Web Data Integration
Open Music Data Integration

Anh-Nhat Nguyen[2034311], Ching-Yun Cheng[2112322], Shamalan Rajesvaran[2115475], Yen-An Chen[2113612], and Phelan Lee Yeuk Bun[2053019]

Team 1

1 Introduction

In this Web Data Integration project, we aim to consolidate and analyze standardization from multiple sources to gain comprehensive insights into music streaming trends, track performance, and audience preferences. The project leverages three key datasets:

1. Million Song Dataset with Spotify and Last.fm Features Dataset (csv) [4]: This dataset is an enriched version of the Million Song Dataset, a large-scale music database containing detailed metadata and audio features for over 50,000 tracks with 21 attributes. It integrates additional attributes from Spotify and Last.fm, including audio features like danceability, energy, loudness, and popularity metrics such as tags (list attribute), preview URLs, and genre classifications. The merging of these three data sources provides a comprehensive view of each song, making it suitable for analyzing music trends, listener behaviors, and track popularity across platforms.

2. Apple Music Tracks (csv) [2]: This dataset contains detailed information on 10,000 tracks with 24 attributes sourced from Apple Music. It includes attributes such as artist names, album titles, track features (e.g., tempo, key, mode), and genre classifications. In addition to audio metadata, the dataset provides insights into song popularity metrics and trends across the platform. It is ideal for exploring the characteristics of songs on Apple Music, understanding artist performance, and analyzing trends in genres and musical features.

3. Openmusic API Dataset (json) [3]: This dataset offers details about over 5,500 tracks via web APIs (retrieved by /explore & /album?id=<AlbumID>), including track and album metadata, artist information, and playback types (clean/explicit). The use case for this dataset revolves around leveraging real-time API data for in-depth analysis of track consumption patterns, artist popularity, and changes in audience preferences over time.

Together, these datasets will help provide a holistic view of how various musical, commercial, and audience factors contribute to the success of music tracks on different platforms based on various algorithm approach [1].

2 Data Collection and Data Translation

2.1 Data Collection and Dataset

Overview of the Datasets The dataset was obtained from Kaggle in the form of *csv* and OpenMusic API in the form of *csv* and *json*. An overview of dataset attributes is presented in Table 1 below.

Table 1. Dataset structure

Dataset	No Entities	No. Attributes	Attributes
Million Song Dataset with Spotify and Last.fm Features	50,683	21	Track ID, Name, Artist, Spotify Preview URL, Spotify ID, Tags, Genre (*MV 56%*), Year, Duration MS, Danceability, Energy, Key, Loudness, Mode, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, Time Signature
Apple Music Tracks	10,000	24	Artist ID, Artist Name, Collection Censored Name, Collection ID, Collection Name, Collection Price, Content Advisory Rating (*MV 85%*), Country, Currency, Disc Count, Disc Number, Is Streamable, Kind, Preview URL, Primary Genre Name, Release Date, Track Censored Name, Track Count, Track Explicitness, Track ID, Track Name, Track Number, Track Price, Track Time (Milliseconds)
Open Music	5,558	18	ShelfTitle, AlbumId, AlbumName, AlbumArtwork, AlbumType, AlbumYear, ArtistId, ArtistName, ArtistProfilePhoto, ArtistSubscribers, TrackId, TrackTitle, TrackPlaybackClean, TrackPlaybackExplicit, TrackLength, TrackIndex, TrackViews, TrackFeatures

Data Preprocessing Data preprocessing steps included filtering irrelevant attributes and normalizing data types.

2.2 Schema Mapping

Design of the Integrated Schema The three datasets contain multiple overlapping attributes as seen in Table 2 below. There are 5 attributes within our integrated schema that overlap across at least 2 of 3 input schemata, namely "Artist", "Track Name", "Genre", "Track Duration" and "Release Date".

Table 2. Table of Integrated Schema Attributes

Attribute Name	Datatype	Datasets in which attribute found
Track	string	Spotify Musicality, Spotify Streaming Statistics, Openmusic
Artist	string	Spotify Musicality, Spotify Streaming Statistics, Openmusic
Album	string	Spotify Musicality, Spotify Streaming Statistics, Openmusic
Youtube Views	decimal	Spotify Streaming Statistics, openmusic
YouTube Likes	decimal	Spotify Streaming Statistics
Release Date	datetime	Spotify Streaming Statistics, openmusic
Danceability	decimal	Spotify Streaming Statistics
Energy	decimal	Spotify Streaming Statistics
Key	decimal	Spotify Streaming Statistics
Loudness	decimal	Spotify Streaming Statistics
Speechiness	decimal	Spotify Streaming Statistics
Acousticness	decimal	Spotify Streaming Statistics
Instrumentalness	decimal	Spotify Streaming Statistics
Liveness	decimal	Spotify Streaming Statistics
Valence	decimal	Spotify Streaming Statistics
Tempo	decimal	Spotify Streaming Statistics
TrackPlaybackClean	string	openmusic
TrackPlaybackExplicit	string	openmusic

Tools and Challenges Tools used include Altova MapForce. Challenges encountered during schema mapping included aligning attributes and handling missing data.

Conversion to Target Schema Datasets were converted to the target schema resulting in XML files.

3 Phase II: Identity Resolution

3.1 Initiate Gold Standard

Method for Building the Gold Standard The gold standard was built using a sampling method ensuring matches, non-matches, and corner cases.

3.2 Challenges with Gold Standard and Improvement

Edit-Distance with Single Key Edit-distance with a single key (Track) did not have high coverage. Solutions included using multiple keys and advanced matching techniques.

Selection Bias Addressed selection bias by ensuring diverse and representative samples.

Performance Improvement Implemented blocking, Bloom Filtering, and used LLM for performance improvement.

3.3 Matching Strategies

Similarity Metrics Used Levenshtein for names and numeric thresholds for streams.

Blocking Techniques Blocking techniques improved efficiency by reducing the number of comparisons.

3.4 Evaluation

Metrics and Analysis Metrics used included precision, recall, and F1 score. Analysis of results highlighted challenges such as noisy data and near-duplicate names.

Benchmark Table A benchmark table was created to compare different matching strategies.

4 Data Fusion

4.1 Fusion Rules

Conflict Resolution Strategies Conflict resolution strategies included prioritizing reliable datasets, averaging numeric attributes, and using union for lists.

Specific Fusion Rules Specific fusion rules for key attributes included taking the shortest string for "Track Name" and averaging for "Streams".

4.2 Fused Data Output

Post-Fusion Dataset Post-fusion dataset size and density improvements were noted. Examples of fused records were provided.

4.3 Quality Evaluation

Metrics for Evaluation Metrics for evaluating the quality of the integrated dataset included accuracy, consistency, and density. Summary of improvements achieved through data integration was provided.

4.4 Challenges and Lessons Learned

Addressed issues like conflicting values, missing data, or incorrect matches from identity resolution.

5 Conclusion and Future Work

5.1 Limitations of the Project

Discussed limitations such as incomplete data and computational constraints.

5.2 Recommendations for Future Improvements

Recommendations included using advanced ML models for identity resolution.

References

1. AnHai Doan, Alon Halevy, and Zachary Ives. *Principles of Data Integration.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, June 2012.
2. Kanchana1990. Song dataset: 10,000 apple music tracks. URL: https://www.kaggle.com/datasets/kanchana1990/apple-music-dataset-10000-tracks-uncovered.
3. OatsCG. OatsCG/Openmusic-Server-Specs, September 2024. original-date: 2024-01-10T01:36:38Z. URL: https://github.com/OatsCG/Openmusic-Server-Specs.
4. UndefineNull. Million song dataset + spotify + last.fm. URL: https://www.kaggle.com/datasets/undefinenull/million-song-dataset-spotify-lastfm.