LIS 546A: Data Curation II

1. Project Definition, Scope, and Audience

1.1. Group Name: Datavision

1.2. Group Members: Javier Martinez Rodriguez, Bubbles Cram

1.3. Topic:
The project aims to develop a curation protocol for analyzing and comparing predictions of placement at the Eurovision Song Contest made by different predictors. The Eurovision Song Contest is an internationally televised songwriting competition featuring participants from across Europe and beyond. The project will curate relevant data to determine the most successful predictors by contrasting their predictions with the actual final results. The curation protocol will be designed to continuously update the data as new contests occur.

1.4. Target Audience:
The target audience for this curation protocol includes the Eurovision fandom, which consists of fans of the contest, bloggers, content creators, and media outlets dedicated to the contest. These individuals and their audiences are increasingly interested in analyzing data related to the Eurovision Song Contest to gain insights and predictions about the results.

1.5. Collective Goals:
The main goal for our group is to understand the data curation process with a focus on data integration. The main task for this project is to collect and integrate contest and prediction data into distinct datasets that can then be replicated for future editions. An important goal for our team, considering the nature of our data, is to become more familiar with data integration and packaging. Finally, we would like to understand the many components of a data curation protocol.

1.6. Relevant Data Sources:
1.6.1. Eurovision Grand Final Official Results: The actual results of the contest as published and verified by the European Broadcasting Union. This would be the benchmark data to determine success (https://eurovision.tv/event/liverpool-2023/grand-final). The website EurovisionWorld also offers a breakdown of the official results (https://eurovisionworld.com/eurovision/2023).

1.6.2. OGAE Poll Results: OGAE is the biggest fan club for followers of the Eurovision Song Contest. They have chapters in each participating country and one chapter for fans around the rest of the world. They conduct an annual poll with points assigned per chapter of the club. This is regarded as an indicator of the favorite songs by fans of the contest (https://ogaeinternational.org/2023-ogae-poll/).

1.6.3. Eurovoix's Euro Jury Results: this is a poll conducted by Eurovision media outlet Eurovoix, According to their website: "Participating jury members are asked to rank their top ten songs of this year's Eurovision Song Contest, which will then be combined with the results of other jurors from their

country to make up one national result. This national result, when added to those of other countries, will produce the final 'jury vote'." (https://eurovoix.com/category/theeurojury/)

1.6.4. EurovisionWorld's Odds Summary: The results as predicted by bookmakers. This data is presented and separated by the bookmaking organization and an aggregate is available for each competing country to determine their position in the scoreboard (https://eurovisionworld.com/odds/eurovision).

2. User Stories

The following User Stories were envisioned as a way to guide the creation of this protocol:

Goals	**User Stories**
Access a clean dataset.	As a Eurovision content creator, I want access to a curated dataset that includes the official contest results, bookmakers' predictions, and the results of the OGAE Poll. This will allow me to analyze the data and provide accurate and up-to-date predictions to my audience.
Visualize compiled data and understand how it was compiled.	As a Eurovision enthusiast, I want a user-friendly interface and documentation that allows me to access and utilize the curated data. This will help me explore the correlations between predictions and actual results, enhancing my understanding of the contest.
Obtain all necessary data for analysis.	As a Eurovision content creator, I need a methodology or protocol for analyzing predictions and evaluating the performance of different predictors. This will enable me to assess the accuracy and reliability of various prediction methods and incorporate them into my analysis.
Contrast prediction and actual results in one package.	As a Eurovision enthusiast, I want to be able to compare the predictions made by different predictors against the actual final results. This will help me identify the most successful predictors and learn from their approach.
Find data for new editions in the same place and format.	As a Eurovision content creator, I need the curated data to be regularly updated with new contest results. This will ensure that I have access to the latest information for my analysis and predictions.
Make informed predictions based on previous trends.	As a Eurovision enthusiast, I want to navigate and understand the data effectively, enhancing my ability to engage with the Eurovision community and make informed predictions.

Find all relevant data in a single package without navigating to many different sources.	As a Eurovision fan, I want the convenience of accessing the relevant data (scores, placements, etc.) from a single source of information rather than browsing through multiple websites or sources. The relevant data in a single source will save me time and frustration, and allow me to easily find the information I desire about the contest without investing excessive amounts of time.

3. Collection Policies:

3.1. Deposit Processes:

Datavision uses a combination of an opportunistic approach and a submission process.

- Data on subsequent editions of the contest will generally be added to the collection as a contribution from the curating team at Datavision.
- Considering that predictions and poll data may exist for past editions, Datavision accepts submissions of data on previous editions of the contest. Submissions must include
    - Contact information for the contributor of data.
    - Values for the following metadata elements for each of the submitted datasets, submitted through a structured form:
        - Title according to our naming conventions: 'Year_FinalPredictions' or 'Year_OrganizationName_Poll'.
        - Date in ISO 8601 format of the creation of each prediction, particularly for odds data.
        - A description of the dataset that includes an overview of the organizations included in it (1000 character limit).
        - Names of all publishers included in the dataset, including the European Broadcasting Union.
        - URL's or any other source or citation for all information included in the dataset.
    - The contributor must include a statement certifying that they have all necessary permissions for the submission of data including compliance with any applicable licensing requirements from the creators of the raw data.
    - The contributor must include a statement granting permission to Datavision for the storage and archiving of the dataset or datasets.

3.2. Inclusion Criteria:

Datasets will be included in the collection provided they comply with the following requirements:

- Content
    - All datasets must contain results and predictions of the results of a single edition of the Eurovision Song Contest. Datasets where predictions and results are not from a single year will be rejected.
    - In order to meet the above requirement, each dataset must include the official results and at least one set of prediction data. Datasets including only results will be rejected.

- Type
    - All datasets must include numeric data for all applicable rankings and scores except where information is missing (e.g. a country appears in the final ranking but not in a prediction).
    - Datavision does not accept images, videos, or any file type that is not alphanumeric values.
- Format
    - Only electronic files will be accepted.
    - Only tabular data will be accepted, specifically in CSV (Comma-Separated Values) format.
    - Files in other formats may be accepted at the sole discretion of Datavision given that they are easily transformed into CSV format.
    - All datasets in formats deemed ineligible by the Datavision team will be rejected.
- Size
    - Submissions may not exceed 1GB in size.

3.3. Preservation Policies:

The following preservation actions will be taken for all files in the collection:

- bitstream maintenance
- persistent, permanent identifier
- preservation metadata
- onsite and offsite backup copies
- regular virus and file corruption checks
- periodic refreshments to new storage media
- Monitor file format for changes that might warrant transformation or reassessment
- Migration of document to successive format when necessary

Note that all files are required to be stored in CSV format. The homogeneity of file formats allow us to have a consistent preservation policy for all files in the collection.

3.4. Removal and Archiving:

Files in the collection will be available for a period of 10 years prior to their inclusion in a data archive. The Datavision team will take different metrics into consideration prior to making any archiving decisions, such as:

- Traffic to the dataset
- Number of times a dataset has been downloaded
- Number of times we have been credited for a particular dataset
- General user demand for a particular dataset

The metadata for all files will remain available in the dataset page and contact information will be available for all requests.

4. Transformations and Quality Criteria:

4.1. Data Transfer:

The following outlines the processes for data transfer into the Datavision Collection provided that the submission has been approved by the Datavision team:

- Datavision will use a Google Form as a way to upload all files for a specific country.
- Please note that naming conventions are essential for the identification of files and their appropriate ingestion into the collection.
- A different form must be used for uploads of datasets related to each separate edition of the contest.
- All datasets will first be analyzed by virus detection software.
- As part of our quality assurance efforts, we will perform a spot check of the official rankings of the edition against official sources. Additionally, at least one set of prediction data will be spot checked for accuracy.
- No submitted files should include confidential or sensitive information. Files including any such information will be immediately excluded from the collection.
- Each Dataset will be assigned a DOI identifier by the curation team.
- All necessary metadata, per our outlined metadata profile, will be included at the point of data transfer by the curation team. This includes the aforementioned DOI identifier.

4.2. Data Transformation:

4.2.1. Files/Format:

- All files will be stored as CSV files, this is the same file format in which it will be made available to users. This will also facilitate future data visualization efforts for the collection.
- As outlined in our data collection and data transfer policies, datasets must be submitted as CSV files. No proprietary files will be accepted for ingestion into the collection.
- It is not expected that datasets will undergo any major changes given the nature of the compiled data. Namely, this data is published once and does not undergo any changes or updates after its publication. Because of this, no updates are expected after files have been released for reuse. In the case that this happens, the name will be updated to include a version number (e.g. 2023_FinalPredictions_1.0).

4.2.2. Naming Conventions

- All file names will begin with the year of the edition to which the data refers (e.g. 2023)
- All final predictions files will then include this information (e.g. FinalPredictions)
- In the case that there are multiple versions of a dataset, this will be indicated as a number in the file name
- All elements of the name of the file will be separated by an underscore character “_”
- A final prediction file name example: 2023_FinalPredictions_1.0
- All top 10 prediction files will include the name of the organization compiling the top 10 predictions
- All top 10 prediction files will include the word “Poll” as a final element.
- A top 10 prediction file name example: 2023_OGAE_Poll_2.0

- Folders will be named in accordance with the year of the edition to which the data corresponds and the phrase “PredictionData” (e.g. 2023_PredictionData).
- All self-explanatory files will be given the name of the subject of the file and, if applicable, the master file to which they belong (e.g. 2023_PredictionData_ReadMe)

4.2.3. Data Values

The following data values transformations are necessary for the ingestion of a dataset into the collection:

- Special characters will be removed from all datasets
- Rows will be organized based on the countries column, which will be organized in alphabetical order from A to Z.
- All country names will be spelled out and articles will be removed (e.g. ‘Netherlands’ used for ‘The Netherlands’)
- Variable names will be changed based on our standard naming convention (NameOfOrganization_Ranking OR NameOfOrganization_Score)
- All missing values will be standardized as blank
- Information such as URL’s directing to the origin source will be moved to the file’s metadata
- Removal of all formulas for calculation of totals, only values will remain

For Final Predictions files, the following changes may be made:

- Trimming of rankings to a number in accordance with the official result. The countries in excess of the number of countries participating in the final will be removed from the dataset. The blank value may be interpreted as a failure to predict a specific country’s participation in the grand final.

For Top 10 predictions by individual clubs or juries the following changes may be made:

- A ranking value will only be provided for the top 10 countries in the official ranking to indicate a benchmark for the juries or clubs.
- The total official scores for the top 10 countries will be transformed into a 12 to 1 score in accordance with Eurovision Song Contest rules to facilitate comparison between predictions and rankings.
- All variable names will be transformed using the following format: Name of Organization followed by name of voting country (e.g. OGAE_Albania or Eurovoix_UnitedKingdom)

5. Metadata Application Profile

5.1. General Considerations:

There were two main considerations when selecting a set of metadata terms for this project: user needs and previous practices. User needs considerations are based on our user stories which, in their majority, indicate that users were consulting several different sources to find this information. The necessity to create an aggregated dataset that users can consult is a driving factor of this project. To accommodate this need, our selected metadata terms include information about the year of the contest edition being

predicted, it also includes information about the sources for the prediction information as well as the source for the official results.

The second consideration was more difficult to accommodate because it has to do with previous practice. Although there is certainly previous practices related to prediction markets, we could not find any antecedents using Eurovision data. Moreover, the data that we have integrated would represent one of many factors to consider when attempting to make an accurate prediction for Eurovision. Because of this, and to ensure that enthusiasts of the contest can access and work with the data, we have aimed for simplicity and ease of use.

Considering all the above, we have chosen to leverage the terms created as part of the Dublin Core Metadata Initiative (DCMI) (https://www.dublincore.org/). Dublin Core is a highly flexible schema that can be used to accurately represent the characteristics of the resources compiled as part of this project. DCMI is an established standard that is maintained and updated regularly. Moreover, after our assessment of the necessary elements for our project, we determined that Dublin Core accommodated all our needs. These elements are, in alphabetical order:

- Created
- Date
- Description
- Format
- Is Part Of
- Identifier
- Publisher
- Source
- Title
- Type

We were able to apply these terms, as defined by DCMI, without little to no modifications. One of the modifications was to restrict the 'created' date to only the year as this would suffice in identifying the year to which datasets refer. Another motivator was to avoid the use of placeholder days and months as values within this element. We also chose to elaborate on requirements for certain elements. For instance, we have included a restriction to the amount of allowable characters in 'description'. Similarly, we have leveraged the 'format' term to include both the file format and the size of the file as this is important information to provide to users. Finally, we have chosen XML as the encoding scheme for our records.

Detailed information regarding the selected metadata elements for this collection can be found in Metadata_Elements.xlsx.

5.2. Controlled Vocabularies:

The use of controlled vocabularies was generally omitted due to the homogeneity of the datasets in terms of their topic, subject, or theme. There are only restrictions for the following terms: date, isPartOf, Publisher, Title, and Type. The following table outlines the respective restrictions:

Term	**Restriction**
date	Only the year of the contest is to be included. Refrain from included month and day information.
isPartOf	The values in this field must be identical to the title value of a separate dataset of which the indexed dataset must be a logical part.
title	We have established a naming convention for files as outlined in the collection policies.
type	The type of dataset must be indicated based on the benchmark used for results: 'Final Ranking' or 'Top 10'. This is the closest we need to a controlled vocabulary and each value indicates whether or not a dataset includes data for the top 10 or for the final ranking for all participating countries (the number of countries varies by year).

6.0 Licensing 

Notice: Database License: Attribution-NonCommercial (CC BY-NC) This notice serves to inform users that the database available on our platform is licensed under Attribution-NonCommercial (CC BY-NC). This license allows others to remix, adapt, and build upon the database for non-commercial purposes. While any derivative works must acknowledge the original source and be non-commercial, there is no requirement for the derivative works to be licensed under the same terms. Please note that the publishers of the original raw data are welcome to contact our team if they wish to discuss or request a change in the license of our database. We value open collaboration and are open to considering modifications to ensure the appropriate licensing of the data. If you have any questions or require further information regarding the database license or any other related matter, please reach out to our team. We are here to assist you. Thank you for your understanding and cooperation.
Sincerely, Datavision