|  |
| --- |
|  |
| Introduction To Database |
| Project Report |
| Group 3 :  Joanna Salathé |
| Eliéva Pignat  Sacha Bron |
|  |

|  |
| --- |
| This document contains the 3 deliverables of the introduction to database project. |

Table des matières

[Deliverable 3 1](#_Toc389289343)

[Modification of deliverable 2 1](#_Toc389289344)

[Query H to S 1](#_Toc389289345)

[Index performance 1](#_Toc389289346)

[Graphic Interface 1](#_Toc389289347)

[Deliverable 2 4](#_Toc389289348)

[Importation 4](#_Toc389289349)

[New ER-Model 4](#_Toc389289350)

[Changes in Table Creation 5](#_Toc389289351)

[Search Functionality 7](#_Toc389289352)

[SQL Queries 8](#_Toc389289353)

[Queries A to G 8](#_Toc389289354)

[Interface 12](#_Toc389289355)

[Deliverable 1 14](#_Toc389289356)

[ER-Model Choices 14](#_Toc389289357)

[Constraints Explanation 14](#_Toc389289358)

[Areas 14](#_Toc389289359)

[Tracks 14](#_Toc389289360)

[Recordings 14](#_Toc389289361)

[Releases 15](#_Toc389289362)

[Schemas 16](#_Toc389289363)

[Tables Creation 17](#_Toc389289364)

Deliverable 3

# Deliverable 3

## Modification of deliverable 2

After the feedback, the deliverable 2 was modified in the following way :

1. Some queries had to be modified and corrected. For example we add DISTINCT when we count. And we GROUP BY first by ID and after by Name
2. The query B was completly modified.
3. A short explanations of the query was added

An explanation about the technical part of the graphic interface is added in the deliverable 3 chapter.

## Query H to S

## Index performance

## Graphic Interface

# Deliverable 2

## Importation

We were confronted to some problems during the import part. The « real » data which we have at our disposal are far from the idealized schema that we made for the first part. To reduce those differences, some modifications had to be made.

### New ER-Model

Because some data are missing, artist can be associated with one or zero area and release can be associated with one or zero medium.

### Changes in Table Creation

Weaken constraints :

Some constraints had to be weaken by deleting some FOREIGN KEY. Indeed, some constraints that seem natural during the design part of the database, were finally not well adapted to the available data. For example, in the artist table the FOREIGN KEY on the area ID had to be suppressed because some artists are associated with an area ID that does not exist in the area table. For the same reason the FOREIGN KEY on the artist ID in the artist\_genre table had to be suppressed, as well as the FOREIGN KEY on the track ID in the artist\_track table and the FOREIGN KEY on the Medium ID in the Track table.

Merges for participation constraints :

To capture the constraints that an artist is associated with at most one area, the artist\_area table is merged with the artist table. The same is made for the physical\_song and the medium.

Fields types more accurate :

We also decided to change every field type for the tables creation (before every fields were just CHAR).

Every field for the creation has to be modified. IDs are only integer, thus use integer instead of CHAR are more suitable. For other fields, VARCHAR are used instead of CHAR because the lengths of the field are really variable thus VARCHAR is more appropriated for the real values of the data.

New table creation sql code :

**CREATE** **TABLE** Area(

ID\_Area INT,

Name VARCHAR2(1000),

**Type** VARCHAR2(60),

**PRIMARY** **KEY**(ID\_Area)

);

**CREATE** **TABLE** Genre

(ID\_Genre INT,

Name VARCHAR2(1000),

**Count** INT,

**PRIMARY** **KEY**(ID\_Genre));

**CREATE** **TABLE** Recording

(ID\_Recording INT,

Name VARCHAR2(2000),

**Length** INT,

**PRIMARY** **KEY**(ID\_Recording));

**CREATE** **TABLE** Release

(ID\_Release INT,

Name VARCHAR2(1000),

**PRIMARY** **KEY**(ID\_Release));

**CREATE** **TABLE** Artist

(ID\_Artist INT,

Name VARCHAR2(1000),

**Type** VARCHAR2(60),

Gender VARCHAR2(20),

ID\_Area INT,

**PRIMARY** **KEY**(ID\_Artist)

);

**CREATE** **TABLE** Medium

(ID\_Medium INT,

Format VARCHAR2(60),

ID\_Release INT,

**PRIMARY** **KEY**(ID\_Medium)

);

**CREATE** **TABLE** Track

(ID\_Track INT,

**Position** INT,

ID\_Medium INT,

ID\_Recording INT,

**PRIMARY** **KEY**(ID\_Track),

**FOREIGN** **KEY**(ID\_Recording) **REFERENCES** Recording

);

**CREATE** **TABLE** Artist\_Genre

(ID\_Artist INT,

ID\_Genre INT,

**PRIMARY** **KEY**(ID\_Artist, ID\_Genre),

**FOREIGN** **KEY**(ID\_Artist) **REFERENCES** Artist,

**FOREIGN** **KEY**(ID\_Genre) **REFERENCES** Genre);

**CREATE** **TABLE** Artist\_Track

(ID\_Artist INT,

ID\_Track INT,

**PRIMARY** **KEY**(ID\_Artist, ID\_Track),

**FOREIGN** **KEY**(ID\_Artist) **REFERENCES** Artist);

## Search Functionality

We decided to implement the search functionality as follow:

1. We chose not to allow the user to search on any column of a table. The reason is we thought all the columns were not really interesting to run queries on. For example in the GENRE table there is a count column, but we believe this is not a useful query to be run for a lambda user. That’s why the only search that can be done on our UI is about the NAME column of the table.
2. We chose to allow the user only to search in table where there is a NAME column. The reason follow from the discussion in the first point previously mentioned. Thus the available tables for search are: AREA, ARTIST, GENRE, RECORDING and RELEASE. In the UI, this table selection can be done through a drop-down list (in order to not let the user write junk as a table name).
3. We construct our queries by using the LIKE operator of SQL using the keyword the user inputted. More formally the query was designed as follow:

"SELECT \* FROM "**+**table**+**" WHERE LOWER(name) LIKE lower('%" **+** keyword **+** "%')"

1. The resulting rows of the simple search query are displayed in a HTML table. The HTML columns are the SQL columns of the table that was searched without id related columns. For example if the search was made on the ARTIST table, the HTML columns are NAME, TYPE and GENDER.
2. For the follow up queries we decided that the user can have more infos about a row by clicking on it. We decided to do so since the resulting queries are made upon the id of the row and not about a specific column.
3. The follow up queries are made using the id of the clicked row on chosen tables that seemed logical to be linked to for us.

More precisely the tables are linked like that:

|  |  |
| --- | --- |
| **ID from table** | **Linked with table(s)** |
| AREA | ARTIST |
| ARTIST | AREA, GENRE, RELEASE |
| GENRE | ARTIST |
| RECORDING | ARTIST, RELEASE |
| RELEASE | ARTIST, RECORDING |

These choices were obvious for us since we thought of it as if we were a user of the UI and we talk between us about “What more infos would want the user x if he choose to search a keyword y?”

1. Concerning the displaying we chose to use tabs. We display one of the linked tables and then the user can click on the other tabs to go to the other linked tables. When the user click on one row of one of the linked table the same behaviour occurs as if these row resulted from a simple search query.

## SQL Queries

We have still some problems for the SQL queries. We test all the following queries with a dummy database that is smaller than the music database and the queries give us the expected results. But with the real database some of the queries take a lot of time to be executed (we kill the process after 5 minutes and when we let the execution run the following error can appear…). Those queries are the query B, C, D, F and G. Thus we think that this problem comes from a lack of optimisation of those queries.



Figure 1: error when the query E is run during a long time

### Queries A to G

--A print the name of artist from switzerland

SELECT A.name

FROM Artist A, Area B

WHERE A.ID\_AREA = B.ID\_AREA AND B.name= 'Switzerland'

--B print the name and the number of female, male and group of the

--area that have the most female, male or group artists.

SELECT \*

FROM

(SELECT B.name AS NAME,

(SELECT COUNT(DISTINCT A.ID\_ARTIST)

        FROM ARTIST A

        WHERE A.GENDER='Female' AND B.ID\_AREA=A.ID\_AREA)

        AS COUNTF,

(SELECT COUNT(DISTINCT A.ID\_ARTIST)

        FROM ARTIST A

        WHERE A.GENDER='Male' AND B.ID\_AREA=A.ID\_AREA)

        AS COUNTM,

(SELECT COUNT(DISTINCT A.ID\_ARTIST)

        FROM ARTIST A

        WHERE A.TYPE='Group' AND B.ID\_AREA=A.ID\_AREA)

        AS COUNTG

FROM AREA B

ORDER BY COUNTF DESC)

WHERE ROWNUM=1

UNION

SELECT \*

FROM

(SELECT B.name AS NAME,

(SELECT COUNT(DISTINCT A.ID\_ARTIST)

        FROM ARTIST A

        WHERE A.GENDER='Female' AND B.ID\_AREA=A.ID\_AREA)

        AS COUNTF,

(SELECT COUNT(DISTINCT A.ID\_ARTIST)

        FROM ARTIST A

        WHERE A.GENDER='Male' AND B.ID\_AREA=A.ID\_AREA)

        AS COUNTM,

(SELECT COUNT(DISTINCT A.ID\_ARTIST)

        FROM ARTIST A

        WHERE A.TYPE='Group' AND B.ID\_AREA=A.ID\_AREA)

        AS COUNTG

FROM AREA B

ORDER BY COUNTM DESC)

WHERE ROWNUM=1

UNION

SELECT \*

FROM

(SELECT B.name AS NAME,

(SELECT COUNT(DISTINCT A.ID\_ARTIST)

        FROM ARTIST A

        WHERE A.GENDER='Female' AND B.ID\_AREA=A.ID\_AREA)

        AS COUNTF,

(SELECT COUNT(DISTINCT A.ID\_ARTIST)

        FROM ARTIST A

        WHERE A.GENDER='Male' AND B.ID\_AREA=A.ID\_AREA)

        AS COUNTM,

(SELECT COUNT(DISTINCT A.ID\_ARTIST)

        FROM ARTIST A

        WHERE A.TYPE='Group' AND B.ID\_AREA=A.ID\_AREA)

        AS COUNTG

FROM AREA B

ORDER BY COUNTG DESC)

WHERE ROWNUM=1;

--C List the name of 10 groups with the most recorded track

/\*order by the number of tracks and select the top 10 for the group\*/

SELECT \*

FROM(SELECT A.NAME

FROM  Artist A, Artist\_Track S

WHERE A.ID\_ARTIST=S.ID\_ARTIST AND A.TYPE='Group'

GROUP BY A.ID\_ARTIST, A.NAME

ORDER BY count(S.ID\_TRACK) DESC)

WHERE ROWNUM <=10

--D List the name of 10 groups with the most release

/\*Order by the number of releases and select the top 10 for the group\*/

SELECT \*

FROM(SELECT **A.NAME**

FROM Artist A, Track T, Artist\_Track S, Medium M, Release R

WHERE A.ID\_ARTIST =S.ID\_ARTIST AND T.ID\_TRACK=S.ID\_TRACK

AND T.ID\_MEDIUM=M.ID\_MEDIUM AND R.ID\_RELEASE = M.ID\_RELEASE

AND A.TYPE='Group'

GROUP BY A.ID\_ARTIST, **A.NAME**

ORDER BY count(DISTINCT **R.ID\_RELEASE**) DESC)

WHERE ROWNUM <=10

--E Print the name of female artist that have the most genres

/\*Order by the number of genre of female artist and select the top 1 artist\*/

SELECT NAME

FROM (SELECT A.NAME AS NAME,

COUNT(DISTINCT G.ID\_GENRE) AS COUNT\_GENRE

FROM ARTIST\_GENRE G, ARTIST A

WHERE A.ID\_ARTIST= G.ID\_ARTIST AND A.GENDER='Female'

GROUP BY A.ID\_ARTIST, A.NAME

ORDER BY COUNT\_GENRE DESC)

WHERE ROWNUM=1

--F Print the name of the cities that have more female artist than

--male artist

/\*Compare the count of the female artist and the count of male artist for the city. And select the cites that have more female artist than male artist\*/

SELECT B.name

FROM Area B

WHERE B.type='City'

AND (SELECT Count(\*)

FROM Artist A

WHERE A.gender='Female' AND A.ID\_AREA= B.ID\_AREA)

     > (SELECT Count(\*)

FROM Artist A1

WHERE A1.gender='Male' AND A1.ID\_AREA=B.ID\_AREA)

--G List the mediums with the most number of tracks

/\*Order the mediums by their track count and select the medium that have their track count egal to the max track count.\*/

SELECT ID, FORMAT, COUNT\_TRACK

FROM

(SELECT M.ID\_MEDIUM AS ID, M.FORMAT AS FORMAT,

COUNT(DISTINCT T.ID\_TRACK) AS COUNT\_TRACK

FROM TRACK T, MEDIUM M

WHERE M.ID\_MEDIUM = T.ID\_MEDIUM

GROUP BY M.ID\_MEDIUM, M.FORMAT)

WHERE COUNT\_TRACK = (SELECT MAX(COUNT\_TRACK)

FROM

(SELECT M.ID\_MEDIUM AS ID, M.FORMAT AS FORMAT,

COUNT(DISTINCT T.ID\_TRACK) AS COUNT\_TRACK

FROM TRACK T, MEDIUM M

WHERE M.ID\_MEDIUM=T.ID\_MEDIUM

GROUP BY M.ID\_MEDIUM, M.FORMAT)

);

## Interface

Figure 2: simple search: the user can choose in which table he wants to do the search

Figure 3 SQL queries: the left table allows the user to select a query and displays the result

****

Figure 4: the user clicked on the row of an artist and obtained different complementary information. Different information are available using the tabs

Figure 5: Mode full screen of the results

Figure 6: the user executed a search on an artist

# Deliverable 1

## ER-Model Choices

Every entity contains only its own attributes (*ID*, *name, etc.* but not the *ID* of other entities). We chose to create relations instead of giving the *ID* of an other entities as an attribute of the entity.

This relation design is especially suitable in this case because if several entities share their *ID* this means that a relation exists between them.

The *Song* relation is a ternary relation that represents a song on a support. It relates a support (*Medium*) with the position on this medium (*Track*) and the recording of the song (*Recording*).

*Physical\_Song* contains all the pairs of *Release* (album, single…) and on which support it is displayed.

Those choices are made in consideration of the given data. For example, for *Physical\_Song* relationship there is an existing relationship between Medium and Release. Medium should contain the ID of Release, so we made a relation of it.

## Constraints Explanation

There is some complex constraints between entities because the musical universe is also complex: collaborations (multiple artists per track), compilations (multiple artists per album), etc.

It is difficult to define more constraints because the available data are often incomplete.

### Areas

An artist can have at most one area. This constraint is given by the available data. Indeed, instinctively we may think that one artist have to come from one or more than one area (for example: New-York (*City*) and America (*Continent*)). But in the given data, the situation is different: an artist can have only one area or no area if the data is missing.

### Tracks

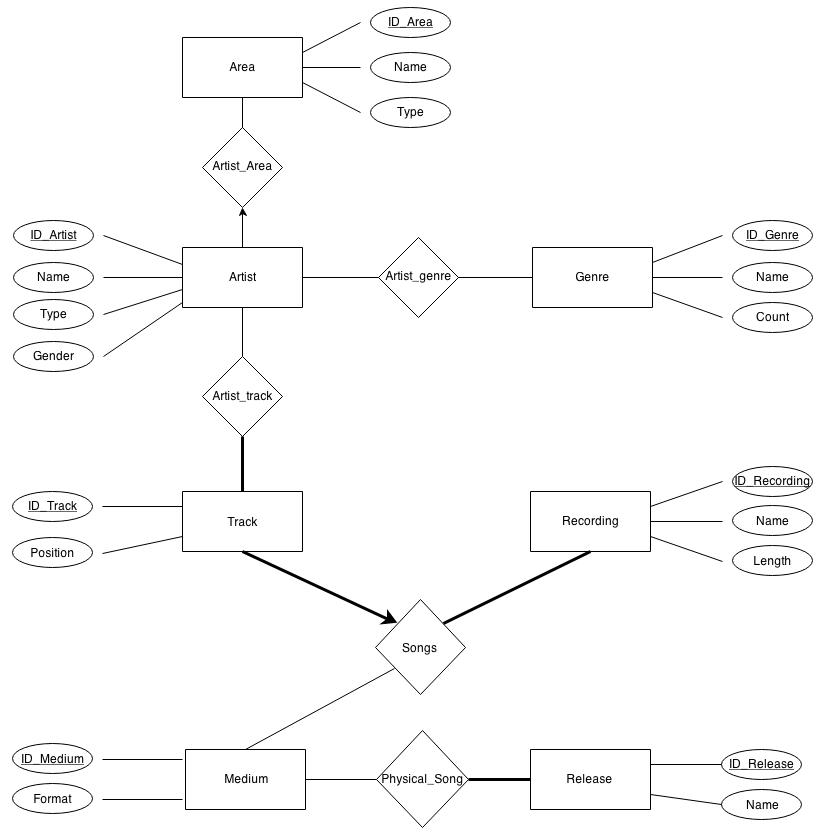
All tracks have to be represented in the relation *Artist\_Track* at least once. It means that all tracks were made by at least one artist.

All tracks have to be represented exactly once in the relation *Song*. It means that all tracks (all positions in a CD for example) are related to exactly one music (i.e there is no position without a music) and one support.

### Recordings

All recordings have to be at least once represented in the *Song* relation. Indeed, recording is the logical song. It does not make sense if it does not stand on any support. But the same recording may be on different supports, and be referred by several tracks etc.

### Releases

All releases must be at least once represented in the *Physical\_Song* relation. For the same reason as for recording. It does not make sense that an album, single … does not stand on any support.

### Schemas

**Area** : ID\_Area, Name, Type

**Artist** : ID\_Artist, Name, Type, Gender

**Genre** : ID\_Genre, Name, Count

**Medium** : ID\_Medium, Format

**Recording** : ID\_Recording, Name, Length

**Release** : ID\_Release, Name

**Track** : ID\_Track, Position

**Artist\_Area** : ID\_Artist, ID\_Area

**Artist\_Genre** : ID\_Artist, ID\_Genre

**Artist\_Track** : ID\_Artist, ID\_Track

**Physical\_Song** : ID\_Medium, ID\_Recording

**Song** : ID\_Track, ID\_Medium, ID\_Recording

## Tables Creation

-- Area :

CREATE TABLE Area

(ID\_Area CHAR(60),

Name CHAR(60),

Type CHAR(60),

PRIMARY KEY(ID\_Area))

-- Artist :

CREATE TABLE Artist

(ID\_Artist CHAR(60),

Name CHAR(60),

Type CHAR(60),

Gender CHAR(60),

PRIMARY KEY(ID\_Artist))

-- Genre :

CREATE TABLE Genre

(ID\_Genre CHAR(60),

Name CHAR(60),

Count CHAR(60),

PRIMARY KEY(ID\_Genre))

-- Medium :

CREATE TABLE Medium

(ID\_Medium CHAR(60),

Format CHAR(60),

PRIMARY KEY(ID\_Medium))

-- Recording :

CREATE TABLE Recording

(ID\_Recording CHAR(60),

Name CHAR(60),

Length CHAR(60),

PRIMARY KEY(ID\_Recording))

-- Release :

CREATE TABLE Release

(ID\_Release CHAR(60),

Name CHAR(60),

PRIMARY KEY(ID\_Release))

-- Track :

CREATE TABLE Track

(ID\_Track CHAR(60),

Position CHAR(60),

PRIMARY KEY(ID\_Track))

-- Artist\_Area :

CREATE TABLE Artist\_Area

(ID\_Artist CHAR(60),

ID\_Area CHAR(60),

PRIMARY KEY(ID\_Artist),

FOREIGN KEY(ID\_Artist) REFERENCES Artist,

FOREIGN KEY(ID\_Area) REFERENCES Area)

-- Artist\_Genre :

CREATE TABLE Artist\_Genre

(ID\_Artist CHAR(60),

ID\_Genre CHAR(60),

PRIMARY KEY(ID\_Artist, ID\_Genre),

FOREIGN KEY(ID\_Artist) REFERENCES Artist,

FOREIGN KEY(ID\_Genre) REFERENCES Genre)

-- Artist\_Track :

CREATE TABLE Artist\_Track

(ID\_Artist CHAR(60),

ID\_Track CHAR(60),

PRIMARY KEY(ID\_Artist, ID\_Track),

FOREIGN KEY(ID\_Artist) REFERENCES Artist,

FOREIGN KEY(ID\_Track) REFERENCES Track)

-- Physical\_Song :

CREATE TABLE Physical\_Song

(ID\_Medium CHAR(60),

ID\_Release CHAR(60),

PRIMARY KEY(ID\_Medium, ID\_Release),

FOREIGN KEY(ID\_Medium) REFERENCES Medium,

FOREIGN KEY(ID\_Release) REFERENCES Release)

-- Song :

CREATE TABLE Song

(ID\_Track CHAR(60),

ID\_Recording CHAR(60),

ID\_Medium CHAR(60),

PRIMARY KEY(ID\_Track),

FOREIGN KEY(ID\_Medium) REFERENCES Medium,

FOREIGN KEY(ID\_Track) REFERENCES Track,

FOREIGN KEY(ID\_Recording) REFERENCES Recording)