**Final Report**

**Data and Metadata Profile**

The Sloan Digital Survey (SDSS) is a large astronomical survey that has collected and analyzed data from millions of celestial objects, including stars, galaxies, and quasars. The survey was initiated in the 1990s and has since produced several major data releases. The SDSS uses a 2.5 meter telescope located in New Mexico to collect data over a large area of the sky. The SDSS data has been used for a wide range of astronomical research, including studies of galaxy evolution, large-scale structure, and the distribution of dark matter and dark energy. The survey has also been used to discover new astronomical objects, such as quasars and supernovae.

The Stellar Classification Dataset—SDSS17 is the most recently released dataset from SDSS and aims to classificate stars, galaxies, and quasars based on their special characteristics. The data consists of 100,000 observations of space taken by the SDSS telescope, and each observation is described by 17 feature columns and one class column which identifies it to be either a star, galaxy, or quasar. The key stakeholders for this data include the SDSS, from where the data originates, as well as astronomers, both professional and amateur, for whom the data is either useful or of interest.

Users can download the data in a singular comma-separated values (CSV) file that includes all 100,000 observations and the 17 feature columns. The data set does not come with any usage restrictions and is available through a public domain license. As the data downloads as a singular CSV file, there is not additional required software to open the files, but because the data includes 100,000 observations, additional software—such as R or Python—may be required to analyze the data.

**Metadata**

The Stellar Classification Dataset—SDSS17 comes with various metadata to describe the features and columns in the dataset of 100,000 observations. As the data includes 17 feature columns identifying each object as either a star, galaxy, or quasar, the metadata explains the meaning of each feature column and provides additional information such as the epoch, the ID used to identify scans and observations, and the date when the data was taken. A full list of the metadata contained within the data includes:

1. obj\_ID = Object Identifier, the unique value that identifies the object in the image catalog used by the CAS
2. alpha = Right Ascension angle (at J2000 epoch)
3. delta = Declination angle (at J2000 epoch)
4. u = Ultraviolet filter in the photometric system
5. g = Green filter in the photometric system
6. r = Red filter in the photometric system
7. i = Near Infrared filter in the photometric system
8. z = Infrared filter in the photometric system
9. run\_ID = Run Number used to identify the specific scan
10. rereun\_ID = Rerun Number to specify how the image was processed
11. cam\_col = Camera column to identify the scanline within the run
12. field\_ID = Field number to identify each field
13. spec\_obj\_ID = Unique ID used for optical spectroscopic objects (this means that 2 different observations with the same spec\_obj\_ID must share the output class)
14. class = object class (galaxy, star or quasar object)
15. redshift = redshift value based on the increase in wavelength
16. plate = plate ID, identifies each plate in SDSS
17. MJD = Modified Julian Date, used to indicate when a given piece of SDSS data was taken
18. fiber\_ID = fiber ID that identifies the fiber that pointed the light at the focal plane in each observation

Some additional metadata for the dataset include keyword tags such as, “Earth and Nature,” “Astronomy,” “Classification,” “Physics,” “Mutliclass Classification,” and “Beginner.” These identifiers act as metadata describing the dataset both as it is and in context to it being a part of the Kaggle community (i.e., it is considered a “beginner” database for use on Kaggle.) The metadata is fully comprehensive and has been graded by the Kaggle community as 10/10 for usability, although it does not follow any particular metadata structure.

There are several ways that the metadata and data could be improved to allow for accessibility and discovery. First, Kaggle does not use a controlled vocabulary for its metadata. While this allows for greater flexibility for users to upload datasets, it also carries the risk of making datasets less discoverable. Similarly, Kaggle could implement context-based recommendations, allowing users to receive recommendations based on the context in which data is being used. This dataset has one publication listed on its page—The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar and APOGEE-2 Data—and this could be used to inform that context. Because the publication listed on the page merely details the publication of the dataset—rather than findings extrapolated from the dataset—an example of how the data may be used could be valuable for new users.

While the dataset as it exists on Kaggle does not reference other datasets in this series, the SDSS website elaborates that this dataset is the 17th and final data release of the fourth phase of the Sloan Digital Sky Survey (SDSS-IV). It further elaborates that this final release is cumulative of the 16 releases that came prior, and “contains new optical and infrared spectra from both Apache Point Observatory and Las Campanas Observatory”. On Google Scholar, searching for publications that cite this dataset returns one article available as open access in The Astrophysical Journal—The Seventheenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data, published March 16, 2022.

**Repository Profile**

The Stellar Classification Dataset—SDSS17 is the most recently released dataset from SDSS and aims to classificate stars, galaxies, and quasars based on their special characteristics. The data consists of 100,000 observations of space taken by the SDSS telescope, and each observation is described by 17 feature columns and one classe column which identifies it to be either a star, galaxy, or quasar. For collaborative projects using this dataset, the SIMBAD Astronomical Database Repository – CDS (Strasbourg) would be the most appropriate.

SIMBAD (Set of Identifications, Measurements, and Bibliography for Astronomical Data) is an astronomical database maintained by the Centre de Données astronomiques de Strasbourg (CDS) in France. It is one of the largest and most comprehensive astronomical databases, containing information on over 7 million celestial objects and more than 1 million bibliographic references. The database provides information on celestial objects such as stars, galaxies, quasars, and planetary systems, as well as data on their properties, such as positions, magnitudes, spectra, and distances. It also includes information on objects’ observations and catalogues, as well as links to other astronomical data sources. Users can search for information on specific objects or perform more complex queries based on specific criteria, such as object type, position, or magnitude. The data can also be displayed graphically, allowing users to view and analyze information in an intuitive way. While there are other astronomical databases available, the SIMBAD Astronomical Database Repository is the most comprehensive repository that allows for raw data.

Any user can submit new data to SIMBAD, and the SIMBAD database is a collaborative project. Data submissions can be made through the online interface provided by the CDS, and their internal team reviews all submissions to ensure the data is accurate and relevant before adding it to the database. This data can be in the form of new observations, updated information on previously known objects, or the discovery of new celestial objects. To allow for the best chance of acceptance, CDS recommends that users reference the article, “Best Practices for Data Publication in the Astronomical Literature,” available via open access in The Astrophysical Journal. Additionally, the SIMBAD database provides detailed guidelines on what should be included in a Submission Information Package (SIP). The SIMBAD team requires SIPs to be comprehensive and properly formatted to ensure that the information can be easily added to the database and accurately used by other users. According to the SIMBAD website, the following information should be included in a SIP:

1. Object name: The name of the celestial object being submitted, following the conventions used by SIMBAD.
2. Coordinates: The accurate position of the object in the sky, including the right ascension and declination.
3. Magnitude: The apparent magnitude of the object, if available.
4. Spectral type: The spectral type of the object, if known.
5. Bibliographic references: A list of relevant bibliographic references, such as articles or catalogues that contain information on the object.
6. Additional information: Any other relevant information about the object, such as its size, luminosity, or distance.

The SIMBAD database requires metadata to be submitted in a specific structure and according to specific standards to ensure that the information can be easily added to the database. The SIMBAD database uses its own data format called the Bibliographical Query Format (BIQ) for bibliographic information. This format provides a standard way of representing bibliographic data including author names, publications titles, and publication dates. For object data, SIMBAD requires that coordinates and magnitudes be specified in standard astronomical units, and that spectral types be specified using the standard spectral classification system. SIMBAD also provides a set of standard object names, which should be used when submitting data on new objects. If users need help, they are directed to contact [cds-question@astro.unistra.fr](mailto:cds-question@astro.unistra.fr).

A login is not required to download data from the SIMBAD Astronomical Database, and it is freely accessible to the public and can be searched and browsed without the need to create an account. However, if you wish to contribute data to the database by submitting a SIP, users will need to create an account with the CDS. To create an account with the CDS, users can visit the [CDS Portal](http://cdsportal.u-strasbg.fr/), click on Register in the top right corner, and follow the instructions. Once users create an account, they can log in to the SIMBAD website to submit SIPs and contribute data to the database. Creating an account also allows users to keep a personal, persistent storage space in the CDS portal where users can upload data files and keep results in their own CDS repository. Additionally, users can post annotations on SIMBAD objects, use the CDS cross-match services, and follow the bibliography of your favorite SIMBAD objects.

The database provides several different ways to access its data, including:

1. Direct file download: Users can download data files directly from the SIMBAD website in various formats, such as ASCII and VOTable.
2. Database query: Users can perform database queries using the SIMBAD search interface, which allows them to search for objects and retrieve information about them. The results of a query can be displayed in the browser or downloaded in various formats.
3. Automated script: Users can also access the database programmatically using automated scripts. The SIMBAD website provides information on how to access the database using various programming languages, such as Python and IDL.
4. Virtual Observatory (VO) services: SIMBAD data is also available through Virtual Observatory (VO) services, which allow users to perform complex data analysis and visualization using tools such as Aladin, TOPCAT, and others.

The SIMBAD Astronomical Database displays metadata using specific metadata standards and uses the International Astronomical Union (IAU) standards for object names and the Bibliographical Query Format (BIQ) for bibliographic information. Additionally, as stated above SIMBAD provides access to its data through VO services which use metadata standards such as the VO Table format to represent astronomical data.

The Dissemination Information Package (DIP) in the SIMBAD Astronomical Database typically includes the following information:

1. Object data: This includes information on specific astronomical objects, such as their names, coordinates, magnitudes, and spectral types.
2. Bibliographic information: This includes information on the sources of the object data, such as author names, publication titles, and publication dates.
3. Cross-referenced data: This includes links to related information in other astronomical databases, such as the NASA/IPAC Extragalactic Database (NED) and the VizieR catalogue service.
4. Link to images: This includes links to images of astronomical objects, where available, from sources such as the Digitized Sky Survey (DSS) and other astronomical image archives.
5. Other related information: This may include additional information on astronomical objects, such as their morphological type, redshift, and variability.

Repository:

[SIMBAD Astronomical Database – CDS (Strasbourg)](http://simbad.cds.unistra.fr/simbad/)

**Recommended Data Citation**

Albareti, F.D., Allende Preto, C., Almeida, A., et al. (2017). Stellar Classification Dataset—SDSS17. Sloan Digital Sky Survey. https://www.sdss.org/

**Considerations for Long-Term Preservation**

When it comes to long-term preservation, the file formats and software needed to access the data are important considerations. In the case of the Stellar Classification Dataset—SDSS17, it’s worth noting that the data is stored in the SDSS Spectroscopic Data Model (SDM) format, which is a proprietary format specific to the Sloan Digital Sky Survey.

While the SDM format is well-documented and widely used within the astronomy community, it’s possible that it may become obsolete or unsupported in the future. To ensure long-term preservation of the data, it’s therefore recommended that the data be converted to open, widely-used file formats such as FITS (Flexible Image Transport System\_ or HDF5 (Hierarchical Data Format version 5) whenever possible.

Additionally, it’s important to ensure that any software needed to access and analyze the data remains available and compatible over time. In the case of the SDSS, the data is typically accessed using software tools such as the SDSS Science Archive Server (SAS) or the SDSS SkyServer, which are publicly available and well-supported by the SDSS collaboration. However, it’s always possible that changes in technology or funding could impact the long-term availability of these tools. To mitigate this risk, it’s recommended that the data be accompanied by detailed documentation outlining the software and tools needed to access and analyze the data, and that efforts be made to ensure that open-source software alternatives exist to provide ongoing support for the data over time.

**Copyright License Statement**

An appropriate copyright license for the Stellar Classification Dataset—SDSS17 would be one that allows for free and open access to the data while also protecting the intellectual property rights of the original creators. One such license is the Creative Commons Attribution (CC-BY) license, which allows for free use, distribution, and adaptation of the data if proper attribution is given to the original creators.

The CC-BY license is a popular choice for scientific data sets because it enables researchers and other users to freely access and analyze the data while also ensuring that the creators of the data receive credit for their work. It also allows for the data to be reused in other projects, including commercial projects, if proper attribution is given.

**Human Subject Considerations**

The Stellar Classification Dataset—SDSS17 does not contain any personally identifiable data about people, as it consists of astronomical data obtained from observations of stars. Therefore, there are no specific human subject considerations that apply to this dataset.