Databases Project – Spring 2017

Prof. Anastasia Ailamaki

Team No: **14**

Names: Perrotta Lucie, Phan Hoang Kim Lan, Nguyen Tim

# Deliverable 2

## Assumptions

Parsing leads to lots of assumptions. Here the ones we had to assume :

* Names are full of information between parenthesis or brackets (such as (signed), (translator)…). We delete them to be able to rely them (instead of having twice the same author).
* Dates are difficult to retrieve from given csv file, because of its non-uniform format. We assume:
  + “1870’s” become “1870”
  + “July 10 1870” become “1870”
  + “1870-07-10” become “1870”

## Parsing

We choose to use a local database, using **wamp server** (windows local server) and **phpmyadmin**.

We used the following php commands to do the parsing. This is the “*functions.php*” file, used in every php parsing file.

<?php

function **parseNullValue**(*$s*) {

**if**(**empty**(*$s*) **&&** *$s***!=**'0') **return** *true*;

**if**(*$s***==***0*) **return** *false*; // handle 0 values like id

*$nullValues* **=** ['/NULL/', '/\[nn\]/', '/\[none\]/','/\?(?![a-z])/', '/none/'];

**foreach**(*$nullValues* **as** *$n*) {

*$matches*;

**preg\_match** (*$n*, *$s*, *$matches*, *PREG\_OFFSET\_CAPTURE*);

**if** (sizeof(*$matches*) **!=** *0*) {

**return** *true*;

}

}

**return** *false*;

}

function **parseNullValueWebsite**(*$s*) {

**if**(**empty**(*$s*)) **return** *true*;

*$nullValues* **=** ['NULL', '[nn]', '[none]','?', 'none', 'url']; // url comes from first line of csv

**foreach**(*$nullValues* **as** *$n*) {

*$res* **=** **strpos**(*$s*, *$n*);

**if** (*$res* **!==** *false*) {

**return** *true*;

}

}

**return** *false*;

}

function **parseDoubleQuote**(*$s*) {

**if**(parseNullValue(*$s*)) **return** "NULL";

*$res* **=** **str\_replace**('"', '\"', *$s*);

**return** '"'**.***$res***.**'"';

}

function **getDateFromYear**(*$year*) {

**if**(parseNullValue(*$year*)) **return** "NULL";

// remove any char, ], [, ', ", whitespace

// should get 4 digits group

*$res* **=** **preg\_replace**("/[^\d\s-]/" , "", *$year*);

*$res1* **=** **preg\_split**("/[\s-]/" , *$res*);

**for** (*$i* **=** *0*; *$i* **<** sizeof(*$res1*); *$i***++**) {

**if**(**strlen**(*$res1*[*$i*])**==***4*) {

*$res* **=** *$res1*[*$i*];

}

}

*$month***=***1*; *$day***=***1*;

*$hour***=***0*; *$minute***=***0*; *$second***=***0*;

**return** '"'**.***$res***.**'-'**.***$month***.**'-'**.***$day***.**'"';

}

function **getInt**(*$i*) {

**if**(parseNullValue(*$i*)) **return** "NULL";

// suppress [ ] char

*$i* **=** **preg\_replace**("/\[**\***\]**\***/" ,"", *$i*);

**return** *$i*;

}

// return last index used in csv - useful for assigning id

function **getLastIndex**(*$file*) {

*$index*;

**while**((**!feof**(*$file*)) **&&** (*$val* **=** **fgetcsv**(*$file*))){

*$index* **=** *$val*[*0*];

}

**return** (**empty**(*$index*)) ? *0* : *$index***+***1*;

}

// return words separated by delimiter

function **parseNames**(*$s*, *$delimiter***=**";"){

// get rid of first and last double quotes

*$string* **=** **substr**(*$s*, *1*, **-***1*);

*$array* **=** **explode**(*$delimiter*, *$string*);

**for**(*$i* **=** *0*; *$i***<** sizeof(*$array*); *$i***++**){

*$array*[*$i*] **=** **ltrim**(*$array*[*$i*]);

}

**return** *$array*;

}

// return true if $s in contained in $csv file (opened) (at position $pos), false otherwise

function **isInCsv**(*$file*, *$s*, *$pos*){

**rewind**(*$file*);

**if**(**empty**(*$s*)) {

**return** *false*;

}

**while**(**!** **feof**(*$file*)){

*$val* **=** **fgetcsv**(*$file*);

**if**(*$val*[*$pos*]**==***$s*) {

**return** *$val*[*0*];

}

}

**return** *false*;

}

// return true if $s in contained in $csv file (opened) (at position $pos), false otherwise

// @TODO parse "?"

function **isInCsvName**(*$file*, *$s*, *$pos*){

**rewind**(*$file*);

**if**(**empty**(*$s*)) {

**return** *false*;

}

*$string* **=** parseToCompare(*$s*);

**while**(**!** **feof**(*$file*)){

*$val* **=** parseToCompare(**fgetcsv**(*$file*)[*$pos*]);

**if**(*$val***==***$string*) {

**return** *true*;

}

}

**return** *false*;

}

// modify string so that it can match even with the following differences : whitespaces, dot, dash

// in csv -> will keep first occurence

// @todo match entries when [as ...] is defined

function **parseToCompare**(*$s*){

*$res* **=** **preg\_replace**("/\s/", "", *$s*);

*$res* **=** **preg\_replace**("/\-/", "", *$res*);

*$res* **=** **preg\_replace**("/\./", "", *$res*);

*$res* **=** **strtolower**(*$res*);

**return** *$res*;

}

// delete from $s all content between () or []

function **parseComments**(*$s*) {

*$res* **=** **preg\_replace**("/\[.**\***\]/", "", *$s*);

*$res* **=** **preg\_replace**("/\(.**\***\)/", "", *$res*);

*$res* **=** **trim**(*$res*);

**return** *$res*;

}

We had to parse multiple times the csv’s given to us.

For the beginning, we created scripts which create sql commands from csv data. During the process, we get rid of null values such as “Null”, “none”, “[nn]”, “?”, etc. We also had to change some column types such as:

* isbn , rating, number of issue becoming varchar
* synospsis of story becoming text

**Website** : Websites come from publisher, indicia\_publisher and brand\_group csv file. We get the url values, parse null values, and add it in a new website table. We never add twice the same website. Once done, we process publisher, indicia\_publisher and brand\_group csv file and change their url values into website id (which become a foreign key referencing the website table).

<?php

**include**("db.php");

**include**("functions.php");

*$file* **=** **fopen**("comics/website.csv","r");

*$mysql* **=** **fopen**("website.sql", "w"); // write into this sql to import

*$min* **=** *0*;

*$max* **=** *100000*;

*$i* **=** *0*;

**while**(**!** **feof**(*$file*)){

*$i***++**;

*$val* **=** **fgetcsv**(*$file*);

**if**(*$i* **>** *$min*){

*$id* **=** getInt(*$val*[*0*]);

*$url* **=** parseDoubleQuote(*$val*[*1*]);

*$query* **=** '**INSERT INTO** website(id,url) **VALUES**(

'**.***$id***.**', '**.***$url***.**'

);';

**fwrite**(*$mysql*,*$query*);

*$s1* **=** *$con***->**query(*$query*);

**if**(*$i***==***$max*){

**break**;

}

}

}

**fclose**(*$file*);

?>

**Artist** : Artists come from story csv file. We get names, parse null values, and add it in a new artist table. To not add an artist twice, we construct a comparative string, which is the name of author without spaces, dot and hyphen. Unfortunately, getting rid of construction like [as name] give us information about full name or nickname of artists. These can permit us to avoid duplicata or obtain full names of artists (our actual version uniquely keep first found entry).

The idea is then to add id of artist in has\_ table (same idea as used for websites).

<?php

**include**("db.php");

**include**("functions.php");

*$file* **=** **fopen**("comics/story.csv","r");

*$csv* **=** **fopen**("comics/artist.csv", "a+");

**echo** "<h1>Artists</h1>";

*$index* **=** getLastIndex(*$csv*);

*$min* **=** *0*;

*$max* **=** *200*;

*$i* **=** *0*;

**var\_dump**(**fgetcsv**(*$file*));

**while**(**!** **feof**(*$file*)){

*$i***++**;

*$val* **=** **fgetcsv**(*$file*);

**if**(*$i* **>** *$min*){

*$script* **=** parseDoubleQuote(*$val*[*4*]);

*$pencils* **=** parseDoubleQuote(*$val*[*5*]);

*$inks* **=** parseDoubleQuote(*$val*[*6*]);

*$colors* **=** parseDoubleQuote(*$val*[*7*]);

*$letters* **=** parseDoubleQuote(*$val*[*8*]);

// use lastp to speed up the query, since artists are close to each other in the csv file

**if**(*$script***!=**"NULL"){

*$script\_array* **=** parseNames(*$script*);

*$lastp* **=** "";

**foreach** (*$script\_array* **as** *$p*){

*$p* **=** parseComments(*$p*);

**if**(*$p***==***$lastp*) **continue**;

**if**(**!**isInCsvName(*$csv*, *$p*,*1*)) {

*$add* **=** *$index* **.** ","**.***$p***.**"\n";

**fwrite**(*$csv*, *$add*);

*$index***++**;

*$lastp* **=** *$p*;

}

}

}

**if**(*$pencils***!=**"NULL"){

*$p\_array* **=** parseNames(*$pencils*);

*$lastp* **=** "";

**foreach** (*$p\_array* **as** *$p*){

*$p* **=** parseComments(*$p*);

**if**(*$p***==***$lastp*) **continue**;

**if**(**!**isInCsvName(*$csv*, *$p*,*1*)) {

*$add* **=** *$index* **.** ","**.***$p***.**"\n";

**fwrite**(*$csv*, *$add*);

*$index***++**;

*$lastp* **=** *$p*;

}

}

}

**if**(*$inks***!=**"NULL"){

*$inks\_array* **=** parseNames(*$inks*);

*$lastp* **=** "";

**foreach** (*$inks\_array* **as** *$p*){

*$p* **=** parseComments(*$p*);

**if**(*$p***==***$lastp*) **continue**;

**if**(**!**isInCsvName(*$csv*, *$p*,*1*)) {

*$add* **=** *$index* **.** ","**.***$p***.**"\n";

**fwrite**(*$csv*, *$add*);

*$index***++**;

*$lastp* **=** *$p*;

}

}

}

**if**(*$colors***!=**"NULL"){

*$colors\_array* **=** parseNames(*$colors*);

*$lastp* **=** "";

**foreach** (*$colors\_array* **as** *$p*){

*$p* **=** parseComments(*$p*);

**if**(*$p***==***$lastp*) **continue**;

**if**(**!**isInCsvName(*$csv*, *$p*,*1*)) {

*$add* **=** *$index* **.** ","**.***$p***.**"\n";

**fwrite**(*$csv*, *$add*);

*$index***++**;

*$lastp* **=** *$p*;

}

}

}

**if**(*$letters***!=**"NULL"){

*$letters\_array* **=** parseNames(*$letters*);

*$lastp* **=** "";

**foreach** (*$letters\_array* **as** *$p*){

*$p* **=** parseComments(*$p*);

**if**(*$p***==***$lastp*) **continue**;

**if**(**!**isInCsvName(*$csv*, *$p*,*1*)) {

*$add* **=** *$index* **.** ","**.***$p***.**"\n";

**fwrite**(*$csv*, *$add*);

*$index***++**;

*$lastp* **=** *$p*;

}

}

}

**if**(*$i***==***$max*){

**break**;

}

}

}

**fclose**(*$file*);

**fclose**(*$csv*);

?>

## Query Implementation

Here are the 8 queries we were asked to implement in SQL.

For this first query we first select all the brand name possessing at least one indicia publisher from Belgium. We choose to get the name of the brand\_group and the number of Belgian indicia it possesses. Then we simply sort the resulting table by number of indicia and get the name only.

*-- a)*

**SELECT** **T**.**name**

**FROM** (

**SELECT** **B**.**name**,

**COUNT**(**\***) **AS** bid

**FROM** brand\_group B,

indicia\_publisher I,

publisher P,

country C

**WHERE** **C**.**name** **=** 'Belgium' **AND**

**C**.**id** **=** **I**.**country\_id** **AND**

**I**.**publisher\_id** **=** **P**.**id** **AND**

**P**.**id** **=** **B**.**publisher\_id**

**GROUP BY** **B**.**name**

) **AS** T

**ORDER BY** **T**.**bid**

For the query b), we simply use the chain AND rule to get all publishers from Denmark, in a straightforward fashion.

*-- b)*

**SELECT** **P**.**id**, **P**.**name**

**FROM** publisher P,

country C,

series S

**WHERE** **C**.**name** **=** 'Denmark' **AND**

**S**.**country\_id** **=** **C**.**id** **AND**

**S**.**publisher\_id** **=** **P**.**id**

For querry c), the fashion is similar to b), we simply apply the chain rule, to get series from Switzerland and published in a magazine.

*-- c)*

**SELECT** **S**.**name**

**FROM** series S,

country C,

series\_publication\_type T

**WHERE** **T**.**name** **=** 'magazine' **AND**

**T**.**id** **=** **S**.**publication\_type\_id** **AND**

**S**.**country\_id** **=** **C**.**id** **AND**

**C**.**name** **=** "Switzerland"

Here in d) we simply want to get all issues from 1990, sorted by year. Note that for simplicity purposes, all the dates have been converted to years stored as integers, since most of the dates simple consists as a year, often followed by text such as “circa”. It is hence easier to work with int(4) formatted years.

*-- d)*

**SELECT** **COUNT**(**\***)

**FROM** issue I,

**WHERE** **I**.**publication\_date** **>=** 1990

**GROUP BY** **I**.**publication\_date**

We still are not sure about the query e), our result returns the number of series that have been published by some publisher, instead of separating them by indicia\_publisher. The correct query should be quite similar but we had no time yet to figure out what’s wrong.

*-- e) almost OK (now returns all series by publisher, not by indicia)*

**SELECT** **T**.**name**, **COUNT**(**\***)

**FROM** (

**SELECT** **I**.**name**

**FROM** indicia\_publisher I,

publisher P,

series S

**WHERE** **I**.**publisher\_id** **=** **P**.**id** **AND**

**S**.**publisher\_id** **=** **P**.**id** **AND**

**I**.**name** **LIKE** '%DC\_\_omics%'

) **AS** T

**GROUP BY** **T**.**name**

For query f), we first select all stories that have been reprinted at least once, and then regroup them by original story. Finally, we count how many times each original story has been reprinted and sort them according to that. We only print the names of the stories for aesthetical purposes.

*-- f)*

**SELECT** **S**.**title**

**FROM** story S,

story\_reprint R

**WHERE** **S**.**id** **=** **R**.**origin\_id**

**GROUP BY** **R**.**origin\_id**

**ORDER BY** **COUNT**(**R**.**origin\_id**)

This query (g) was interesting since it uses the chain rule in a particular fashion. What we are seeking are artists who contributed to every part of the making of some story. That is, we want all artist who did color, write, draw and ink a story. We simply want an artist who did all 4 on a same story and a story who had all 4 done by a single artist.

*-- g)*

**SELECT distinct** **A**.**name**

**FROM** artist A,

has\_script SC,

has\_pencils P,

has\_color C,

has\_inked I,

story S

**WHERE** **A**.**id** **=** **SC**.**artist\_id** **AND**

**A**.**id** **=** **P**.**artist\_id** **AND**

**A**.**id** **=** **C**.**artist\_id** **AND**

**A**.**id** **=** **I**.**artist\_id** **AND**

**S**.**id** **=** **SC**.**story\_id** **AND**

**S**.**id** **=** **P**.**story\_id** **AND**

**S**.**id** **=** **C**.**story\_id** **AND**

**S**.**id** **=** **I**.**story\_id**

For the last query h), we wanted Batman to be a non-featured character of a non-reprinted story. For that purpose, we seek all stories which were not reprinted, that is, which are not featured in the reprint table. We hence look for the story in the reprint table and expect the number of its occurrences to be 0.

Then, we simply say we want Batman to be in the non-featured characters. Note the utilization of the command LIKE in order to seek for all strings containing “Batman”.

*-- h)*

**SELECT** **S**.**title**

**FROM** story S,

character C,

has\_characters HS

**WHERE** 0 **=** ( **SELECT** **COUNT**(distinct **R**.**origin\_id**)

**FROM** story\_reprint R

**WHERE** **S**.**id** **=** **R**.**origin\_id**

) **AND**

**HS**.**characters\_id** **=** **C**.**id** **AND**

**HS**.**story\_id** **=** **S**.**id** **AND**

**C**.**name** **LIKE** '%Batman%'

## General Comments