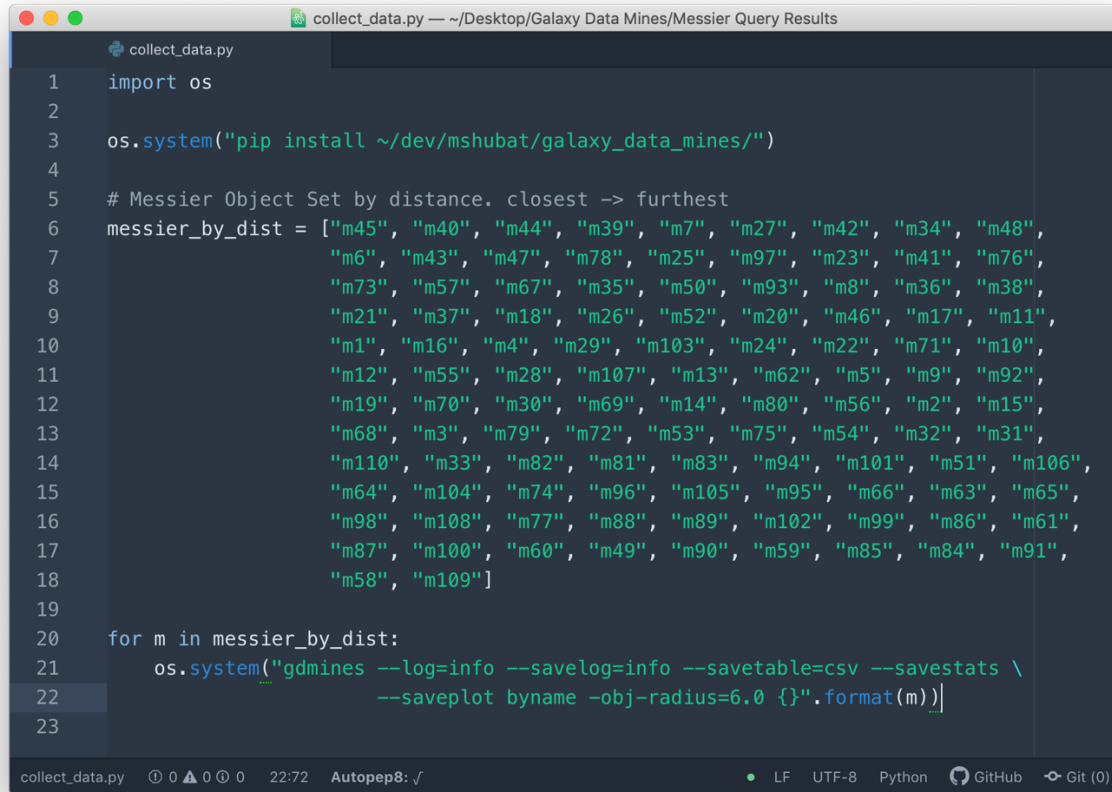
The overall system design was initiated by getting the minimum comparison functionality working on a local data set. After a basic solution was functional, several iterations of feature design and testing commenced to expand the system to the desired capabilities: namely, comparing all object types on remote, queried data from NED and SIMBAD.

The technologies used to develop the system include Python [10] and the SciPy package [11], including Matplotlib; NumPy; and Pandas. A popular astronomy library Astropy [12] was utilized for table processing and its affiliated package Astroquery [13] was employed to query both NED and SIMBAD. Finally, the package Click [14], was used to create the command-line interface for the tool.

NED and SIMBAD Systematic Comparison

In our analysis, 110 Messier “objects” or “regions” were searched via queries. Each region was queried using the gdmimes tool in conjunction with a script to further automate the process. A radius around the query object of 6 arcminutes and the default match tolerance of 1 arcsecond (see glossary) were used for each query. Region queries were done in order of distance from the Earth,

from closest to farthest, to account for any match-type–distance correlation. These distances ranged from 400 light-years (M45) to over 100 million light-years (M109) [4]. Results were saved in the form of logs, tables, statistical summaries and 2D sky plots coloured by match types. For the script used, and Messier Objects queried refer to [4, Fig. 3] below.



```

1  import os
2
3  os.system("pip install ~/dev/mshubat/galaxy_data_mines/")
4
5  # Messier Object Set by distance. closest -> furthest
6  messier_by_dist = ["m45", "m40", "m44", "m39", "m7", "m27", "m42", "m34", "m48",
7                    "m6", "m43", "m47", "m78", "m25", "m97", "m23", "m41", "m76",
8                    "m73", "m57", "m67", "m35", "m50", "m93", "m8", "m36", "m38",
9                    "m21", "m37", "m18", "m26", "m52", "m20", "m46", "m17", "m11",
10                   "m1", "m16", "m4", "m29", "m103", "m24", "m22", "m71", "m10",
11                   "m12", "m55", "m28", "m107", "m13", "m62", "m5", "m9", "m92",
12                   "m19", "m70", "m30", "m69", "m14", "m80", "m56", "m2", "m15",
13                   "m68", "m3", "m79", "m72", "m53", "m75", "m54", "m32", "m31",
14                   "m110", "m33", "m82", "m81", "m83", "m94", "m101", "m51", "m106",
15                   "m64", "m104", "m74", "m96", "m105", "m95", "m66", "m63", "m65",
16                   "m98", "m108", "m77", "m88", "m89", "m102", "m99", "m86", "m61",
17                   "m87", "m100", "m60", "m49", "m90", "m59", "m85", "m84", "m91",
18                   "m58", "m109"]
19
20  for m in messier_by_dist:
21      os.system("gdmimes --log=info --saveplot=info --savetable=csv --savestats \
22                --saveplot byname -obj-radius=6.0 {}".format(m))
23

```

Fig. 3 Script, Query structure, and Messier object data, sorted by distance, used for Queries.

5. Results

5.1. Galaxy Data Mines Tool Results

The gdmtool was able to meet all of the objectives covered in **G1** (listed above). There is more to the tool than will be described here, however the tool can be forked and downloaded from GitHub [3] so it can be further explored. We will, however, provide a brief overview now.

The UML diagram for the system can be seen in Fig. 4 below.

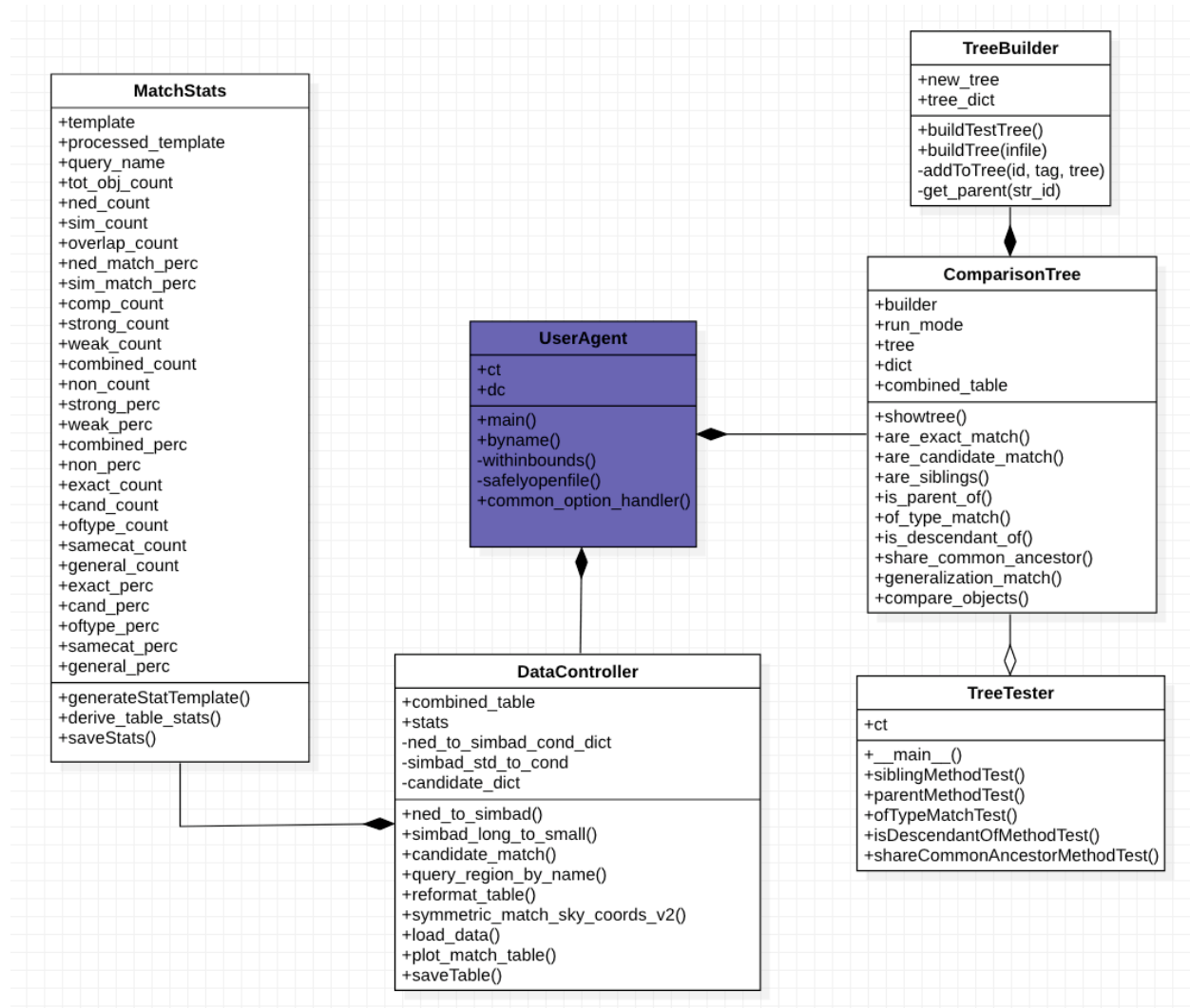
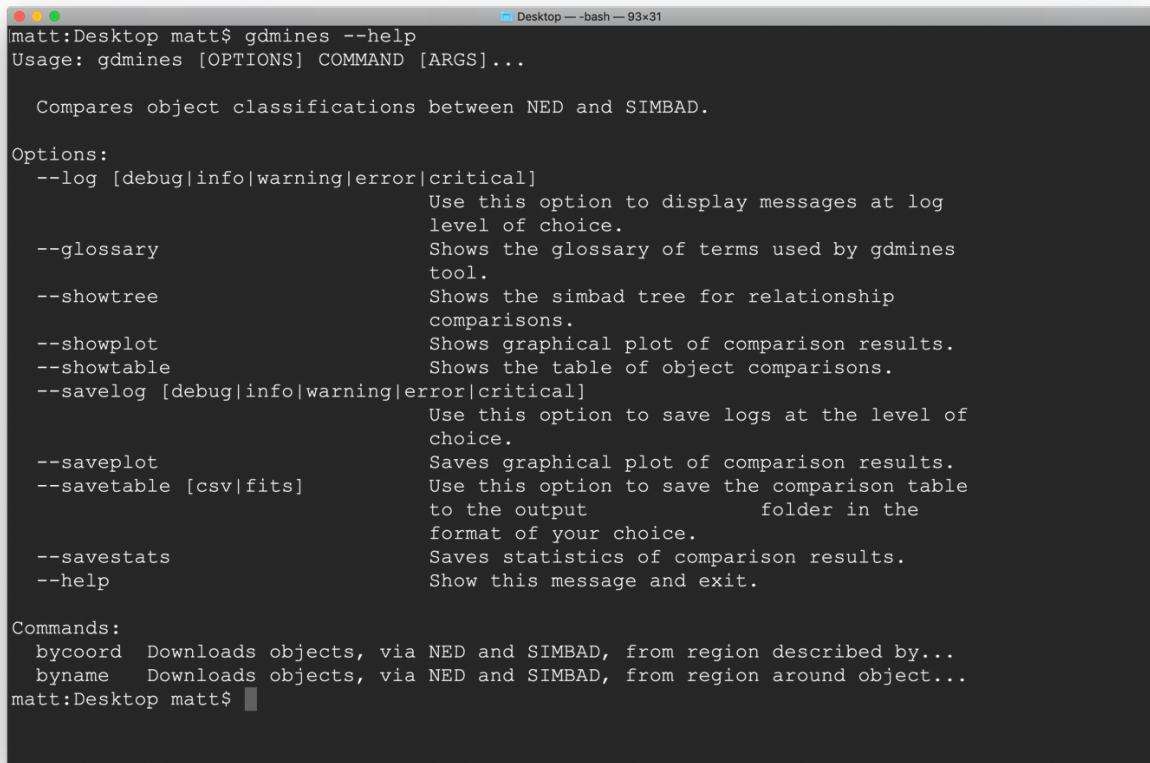


Fig. 4 Basic UML class diagram for the gdmtool.

The user interface of the system is done through “UserAgent”, with a command line entry point which is triggered through the “gdmtool” command. The application works through a command line interface, using the Click project [13] and is installed as a Python package, with the dependencies listed above. It is designed to work on all desktop platforms which support Python, which is to say pretty much all of them. The help screen for the tool can be seen in Fig. 5 below.



```

matt:Desktop matt$ gdmimes --help
Usage: gdmimes [OPTIONS] COMMAND [ARGS]...

Compares object classifications between NED and SIMBAD.

Options:
  --log [debug|info|warning|error|critical]
                                Use this option to display messages at log
                                level of choice.
  --glossary                    Shows the glossary of terms used by gdmimes
                                tool.
  --showtree                   Shows the simbad tree for relationship
                                comparisons.
  --showplot                   Shows graphical plot of comparison results.
  --showtable                  Shows the table of object comparisons.
  --save-log [debug|info|warning|error|critical]
                                Use this option to save logs at the level of
                                choice.
  --saveplot                   Saves graphical plot of comparison results.
  --save-table [csv|fits]      Use this option to save the comparison table
                                to the output folder in the
                                format of your choice.
  --save-stats                 Saves statistics of comparison results.
  --help                       Show this message and exit.

Commands:
  bycoord Downloads objects, via NED and SIMBAD, from region described by...
  byname  Downloads objects, via NED and SIMBAD, from region around object...
matt:Desktop matt$

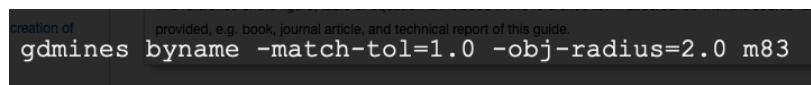
```

Fig. 5 Screenshot of the gdmimes help screen listing options, commands and their usage.

For the development of the system, some initial code, created by Prof. Barmby, was used as a starting point [15]. This code facilitated the matching of common objects from a local dataset, as well as the reading in of the SIMBAD classification scheme. From here, the system was developed and architected to achieve 3 main functions:

Function 1: Fetching the Data – *Primarily Carried out by the “DataController” class.*

- Using Astropy and its affiliated package Astroquery, the gdmimes tool is able to query NED and SIMBAD dynamically for any official astronomical object name.
- Both a radius and match tolerance (see glossary) can also be passed as parameters to further customize the comparisons done in a given region. The command syntax is shown in Fig. 6 below.



```

creation of provided, e.g. book, journal article, and technical report of this guide.
gdmimes byname -match-tol=1.0 -obj-radius=2.0 m83

```

Fig 6. Basic Syntax for M83 Region Query with a match tolerance of 1 arcsecond and an object radius of 2 arcminutes.

- Once objects are retrieved, the data is cleaned and overlapping objects in NED and SIMBAD are computed and saved in a combined table with their matching sky positions.

Function 2: Processing Matches – *Primarily Carried out by the “ComparisonTree” and “TreeBuilder” classes.*

- For each overlapping object pair, the classification given by NED is then mapped to SIMBAD’s hierarchical classification scheme, which is effectively a superset of the one provided by NED.
- Each object classification is then compared and assigned to one of the following match categories: Strong Matches (“Exact”, “Candidate”, and “Of-Type”), Weak Matches (“Shared Category” and “Generalization”) and “Non-Matches”. Further details on these match types and how they are computed can be found in the glossary.
- Internally the match results are represented by a table like the one shown in Fig. 7 below.

Index	RA(deg)	DEC(deg)	Type_N	Type_N_Ana	RA_d	DEC_d	Type_S	Exact Match	Candidate	Of-Type	Match	Shared Cate.	Generalization	Non Match
1	204.24	-29.86	IrS	IR	204.24	-29.86	CI*	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
2	204.24	-29.87	HII	HII	204.24	-29.87	G	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
3	204.24	-29.86	XrayS	X	204.24	-29.86	X	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	204.24	-29.86	*CI	CI*	204.24	-29.86	CI*	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	204.24	-29.86	XrayS	X	204.24	-29.86	X	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	204.24	-29.87	XrayS	X	204.24	-29.87	X	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	204.24	-29.85	*CI	CI*	204.24	-29.85	C?*	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
8	204.24	-29.86	XrayS	X	204.24	-29.86	X	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
9	204.24	-29.87	IrS	IR	204.24	-29.87	CI*	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
10	204.24	-29.86	*CI	CI*	204.24	-29.86	C?*	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
11	204.24	-29.88	XrayS	X	204.24	-29.88	X	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
12	204.24	-29.88	XrayS	X	204.24	-29.88	X	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
13	204.24	-29.85	XrayS	X	204.24	-29.85	SNR	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
14	204.24	-29.85	*CI	CI*	204.24	-29.85	C?*	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
15	204.25	-29.85	*CI	CI*	204.25	-29.85	C?*	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
16	204.25	-29.85	IrS	IR	204.24	-29.85	CI*	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
17	204.25	-29.87	SNR	SNR	204.25	-29.87	SNR	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
18	204.25	-29.85	*CI	CI*	204.25	-29.85	C?*	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
19	204.25	-29.85	RadioS	Rad	204.25	-29.85	Rad	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
20	204.25	-29.86	HII	HII	204.25	-29.86	CI*	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE

Fig 7. Condensed Table of Object Match Results for Query.

Function 3: Outputting the Results in a meaningful format – *handled by the “DataController”, “MatchStats”, and “UserAgent” classes.*

- Comparison results are provided in a variety of formats: a match table, statistical summary, 2D sky plot coloured by match type, and detailed log file can all be outputted and saved by the gdmimes tool.

An instance of the plot and statistical summary, produced by the gdmimes tool, are shown in Fig. 8 and Fig. 9 below.

Example of gdmimes plot output.

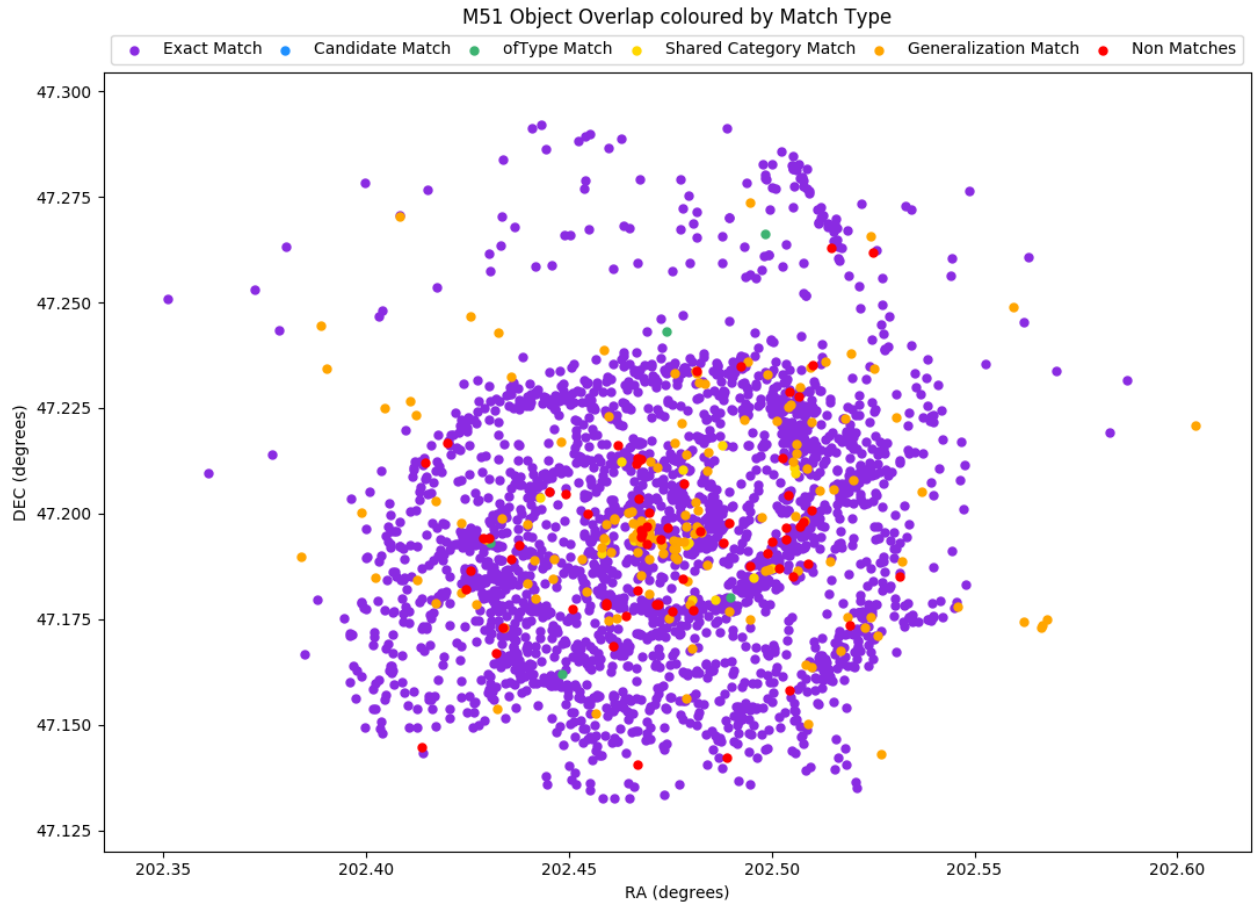
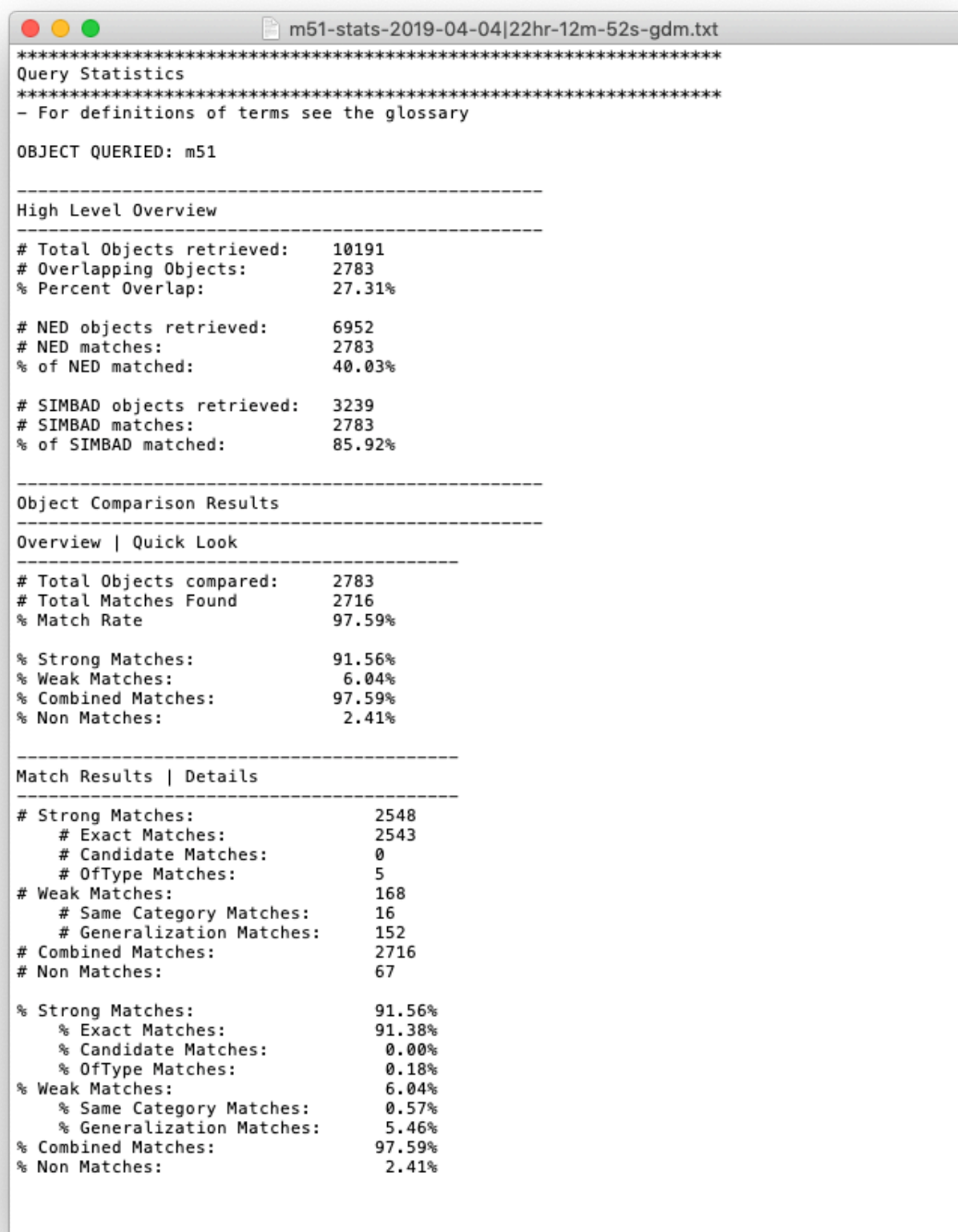


Fig. 8 – 2D Plot of NED and SIMBAD overlapping objects for the M51 “Whirlpool Galaxy” Coloured by Match Type.

Example of gdmstats statistical summary.



```

m51-stats-2019-04-04|22hr-12m-52s-gdm.txt
*****
Query Statistics
*****
- For definitions of terms see the glossary

OBJECT QUERIED: m51

-----
High Level Overview
-----
# Total Objects retrieved:    10191
# Overlapping Objects:       2783
% Percent Overlap:           27.31%

# NED objects retrieved:     6952
# NED matches:               2783
% of NED matched:           40.03%

# SIMBAD objects retrieved:   3239
# SIMBAD matches:            2783
% of SIMBAD matched:         85.92%

-----
Object Comparison Results
-----
Overview | Quick Look
-----
# Total Objects compared:    2783
# Total Matches Found        2716
% Match Rate                 97.59%

% Strong Matches:            91.56%
% Weak Matches:              6.04%
% Combined Matches:          97.59%
% Non Matches:               2.41%

-----
Match Results | Details
-----
# Strong Matches:            2548
#   Exact Matches:          2543
#   Candidate Matches:       0
#   OfType Matches:         5
# Weak Matches:              168
#   Same Category Matches:   16
#   Generalization Matches:  152
# Combined Matches:          2716
# Non Matches:               67

% Strong Matches:            91.56%
%   Exact Matches:          91.38%
%   Candidate Matches:      0.00%
%   OfType Matches:         0.18%
% Weak Matches:              6.04%
%   Same Category Matches:   0.57%
%   Generalization Matches:  5.46%
% Combined Matches:          97.59%
% Non Matches:               2.41%

```

Fig. 9 – Statistical Summary of a Query to NED and SIMBAD for the M51 “Whirlpool Galaxy”.

5.2. NED and SIMBAD Classification Comparison Results

In our analysis of the Messier object set over 340,000 objects were returned, ~38,000 of which were overlapping objects. After systematically comparing the classifications of these 38,000 overlapping objects, as well as saving and searching through all of the outputs of the gdmimes tool previously mentioned, we were able to establish some key results and notice some clear trends.

1. First, with respect to the sample retrieved, of the 340,000 objects returned ~263,000 (~77%) were from NED and ~ 79,870 (~23%) were from SIMBAD, reflecting the relative size of each database.
2. Related to the previous observation: the overlap rate of each of the systems is defined as the proportion of objects which are overlapping of all the objects returned. *For example, if a query to NED returned 100 objects and 50 of them overlapped with the objects returned from SIMBAD for the same query, then the NED overlap rate would be 50%.* Given this definition we found three results:
 - a. NED was found to have an overlap rate of ~13%.
 - b. SIMBAD was found to have an overlap rate of ~60%.
Again, reflecting the relative sizes of the datasets.
 - c. Both systems overlap percentages had no observable correlation with the queried object's distance from the Earth up to the maximum distance queried (100 Mly). Refer to Fig. 10 and Fig. 11 below.

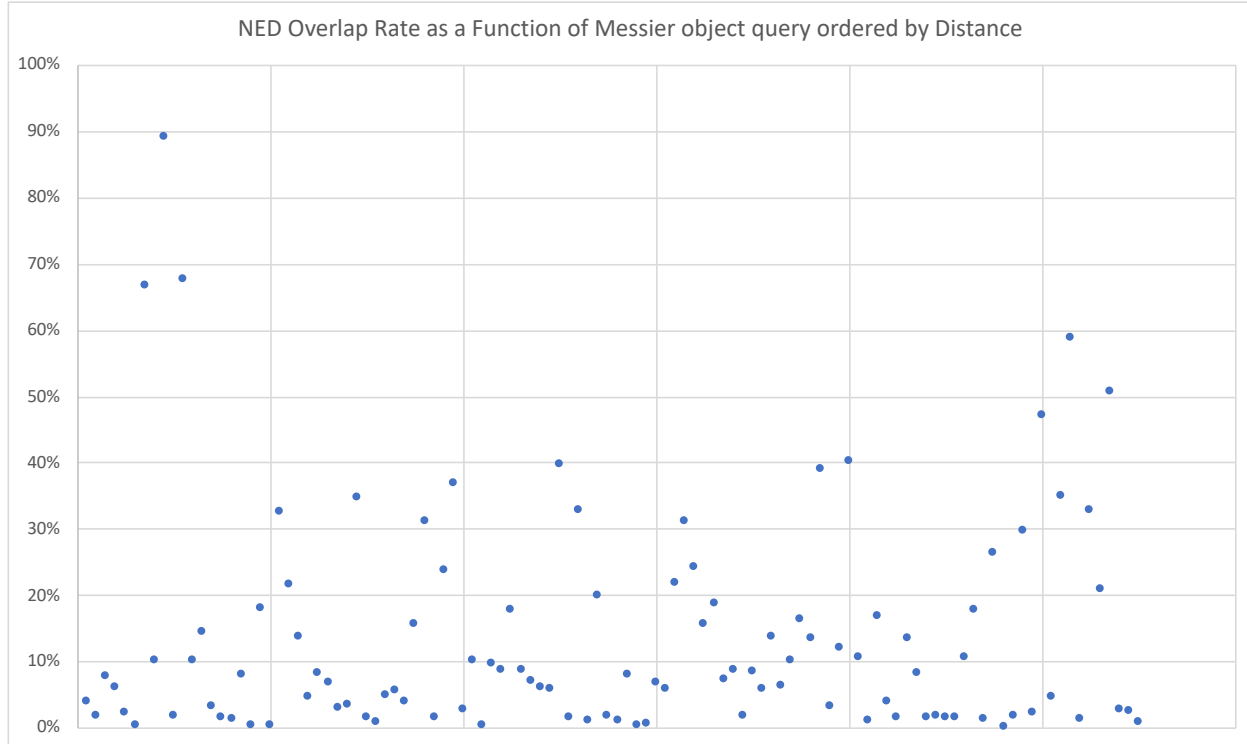


Fig. 10 – Ned Overlap Rate for Messier Queries ordered by Distance.

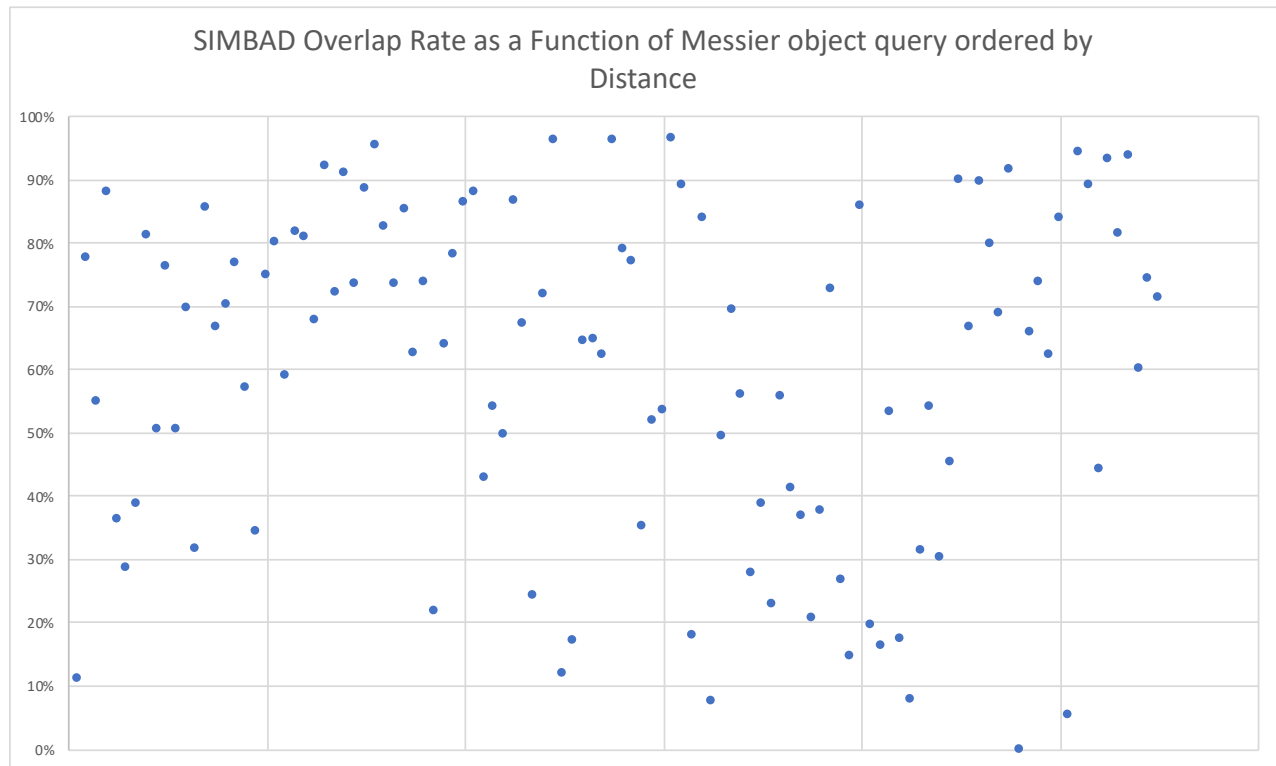


Fig. 11 – SIMBAD Overlap Rate for Messier Queries ordered by Distance.

As mentioned, no observable correlation is present.

3. When taking into account both Strong and Weak match categories, NED and SIMBAD were found to have an average classification overlap greater than 90%. That is to say, over 90% of the 38,000 overlapping objects compared were given either Strong or Weak match classifications between the two systems.
4. Combined Match Percentage (combination of “Strong” and “Weak” matches) did not meaningfully depend on the distance of the queried object from Earth, see Fig. 12 below. *Though there is a slight dispersion of combined matches around the 100 kly mark which is likely caused by the increase in Strong Matches, thus decrease in Weak matches, which are more “forgiving” due to categories like generalization matches.*

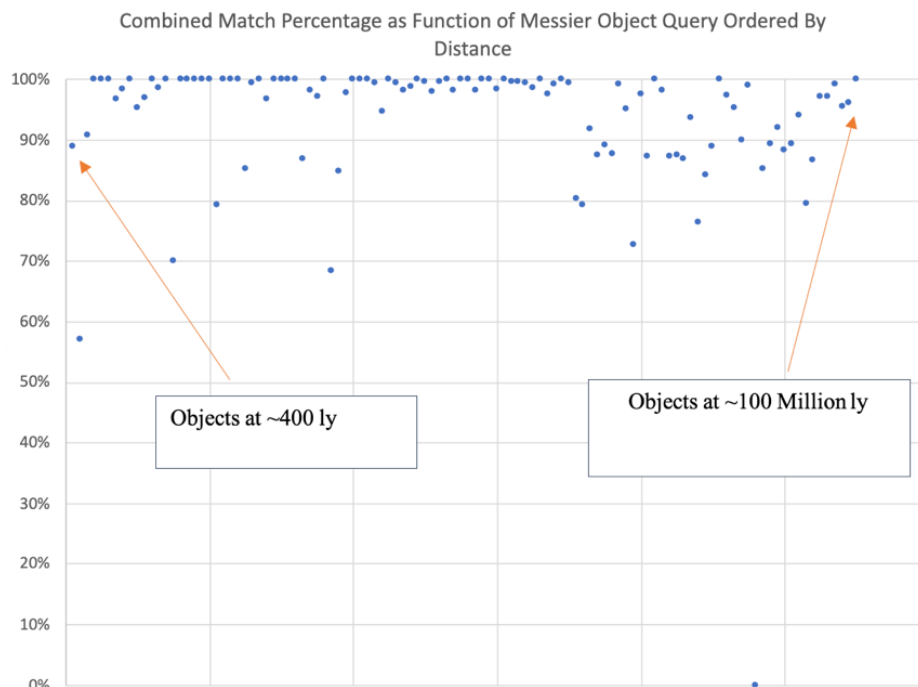


Fig. 12 – Combined Match % as a Function of Object Query Ordered by Distance.

5. Closer objects (<100 kly from Earth) tended to have a higher proportion of weak matches when compared to objects at larger distances (>100 kly). See Fig. 13 below.

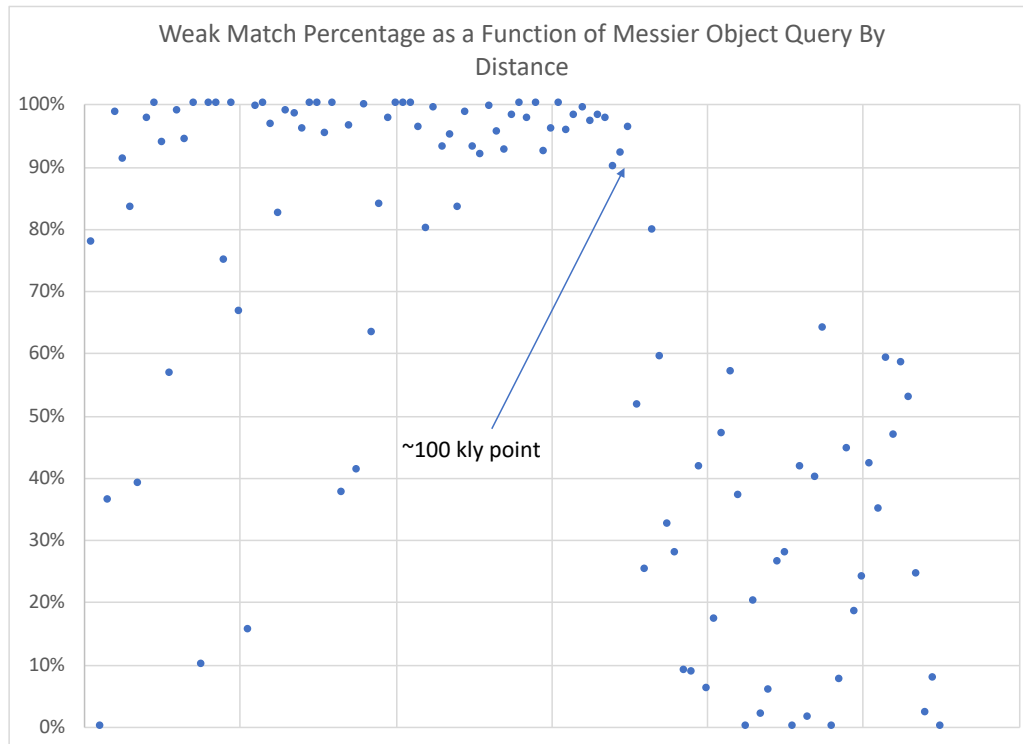


Fig. 13 – Weak Match Percentage as a Function of Messier Object Query by Distance. *Here it can be observed that for objects closer to earth, there is a high percentage of “Weak” matches. But at larger distances (>100 kly) there is a significant drop in weak matches.*

Conversely, at distances greater than 100 kly strong matches made up a higher proportion of the total combined matches, as shown in Fig. 14 below.

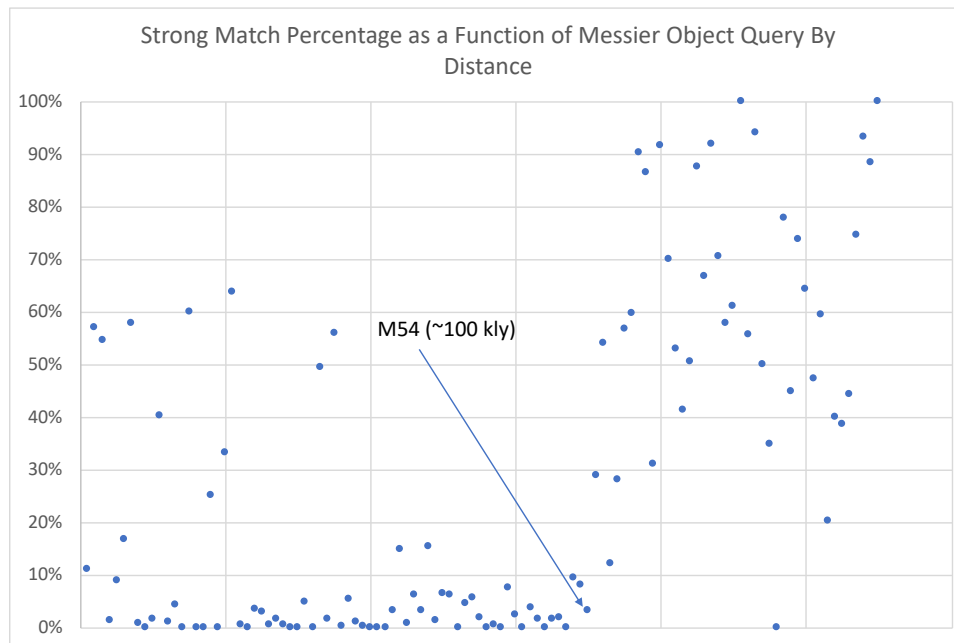


Fig. 14 - Strong Match Percentage as a Function of Messier Object Query by Distance.

6. When averaged across all Messier object queries, it was found that ~65% of “Strong” matches were “Exact” matches; ~33% of Strong matches were “Of-type” Matches, and only ~2% were Candidate matches, refer to Fig. 15 below.

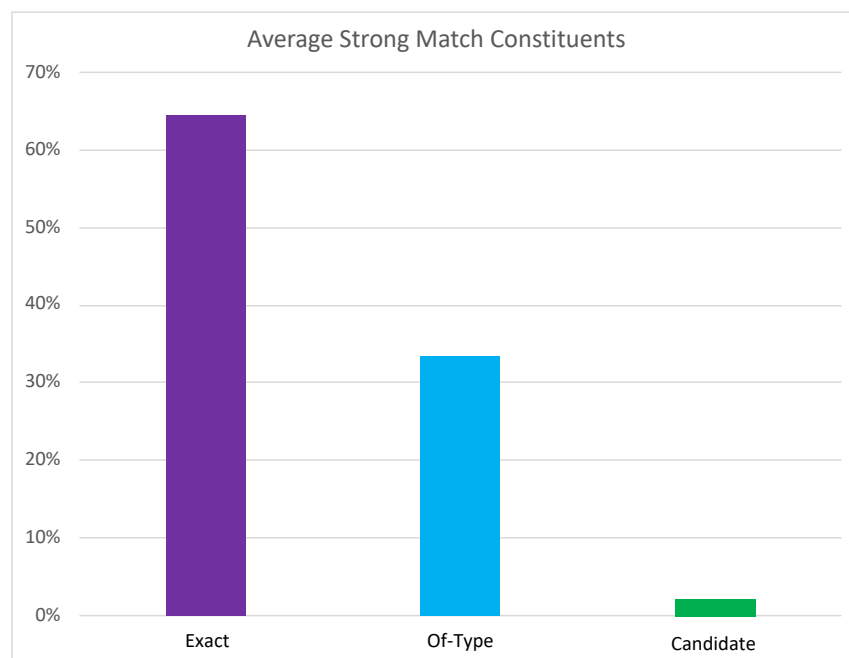


Fig 15. Average values of Strong Match constituents across all 110 Messier queries.

7. When averaged across all Messier object queries, it was found that ~93% of “Weak” matches were “Generalization” matches and only ~7% of weak matches were “Shared Category” Matches. See Fig. 16 below.

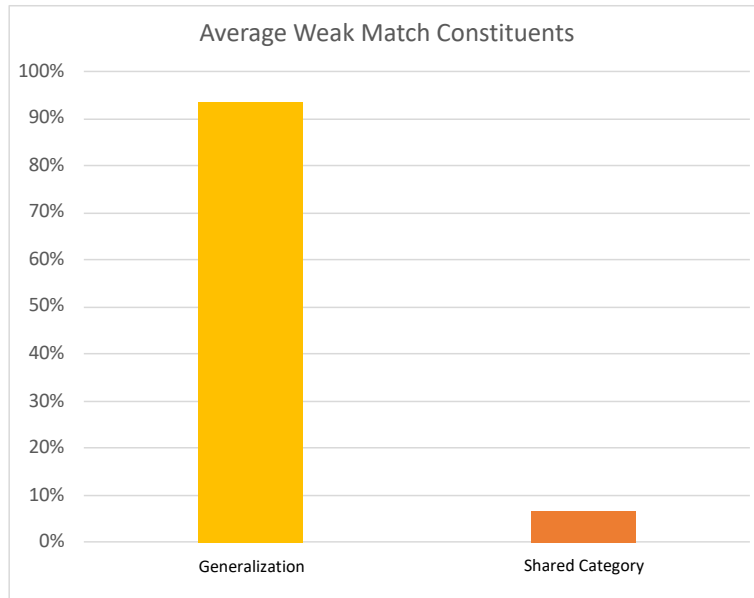


Fig 16. Average values of Weak Match constituents across all 110 Messier queries.

6. Discussion

Key takeaways from the results

- When taking into account both strong and weak match types, NED and SIMBAD had a classification agreement given to common objects of over 90%.
- Match strength tended to increase with the distance of the Messier region queried—specifically around the threshold of 100 kly—and/or the size of the object queried.
- Weaker matches, which were found in regions closer to the Earth, were made up primarily of Generalization Matches, where general types (i.e. VisS) were often given by NED. *This result is likely due to the extragalactic focus of NED.*
- Stronger matches, which were typically found in regions further from the Earth, were primarily composed of “Exact” matches and some “Of-Type” matches, but few “Candidate” matches

Use of the results

- For nearby objects (<100 kly from Earth) it would be advisable to use SIMBAD as the data source; SIMBAD generally gives more specific classifications to objects within this radius, whereas NED often gives “generalized” object classifications in this range (i.e. XrayS).
- For objects of interest outside of the radius of 100 kly, it is advisable to use either SIMBAD or NED, since both shared a high percentage of strong matches for overlapping objects beyond this radius. *This is likely due to the 100 kly point roughly being at the edge of what*

is considered to be part of our galaxy, thus where NED would consider an object to the extragalactic.

- For these further objects SIMBAD still often gives more specific object classification due to the nature of its classification scheme.
- NED may however often prove to be the best choice as it consistently returned more objects for each of our region queries, and often contains objects which SIMBAD does not.
- The two systems were highly compatible with a combined match rate of over 90% so use of either system is likely to be of similar validity.

Threats to validity and limitations.

- Weak match types such as “Generalization” and “Same Category” matches, although generally safe, could in fact be referring to different objects.
- Messier objects may not be representative of all objects housed within SIMBAD and NED.
- Limited sample size was used for comparison data.
- Overlapping objects were calculated with some degree of uncertainty, thus some computed matches could have been marked as Non-Matches when actually referring to different objects.
- Non-Matches do not necessarily guarantee object classifications are incompatible, but rather that no match type is defined for the given object pair.

7. Conclusions

In this paper we described the current classification challenge researchers face when working between the two popular astronomical “data mines” NED and SIMBAD. We compared SIMBAD’s [9, Fig. 2] hierarchical scheme to NED linear system [7, Fig. 1].

To address this challenge and seek clarity into how these two systems’ object classifications compare, we proposed an automated tool in G1 to help relate these two systems together so meaningful comparisons could be made. In G2 we specified our methodology for exploring how these two systems compared.

Using the famous Messier object set [4] we queried all 110 regions in order of distance and computed overlapping objects. From here we compared these common objects classifications using the gdmies tool and gathered plots, statistics, and tables of all comparison data.

From this analysis we were able to gain some insights into how these systems relate to one another. We found that for regions closer to Earth, less the 100 kly in distance, classification matches of overlapping objects tended to be much weaker than classifications beyond the 100 kly mark. We also found that the majority of weak matches are Generalization matches, which often reflected the difference in priorities of the two systems at closer range. Furthermore, we found that Strong matches were primarily composed of Exact matches, and Of-Type Matches showing a strong agreement among objects at further range, and a tendency for greater classification specificity from SIMBAD.

If researchers must choose a system to gather their data, our results can provide some insight to help them make this decision. If objects of interest are within our galaxy, with ranges at or below the 100 kly mark, then SIMBAD may be a better option for more specific object classifications.

At larger distances SIMBAD may still be a good option, as overlapping objects at this range had primarily strong match types with some Of-type matches, showing that SIMBAD was again more specific than NED. However, NED may often be the only possible choice for many objects, especially extragalactic, as it was found to consistently return significantly more results for each region query submitted.

Overall, both systems were found to have a high agreement with one another, with a combined match rate of over 90%. This makes either system a comparatively reliable source. One just tends to be more specific in its classifications, whereas the other tends to be more complete.

For researchers, working with both systems, the gdmtool may also provide some use in comparing objects in regions of interest to get a better sense of what is present, and how each system's data compares.

8. Future Work and Lessons Learnt

Object Class Comparisons

- A greater sample size would help to further validate our findings and improve the usefulness of our results.
- Additional queries could be made to expand outside of the Messier data set.
- Queries could also control for distance and object type separately to further elucidate whether or not the object type is more important than the distance in determining match strength.
- Match types could be further broken down and defined to give more accurate view of "strength".

Galaxy Data Mines Tool

- Querying regions by coordinates and scanning through the sky could be automated by the gdmtool and could provide an overall "big picture" of classification matches.
- The gdmtool could have greater flexibility and options for specifying queries.
- An improved graphical display of match relationships could provide a faster means of understanding the match distribution for a query.
- Relationship "Matches" and algorithms could be further refined to determine with greater accuracy the relatedness of the two systems.
- Taking into account the sources of the information used by NED and SIMBAD could also help to account for any similarities or differences found between objects returned and overlapping objects classifications.
- There may be an even better means of determining a match or non-match, for example explicitly defining for each object match whether or not the object is a match or non-match. This would, however, give many different possibilities.

9. Glossary

9.1. General Terms

NED – NASA Extragalactic Database

SIMBAD – Set of Identifications, Measurements, and Bibliography for Astronomical Data

Messier Objects – a famous set of 110 astronomical objects catalogued by the French astronomer Charles Messier. The set includes a diverse range of astronomical objects: star clusters, nebulae, galaxies, and more.

Overlapping Objects – Common objects, i.e. share the same sky coordinates, contained in both NED and SIMBAD.

Candidate Object - In astronomy, candidate objects are objects which are possibly in a particular class, but this is not known for certain.

Arcminute – Unit to measure very small angles, one arcminute is equivalent to $1/60$ of 1 degree.

Arcsecond – Unit to measure extremely small angles, one arc second is $1/60$ of an arcminute: thus, $1/3600$ of a degree.

Right Ascension (RA) – One of the two necessary measurements needed to specify a point on the celestial sphere in the equatorial coordinate system. RA is a measure of angular distance of a point measured eastward from the celestial equator.

Declination (DEC) – One of the two necessary measurements needed to specify a point on the celestial sphere in the equatorial coordinate system. The declination angle is measured north or south of the celestial equator.

Object Radius – the radius around the astronomical object being queried measured in arcseconds. The default value is 1.0 arcminute. A value between 0 and 60 arcminutes may be provided.

Match Tolerance – the threshold angle that an object in NED and SIMBAD must fall into in order to be considered the same object. The default value is 1.0 arcsecond. A value between 0 and 60 arc minutes may be provided.

Light-year – a unit of astronomical distance equal to the distance that light travels (in vacuum) in one year. One light-year is equivalent to $9.461 \times 10^{15} \text{ m}$.

kly – kilo-light-year – 1000 light-years ($9.5 \times 10^{18} \text{ m}$).

Mly – Mega-light-year – 1 million light-years ($9.5 \times 10^{21} \text{ m}$).

Galaxy - a system of millions or billions of stars, together with gas and dust, held together by gravitational attraction.

gdmimes – “Galaxy Data Mines” - the command line tool created for this project which automates the comparison of overlapping objects between NED and SIMBAD.

9.2. Match Types

Strong Matches – *Match Category* - Objects with a strong match between NED and SIMBAD have classification which are identical or basically identical. These are confident match types.

Weak Matches – *Match Category* - Objects with a weak match between NED and SIMBAD have classifications which are related, but not as closely. These are less confident match types, but still compatible.

Exact Match (*Strong*)– These are matches in the truest sense of the word. These matches occur when NED and SIMBAD both give the same class to an object. *Syntactic differences between corresponding classes are ignored.*

Example:

Match Type	NED Classification	SIMBAD Classification
Exact Match	* (star)	* (star)
Exact Match (ignoring syntactic difference)	*Cl (star cluster in NED matches a)	Cl* (cluster of stars in SIMBAD)

Candidate Match (*Strong*) – NED does not differentiate between candidate and non-candidate objects, thus a “Candidate Match” is a match between NED’s non-candidate category and SIMBAD’s candidate category.

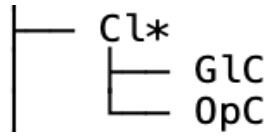
Example:

Match Type	NED Classification	SIMBAD Classification
Candidate Match	* (star)	*? (star)
Candidate Match	GClstr (cluster of galaxies)	C?G (possible cluster of galaxies)

This suggests SIMBAD is less confident about the object’s true classification. Although not as strong as an “Exact Match”, this is still considered a “Strong Match”; both NED and SIMBAD are likely to agree about the nature of the object being classified.

Of-Type Match (Strong) – These are matches where one object is a descendant of the other in the classification hierarchy.

Example:

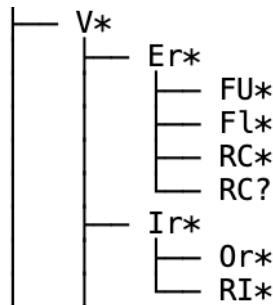


Cl and GLC would be an “Of-Type” Match, since GLC (globular cluster) is a descendant of Cl*.*

This is still considered a “Strong Match” as the two systems fundamentally agree on what the object is. One system is just being more specific.

Shared Category Match (Weak) – these matches share a common ancestor in the classification hierarchy, and therefore are in the same object category.

Example:



Er (eruptive star) and Ir* (irregular star) would have a “Shared Category” Match type, since Er* and Ir* are both a type of V* (variable star).*

Generalization Match (Weak) – a generalization match occurs when NED or SIMBAD give an object a very general definition, which is not necessarily incompatible with other classifications, but it is much too broad to say a stronger match could be made.

Example:

Match Type	NED Classification	SIMBAD Classification
Non-Match	VisS (star)	G (galaxy)

NED calls an object a **VisS** (visual source), whereas SIMBAD gives the same object the class * (star).

*Although these classifications are not incompatible, many different objects can be a source of visible light. Thus, we cannot give this match a Strong match type, but we cannot call it a “Non-Match” either. Thus, it is categorized as a “Weak” generalization match. Similar operations are also done with *XrayS*, *IrS* etc.*

Non-Match – this is the label given to overlapping objects for which no match category has been defined.

Example:

Match Type	NED Classification	SIMBAD Classification
Non-Match	* (star)	G (galaxy)

It is important to note that a “Non-Match” classification given to a pair of objects does not guarantee the given classifications disagree, it just means that currently there is no defined match type for the given relationship.

10. Acknowledgements

Thank you to my thesis advisor Professor Pauline Barmby, who developed the original idea for this project, contributed some foundational elements to the codebase, and provided guidance. The project would not exist without her contributions.

Thank you to the Nearby Galaxy Group [15] for providing astronomical knowledge, support, and input.

Thank you to Professor Nazim Madhavji for instructing the “CS4490 - Thesis” course and resolving my questions and concerns along the way.

Thank you to all the developers of the tools, technologies and libraries I used in this project. It would not have been possible otherwise.

11. References

- [1] J. M. Mazzarella, "Evolution of the NASA/IPAC Extragalactic Database (NED) into a Data Mining Discovery Engine," *Proceedings of the International Astronomical Union*, vol. 12, no. S325, pp. 379-384, 2016.

- [2] D. Severín, "Cross-identification of stellar catalogs with multiple stars: Complexity and Resolution," *Electronic Notes in Discrete Mathematics*, vol. 69, pp. 29-36, 2018.
- [3] M. Shubat and P. Barmby, "Galaxy Data Mines," April 2019. [Online]. Available: https://github.com/mshubat/galaxy_data_mines.
- [4] "The Messier Catalog," SEDS, 2008. [Online]. Available: <http://www.messier.seds.org>. [Accessed 2019].
- [5] J. Mink, "Astronomical Data Sources on the Web," 2 November 2018. [Online]. Available: <http://tdc-www.harvard.edu/astro.data.html>. [Accessed April 2019].
- [6] Jet Propulsion Laboratory, California Institute of Technology, "Database holdings for release 28.11.1," [Online]. Available: <https://ned.ipac.caltech.edu/CurrentHoldings>. [Accessed 6 April 2019].
- [7] "List of Object Types," [Online]. Available: <https://ned.ipac.caltech.edu/?q=help/srcnom/list-objecttypes&popup=1>.
- [8] Centre de Données astronomiques de Strasbourg (CDS), "SIMBAD Astronomical Database - CDS (Strasbourg)," 6 April 2019. [Online]. Available: <http://simbad.u-strasbg.fr/simbad/>.
- [9] "Object Classification in SIMBAD," 7 Nov 2013. [Online]. Available: [http://vizier.u-strasbg.fr/cgi-bin/OType?\\$1](http://vizier.u-strasbg.fr/cgi-bin/OType?$1).
- [10] "Welcome to Python.org," 2019. [Online]. Available: <https://www.python.org>.
- [11] S. developers, "SciPy.org," 2019. [Online]. Available: <https://www.scipy.org>.
- [12] T. A. Collaboration, "Astropy: A community Python package for astronomy," *Astronomy & Astrophysics*, vol. 558, p. A33, 2013.

[13] A. Ginsburg, "Astroquery," 2019 03 April. [Online]. Available:

<https://astroquery.readthedocs.io/en/latest/>.

[14] "Click | The Pallets Projects," April 2019. [Online]. Available:

<https://palletsprojects.com/p/click/>.

[15] "Western Research on Nearby Galaxies - Assembling the big picture," [Online]. Available:

<https://nearby-galaxies.github.io>.

[16] X. Chen, "treelib," 2018. [Online]. Available:

<https://treelib.readthedocs.io/en/latest/index.html>.

[17] O. Anaïs, "Categorisations of object types in SIMBAD," *EPJ Web of Conferences*, vol.

186, p. 12009, 2018.

[18] D. Muna, "The Astropy Problem," *arXiv:1610.03159 [astro-ph, physics:physics]*, 2016.

[19] "About NED," [Online]. Available: <https://ned.ipac.caltech.edu/Documents/Overview>.