

# Visual Analytics for the Interpretation of Fluency Tests during Alzheimer Evaluation

Antoine Lhuillier  
ENAC  
University of Toulouse  
Toulouse, France  
antoine.lhuillier@enac.fr

Christophe Hurter  
ENAC  
University of Toulouse  
Toulouse, France  
christophe.hurter@enac.fr

Christophe Jouffrais  
IRIT  
University of Toulouse  
Toulouse, France  
jouffrais@irit.fr

Emmanuel Barbeau  
CERCO  
University of Toulouse  
Toulouse, France  
barbeau@cerco.ups-tlse.fr

Hélène Amieva  
INSERM  
University of Bordeaux  
Bordeaux, France  
helene.amieva@isped.u-bordeaux2.fr

## ABSTRACT

A possible way to evaluate the progress of Alzheimer disease is to conduct the Isaac test [13,14]. In this activity, patients are asked to cite the largest possible number of city names within a minute. Since the city names are handwritten very quickly by a medical practitioner some cities are abbreviated or poorly written. In order to analyze such data, medical practitioners need to digitize the notes first and clean the dataset. Because these tasks are intricate and error prone we propose a novel set of tools, involving interactive visualization techniques, to help medical practitioners in the digitization and data-cleaning process. This system will be tested as part of an ongoing longitudinal study involving 9500 patients.

## CCS Concepts

• **Human-centered computing~User centered design;** • **Human-centered computing~Visual analytics;** • *Human-centered computing~Visualization techniques;* • **Social and professional topics~Personal health records;** • *Information systems~Data cleaning;* • *Applied computing~Document capture*

## Keywords

Alzheimer Disease, 3C Dataset, Electronic Health Records, Information Visualization, Data cleaning.

## 1. INTRODUCTION

According to a ministerial review from 2004, approximately 860,000 people are affected by Alzheimer's disease in France. This figure is estimated to reach 1.3 million in 2020 and 2.1 million in 2040. This is why the study of Alzheimer's disease has been identified as a major challenge for society. In the 3C cohort study [1], 9500 elderly people performed a set of tasks (lexical, cognition evaluation, sensory, and motor assessments). They were longitudinally followed during a 22 year period and some of them developed Alzheimer's disease (AD). Hence, the 3C project contains Electronic Health Record (EHR) that has been used to extract relevant information [2,8] regarding personal health.

© 2015 Association for Computing Machinery. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of a national government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.

VAHC '15, October 25, 2015, Chicago, IL, USA

© 2015 ACM. ISBN 978-1-4503-3671-0/15/10...\$15.00

DOI: <http://dx.doi.org/10.1145/2836034.2836037>

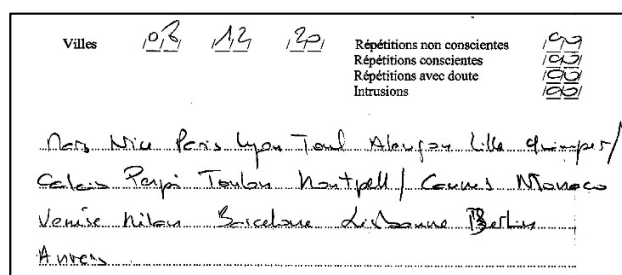


Figure 1. Sample of a fluency test result. The slashes represent timestamp separators (15, 30, and 60 seconds).

Our study focuses on one of the tasks performed by the patients in the 3C project: the lexical evocation task (called a fluency task). This fluency task is a part of the Isaac test set [13,14]. In this task, the patient is asked to cite the largest possible number of city names within a minute.

The lexical evocation task is directly connected to the concept of cognitive maps introduced by Tolmann [28] and echoed by O'Keefe and Nadel [22]. A cognitive map can be considered as a conceptual matrix in which episodes of life are recorded to be later accessed by mental processes (e.g., in context of a fluency test). Moreover, according to O'Keefe and Nadel, hippocampal lesions distort spatial 'imaging'. Since AD impacts the hippocampal region, we believe that analyzing the patient's cognitive map helps to detect the beginning of AD at an early stage.

We hypothesize that being able to spatially and temporally map cities cited during the test will help neuropsychologists to analyze the evolution of the patient's cognitive map. Therefore, performing the digitization and cleaning of the fluency test notes is a great opportunity to study how aging affects the human brain.

The digitization process is not an easy task and requires special care. Since notes are taken by hand by a medical practitioner, some city names are abbreviated or the handwriting is hard to read (a sample is shown in figure 1). The first step toward analyzing this data is the digitization step. This process involves transforming the handwritten information into a list of cities with temporal timestamps. Subsequently, one needs to clean the dataset from digitization errors and to remove ambiguities such as misspelled cities or any other kind of confusion of data. Due to the complexity of these two tasks, no automatic method can be applied. Furthermore, some errors in the list of cities can also contain relevant information regarding the cognitive state of the patient

(e.g., some of them are called *intrusion* and will be detailed in Section 5).

We used a User Centered Design (UCD) [27] process and developed a set of tools (like an auto-completion algorithm, a confidence-based algorithm to support the data cleaning process, and finally visualization tools to help knowledge extraction) to help medical practitioners during the digitization and cleaning process.

The main contribution presented in this paper is our interactive visualization system to help practitioners clean the patient records, trace modifications, and remove ambiguities between digitization error and actual insights.

In the following section we present relevant the related work. Then we list the design requirements to support the digitization and cleaning tasks. Next, we describe our software features and justify our implementation choices. Finally, we outline the strength of our design with specific cleaning scenarios.

## 2. RELATED WORK

Data cleaning, also called data cleansing, or data scrubbing is the process of correcting data by removing errors. The quality of the cleaning results is highly dependent of the intervention by a human domain expert [9]. This is also the case with our dataset, where only an expert can accurately perform the cleaning process.

Many tools focus on the data cleaning process, such as TimeCleanser [11] which is specialized in time oriented data. Our data set is also time oriented since the city citations are ordered by time, but the majority of the cleaning process will operate on the city names and not on the temporality of the city citations. In our case, the cleaning process will deal mostly with the geographical aspect of the dataset.

Different taxonomies of data quality problems exist [3,6,15, 21,23,26]. However, none of them propose a solution in which inconsistent spatial data can automatically be cleaned.

Among the available data cleaning tools, AJAX [10] provides a valuable framework including expressive and declarative language specifications to perform the data cleaning process. For instance, it supports mapping, matching, clustering, and merging of data entries. However in our study, very few errors are generic and only the individual investigation of the records can help to determine if they contain errors or ambiguities.

We investigated text recognition algorithms like A2iA [19], but the handwriting process with time restriction produced too many abbreviations and poor quality trails which hinder and prohibit the usage of automatic text recognition.

In their toponym disambiguation Bensalem et al. [4] and Buscaldi et al. [7] base their analysis on the geographical as well as on the hierarchical relations (e.g., Paris is the capital of France and France is a country of Europe). In their work, the hierarchical relations are based on the WordNet ontology [20]. However, the WordNet ontology was not precise enough to help us disambiguate toponyms of small villages of France.

## 3. DEFINITIONS

In this section we define important medical and technical terms.

**Fluency Test:** This is a task part of the Isaac test set [13,14]. In the fluency test, patients are asked to cite the largest possible number of city names within a minute. As the patient lists city names, the practitioner writes down the cited cities and adds a marker at 15, 30 and finally 60 second intervals (the end of the test).

**Citation:** We distinguish two types of citations: *Verbal* and *Digitized* citations. Verbal citations are cited by the patient and then noted on paper. Digitized citations are verbal citations that have been digitized.

**Intrusion:** A word that is definitely not a city (e.g., Banana). This is one of the errors that can be present in verbal and/or digitized citations. More details regarding errors and ambiguities are provided in Section 5.

**Paraphasia:** A speech defect characterized by incoherence in the arrangement of a syllable or word. There are different types of paraphrases depending on the type of incoherence. (e.g., semantic, verbal, or phonemic). For example saying 'tiger' instead of 'lion' is a semantic paraphasia.

**Patient:** The subject who performs the fluency test. The patient may be diagnosed as having dementia or not, as a result of medical testing processes which are outside of the scope of this paper.

**Medical Practitioner:** The person who supervised the fluency test and noted the cited city names on paper. He/She is a medical expert.

**Digitizer:** The person who takes the paper based record of the fluency test and enters the list of city citations into the software. This software contains many features to help the digitizer during the digitization process (these features will be explained in Section 7.1).

**Cleaner:** The person who employs our tools to perform data cleaning. The cleaner may as well be the practitioner, however, it is important that he/she has specific knowledge about cognitive processes. He/She must be an expert to determine whether the investigated record contains errors or ambiguities and finally he/she needs to perform possible corrections.

**Inspector:** The person in charge of checking the correctness of the cleaned data. The inspector is involved in our process after the cleaner. He/She must be a medical expert regarding cognitive processes (e.g., a neuropsychologist).

## 4. USER CENTERED DESIGN

During design and implementation of our digitization and cleaning tools, we applied a user centered design process (UCD) [27]. In this design process, we involved users as much as possible in the tool developments.

Different kinds of users were involved in this project: the practitioners who ran the tests, the patients who took the tests, the digitizer who carries out the digitization, the cleaner who cleans the data, the inspector who validates the cleaned data, and finally the neuropsychologist who analyzes the results. We recorded interviews with practitioners and digitizers, and we had numerous discussions with one specific inspector and one neuropsychologist.

In total we spent one day with the practitioners located in Bordeaux, and we conducted many discussions with three practitioners, two cleaners, and one inspector. We spent up to twenty hours of discussion time on the phone to assess our tools. We conducted one brainstorming with two HCI experts, one cognitive science expert (co-author of this paper), a practitioner, and a cleaner. We conducted one design walkthrough (i.e., human computer interface validation) with one cleaner. Finally, we spent three hours installing and detailing our tools for one cleaner.

These interviews and brainstorming sessions helped us to analyze errors and to refine the design requirements of our solutions.

## 5. ERRORS AND UNCERTAINTIES

Exploring, analyzing, and building statistics on a large cohort of patients requires taking into account errors during the digitization and cleaning process. Therefore, we need to quantify, analyze, and justify each error that may occur during the processes from the digitization to the analysis phase. This led us to define ways to compute a confidence value for digitized cited city names with regards to possible errors.

### 5.1 Types of Error

Thanks to the UCD design process and the numerous discussions with the practitioners, we have identified three different types of errors detailed in table 1.

**Table 1. Error types identified in our study**

Type	Sub-type (with details)
Patient errors	- Semantic paraphasia (lexical substitution) - Verbal paraphasia (phonetic substitution) - Intrusion - Repetition - Undetermined others
Uninterpretable errors	- Unreadable city name (on paper) - Incomplete test (e.g., a whole part of the test is missing)
Digitization errors	- Wrong city (copy error) - Unknown city (city not in our database)

In addition to correcting the digitization and uninterpretable error types, the data cleaning step can also be used to identify patient specific errors. For example, in our particular case, some digitization errors (such as a wrong city) can in fact be a patient error (for example, Toulouse and Toujouse are two existing cities but can also be a verbal paraphasia error).

### 5.2 Confidence Values

The confidence value can be computed using multiple criteria. Thus, in our study we have identified three types of confidence: spelling, phonetic, and Euclidian based distance.

#### 5.2.1 Spelling-based

The first confidence value is spelling. Multiple cities have the same name, for instance, there are more than 10 different cities whose names start by ‘Chalons’ and are orally referred as ‘Chalons’ without the rest of their proper name. There is a large possibility of typing errors (e.g., misspelling). Thus, it is probable that the digitizer made a mistake. From these two errors, it is necessary to remove any ambiguity. Our solution is to define a string-distance-based confidence. This confidence is based upon the Damereau-Levenstein [9,16] distance and allows us to get a list of cities with similar spelling. We define the confidence as such:

$$Card(\{city \mid d(citedCity, city) < thresholdValue\})$$

With the threshold being the minimum allowed string distance between the cited city and another.

#### 5.2.2 Phonetic-based

Relatively similar to the spelling-based confidence, the phonetic-based confidence uses a phonetic comparison between two words. We chose to implement the metaphone algorithm [24,25] suitable for the French language. In this algorithm the listed city names are translated into a key. Thus, we can find phonetically similar cities

by comparing their keys. We define our confidence as the string distance between two city keys.

$$Card(\{CityKey \mid d(citedCityKey, cityKey) < threshold\})$$

With the threshold being the minimum allowed phonetic distance.

#### 5.2.3 Euclidian-distance-based

From our experience, only lexically based confidences may not be sufficient to detect potential errors. Thus, we have defined a third confidence level based on the Euclidian distance of the city from the preceding and the following ones. During the task of verbal fluency, people tend to list cities near to each one’s following routes itinerary as theorized in vista space concepts [18] and revealed during our interviews with practitioners. Thus, if a cited city is far from the previous city, it may highlight a potential digitization error.

Let ‘Center’ be the geographical position between two cities: B, the city cited before the city in question, and F, the city cited after the city in question; while G is the actual geographical position of the city in question. We define our Euclidian based confidence as:

$$\frac{d(G, Center)}{\|BF\|}$$

Normalizing our Euclidian based confidence, allows us to spot a divergent city from the preceding and following ones.

## 6. DESIGN REQUIREMENTS

This section presents the requirements of the digitization and the cleaning tasks. First, we analyze the digitization requirements and then the cleaning requirements. Requirements associated with the digitization process are abbreviated ‘‘Rd’’ and then those with the cleaning process ‘‘Rc’’. All the requirements associated with the two tasks were collected during the UCD process.

### 6.1 Digitization Process Requirements

During the digitization process, the digitizer needs to produce the list of digitized cited cities.

Rd1: Multiple solutions were available to us such as: character recognition, verbal recognition or typing. Due to the poor writing quality of the tests (figure 1), we were not able to use the character recognition. Concerning verbal recognition, as the tests are in French, we needed a French-based grammar recognition. However, we were unable to find an efficient and open source French-based grammar. Thus, finally we choose to digitize the fluency test by manual typing.

Rd2: Manual typing does not solve by itself the need to digitize a correct dataset. Some cities might have the same name (e.g., London, UK and London, CA). To avoid such indetermination, the digitizer must be able to digitize a city to a unique name.

Rd3: The digitizer must specify the separators between cities listed during a period of 0 and 15 seconds, 15 and 30 seconds, and 30 and 60 seconds.

Rd4: The digitizer must be able to specify an intrusion.

### 6.2 Cleaning Process Requirements

Rc1: Each digitized and cleaned fluency test must be validated by an inspector. To enhance his/her validation work, we want the inspector to be able to look through any modification that the cleaner might have made and to validate or reject each test. Thus, we need to provide him/her with the list of each modification as well as allowing him/her to display the original digitized dataset.

The following requirements concern the cleaner tasks:

Rc2: The cleaner must be able to correct a digitized city and specify the error type. Thus, he/she must have interaction means to achieve this task. In the continuity of our digitization process, the cleaner must be able to solve any city-name indetermination problem (i.e., the system must allow the definition of a unique city name).

Rc3: However, allowing the cleaner to fill-in a unique city name, means using a database of existing cities. If a city is missing from such a database, the cleaner needs to be able to enter a name directly.

Rc4: Finally, as some errors are not known in advance, the system must give the cleaner the opportunity to find new errors. Thus, he/she must be able to explore the dataset in order to study prospective analysis methods.

### 6.3 Fluency Test Dataset

The fluency test dataset records the cities listed by a patient during a test (e.g., ‘Paris, Bordeaux, Toulouse, Lyon, etc...’). This list of cited cities is separated into three sub-lists which consists of the cited cities during three sequential intervals (see table 2). Different tests are linked through the patient’s ID.

**Table 2. Fluency test field names and semantics**

Field name	details
TestID	The unique ID of the test
PatientID	The unique ID of a patient
BirthDate	The birthdate of a patient
TestDate	The date of the test
City15	The cited cities between 0 and 15 seconds
City30	The cited cities between 15 and 30 seconds
City60	The cited cities between 30 and 60 seconds

The cited city field includes the cited city name, the real name, and the geographical position (latitude and longitude).

## 7. TOOLS

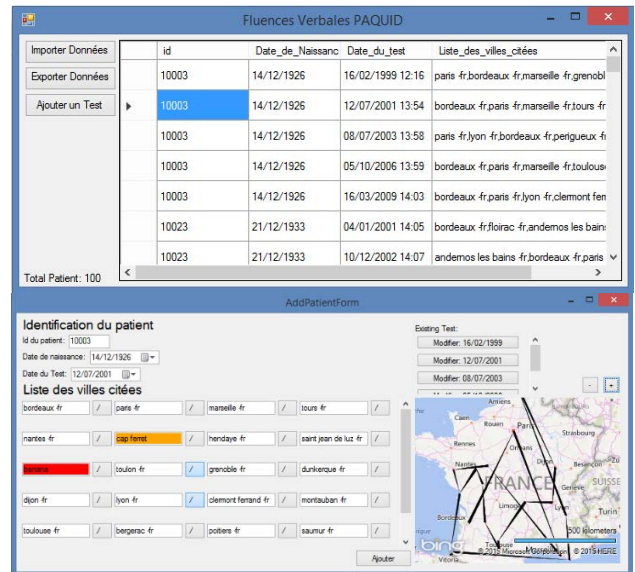
This section details the basic features of the two tools we developed in order to support the previously identified requirements: the digitization tool and the data cleaning tool.

### 7.1 Digitization Tool

The digitization tool has two panels. The main panel shows the list of the fluency test records with values (figure 2 - top). The digitizer can import and export the digitized data at any time.

To add test results, the digitizer uses a form and manually enters the names [Rd1] and fills in the data of the fluency test (see table 2). To allow the digitizer to bind a city name with a unique city, we have implemented an auto-completion algorithm based on a city database [Rd2]. The city database is composed of the major international cities [29] and all French towns [30]. When a town is missing, the digitizer can still override the auto-completion. In that case, the digitizer is warned with an orange feedback. Moreover, he/she can specify an intrusion with a right click on the text box [Rd4].

While entering the listed cities, the digitizer can see the geographical polyline formed by the fluency test (at the bottom-right of figure 2). This tapered polyline shows the location of each



**Figure 2. The digitization tool. On top is the list of tests, below is form to add a fluency test. Each city is typed in combo-boxes. Unknown cities are orange, and intrusions words red.**

city (known in the database) and the direction of the polyline [12]. With this visualization, he/she is able to see whether a city seems erroneous or not (i.e., a city too far from the previously listed ones).

To add the timestamp of the fluency test (each 15, 30, and 60 seconds) the interface provides slash buttons, allowing the digitizer to input the slashes that are present in the paper-based test [Rd3].

Finally, by typing an existing test ID, the digitizer is able to load an existing test and to modify it.

### 7.2 Cleaning Tool

Our cleaning tool (figure 3) is composed of three main panels: the histoSelector panel, the citation panel, and the detail panel. In the following subsections, we detail each component.

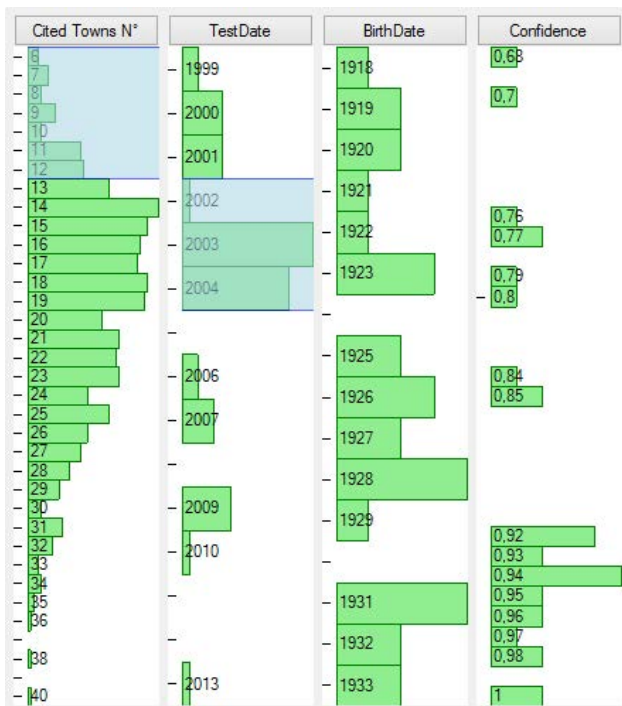


**Figure 3. The cleaning tool. On the left the HistoSelector panel, and on the right the Citation Panel. The cleaner can change the confidence with combo box above the citation panel.**

#### 7.2.1 The HistoSelector Component

To filter and explore the dataset during the cleaning process [Rc4], we display histograms showing the distribution of values of each dataset attribute (figure 4). The cleaner can select the dataset attribute to be displayed by right clicking on the histogram name.

To find new insights or errors in the dataset, we have chosen a selection paradigm based on a brushing interaction [17]. Users



**Figure 4. Histogram Panel.** By drawing boxes on the histograms, a cleaner has selected all patients who revealed only a low number of towns during the 2002-2004 tests. A selection in one histogram updates all histograms to its right.

draw a bounding box (a one dimension brushing technique) and modify it by sliding the bounds of the selection. Multiple boxes can be specified; they are removed by a brush stroke with the right button of the mouse pressed (i.e., the erase mode).

We maintained the classical reading pattern (left to right) in such a way that the selections or modifications of one histogram directly modify its neighbor's values (Figure 4). Nesting multiple histograms allows the user to refine a query depending on multiple dataset attributes. The fluency tests selected through the histograms are displayed in the citation panel.

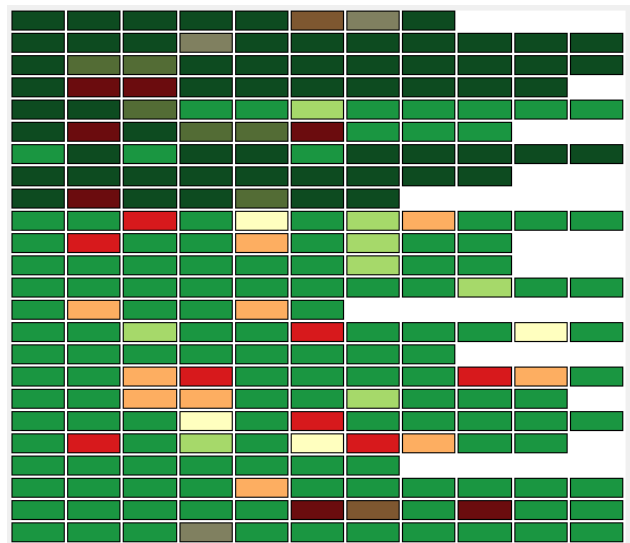
### 7.2.2 Citation Panel

The citation panel is the representation of each fluency test in our cleaning tool (figure 5) [Rc1].

One row corresponds to one fluency test. The row is composed of colored squares, each encoding the confidence value for one cited city by color. Two parameters are represented:

First, the saturation of each square encodes the state of validation of the city. If the city has been validated then it has a low saturation value. On the other hand, if the city was not validated, the saturation is at its full value. As saturation is a dissociative visual parameter [5], encoding the validation state of a city through saturation allows the visual emergence of the not validated cities.

Secondly, the color of each square encodes the confidence value of the city. The associativity of the color improves pre-attentive selection [5] allowing users to instantly locate a group of low confidence cities that need to be reviewed. Our color range defines five levels of confidence with a color generator [31]. The colors are green for confidence values over 80%, then light-green (60-80%),



**Figure 5. Citation panel.** Displays the confidence of each digitized city encoded in five levels of color (red to green). A validated city is de-saturated (top of the figure).

yellow (40-60%), orange (20-40%), and red for confidence values lower than 20%.

Finally, clicking on a colored square opens up the detail panel.

### 7.2.3 Detail Panel

The detail panel allows the user to correct a potential error in a digitized city. This panel is composed of four main parts:

The first part, at the top of the form, displays the values associated with the city (PatientID, BirthDate, TestDate, Confidence value, and the actual digitized cited city). In this first part, the user is able to correct the city name (via the Correct City textbox) and fill-in the error type [2]. Each textbox provides a personalized auto-completion algorithm. The cleaner can add a city to the database via the "City not in DB" button [Rc3]. The second part displays the list of cities lexically similar to the cited one.

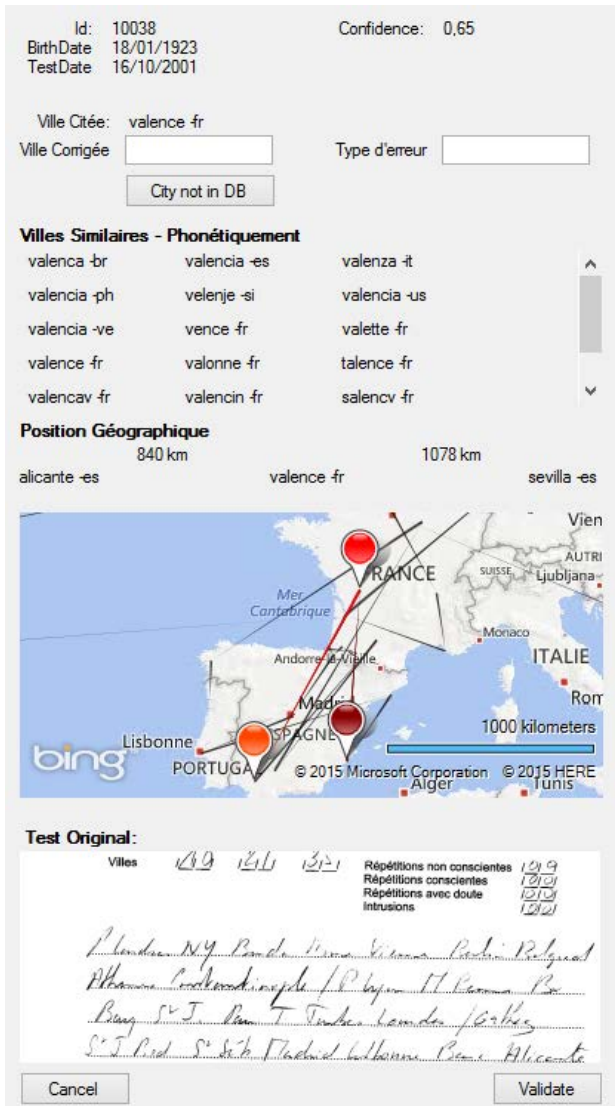
The third part shows the cited cities preceding and following the selected city and displays the distance between them. This geographical distance is enhanced by displaying the polyline formed by the fluency test centered on the selected city. This visualization allows the user to geographically place and check the location of the city and to compare it with the following and preceding cities as well as with the whole fluency test.

The last part of the detail panel (at the bottom) is an image display of the original paper-based test. This allows the cleaner to check if the city was really written or not, especially in the case of a city with a low confidence value but in which the digitized city was indeed cited by the patient (see figure 7 for an example).

Once the correction of the city has been performed, the user can validate or cancel his/her changes.

## 8. SCENARIO

This section describes four scenarios of use. The first two describe a true-positive and a false-positive detection. The third one shows how the cleaner can explore the dataset and find unrevealed erroneous data. Finally the last scenario illustrates the validation process performed by one inspector. The typical meta-scenario is as follows:



**Figure 6. Detail panel. On the top are the test information, then the phonetically similar cities, and the geographical position within the fluency test. Finally, we display the original fluency test.**

While a patient 'V' executes the task of verbally citing city names, a practitioner 'W' notes each cited city on paper. At 15", 30", and 60" the practitioner writes down a slash to timestamp the cited cities. After the test, a digitizer X manually enters each cited city and slashes into our digitization tool as explained in Section 7.1.

Once the data is loaded into our numeric database, a cleaner Y uses our cleaning tool to correct the database. The digitization tools help the cleaner by showing the most error-prone cited cities according to the given type of confidence. The cleaner then has to check whether the detection of an erroneous value by the system was justified (scenario described in Section 8.1) or not (scenario described in Section 0).

Finally, an inspector Z checks each correction made by the cleaner in order to validate or invalidate the changes made to the database. The tracking and validation of these modifications are detailed in Section 8.4.

## 8.1 True-Positive Scenario

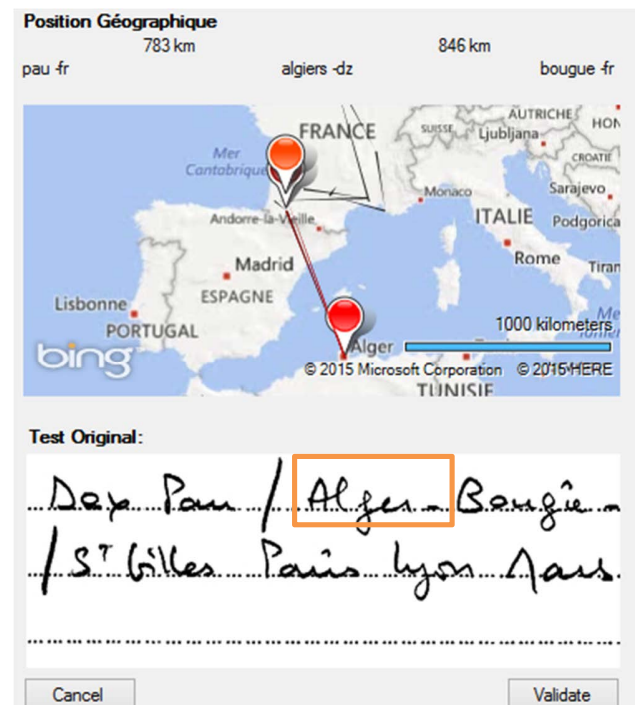
The visualization shown in figure 3 displays the digitized part of our dataset. The cleaner Y wants to validate each digitized fluency test. To do so, Y follows a procedure composed of three steps:

He/She initially decides to correct the tests with a low spelling-based confidence value. To do so, the user selects the lower values in the confidence histogram by brushing them (cursor on figure 3). By selecting those values, only the tests with a low confidence value are displayed in the citation panel. The selected fluency tests are ordered by decreasing level of confidence in the citation panel.

In a second step, Y validates the first row of fluency tests. He/She starts by clicking on the first red rectangle in the fluency test with the lowest confidence (first row of the Citation Panel). This action opens the detail panel to correct this test citation (figure 6).

In the detail panel, the cleaner sees that the digitized city name is 'Valence'. This city is located near Lyon in France. However, the user sees that the preceding city is 'Alicante' and the following one is 'Sevilla', two Spanish cities. The distance between those cities is thus quite large. Moreover, in the list of similar cities (cities at a small distance from the original one), the cleanser finds a Spanish city named 'Valencia'. Thanks to these two insights, the cleaner can infer that the intended city is rather not 'Valence' in France but 'Valencia' in Spain. Y enters the corrected city name in the Corrected City textbox. Then, the cleaner specifies the type of error; in this case a digitization error.

This example illustrates how the cleaner can correct the tests that are likely to be erroneous according to our spelling-based confidence algorithm. Moreover, it shows how one can help the cleaner to solve a decision problem and gain more insights to our database by the types of errors.



**Figure 7. Detail panel showing a false positive case. Although the cited city (Alger) is far from its neighbors (Pau and Bougue), we see on the picture of the original test that the patient did cite it (in the orange box).**

## 8.2 False Positive Scenario

In a similar manner to the previous use-case, the cleaner Y can use distance-based confidence to enhance his/her task of data cleaning. In this use-case, he/she wants to validate each digitized fluency test. Y follows the same steps as in the spelling distance scenario.

By selecting the right criteria, he/she can display the tests to be cleaned, thus allowing him/her to focus on those tests. Then, the cleaner selects the first red square showing a very low confidence value on a citation.

In the detail panel, the cleaner sees the information concerning the citation (Figure 7). The digitized citation shows the city 'Algiers' in Algeria. The previous cited city is 'Pau' and the following is 'Bougue'. Thus, the Euclidian distance is relatively high.

To confirm the inaccuracy of the digitized citation, the cleaner takes a look at the original citation at the bottom of the detail panel in figure 7. However, Y sees that the city that was written down on paper during the test is Algiers. He/She concludes the low confidence level was a false-positive one. Finally, he/she validates the correctness of the digital citation by clicking the validation button.

Since the confidence based criteria cannot always be true, this use-case highlights the way a cleaner can detect and resolve a false-positive result.

## 8.3 Detect Outliner Subsets Scenario

In this use-case the cleaner Y wants to detect a subset of outliers in the data to be cleaned. In this particular use-case, he/she wants to investigate the relationship between the length of the fluency tests and the year of the test.

Y initially decides to bind a first histogram to the fluency test length (the number of cited cities). By clicking on the tag name of the histoSelector component, he/she is able to change its binding. Then, he/she repeats the operation on the next histoSelector component (directly on the right of the previous one) binding it to the year of the test. With this configuration the cleaner is able to visualize the distribution of the test years depending on the selected fluency test lengths (figure 4).

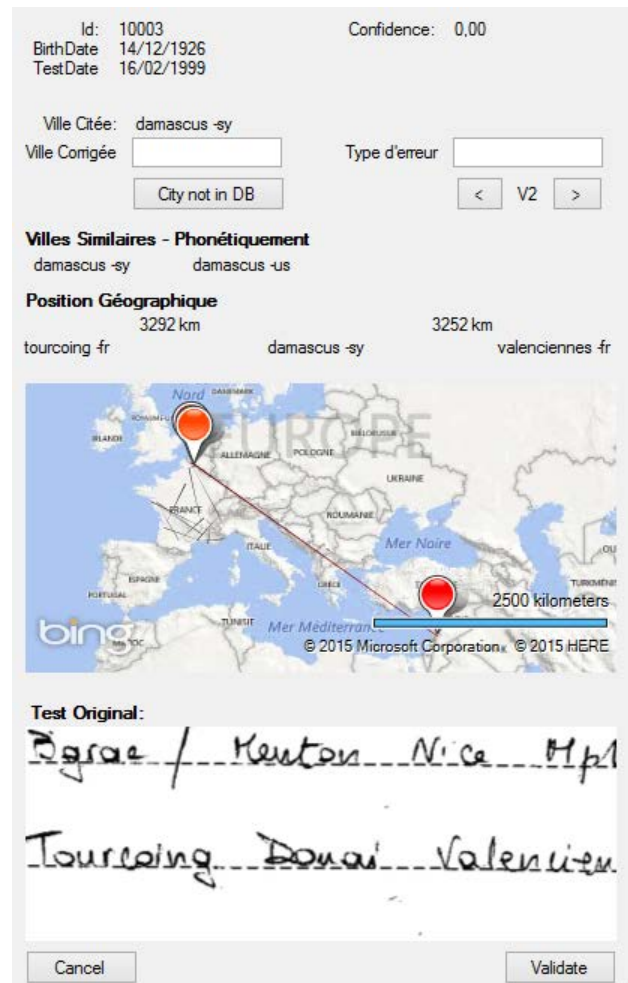
Then he/she selects the fluency test with a low number of listed cities by brushing the histoSelector component. During the brushing sequences, the cleaner can directly see the distribution of corresponding test dates. Y discovers that the majority of tests with a low number of listed cities were performed in 2003-2004. Moreover, the cleaner recognizes that all these low numbered tests are missing the 30 seconds to 60 seconds parts. Thus, the cleaner can conclude the presence of an error in the test protocol during the tests in the period from 2003 to 2004.

This use-case illustrates how the cleaner can detect outlier subsets through exploration of data properties. Exploring and interacting with the dataset revealed unpredicted errors.

## 8.4 Track Modifications Scenario

We detail a scenario where an inspector Z wants to validate corrections made by the cleaner Y during the cleaning process.

Once the cleaner has finished the cleaning process, the inspector can validate or invalidate these modifications. To achieve this objective, he/she starts binding the first histoSelector to the modified attribute and the second to the validated attributes. Then by selecting the modified but invalidated fluency tests (via the two



**Figure 8. Validation of a cleaned test. In this case, the cleaned city is Damascus but the original cited city was Douai. The expert can invalidate the modification and make a new correction to the data.**

configured histograms) Z displays the list of modifications requiring validation in the citation panel.

Within these tests, the inspector can select and explore each test. He/She chooses to click on the first test and to check the digitized cited city. Thus, Z opens the detail panel. In the detail panel he/she sees that the corrected city name is 'Damascus'. By clicking on the button next to the versioning number (left arrow '<' in figure 8), he/she sees that the initial value of the digitized cited city was a city unknown in the database named 'Damas' (the French name for Damascus).

Via the original test view in the bottom of the detail panel, Z checks if 'Damascus' is indeed written in the original test. However, the original test reveals not 'Damas' but 'Douai' a French city near the preceding and following cities. Thus, Z corrects the cited city a second time to the city of 'Douai'. Finally, he/she corrects and validates the change with the Validate button.

Through this use-case, we see how the validation of the inspector can spot errors made during the cleaning process and double check the corrections. This improves the correctness of the cleaned dataset.

## 9. CONCLUSION & FUTURE WORK

The study of Alzheimer's disease is a unique opportunity to work on an unexploited handwritten dataset. In the digitization and cleaning step the main difficulty is to reduce ambiguities between errors induced by the cleaning/digitization process itself and errors made by the patient during the verbal fluency test. Processing and identifying these errors cannot be achieved with an automatic algorithm due to its variability. Thus, it must be carried out by domain experts.

To solve this problem we developed two tools that help experts during the digitization and cleaning steps. As an example, these tools have helped the people in charge of the digitization process by allowing them to directly input known city names into an existing database. In the cleaning process, our tool helps the user to solve the ambiguity problems inherent to city names (having multiple cities with the same name) as well as solving problems from our digitization process (missing city names).

Our work can be further improved and extended in many aspects. For instance, other confidence tests could be investigated: the homonic test or the toponym test are good candidates for extending the available tools to detect issues in our database. We believe that the tools described in this paper can be useful to provide cleaner data and facilitate the study of Alzheimer disease progression.

## 10. REFERENCES

- [1] Alperovitch, A., Amouyel, P., Dartigues, J., et al. Epidemiological studies on aging in France: from the PAQUID study to the Three-City study. *Comptes rendus biologies* 325, 6 (2002), 665–672.
- [2] Auriacombe, S., Helmer, C., Amieva, H., Berr, C., Dubois, B., and Dartigues, J.-F. Validity of the Free and Cued Selective Reminding Test in predicting dementia The 3C Study. *Neurology* 74, 22 (2010), 1760–1767.
- [3] Barateiro, J. and Galhardas, H. A Survey of Data Quality Tools. *Datenbank-Spektrum* 14, 15-21 (2005), 48.
- [4] Bensalem, I. and Kholadi, M.K. Toponym disambiguation by arborescent relationships. *Journal of Computer Science* 6, 6 (2010), 653.
- [5] Bertin, J. *Graphics and graphic information processing*. Walter de Gruyter, 1981.
- [6] Bose, R., Mans, R.S., and van der Aalst, W.M. Wanna improve process mining results? IEEE (2013), 127–134.
- [7] Buscaldi, D. and Rosso, P. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science* 22, 3 (2008), 301–313.
- [8] Catheline, G., Amieva, H., Dillharreguy, B., et al. P2a-2 Variabilité interindividuelle des performances cognitives mnésiques et leurs corrélats neuroanatomiques au cours du vieillissement: Etude sur une cohorte en population (3C). *Revue Neurologique* 165, 10 (2009), 65–66.
- [9] Damerau, F.J. A technique for computer detection and correction of spelling errors. *Communications of the ACM* 7, 3 (1964), 171–176.
- [10] Galhardas, H., Florescu, D., Shasha, D., and Simon, E. AJAX: an extensible data cleaning tool. ACM (2000), 590.
- [11] Gschwandtner, T., Aigner, W., Miksch, S., et al. TimeCleanser: A visual analytics approach for data cleansing of time-oriented data. ACM (2014), 18.
- [12] Holten, D., Isenberg, P., Van Wijk, J.J., and Fekete, J.-D. An extended evaluation of the readability of tapered, animated, and textured directed-edge representations in node-link graphs. IEEE (2011), 195–202.
- [13] Isaacs, B. and Akhtar, A.J. The set test: a rapid test of mental function in old people. *Age and ageing* 1, 4 (1972), 222–226.
- [14] Isaacs, B. and Kennie, A.T. The Set test as an aid to the detection of dementia in old people. *The British Journal of Psychiatry* 123, 575 (1973), 467–470.
- [15] Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., and Lee, D. A taxonomy of dirty data. *Data mining and knowledge discovery* 7, 1 (2003), 81–99.
- [16] Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. (1966), 707–710.
- [17] Li, Q. and North, C. Empirical comparison of dynamic query sliders and brushing histograms. IEEE (2003), 147–153.
- [18] Meilinger, T. The Network of Reference Frames Theory: A Synthesis of Graphs and Cognitive Maps. In C. Freksa, N.S. Newcombe, P. Gärdénfors and S. Wölfl, eds., *Spatial Cognition VI. Learning, Reasoning, and Talking about Space*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, 344–360.
- [19] Menasri, F., Louradour, J., Bianne-Bernard, A.-L., and Kermorvant, C. The A2iA French handwriting recognition system at the Rimes-ICDAR2011 competition. International Society for Optics and Photonics (2012), 82970Y–82970Y.
- [20] Miller, G.A. WordNet: a lexical database for English. *Communications of the ACM* 38, 11 (1995), 39–41.
- [21] Müller, H. and Freytag, J.-C. *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. Für Informatik, 2005.
- [22] O'keefe, J. and Nadel, L. *The hippocampus as a cognitive map*. Clarendon Press Oxford, 1978.
- [23] Oliveira, P., Rodrigues, F., and Henriques, P.R. A Formal Definition of Data Quality Problems. (2005).
- [24] Philips, L. Hanging on the metaphone. *Computer Language* 7, 12 (December) (1990).
- [25] Philips, L. The double metaphone search algorithm. *C/C++ users journal* 18, 6 (2000), 38–43.
- [26] Rahm, E. and Do, H.H. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.* 23, 4 (2000), 3–13.
- [27] Schuler, D. and Namioka, A. *Participatory design: Principles and practices*. CRC Press, 1993.
- [28] Tolman, E.C. Cognitive maps in rats and men. *Psychological Review* 55, 4 (1948), 189–208.
- [29] GeoNames world cities over 15000 citizens database. *GeoNames*. <http://download.geonames.org/export/dump/cities15000.zip>.
- [30] GeoNames French cities database. *GeoNames*. <http://download.geonames.org/export/dump/FR.zip>.
- [31] 5 Class color scheme from red to green. *ColorBrewer*. <http://colorbrewer2.org/?type=diverging&scheme=RdYlGn&n=5>.