# All that glitters is not gold!

Pedro Cunha (pedro.cunha@ine.pt), Statistics Portugal/Information Infrastructure
Sónia Quaresma (sonia.quaresma@ine.pt), Statistics Portugal/Information Infrastructure
Jorge Magalhães (jorge.magalhaes@ine.pt), Statistics Portugal/Methodology and Information System

## Abstract

In Portugal, it is possible since last March to change our personal address using the citizen portal, not only within public entities like the national registry, the tax authority or the employment agency and even some private companies like the Highways Green Card, whenever we choose to do so. Having this ability is great but it relies on the citizen's skill to provide the correct address thus having a limited impact in the construction of a national address database.

In the second quarter of 2015, the *Base Adresse Nationale* was inaugurated in France, comprising 25 million addresses compiled with the collaboration of the citizens, public authorities, state authorities, public operators and firms. This database provides open data and aims to quickly identify and manage 200-300000 new addresses created each year.

Last April, during the U.S. National Address Database Summit, the U.S. recognized the importance of a usable, unified set of addresses as a piece of national infrastructure. But there, as well as in Portugal, a combination of bureaucratic, legal and organizational challenges have stood in the way of fixing this problem, with the main concern being the security of personal information.

During the last census operation, Portugal also performed a housing census integrating geospatial information which could constitute our initial National Households Register but has to be updated to be of use as an instrument for sample design for the surveys of the NSI as well as a base for the next census operation.

We propose to show how we guarantee the enhancement and the measurement of the file quality, through its update and maintenance, using administrative data and surveys.

Furthermore, we suggest and discuss models of partnership and collaboration between public and private organizations (ADENE and EDP) to secure and assess the quality management of a National Address Database.

**Keywords:** National Address Database; Information Management and Maintenance; Partnership between Public and Private Organizations regarding information collection

## 1. Introduction

Following the 2011 Census operation, a database storing micro-data relative to all households in the national territory was created. From that moment forward, it became imperative to update it, in a correct, complete and efficient way. The "National Dwelling Registry" (FNA) supports all the sampling for the family statistical operations, making its completeness and accuracy crucial for the Statistics Portugal's activity.

Although we receive information from several internal (surveys) and external (administrative) sources, their quality varies not only across sources but also over time. Not all that glitters is gold and this project took upon it to separate the wheat from the chaff. This means that, given a specific data source, we should be able to evaluate whether it can be used on a regular basis to keep FNA properly updated.

In the following sections, we present a brief historical overview of the situation that motivated this journey through the constitution of FNA and the construction of its feeding procedures, while at the same time measuring quality at each step and trying to be proactive in the anticipation of problems. We conclude by discussing future directions and developments of a project in this field.

## 2. How it all began?

Since 1979 until 2012, the sampling of family statistical operations (SO) at Statistics Portugal has been based on a large sample named "Master Sample" (AMAE). The MS is built after the conclusion of each Census of Population and Housing operation and is maintained for a decade and updated based on current SO or through specific field work. This method proved to be inefficient, time consuming and very expensive leading to a lower quality of collected data.

The 2011 Census, the geo-referencing of buildings, access to different administrative sources (with different attributes, key fields and record drawings), the EURADIN project (European Address Infrastructure) and INSPIRE (Infrastructure for Spatial Information in the European Community) were, on the whole, an opportunity to change the strategy in defining the sampling frames of SO directed to families.

The new strategy is supported by the implementation of a national dwelling/housing register (FNA), originally created on the basis of the 2011 Census micro-data, and updated

based on different sources. This register sets up a reference universe from which the different base samplings are extracted.

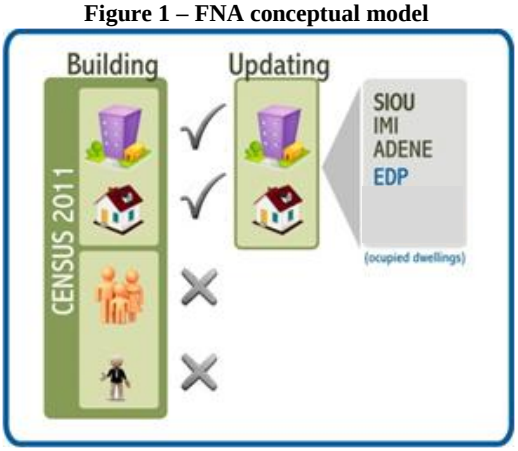
Updating FNA is the next and critical stage to guarantee the enhancement of the data quality. Achieving this goal is only possible through the use of relevant information, with quality and contemporary data sources.

There are several sources to be considered for FNA's updating. Within Statistics Portugal activity, the Indicators System of Urban Operations (SIOU) is an appealing possibility but other statistical operations can be used to feed FNA. As external sources, we highlight the data systems associated with the Local Tax on Real Estate (IMI), the Agency for Energy Certification (ADENE), the Personal Income Tax (IRS), Social Security (SS), the Civil Identification Database (BDIC/CCIC) and Portugal Energies (EDP).

This model follows Statistics Portugal's experience in the field of business statistics, in particular regarding the Global Survey Management System (SIGINQ), which includes subsystems for the Register of Statistical Units (FUE) and Universes and Samples Management System (SIGUA).

The conceptual model of FNA includes two stages: the creation, using the 2011 Census data, and the updating, using both internal and external sources.

**Figure 1 – FNA conceptual model**




In the first stage – creating the register – only information about buildings and dwellings drawn from the 2011 Census operation were used

The regiter contained localization variables (e.g. administrative division codes, geographical coordinates and the address), characterization variables (e.g. socioeconomic variables and the dwelling status) and internal variables (e.g. the dwelling availability, the update date and update source).

Specifically, the geographical component of FNA is largely based on the Buildings Geographic Base (BGE), one of the main components of Statistics Portugal's Spatial Data Infrastructure (IDE).

The first version of BGE refers to the census moment and integrates all residential buildings geo-referenced in the 2011 Census in which the FNA dwellings are located.

In addition to the residential buildings of BGE, the FNA will include, as the reference geography, the official European Kilometric GRID – Grid_ETRS89_LAEA_1K – developed by EUROSTAT for the European territory. This GRID contains, for Portugal, 94.265 rectangular cells of 1 $km^2$, with each cell having a unique identifier.

The use of GRID allows for the use of a uniform geography and enables the geographic location of buildings in a harmonized and interoperable manner regardless of changes that may occur in the administrative boundaries (in the Portuguese case, municipalities and parishes) or in the boundaries set specifically for statistical purposes (statistical sections and subsections).

Despite the use of GRID, the buildings in FNA keep, albeit indirectly, the relationship with the administrative geography of the Official Administrative Map of Portugal (CAOP).

## 3. How can it be fed?

As previously mentioned, currently, the major source for updating FNA is SIOU. This source allows update of units (buildings or dwellings) in terms of its demography (born and death of units). This system is based fundamentally on the use of administrative information associated with the new legal framework that distinguishes the different forms of procedure regarding real estate interventions.

The areas covered by this system are the allotments, land redevelopment work, the works of construction and demolition of buildings and the use change of buildings. To collect all this information, the following tools were created:

Q1 – Survey on the Allotment of Urban Operations;
Q2 – Survey on Land Refurbishment Works;
Q3 – Survey on Building Works Projects and Building Demolition;
Q4 – Survey on the Use of Completed Buildings;
Q5 – Survey on Work Completion; and,
Q6 – Survey on the Buildings Use Changes.

Out of the six instruments collection, only Q5 is a Statistics Portugal initiative survey. The others fall within each municipality administrative responsibility. In these five cases, the information is monthly received by Statistics Portugal. Specifically, for the purpose of maintaining FNA, only information about Q3, Q4 and Q5 is considered.

The Q5 survey is conducted by post mail and responded by post mail or telephone among the project promoters on the scheduled date of completion of work but only for those where there is no information that the work has already finished (because if the work has been completed, a Q4 is acquired via the municipality).

The micro-data of SIOU constitutes a dynamic file, with the registration opening being set by the Q3 input, with the subsequent registration of the closure being given by the input of the corresponding Q4 or Q5. Only with the reception of the Q4 or Q5 information, there is confirmation of the execution of the work.

Generally, it can be said that the Q3 is the intention of constructing from scratch (or demolish, enlarge, alter, reconstruct) while the Q4 and Q5 represent the effective completion of the work.

As there are two possible surveys to complete the same work, SIOU establishes a priority system: if the work was completed by Q4, the license cannot ever be changed; but if completed by Q5 and afterwards a Q4 is issued by the municipality, the original date of conclusion is preserved (obtained by Q5) but the characteristics of the work (address, promoter, coordinates, etc.) are updated by Q4.

From what was described, SIOU is recognized as a major source for updating FNA. The first priority will be given to the completion of works (Q4 and Q5) that generate new dwellings that should be integrated in the upgrading of FNA.

By the analysis made, the updating process is different depending on the type of work of the license:

- new construction will give rise to new buildings and the corresponding new dwellings;
- expansions, reconstructions or refurbishing involving dwellings give rise to changes in the characteristics of buildings or to the insertion of new dwellings or to the exclusion of the demolished ones;
- total demolition leads to the elimination of the building and the corresponding dwellings in FNA.

## 4. Building the machine to operate the system

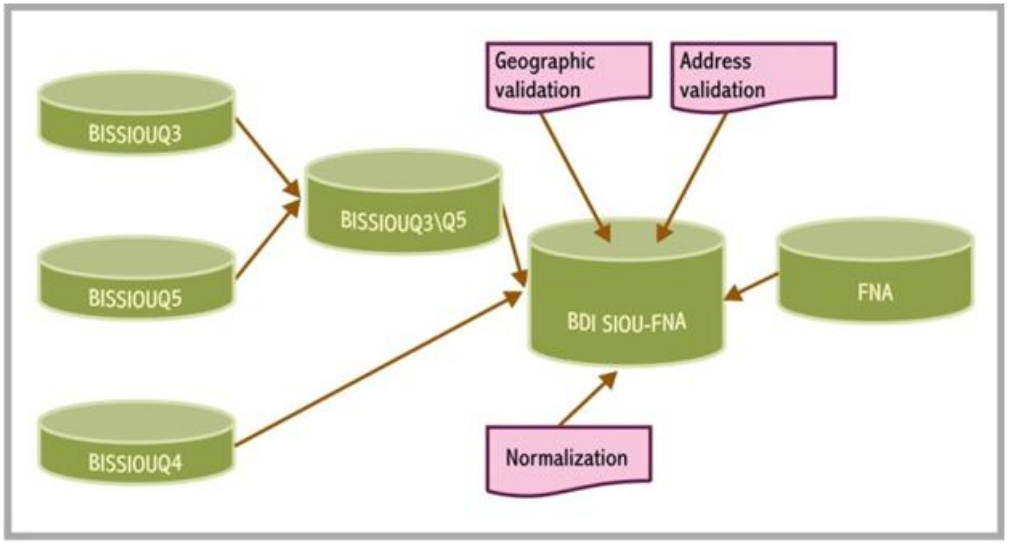
Up to now, Statistics Portugal deals with two different systems – FNA and SIOU - which should work jointly but are somehow incompatible because there is not any link between them. For that matter, Statistics Portugal decided to use the in-house knowledge of data warehouse and the already familiar analysis tools to overcome that caveat.

To link the two systems, it is necessary to store in FNA the original key of SIOU whenever a building or dwelling is created, updated or marked as demolished with information that was originated by SIOU.

Currently, the two systems are connected but we need to move forward and control the whole system: Which buildings and dwellings were inserted in FNA? Which buildings were not inserted in FNA? Why? What is the status of a process? These are some questions that require an answer, highlighting the need to build a data warehouse project with the ability of centralizing and controlling the operations on the system.

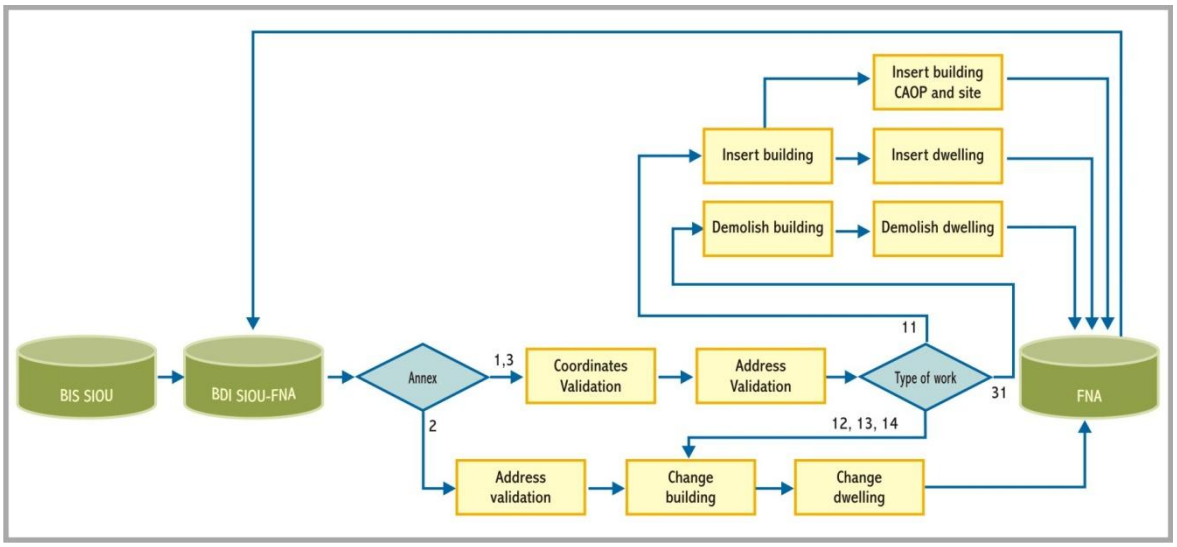
This project was built linking the two systems and integrating the information. The information was divided taking in account the type of survey – either Q4 or Q3 in conjunction with Q5 – and the type of work. It is updated every night or whenever it is needed. This data warehouse project is powered by the existing databases that support the SIOU (Q3, Q4 and Q5) and FNA. In addition, it also benefits from several input and validation files produced both by the infrastructure unit and the geography unit.

**Figure 2 – Project sources**



The global vision of the system is depicted in Figure 3.

**Figure 3 – Global vision of the system**




As a rough idea, it is possible to summarize the entire process as follows: (1) obtaining buildings that are not dealt with yet; (2) validating the coordinates; (3) normalizing the data; (4) validating the addresses, and finally (4) inserting, updating or marking as demolished buildings and dwellings in FNA.

Different actions must be taken depending on the type of work permit. So, the system must respond accordingly if it is the case of a new construction, a demolition or an enlargement, a reconstruction or a refurbishing. The actions to be taken are portrayed in the Figure 4.

**Figure 4 – The system's actions**


1 – The elements remained unchanged (and nothing happens);

2 – New elements should be inserted;

3 – The elements should be removed (when buildings are totally demolished);

4 – The geographical coordinates of the element changed (should be corrected);

5 – The characteristics of the elements should be changed

As mentioned previously, new constructions can lead to new buildings and new dwellings in FNA. The enlargement, reconstruction or refurbishing of dwellings may lead to new housing and demolitions may lead to the exclusion of existing ones.

However, we must ensure that the building to expand, rebuild or refurbish already exists in the FNA by analyzing the coordinates and the existing address. If the building is in FNA, that may involve changes to the accommodations. If there is a reconstruction and the building doesn't exists in FNA, we have to consider the creation of the building and corresponding dwellings. However, if it is an enlargement work and it was not possible to detect the building in FNA, the building and the related dwellings cannot be inserted since only information about the features to expand are provided and not the characteristics of the whole building.

The demolitions may lead to the removal of buildings and dwellings from the FNA. In this type of work, a check for the building is needed by analyzing the coordinates or the address indicated. If the building does not exist in FNA, there is no action. If there is, it must be removed.

Some of the actions needed to manage the system are described in the next figures. The first step is to obtain the data on the buildings that were never processed and making the coordinates validation and coordinate system conversion.

Next, for those buildings with the correct coordinates, it is necessary to normalize and to validate the address. When this process is done, duplication verification is needed to ensure that we neither insert buildings that already exist in FNA nor remove or change the incorrect ones.

When this verification is done, buildings and dwellings are ready to be inserted (or removed or updated) in FNA.

**Figure 5 – Coordinate validation – GEO unit**

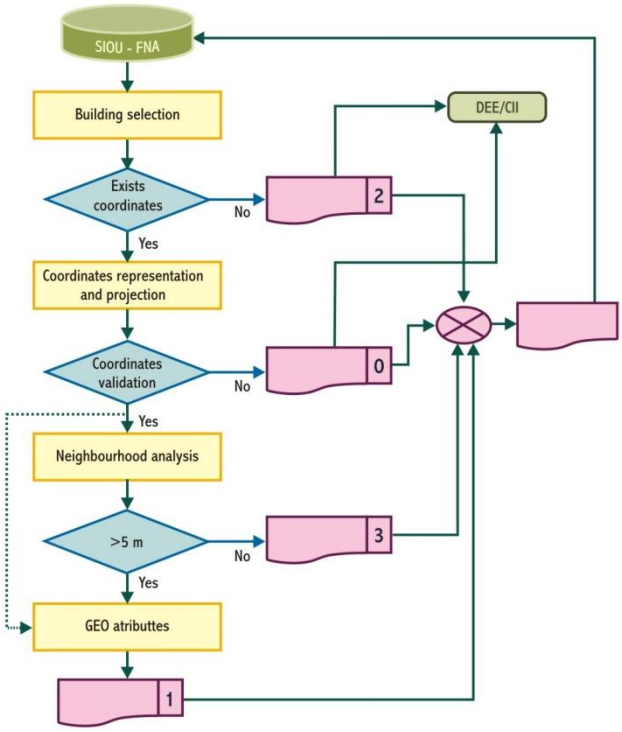



**Figure 6 – Normalization and address validation**

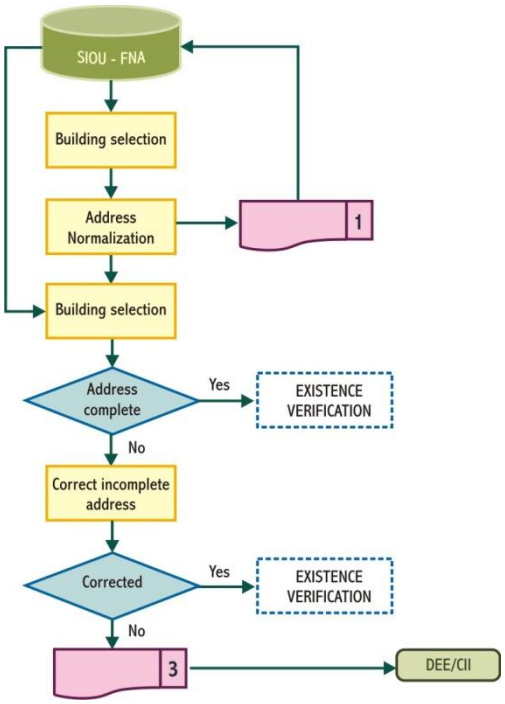



**Figure 7 – Address duplication check**

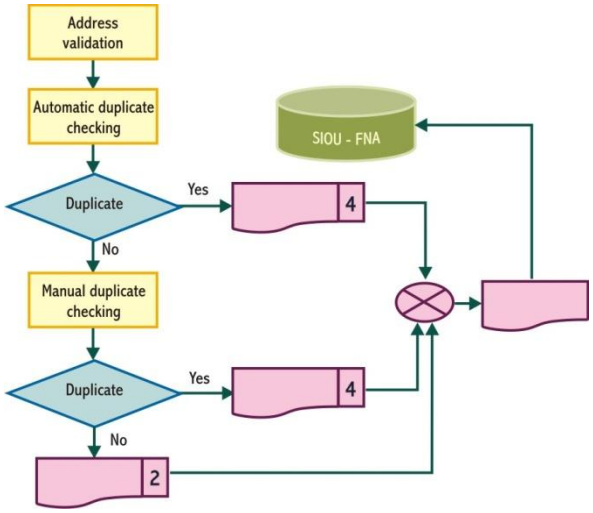
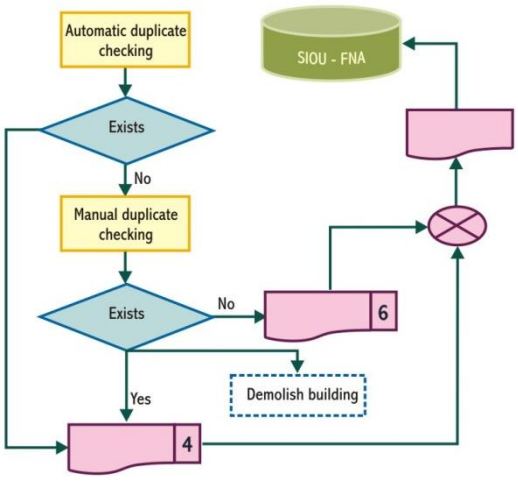



**Figure 8 – Address existence check prior to demolition**




Two variables were created to ensure that the permit is in its right path: One variable keeps track of the geography unit involvement and another keeps track of the infrastructure unit intervention. During this process these two variables are updated in each step on the basis of specific files that are produced containing the new status.

**Figure 9 – Geographical variable flow**

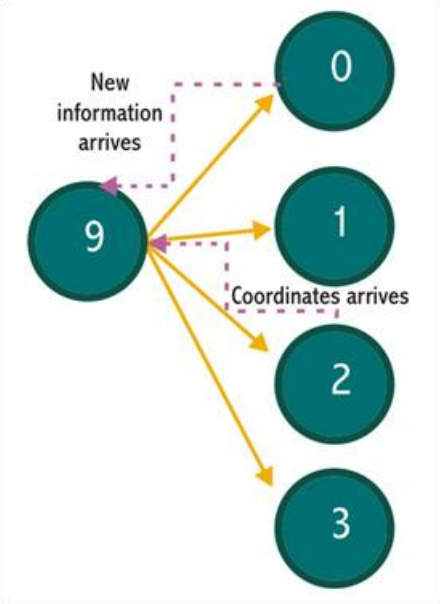



**Table 1 – Geographical variable status**

| Id | Meaning – Geographical variable |
|---|---|
| **9** | Not validated |
| **0** | Incorrect coordinate |
| **1** | Valid and correct coordinate |
| **2** | Non existing coordinate |
| **3** | Coordinate less than 5m from another |

**Figure 10 – Infrastructure department variable flow**

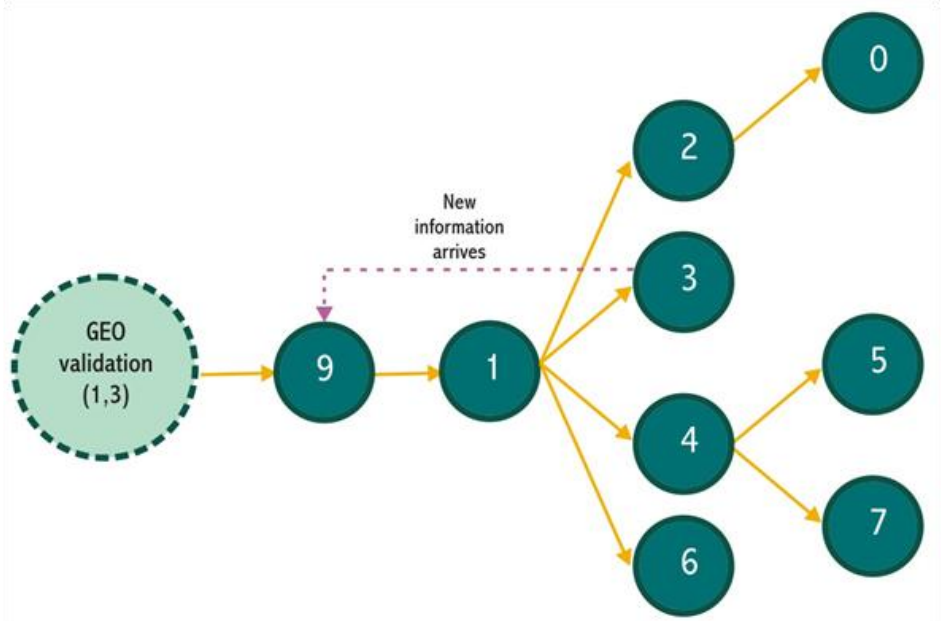



**Table 2 – Infrastructure variable status**

| Id | Meaning – Infrastructure variable |
|---|---|
| **9** | Not validated |
| **1** | Normalized address |
| **2** | Address verified and correct |
| **3** | Incomplete or incorrect address |
| **4** | Existing address |
| **5** | Marked as demolished in FNA |
| **6** | Not demolished or updated due to non existing address |
| **7** | Verified and updated in FNA |
| **0** | Inserted in FNA |

At this stage, it becomes easy to define some filters to obtain the information we need:

- Coordinate validation by the geographic unit: *Geo Validation = 9*
- Addresses to normalize: *Geo Validation* in (1,3) and *II Validation=9*
- Addresses for validation and duplication verification: *II Validation=1*
- ***Building to be inserted in FNA*****:** *II Validation=2*
- *Building to be updated in FNA*: *II Validation=4*
- ***Dwelling to be inserted in FNA***: *II Validation=0* and dwelling not inserted
- *Dwelling to be updated in FNA*: *II Validation=4* and dwelling not updated and type of work is not demolition
- Building and dwellings to be removed from FNA: *II Validation=4* and type of work is demolition

## 5. Measuring quality

The FNA dwellings address profile is in line with the INSPIRE Directive (European Union initiative to build a European infrastructure for spatial information) and was used in the 2011 Census operation. The INSPIRE address profile integrated the contributions of the EURADIN European project, a consortium of 30 participants from 16 countries for the harmonization and development of a European infrastructure addresses.
The address in FNA is defined in everyday language, as a geographic object / element that identifies the location of a dwelling. It is characterized by covering a set of alphanumeric information, organized hierarchically with an increasing level of detail.

The address plays a crucial role in FNA as it sets, so far, the only connection key to the appropriate administrative data from various external sources which will update FNA. Normalization of address, both internally in the various SO, and externally, in the various bodies of public administration, appears as a determining factor for the success of FNA's updating.

Semantic and syntactic properties are the two main areas to assess the quality of an address. Some tools are available to ensure the quality of semantics: postal code file with addresses by postal code and parish; road segment base file; and a future global address file to be implemented by Statistics Portugal assigning a unique code to each address.

A variable called Address Quality Degree (GQE) was build to ensure the syntax quality of the address field. This variable applies to addresses already inserted in FNA and to address to be inserted in FNA trough SIOU and allows for the monitoring and assessment of the

dwelling address status and for the identification of the dwellings that require a more thorough analysis of the address and eventually its correction and updating. This variable can also be used as an input variable for sample selection excluding poor quality addresses.

## 6. Anticipating problems

To anticipate future needs and keep FNA as up to date as possible, information on new construction licenses (Q3) should be incorporated in FNA marked as "Projected". As soon as data on the works conclusion arrives, its status changes to "Functional". Whenever information is provided indicating the cancelling of a license, the building's register in FNA is labelled as "Non existing" but not removed physically.

## 7. Next developments. Does anyone live here?

Solving the insertion of new data in FNA, and updating it with administrative data does not end the challenge. Some questions have yet to be answered in order to characterize every register appropriately. We must know the status of the dwelling: if it is occupied, as a primary or secondary dwelling, or if it is vacant.

The main administrative data source that we are studying is the electricity consumption of dwellings. We are currently in the stage of linking the two files on the basis of the address and establishing a threshold that could determine the dwelling occupation. Since we do not have a unique identifier for buildings, and there are legal restrictions about the exchange of personal data, an address comparison must be made using both stochastic and deterministic algorithms. This phase is crucial for the success of the operation.

## 8. Conclusion

Maintaining FNA with new data, that is current and accurate, is a key factor for the quality of social surveys. However this endeavour can be undermined if administrative data provided hasn't the needed quality. Understanding this question and to insure that all public entities keep address databases in an uniformed way, a special group at national level was created to establish norms and rules for addresses and to provide a unique building identifier. Statistics Portugal is a member of the group and having accumulated the experience on building such a registry can provide an added value to the discussion.

We have shown how through a necessity we came to develop a National Dwelling Registry that can go further than the simple support of the statistical operations. Furthermore we addressed the register feeding procedures and automatic maintenance while not discarding the quality issues that can transform address data on dwellings information. Future directions and the obstacles that they may encompass are a constant concern and classify the dwelling as to its occupation status, as already mentioned (chapter 7) is one of the main constraints we are facing. For these reasons besides all the work already developed, and still to be undertaken with electricity consumption, other administrative data sources must be studied.

## References


[1] INE/DMSI (2012), O Ficheiro Nacional de Alojamentos e os Inquéritos às Famílias – Uma nova abordagem para a obtenção de Universos, Bases de Amostragem e Amostras.

[2] INE/DMSI (2013), Modelo de atualização do Ficheiro Nacional de Alojamentos.

[3] INE/DMSI (2014), Especificações do Sistema Global de Gestão de Inquéritos – Inquéritos por Entrevista (SIGINQ-IE).

[4] INE/DMSI (2014), Regras para Normalização das Moradas.