Musicbrainz + Discogs Datasets

Eiler Schiötz and Nicole Currens
(Jeffersonballers)

Supplementing Music Encyclopedias

Interest: Austin has lots of small bands. They can be more successful if they are easily searchable by their desired audience.

Question: Can we improve music datasets by combining them?

Musicbrainz Overview

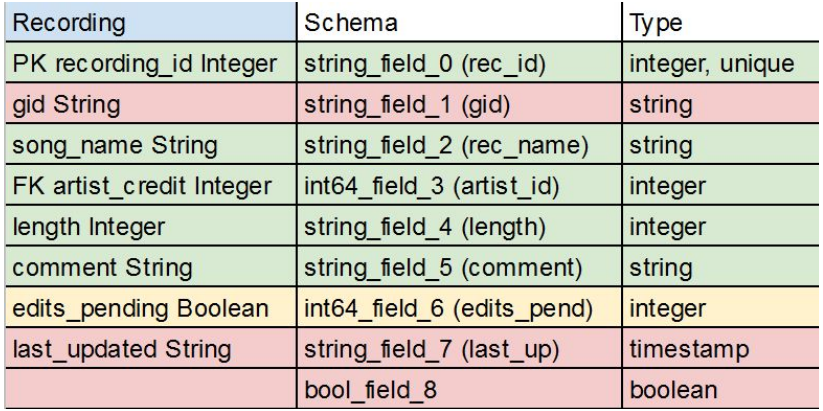
- JSON Data
- 19 Tables ranging from 4 - 2,434,536 rows
- Well-modeled but a work in progress
- User-built data

Discogs Overview

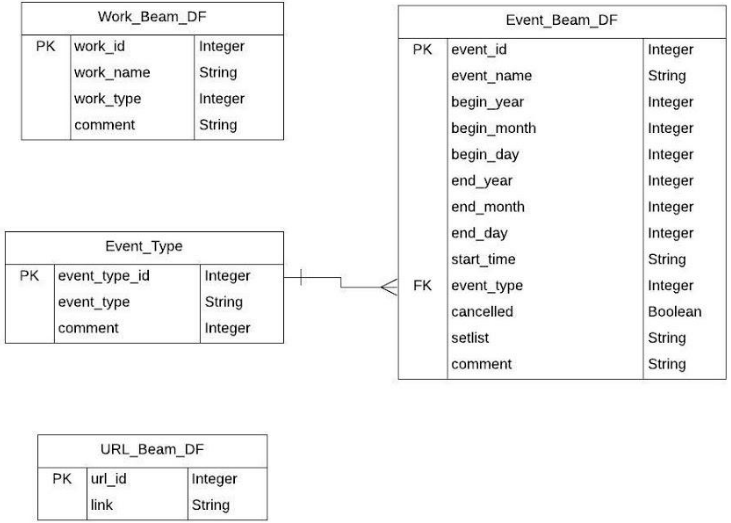
- XML Data (Nested Structure)
- Artist Table (6,789,773 rows)
- Label Table (1,506,930 rows)
- Includes name variations, urls, addresses, etc.
- User-built data

MusicBrainz Modeling

- Initially well-modeled, included entity-type tables
- Outdated MusicBrainz documentation
  - Missing columns names
  - Deleted, column headers, tables
- Disconnected tables
  - Work, URL, Events



Recording	Schema	Type
PK recording_id Integer	string_field_0 (rec_id)	integer, unique
gid String	string_field_1 (gid)	string
song_name String	string_field_2 (rec_name)	string
FK artist_credit Integer	int64_field_3 (artist_id)	integer
length Integer	string_field_4 (length)	integer
comment String	string_field_5 (comment)	string
edits_pending Boolean	int64_field_6 (edits_pend)	integer
last_updated String	string_field_7 (last_up)	timestamp
	bool_field_8	boolean



Work_Beam_DF		
PK	work_id	Integer
	work_name	String
	work_type	Integer
	comment	String

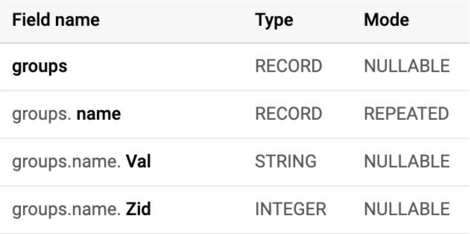
Event_Beam_DF		
PK	event_id	Integer
	event_name	String
	begin_year	Integer
	begin_month	Integer
	begin_day	Integer
	end_year	Integer
	end_month	Integer
	end_day	Integer
	start_time	String
FK	event_type	Integer
	cancelled	Boolean
	setlist	String
	comment	String

Event_Type		
PK	event_type_id	Integer
	event_type	String
	comment	Integer

URL_Beam_DF		
PK	url_id	Integer
	link	String

Discogs Modeling

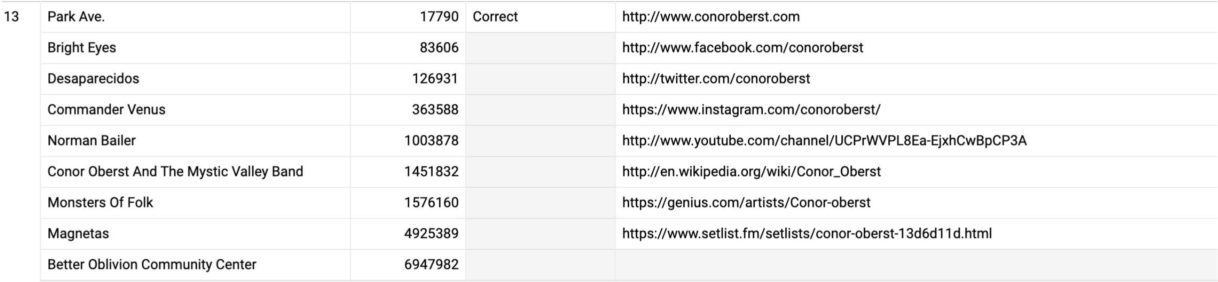
- Initial difficulty with nested XML data - required cleaning to remove wrapper columns



Field name	Type	Mode
groups	RECORD	NULLABLE
groups. **name**	RECORD	REPEATED
groups.name. **Val**	STRING	NULLABLE
groups.name. **Zid**	INTEGER	NULLABLE

- Initial tables

  not modeled



13	Park Ave.	17790	Correct	http://www.conoroberst.com
	Bright Eyes	83606		http://www.facebook.com/conoroberst
	Desaparecidos	126931		http://twitter.com/conoroberst
	Commander Venus	363588		https://www.instagram.com/conoroberst/
	Norman Bailer	1003878		http://www.youtube.com/channel/UCPrWVPL8Ea-EjxhCwBpCP3A
	Conor Oberst And The Mystic Valley Band	1451832		http://en.wikipedia.org/wiki/Conor_Oberst
	Monsters Of Folk	1576160		https://genius.com/artists/Conor-oberst
	Magnetas	4925389		https://www.setlist.fm/setlists/conor-oberst-13d6d11d.html
	Better Oblivion Community Center	6947982		

Beam Cleaning & SQL Transformations

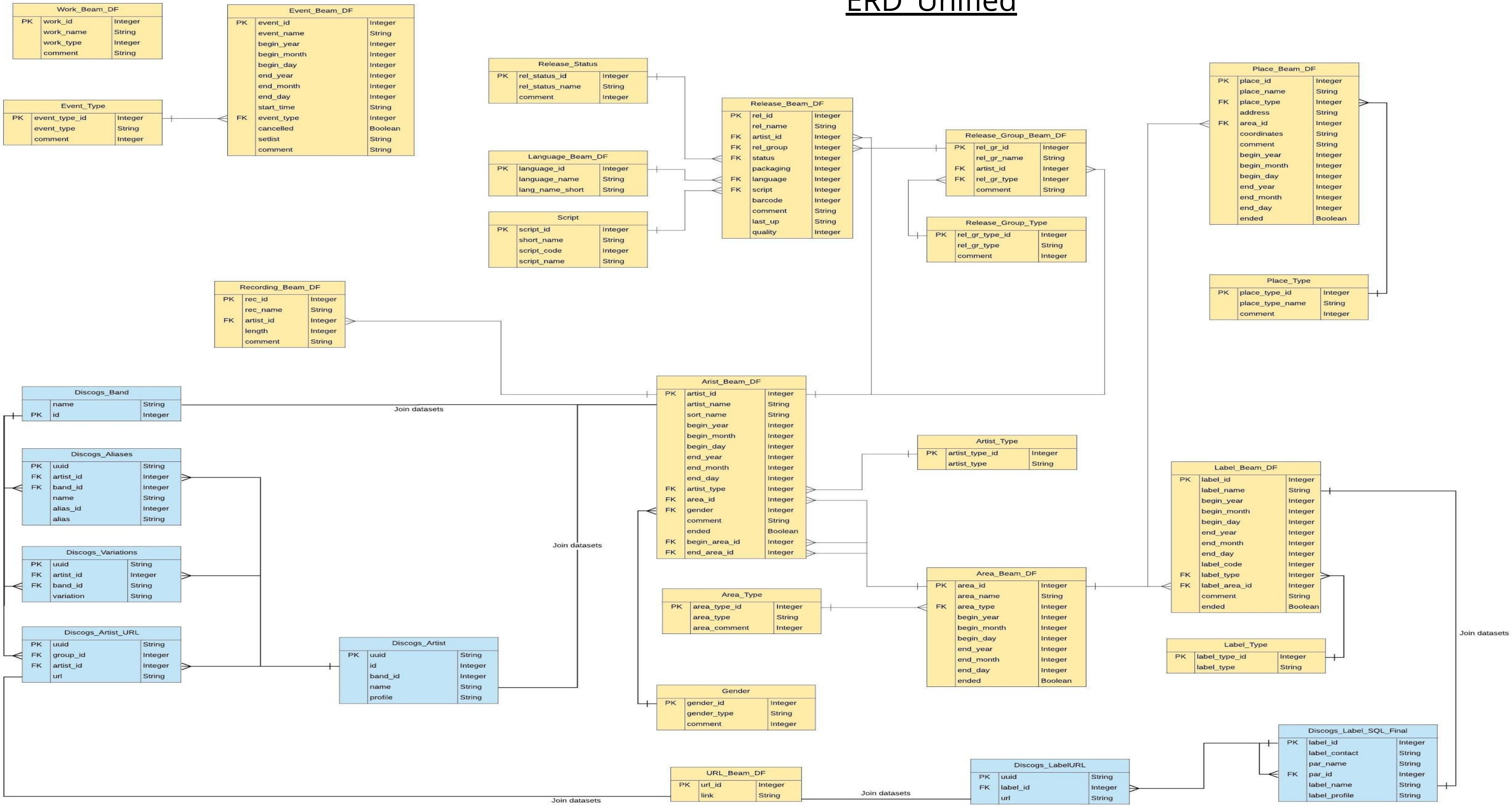
MusicBrainz Beam Cleaning

- Changed blank cells to nulls
- Typecasting
  - ‘T’ and ‘F’ to true and false
  - Strings to Integers

Discogs SQL Transformations

- Removed records with “Entirely Incorrect” data quality
- SQL transform to get the label to refer to the parents

ERD Unified



Work_Beam_DF

PK	work_id	Integer
	work_name	String
	work_type	Integer
	comment	String

Event_Type

PK	event_type_id	Integer
	event_type	String
	comment	Integer

Event_Beam_DF

PK	event_id	Integer
	event_name	String
	begin_year	Integer
	begin_month	Integer
	begin_day	Integer
	end_year	Integer
	end_month	Integer
	end_day	Integer
	start_time	String
FK	event_type	Integer
	cancelled	Boolean
	setlist	String
	comment	String

Release_Status

PK	rel_status_id	Integer
	rel_status_name	String
	comment	Integer

Language_Beam_DF

PK	language_id	Integer
	language_name	String
	lang_name_short	String

Script

PK	script_id	Integer
	short_name	String
	script_code	Integer
	script_name	String

Release_Beam_DF

PK	rel_id	Integer
	rel_name	String
FK	artist_id	Integer
FK	rel_group	Integer
FK	status	Integer
	packaging	Integer
FK	language	Integer
FK	script	Integer
	barcode	Integer
	comment	String
	last_up	String
	quality	Integer

Release_Group_Beam_DF

PK	rel_gr_id	Integer
	rel_gr_name	String
FK	artist_id	Integer
FK	rel_gr_type	Integer
	comment	String

Release_Group_Type

PK	rel_gr_type_id	Integer
	rel_gr_type	String
	comment	Integer

Place_Beam_DF

PK	place_id	Integer
	place_name	String
FK	place_type	Integer
	address	String
FK	area_id	Integer
	coordinates	String
	comment	String
	begin_year	Integer
	begin_month	Integer
	begin_day	Integer
	end_year	Integer
	end_month	Integer
	end_day	Integer
	ended	Boolean

Place_Type

PK	place_type_id	Integer
	place_type_name	String
	comment	Integer

Recording_Beam_DF

PK	rec_id	Integer
	rec_name	String
FK	artist_id	Integer
	length	Integer
	comment	String

Discogs_Band

	name	String
PK	id	Integer

Discogs_Aliases

PK	uuid	String
FK	artist_id	Integer
FK	band_id	Integer
	name	String
	alias_id	Integer
	alias	String

Discogs_Variations

PK	uuid	String
FK	artist_id	Integer
FK	band_id	String
	variation	String

Discogs_Artist_URL

PK	uuid	String
FK	group_id	Integer
FK	artist_id	Integer
	url	String

Discogs_Artist

PK	uuid	String
	id	Integer
	band_id	Integer
	name	String
	profile	String

Arist_Beam_DF

PK	artist_id	Integer
	artist_name	String
	sort_name	String
	begin_year	Integer
	begin_month	Integer
	begin_day	Integer
	end_year	Integer
	end_month	Integer
	end_day	Integer
FK	artist_type	Integer
FK	area_id	Integer
FK	gender	Integer
	comment	String
	ended	Boolean
FK	begin_area_id	Integer
FK	end_area_id	Integer

Artist_Type

PK	artist_type_id	Integer
	artist_type	String

Area_Type

PK	area_type_id	Integer
	area_type	String
	area_comment	Integer

Area_Beam_DF

PK	area_id	Integer
	area_name	String
FK	area_type	Integer
	begin_year	Integer
	begin_month	Integer
	begin_day	Integer
	end_year	Integer
	end_month	Integer
	end_day	Integer
	ended	Boolean

Gender

PK	gender_id	Integer
	gender_type	String
	comment	Integer

Label_Beam_DF

PK	label_id	Integer
	label_name	String
	begin_year	Integer
	begin_month	Integer
	begin_day	Integer
	end_year	Integer
	end_month	Integer
	end_day	Integer
	label_code	Integer
FK	label_type	Integer
FK	label_area_id	Integer
	comment	String
	ended	Boolean

Label_Type

PK	label_type_id	Integer
	label_type	String

URL_Beam_DF

PK	url_id	Integer
	link	String

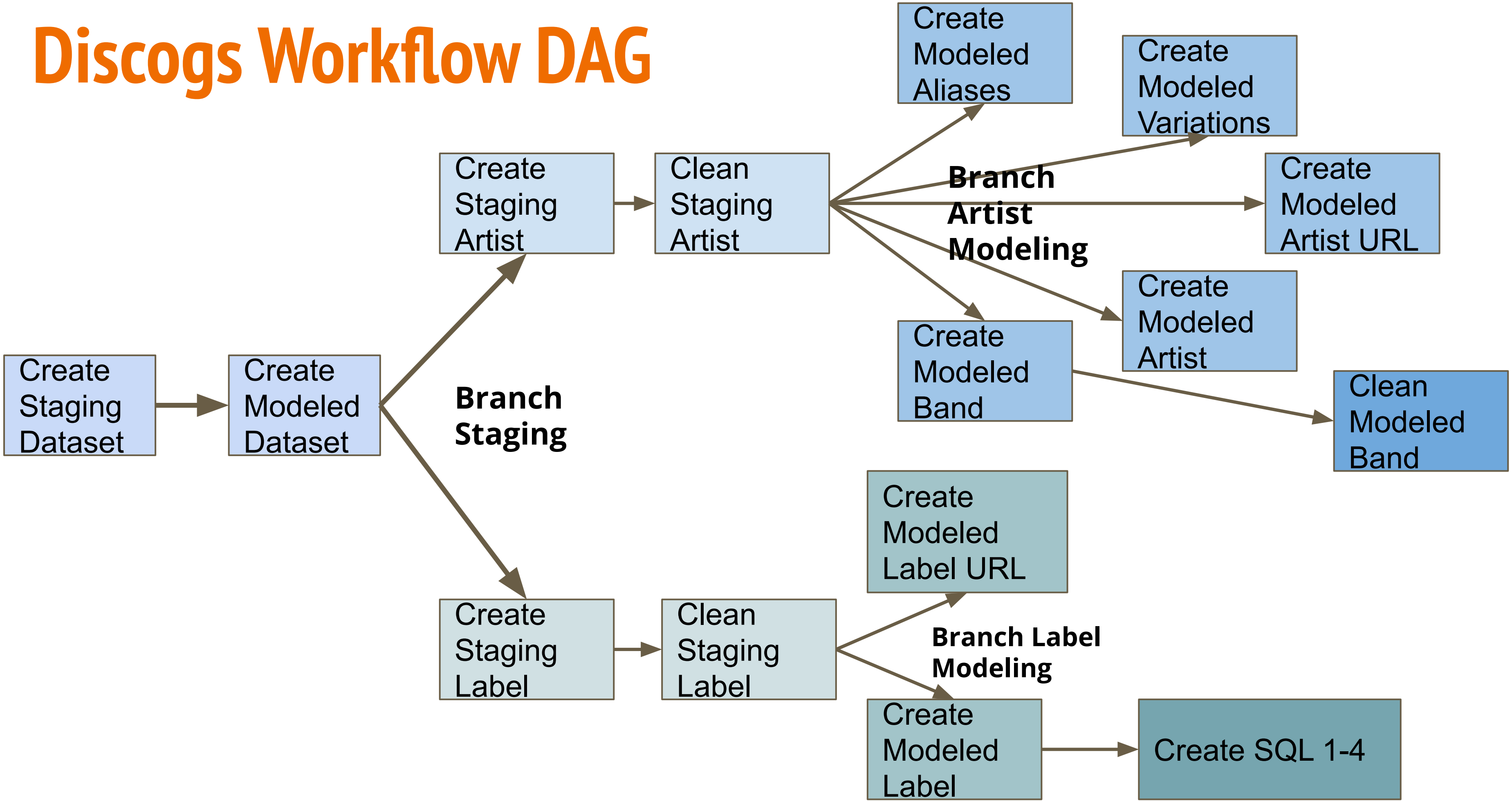
Discogs_LabelURL

PK	uuid	String
FK	label_id	Integer
	url	String

Discogs_Label_SQL_Final

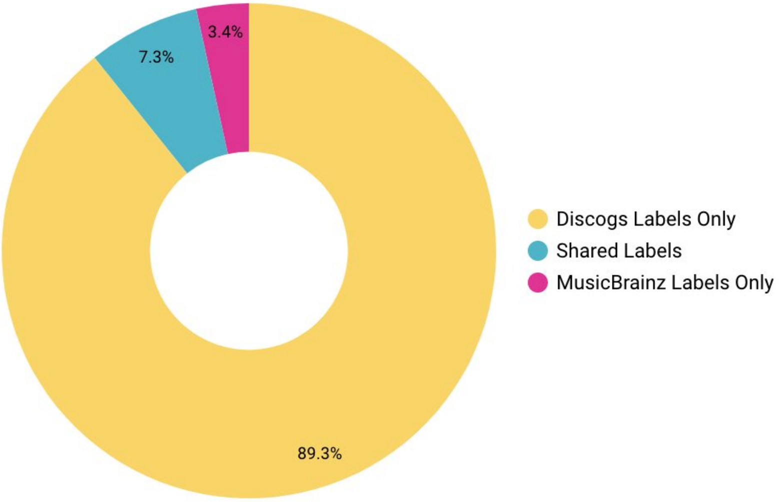
PK	label_id	Integer
	label_contact	String
	par_name	String
FK	par_id	Integer
	label_name	String
	label_profile	String

Discogs Workflow DAG

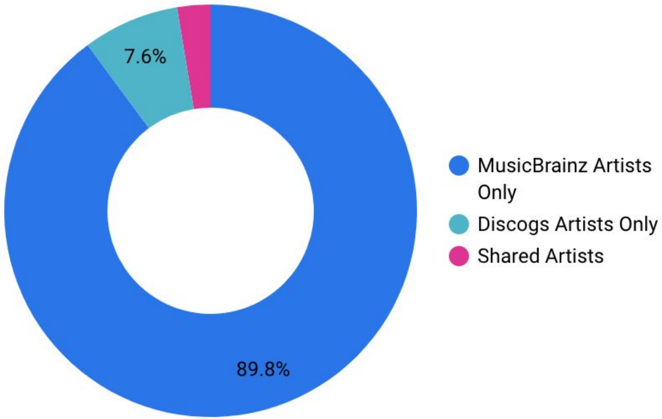


Cross Dataset Queries

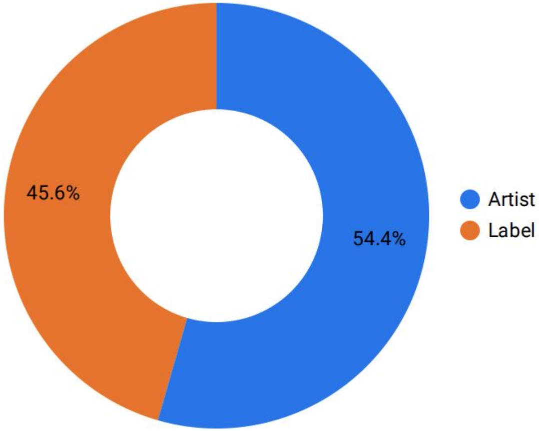
Percentage of Labels From MusicBrainz and Discogs



Percent of Artists from MusicBrainz and Discogs



Types of URLs in both datasets



Demo

SQL Transformations for cross dataset queries

SQL Transformations for Discogs Label table modeling

Future Work

- Link more artists with URLs and check validity
- Could not change Year, Month, Day to datetime in musicbrainz
- Disconnected Work table in musicbrainz

Discogs®

MusicBrainz