The Use of Data Cleansing in Mobile Devices

María del Pilar Angeles, Francisco García-Ugalde
Facultad de Ingeniería
Universidad Nacional Autónoma de México
México, D.F.
pilarang@unam.mx, fgarciau@servidor.unam.mx

David Alcudia-Aguilera
Facultad de Ciencias
Universidad Juárez Autónoma de Tabasco
Tabasco, México
072H3024@alumno.ujat.mx

***Abstract*—People receive information on PDA, mobile phone or mp3 player. If such information was correct, current, useful and usable, users would be able to use it immediately from any mobile device with no requirement of a personal computer to assess, clean and integrate data. The present research proposes the utilization of a data cleansing framework for the improvement of data quality in mobile devices.**

Keywords-data quality;mobile devices; data cleansing.

I. INTRODUCTION

Nowadays, people utilize huge amount of data coming from a range of mobile devices under different data storage and data formats in order to make business. It is very well known that companies having useful information have better possibilities to exploited it, and make better informed decisions. Companies establishing better strategies are able to become leaders and business competitive. For instance, employees are more productive by keeping in touch through text messages to customers or colleagues while they are attending a meeting out of the office.

However, since information come from a number of data sources, such data sources are structured, unstructured or semi structured and the process of information integration is not a trivial task, due to semantic and syntactic heterogeneities degrading data quality as a consequence.

Typical causes of poor data quality are data entry errors, wrong or unspecific metadata, lack of enforcement or not defined appropriately integrity constraints [1], [2].

People are used to edit information from the songs they legally downloaded from the internet, or type personal information as business contacts within mobile phones o personal digital assistants (PDA) in order to make phone calls, send text messages or emails to arrange a meeting for business. For instance, in the case of an mp3 player, hundreds or thousands of songs can be stored (on IDv3 tags) under different genres, albums, singers, etc. The management of information becomes chaotic, tiring and annoying. Unfortunately, having duplicated or obsolete personal details stored in an agenda is not unusual and could be the cause of losing business opportunities. The problem increases after a software upgrade or data migration to other mobile phone**.**

There are a number of software tools for the edition of audio files tags, or for the management of contacts in mobile phones or electronic organizers namely Personal Information Management (PIM). However, such tools are not very practical when there are hundreds of rows to manage. Users have to select by hand all the records they want to update.

Furthermore, if there are 200 songs of the same genre, but this field has been captured in 20 different forms, the genres are semantically equal but syntactically different, for instance, ROCK, rock, RoCK, etc. The readability and usability of the mp3 player is affected.

Data quality patterns and data matching have been developed recently for the detection and correction of data errors within the process of data cleansing [6], [7]. Data cleansing has being widely used on data warehousing [3], [4], [5], but not on data stored in mobile devices such as mobile phones, electronic organizers o audio players.

The following section is aimed to explain the problem of poor data quality during integration process within mobile devices. The third section provides the bases of context of Data Quality, data cleansing and data quality patterns. The fourth section presents a Data Cleansing Framework for mobile devices. The fifth section concludes with the main findings and implications for future research.

II. POOR DATA QUALITY WITHIN MOBILE DEVICES

A. Poor data quality on mp3 files

In this section, we analyze the problem of data cleansing in mp3 files. It is common to have a large list of mp3 files in a directory with several files wrong documented, repeated or incomplete.

- Problem description

"I upgraded my mp3 player to the current version, and I am enjoying the new features. However, my album folders are not sorted properly (even though they appear as intended in the mp3 organizer on my PC).

Some songs are not with the album they are part of. Other albums are not where they should be chronologically, numerically, or alphabetically."

An ID3 tag is a data container within an MP3 audio file stored in a prescribed format. This data commonly contains the Artist name, Song title, Year and Genre of the current audio file. However, even if the music titles were rightly spelled and correct, if they were longer than 17 characters, users are unable to identify which track will be played because display restrictions of some mp3 players that only 16 characters would be displayed.

Large repositories of information are frequently incomplete, redundant or corrupted. The same problem is arising within small data sets. For instance, is common to have a music folder in our computer containing same song several times. In addition, is usual to have non descriptive or invalid filenames.

Trying to resolve this problem "by hand" is time consuming as presented in the following section.

 ISBN:978-1-61208-115-1

- Typical solution

MP3 organizers allow users a number of features such as localize songs performed by a main artist under one name, preferably with the correct spelling within one folder. Make sure album titles are the same and correctly-spelled throughout the album. Add the album artwork to the album if you wish. Find out the original albums of songs in "Greatest Hits" to avoid having two pointless albums. Put any soundtracks in the genre "Soundtrack" for easy access. Edit and check spelling other genre classifications (Rock, Pop, Indie). Organize music by time. Find duplicate songs. Select a song and get info. Type in the correct artist name and the already mentioned info, Edit multiple songs by the same artist, by clicking the first song, then hold down CTRL and click the others by that artist, right-click one of them and select get info. The mp3 organizer program will ask for confirmation to edit multiple songs information. The music organizer programs might offer a full range of features. However, if they are all manual, they are not practical approaches in the case of hundreds or thousands of contacts or music files.

B. Poor data quality within Smart phones or Personal Digital Assistants

- Problem Description

In the late 90`s the PDA offered users the possibility to become mobile professionals enable to better manage information through a personal organizer on the go. Years later smart phones appear and the contacts information were migrated from one platform to other, as the software were not compatible, one contact with three phone numbers (home, job, mobile) on the PDA became three entries under the same contact name with one phone number each. Some entries were not identified containing the phone number only. Other problem was regarding ciphering, due to incompatibility between Pocket PC o Smartphone and mobile phones because management software does not support ciphering, making data corrupted or even lost.

- Typical Solution

Users spent hours trying to find which phone number belong to which person and if it was office number or mobile phone number. Therefore users definitely delete extra entries. Most people just delete the entries and keep only the most known contact numbers missing business opportunities. The migration process requires software installed on the personal computer. If the contact information was barely captured, the migration would perform with no prove. However, if the contact lists contains for instance '+' for international phone calls, such character would not be reflected in the new contact list and changes have to be performed by hand again.

There are open software that provides mobile synchronization and push email solutions to mobile phones. It includes a mobile server that pushes email to mobile phones from a number of mail servers. It enables users to synchronize contacts, calendar, tasks and notes. Offering a portal that lets anyone get email on their mobile phone and that syncs PIM data to their devices. Users are required to provide personal details such as email user and password and their phone number. The main issue here is data security and user information confidentiality.

III. THE CONTEXT OF DATA QUALITY

The subjective nature of the term Data Quality (DQ) has allowed the existence of general definitions such as "fitness for use" in [18], which implies that quality depends on customer requirements.

The definition established by Redman [11], suggests that data quality can be obtained by comparing two data sources. "A datum or collection of data X is of higher or (better) quality than a datum or collection of data Y if X meets customer needs better than Y".

Recently, data quality has been defined as "the capability of data to be used effectively economically and rapidly to inform and evaluate decisions" [15]. Such definition considers data quality not as the end but the means for making informed decisions.

Data quality is characterized by quality criteria or dimensions such as accuracy, completeness, consistency, and timeliness in several approaches such as [1], [5], [16], [17], [18] mainly because classification facilitates the characterization and definition of an overall quality.

In the case of mobile applications, there is not massive information like in data warehouse environments but sufficiently relevant for making business. For instance, a mobile application developed to collect census data is detailed in [9].

The quality properties considered in this research are accuracy, amount of data, Format appropriateness, format precision, representation consistency, uniqueness, completeness, usability, usefulness, which will be explained in detail as follows:

Accuracy has been considered in [16] as the "measure of the degree of agreement between a data value and the source agreed to be correct".

The quality property amount of data refers to the extent to which the volume of data is appropriate for the task at hand.

Format appropriateness refers to the capability of a data format to be more appropriate than other because it is better suited to users' needs [11].

The Format precision refers to the set of symbolic representations are sufficiently precise to distinguish among elements in the domain that must be distinguished by the users, there are values correctly represented, values not represented (missing) and values that do not correspond with real world. The data cleansing process should support several character sets in order to be suitable for different languages.

Representation consistency refers to whether physical instances of data are in accord with their format; the constraints are posed in terms of membership in the set of a symbolic representation [11], or in terms of conformance to a format standard.

The Uniqueness property is the extent where an entity from the real world is represented once.

Usability is the extent to which data are used for the task at a hand with acceptable effort.

The Usefulness property refers to the degree where using data provides benefit on the performance on the job, in other words the extent to which the user believes data would be useful for the task at a hand. The data cleansing process shall be effective enough to produce clean data and useful.

In order to correct, standardize and consequently, to improve data quality, data cleansing has emerged to define and determine error types, search and identify error instances, and correct the errors.

A. Data cleansing

"Data cleansing is applied especially when several databases are merged. Records referring to the same entity are represented in different formats in different data sets or are represented erroneously. Thus, duplicated records will appear in the merged database. This problem is known as merge/purge problem." [3].

According to [14] the most common methods utilized for error detection are: statistical methods through standard deviation, quartile ranges, regression analysis, etc. [13], [12]; clustering is a data mining method to classify data in groups to identify discrepancies; pattern recognition based methods to identify records that do not fit into a certain specific pattern and association rules to find dependencies between values in a record [3].

Data cleansing is commonly performed in offline time, which is unacceptable for operational systems. Therefore, cleansing is often regarded as a pre-processing step for Knowledge Discovery in Databases and Data Mining systems during the Extraction Transformation and Load (ETL) process. However, it is still a very time consuming task, "The process of data cleansing is computationally expensive on very large data sets and thus it was almost impossible to do with old technology" [3]. The main objective within the present research is to use the data quality pattern recognition and record matching as useful data cleansing tools within mobile devices. The main advantage is the feasibility of using data cleansing in small quantity of data contained in mobile devices.

Previous works and approaches related to the subject of this paper try to resolve and propose methods for the issues of data cleansing, data quality, and resolution of data inconsistencies. Some of them are presented as follows:

Loshin [2], approached the problem of Data Quality. Their objectives were to develop an automated procedure to assess the quality of civil infrastructure monitoring data and to explore how effectively the data can be cleansed using the assessments results. Loshin proposed seven different cleansing techniques. (1) the do nothing technique leaves the data set as is, (2) the correction technique applies a constant additive or multiplicative factor to the data, (3) the replace technique replaces the original value with a new value; the new value might be an average over the attribute values or a value generated by a prediction algorithm, (4) the split/combine technique splits a single record into two records or combines two records into a single record, (5) the insertion technique duplicates data from one part of the data set and inserts the duplicate data into a part where data are missing, (6) the remove technique removes data from the data set, and (7) Meta-data techniques add new information to the data set, in the form of confidence intervals, weights, or flags.

Maletic and Marcus [3] confront the problem of data cleansing and automatically identifying potential errors in data sets. They presented three phases in the process of data cleansing: (1) define and determine error types, (2) search and identify error instances, and (3) correct the uncovered errors.

In addition, Hernandez and Stolfo [4] addressed the problem of merging multiple databases of information about common entities (the Merge/Purge Problem). They developed a system for accomplishing this data cleansing task.

The next section presents the use of pattern recognition to identify records that do not fit into certain specific problem and correct them by data standardization as part of the data cleansing process.

B. Detection of Data Quality Patterns

The detection of data quality patterns aims to the improvement of data quality by ensuring data consistency. Once the data pattern is detected, the metadata is updated with the right definition of data format. For instance, the integrity constraint is updated or defined according to the data pattern. Therefore, the integrity constraint is forced into the data and data become consistent.

Data quality issues are most sever when information is scattered across isolated and heterogeneous data stores. To turn information into insight and to leverage its significant value, data quality needs to be addressed by applying data cleansing consistently, using consistent cleansing rules throughout the enterprise, not only in the database layer but also in the application and process layers as mentioned by Sautter and Mathews [7].

The Pattern analysis is aimed to understand which formats are valid for a particular field by examining the distribution of character types found within the column. If we analyze the patterns over several thousand records we can start to uncover the data quality rules for this field as the patterns that occur with the most frequency are typically the valid rule and the ones which occur least frequently are often in error.

The main advantages of data cleansing patterns are the increment of data quality, the reusability of the rules in order to update data that are not under the specified pattern and therefore inaccurate, besides of lower costs of data maintenance [7].

C. Record Matching

Deterministic matching systems use a combination of algorithms and business rules to determine when two or more records match. Algorithms catch simple common errors such as typos, phonetic variations and transpositions. Either the records match the requirements of the business rule or they do not match. However, deterministic systems lack scalability. When the amount of records increases, deterministic matching requires expensive customization and

business rule revision, impacting performance and cost as a consequence.

IV. FRAMEWORK OF DATA CLEANSING WITHIN MOBILE DEVICES

In the case of mobile applications, the information most frequently used is concerned with personal information such as contact details, audio data, e-mails, notes, tasks to do, etc. There is not massive information like in data warehouse environments but sufficiently relevant for making business.

This section details a Framework proposed and implemented for Data Cleansing within mobile devices. The main objective of this Framework is to cover the main issues of data quality under a very pragmatic focus. Fig. 1 shows the most relevant steps.

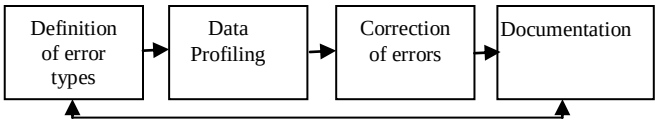


Figure 1. Data Cleansing Framework.

A. Define and determine error types

Regarding the quality properties within mobile devices, we present the most frequent problems of accuracy, completeness, usability, usefulness. The definition of such quality properties has been pointed in section III. Therefore, the corresponding error types are mentioned as follows:

a) Accuracy: Within the context of data stored in mobile devices, inaccurate data would be a misspelled contact name, which in fact may incur in a business failed, or a misspelled name song.

b) Amount of data: In the case of mobile devices, the possibility to detect and correct wrong data by hand becomes impractical at the range of hundreds.

c) Format appropriateness: In the case of IDv3 tag would be better to homogenize format by detecting and forcing pattern specified by users.

d) The Format precision there are values correctly represented, values not represented (missing) and values that do not correspond with real world. The data cleansing process should support several character sets in order to be suitable for different languages.

e) Representation consistency: In the case of contact details there must exist data integrity constraints that force the corresponding format or data pattern in order to avoid data errors. For instance, in the case of songs, the author name should be an aggregation of first name, last name, etc.

f) The Uniqueness property: Frequently after and upgrade or migration of contacts we find our data duplicated because of different formats between one mobile device and the new one. A similar situation is in the case of downloaded mp3 songs, because there are different versions of IDv3tags.

g) Usability: Nowadays is usable a software that allows to edit details of a business contact or the details of a mp3 song, but when the manipulation of data is in terms of hundreds, the software becomes unusable at all.

h) Usefulness: Wrong data would not be useful for the task at a hand.

B. Data profiling

This process required analysis and clustering phases for parsing and searching in order to identify data quality patterns and error instances in the source files and then isolation of these data elements in target files.

- Analysis

Given a data set of a number of mp3 files the prototype gets the ID3 tag from the end of each mp3 file and returns whether there was an ID3 tag on the file, throws unsupportedEncodingException if system cannot handle ASCII or throws IOException if an error occurs reading the file.

- Clustering

The prototype examines file by file and extracts all metadata of each file such as title, artist, year, album, comment and filename then save these values into a text file.

The process of extracting metadata of mp3 files in IDv3 tags format was achieved by adjusting the JID3 library proposed in [7]. During this phase the prototype groups the mp3 files by artist name in a number of .txt files and explores which are the fields that have null values to mark them as candidates for data consolidation.

- Data Quality Patterns

Regarding the analysis and detection of data patterns, there has been a considerable work for a number of data repositories such as Excel, Access, Oracle and SQL Server [8],[19] but not for IDv3 tags. Therefore, we have implemented a data quality pattern analyzer in order to obtain the greatest percentage of occurrences for each column of the mp3 files metadata.

The development of a data quality pattern was through the mapping of standard ASCII character set to letters that indicate their character format. For example, the lowercase letters have been mapped to ‘L’ meaning lowercase standard alphabetical character. Uppercase letters have been mapped to ’U’. Numbers were mapped to ‘N’ meaning number.

The simplest form for performing the mappings is to convert the ASCII number of a character to the corresponding mapping character as many of the characters are not visible on a standard editor.

The conversion function converts each character to the new mapping. After obtaining the format pattern for each item, the data quality pattern counts different patterns occurrences in order to identify which pattern is the most frequent utilized and define it as the pattern to be forced by the integrity constraint.

In the case of the IDv3 tags, some of the data quality patterns are shown in Table 1. For instance, the song title will be force to follow the U*WU*WU* pattern because the 45% of the song names are capitalized.

 ISBN:978-1-61208-115-1

The patterns for each field of the IDv3 tag are stored in a metadata file, in order to avoid the analysis of data patterns each time more mp3 files are loaded to the mobile device.

TABLE 1 DATA PATTERNS DETECTED FOR IDV3 TAG

Title_ Sample	Pattern	%
amor de voceador	L*WL*WL*	11
Amor de voceador	UL*WL*WL*	13
AMOR DE VOCEADOR	U*WU*WU*	45
Amor_de_Voceador	UL*SL*SUL*	9
AMOR_DE_VOCEADOR	U*SU*SU*	10
Amor-de-Voceador	UL*DL*DUL*	12

Once the patterns are established for each field of the tag the next step is the correction of errors by the data matching, data consolidation, and data standardization.

C. *Correction of errors*

The process of correction of data errors required sophisticated data algorithms and secondary data sources.

The prototype explores all metadata within each txt file and measures strings distance to detect redundancy and ensure consistency or to detect duplicated records and ensure uniqueness.

- Data Standardization

In order to ensure data consistency, the prototype applies conversion routines to transform data into its preferred (and consistent) format.

For instance, the audio file can be renamed accordingly to the user pattern detected. For instance, Artist – Title (comment), Title (comment), Artist – Title (comment), Title (comment), etc.

- Data Matching:

Data matching is the process of searching and matching records within and across the parsed, corrected and standardized data based on predefined business rules to eliminate duplications [10].

The present work applies the JaroWinklerDistance class [6] in order to identify duplicated songs when there are a relatively strong proximity between their corresponding song titles and album names.

For each artist text file a comparison of each record is done by measuring the distance between titles and albums of all songs if there is similitude of 80% between them or more, we assume that these records are the same entity of the real world.

An example of a simple comparison is shown in the following code:

```
for(int z=x+1;z<canciones.length;z++){
if(!canciones[x].getTitulo().equals("null")
&&!canciones[z].getTitulo().equals("null")){
if(jaroWinkler.proximity(canciones[x].getTitulo(),
canciones[z].getTitulo())>0.8)
```

- Data consolidation:

This is a process of consolidating or merging matched records into one representation. However, in order to achieve completeness and remove redundancy, a comparison between the matching records is carried out in order to preserve existing data values by replacing the corresponding null values for the same field. Finally, the mp3 files are renamed with the complete title of the song following the pattern detected from the user, and with complete and accurate information.

D. *Documentation*

The patterns identified from user regarding the IDv3tag information are documented and stored in order to keep standardization and to restrict and correct information modifying future data entry procedures.

E. *Modify data entry procedures to reduce future errors*

Every time a new song or contact is captured within the mobile device the patterns shall be maintain.

The documentation and the modification of data entry procedures are part of our work in progress.

V. TEST AND EXPERIMENTATION

Regarding the appropriateness of the proposed framework, we have identified representative scenarios according to the data quality properties assessed on specific data sources to test whether the data profiling and the correction of errors are within the expected ranges.

Table 2 shows some tests considering 10, 20, 45 and 77 mp3 files and the execution time the prototype takes according with the number of mp3 files cleaned.

The elapsed time is mainly taken during the process of reading the mp3 files and the process of writing the information in text files.

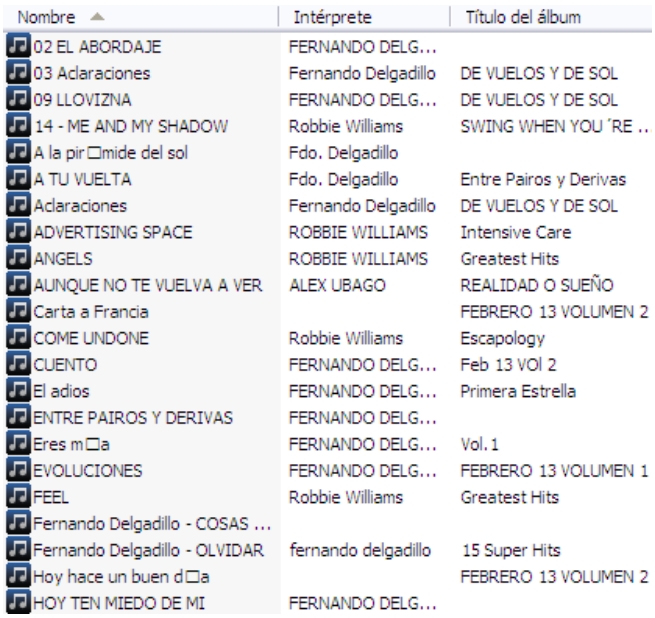
The record matching and cleaning are not taking a long time to process. However, more detailed and exhausted testing of the framework needs to be carried out.

TABLE 2 MP3 FILES AND EXECUTION TIME

Amount of files	ms
10	4984.0
20	13750.0
45	21188.0
77	42078.0

Fig. 2 shows an example of original mp3 files contained within a folder named “My Music”. There were a number of data quality errors: (a) duplicated songs (Aclaraciones), (b) different format representation for Artist name (FDO DELGADILLO, FERNADO DELGADILLO, Fernando Delgadillo, etc.), (c) inaccuracies for song titles (mía, día, pirámide), and (d) incompleteness derived from missing information in album name and artist name fields.

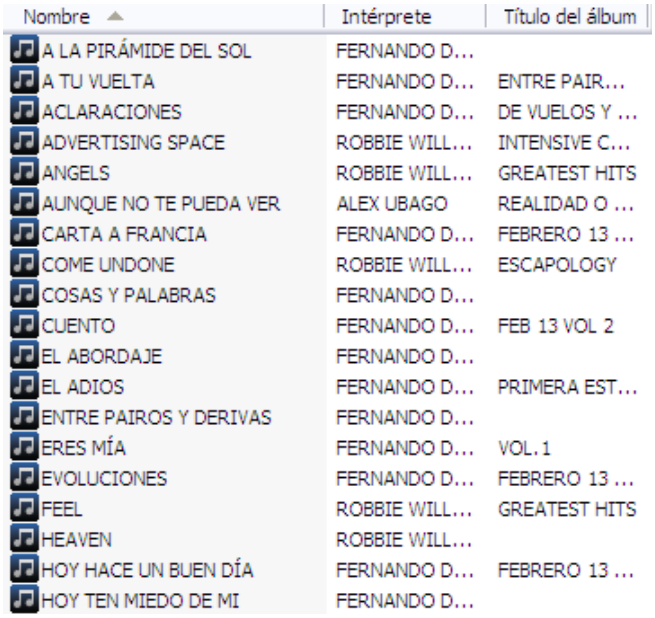
 ISBN:978-1-61208-115-1



Nombre	Intérprete	Título del álbum
02 EL ABORDAJE	FERNANDO DELG...	
03 Aclaraciones	Fernando Delgadillo	DE VUELOS Y DE SOL
09 LLOVIZNA	FERNANDO DELG...	DE VUELOS Y DE SOL
14 - ME AND MY SHADOW	Robbie Williams	SWING WHEN YOU ´RE ..
A la pir□mide del sol	Fdo. Delgadillo	
A TU VUELTA	Fdo. Delgadillo	Entre Pairos y Derivas
Aclaraciones	Fernando Delgadillo	DE VUELOS Y DE SOL
ADVERTISING SPACE	ROBBIE WILLIAMS	Intensive Care
ANGELS	ROBBIE WILLIAMS	Greatest Hits
AUNQUE NO TE VUELVA A VER	ALEX UBAGO	REALIDAD O SUEÑO
Carta a Francia		FEBRERO 13 VOLUMEN 2
COME UNDONE	Robbie Williams	Escapology
CUENTO	FERNANDO DELG...	Feb 13 VOl 2
El adios	FERNANDO DELG...	Primera Estrella
ENTRE PAIROS Y DERIVAS	FERNANDO DELG...	
Eres m□a	FERNANDO DELG...	Vol. 1
EVOLUCIONES	FERNANDO DELG...	FEBRERO 13 VOLUMEN 1
FEEL	Robbie Williams	Greatest Hits
Fernando Delgadillo - COSAS ...		
Fernando Delgadillo - OLVIDAR	fernando delgadillo	15 Super Hits
Hoy hace un buen d□a		FEBRERO 13 VOLUMEN 2
HOY TEN MIEDO DE MI	FERNANDO DELG...	

Figure 2. Example of original mp3 data.

Fig. 3 shows the mp3 data after the data cleansing process. The patterns detected for the song name, artist and title of album have been forced to represent data consistently. Duplicated mp3 files have been merged for completeness and uniqueness.



Nombre	Intérprete	Título del álbum
A LA PIRÁMIDE DEL SOL	FERNANDO D...	
A TU VUELTA	FERNANDO D...	ENTRE PAIR...
ACLARACIONES	FERNANDO D...	DE VUELOS Y ...
ADVERTISING SPACE	ROBBIE WILL...	INTENSIVE C...
ANGELS	ROBBIE WILL...	GREATEST HITS
AUNQUE NO TE PUEDA VER	ALEX UBAGO	REALIDAD O ...
CARTA A FRANCIA	FERNANDO D...	FEBRERO 13 ...
COME UNDONE	ROBBIE WILL...	ESCAPOLOGY
COSAS Y PALABRAS	FERNANDO D...	
CUENTO	FERNANDO D...	FEB 13 VOL 2
EL ABORDAJE	FERNANDO D...	
EL ADIOS	FERNANDO D...	PRIMERA EST...
ENTRE PAIROS Y DERIVAS	FERNANDO D...	
ERES MÍA	FERNANDO D...	VOL. 1
EVOLUCIONES	FERNANDO D...	FEBRERO 13 ...
FEEL	ROBBIE WILL...	GREATEST HITS
HEAVEN	ROBBIE WILL...	
HOY HACE UN BUEN DÍA	FERNANDO D...	FEBRERO 13 ...
HOY TEN MIEDO DE MI	FERNANDO D...	

Figure 3. Mp3 files after data cleansing process.

VI. CONCLUSION AND FUTURE WORK

We have proposed and mostly implemented a Data Cleansing Framework based on previous work in order to deal with poor data quality properties, such as accuracy, amount of data, format precision, representation consistency, duplicated values, and usefulness of data and usability of software.

We present a prototype with the implementation of the method proposed by Maletic and Marcus [3] and the removing technique explained by Loshin [2] applied to mobile devices and enhanced to cope with poor data quality.

Data quality patterns and data matching algorithms have been implemented and enhanced for the detection and correction of data errors within the process of data cleansing in mobile devices.

Data cleansing has being widely used on data warehousing, but not on data stored in mobile devices such as mobile phones, electronic organizers o audio players. The prototype developed within the present research is able to clean mp3 files only with id3tag v1 and v2.

Finally, we are designing a set of test for a full range of mobile devices in order to prove portability and to the correct function of the prototype.

As part of our future work, the cleansing algorithms implemented should support a wider range of formats. Supporting different character sets, a full implementation of data cleansing for all id3tag and vCard versions.

In the case of completeness we require to establish which data sources would be suitable for connection trough internet for further information that could be merged for missing information.

ACKNOWLEDGMENT

This work was supported by a grant from Dirección General de Asuntos del Personal Académico, UNAM.

REFERENCES

[1] P. Angeles M. and F. Garcia Ugalde “A data quality practical approach”. The International Journal on Advances in Software, IARIA.ISSN: 1942-2628, 2009, Vol. 2 No. 2&3, pp. 259-274.

[2] P. Angeles M. and F. Garcia Ugalde, “Assessing Quality of Derived Non Atomic Data by considering conflict resolution Function”, First International Conference on Advances in Databases, Knowledge, and Data Applications,2009, 978-0-7695-3550-0/09 IEEE, pp. 81-86.

[3] S. Chaudhuri and U. Dayal. “An overview of data warehousing and OLAP technology”. SIGMOD Rec. 26, 1 (March 1997), 65-74. DOI=10.1145/248603.248616.

[4] E. Rham and H. H. Do, “Data Cleaning: Problems and Current Approaches” in the Bulletin of the Technical Committee on Data Engineering, December 2000 Vol. 23 No. 4 IEEE Computer Society, pp.3-13.

[5] A. E. Monge “Matching Algorithms Within a Duplicate Detection System”, in the Bulletin of the Technical Committee on Data Engineering, December 2000 Vol. 23 No. 4 IEEE Computer Society, pp.14-20.

[6] D. Loshin “Enterprise knowledge management”, Chapter 3 and 4, pp. 17-37 and pp. 48-52.

[7] J. Maletic and A. Marcus. “Data cleansing: beyond integrity analysis”. Division of Computer Science. The Department of Mathematical Sciences. The University of Memphis. Campus Box 526429.

 ISBN:978-1-61208-115-1

[8] M. Hernandez and S.J. Stolfo. “Real-World data is dirty: data cleansing and The Merge/Purge problem”. Department of Computer Science Columbia University New York, NY 10027.

[9] P. Anokhin and A. Motro, "Fusionplex: Resolution of Data Inconsistencies in the Integration of Heterogeneous Information Sources", Technical Report ISE-TR-03-06, Information and Software Engineering Dept., George Mason University, Fairfax, Virginia, 2003

[10] Class JaroWinklerDistance: http://alias-i.com/lingpipe/docs/api/com/aliasi/spell/JaroWinklerDistance.html (retrieved; November 17, 2010.)

[11] G. Sauter and B. Mathews, “International Bussiness Machine (IBM), Information service Patterns“, Part 3 Data cleansing patern, 2007. http://www.ibm.com/developerworks/webservices/library/ws-soa-infoserv3/ (retrieved; November 17, 2010.)

[12] Mp3Tagger java program Stephen Ostermiller, Copyright © 2007 Free Software Foundation, Inc., http://fsf.org/ (retrieved; November 17, 2010.)

[13] L. Apicella, “Innovative Research Strategies: The Use of Handheld Computers to Collect Data the Population Council“. http://www.popcouncil.org/horizons/ORToolkit/toolkit/pda.html (retrieved; November 17, 2010.)

[14] S. Schumacher, ”Probabilistic Versus Deterministic Data Matching: Making an Accurate Decision”, Information Management Special Reports, January 2007 http://www.information-management.com/specialreports/20070118/1071712-1.html (retrieved; November 17, 2010.)

[15] C. Redman, “Data Quality for the Information Age”, Boston, MA., London : Artech House, 1996, pp. 551-582

[16] R.K. Bock and W. Krischer,”The Data Analysis BriefBook” Springer 1998.

[17] V. Barnett and T. Lewis, ”Outliers in Statistical Data”, John Wiley & Sons, New York, 1984.

[18] R.B. Buchheit, “Vacuum: Automated Procedures for Assessing and Cleansing Civil Infrastructure Data”, PhD Thesis, May 2002

[19] A.F. Karr, A.P. Sanil, and D.L. Banks,” Data Quality: A Statistical Perspective”, Technical Report 151, March 2005, National Institute of Statistical Sciences.

[20] Y. Lee and D. Strong,. “Knowing-Why about Data Processes and Data Quality”, Journal of Management Information Systems, Vol. 20, No. 3, pp. 13 – 39. 2004.

[21] Y. Wand and R. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations", Communications of the ACM, Vol. 39, No. 11, pp. 86-95, 1996.

[22] R. Y Wang and D.M. Strong, “Beyond accuracy: What data quality means to Data Consumers”, Journal of Management of Information Systems, Vol. 12, No. 4 1996, pp. 5 -33.

[23] Data Quality Pro Forum, http://www.dataqualitypro.com/ (retrieved; November 17, 2010.)

 ISBN:978-1-61208-115-1