Merging statistics and geospatial information, 2014 projects - France

Statistics Explained

This article forms part of Eurostat 's statistical report on *Merging statistics and geospatial information: 2019 edition* .

Final report 24 February 2017

Problem

In France, there was significant room to improve the management of addresses as geographical data items. The problem originated, at least in part, from the fact that municipalities (communes) are responsible for naming and numbering roads and streets at a local level (36 000 different municipalities), without any national operator to coordinate, centralise and standardise them.

Objectives

Action 1 : to assess the strengths and weaknesses of various databases that were available for addresses.

Action 2 : to develop a prototype address register to improve and supplement the existing databases within INSEE.

Action 3 : to evaluate the conditions for producing and updating — for statistical purposes — a common address register (RCA).

Method

Action 1 : municipalities (communes) are responsible for naming and numbering roads and streets and there are approximately 36 000 of these in France, without any national address operator to coordinate, centralise and/or standardise addresses. In some (often rural) areas, the addressing system is considered to be under-developed and unsatisfactory.

Many address databases exist, most notably those for: the state postal operator (La Poste); the French national mapping agency (Institut national de l'information géographique et forestière, IGN); the French tax authorities (the Direction Générale des Finances Publiques, DGFiP; which is also responsible for drawing up the cadastral map), and; the statistical office, INSEE (which has an address control list). Some local authorities also have databases for their own use covering local areas, either through their own portals or a shared public portals (based on the Etalab mission), while an association — OpenStreetMap (OSM) France — also developed a publicly available geocoded database.

Numerous non-geocoded databases, mainly of administrative origin, contain address information that may prove useful for the compilation of an address register. A major development within this context was the development of a national address database, set up by La Poste, IGN, the Etalab mission and OSM; this was launched in April 2015.

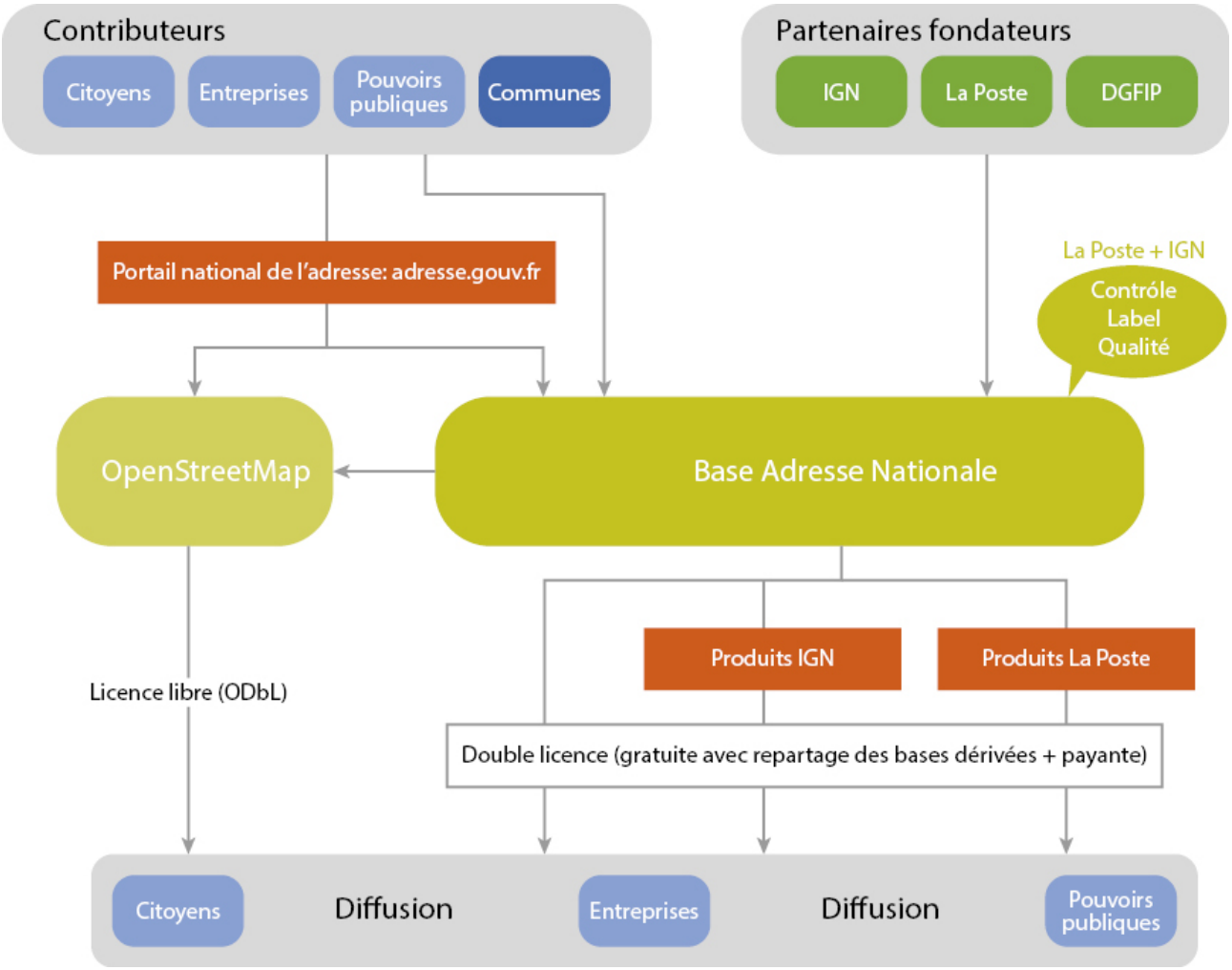

Figure 1: Simplified diagram of the target organisational structure for the BAN

Action 2 : three geocoded address databases were analysed — INSEE's address control list (ACL), the DGFiP's address database (derived from tax sources) and the IGN's address database.

There was no identifying feature that was common to these three databases that could be used to match them. Therefore, a textual comparison of the components of the addresses was required. INSEE had an in-house application that was used to perform this merging operation: it assigned a probability score for potential matches. The information was standardised: for example, the word 'boulevard' was searched for and also checked/recognised from its abbreviations, while insignificant words such as articles (for example, 'le' or 'la') were ignored. The comparison of addresses from three databases showed that:

- 76.9 % of ACL addresses were found without any ambiguity in both the IGN and DGFiP databases;
- 6.8 % of ACL addresses were found without any ambiguity in the DGFiP database only;
- 5.0 % of ACL addresses were found without any ambiguity in the IGN database only;
- 11.3 % of ACL addresses were not found in either the IGN or the DGFiP databases.

The failure to match ACL addresses with the IGN and DGFiP databases may be largely explained by difficulties in recognising the road name as part of an address (81 % of unmatched ACL addresses), while the remainder of the

difficulties were due to house number correspondence problems. Problems matching road names were often related to different types of roads (for example, drive, street, avenue), parts of the name missing (sometimes related to the maximum number of characters, which varied between databases), treatment of dates or numbers in a road name (for example, with ordinal numbers presented differently), or the use of secondary names (when a building had more than one address).

The IGN, DGFiP and INSEE databases also differed in terms of their geographic positioning: tax sources allow for geocoding in the centre of (cadastral) parcels, whereas the IGN database overwhelmingly performed its geocoding along roads, in line with its road map layer, while the INSEE database recorded the position of addresses within its own application (called CICN2); its geometry was not automatically consistent with that used by the IGN. One particular issue concerned addresses with multiple locations, particularly in rural municipalities, in which there were localities corresponding to what may be very large areas in which roads were not numbered or even named. Equally, a single geographic coordinate might correspond to several separate addresses, for example, because of a lack of accuracy.

The prototype database that was developed started by trying to match the information in INSEE's own (ACL) database and the IGN database in terms of the address and/or the geographic positioning and then the location from the IGN's database was adopted. Where this could not be done precisely enough then a match was sought between ACL and DGFiP addresses and in these cases the location of the DGFiP's database was adopted. When neither of the two databases (IGN and DGFiP) provided a sufficient match, then ACL addresses were located manually by INSEE's regional offices.

The second step for producing the prototype database was to extend the coverage beyond that of INSEE's ACL database. Addresses and geocoding for municipalities with fewer than 10 000 inhabitants were taken over from the DGFiP database. For other municipalities, roads in the DGFiP that were not in the ACL database were analysed, checking among other things for possible mismatches (to avoid adding a duplicate).

A final step in this part of the project was to test the information in the prototype register, by using it to geocode the addresses of businesses held in INSEE's statistical business register (SIRENE). In 62.4 % of cases a business address could be geocoded directly, in 7.7 % of cases its position on a (known) road was estimated based on the position of the closest known house number, in 20.6 % of cases it was located in the middle of a (known) road and in 9.3 % of cases the address was located in the centre of the municipality (if the road was unknown). Approximately 100 businesses in three very small municipalities historically not on cadastral plans could not be geocoded. For a sample of businesses, a comparison of INSEE's statistical business register was made with the IGN database, which resulted in a slightly higher frequency of geocoding based on the address, possibly reflecting more information on additional roads or a better road recognition algorithm; this higher frequency may also (potentially) be related to a higher incorrect geocoding rate but this would require more analysis to be verified. This comparison also showed that geocoding was particularly difficult for agricultural holdings, mining and quarrying units (all of which tend to be concentrated in rural environments, where roads are more frequently unnumbered or even unnamed) and for public administration entities (as these are less likely to have been geocoded coherently as there is little interest in such entities for tax purposes).

Action 3 : as part of a longer term and sustainable strategy INSEE also worked on the development of an in-house common address register (RCA), with a common infrastructure. Initially this was designed to include residential addresses in large municipalities for the purpose of collecting census information. Geocoding for 85 % of these addresses come from the IGN with the remainder based on tax sources or manual location by INSEE regional offices. It was proposed to extend this strategy to all addresses, not just those used for the population and housing census. Furthermore, a suggestion was made whereby the information used for topographic base maps should also be combined within this action.

Figure 2: Proposed organisation for INSEE's common address register

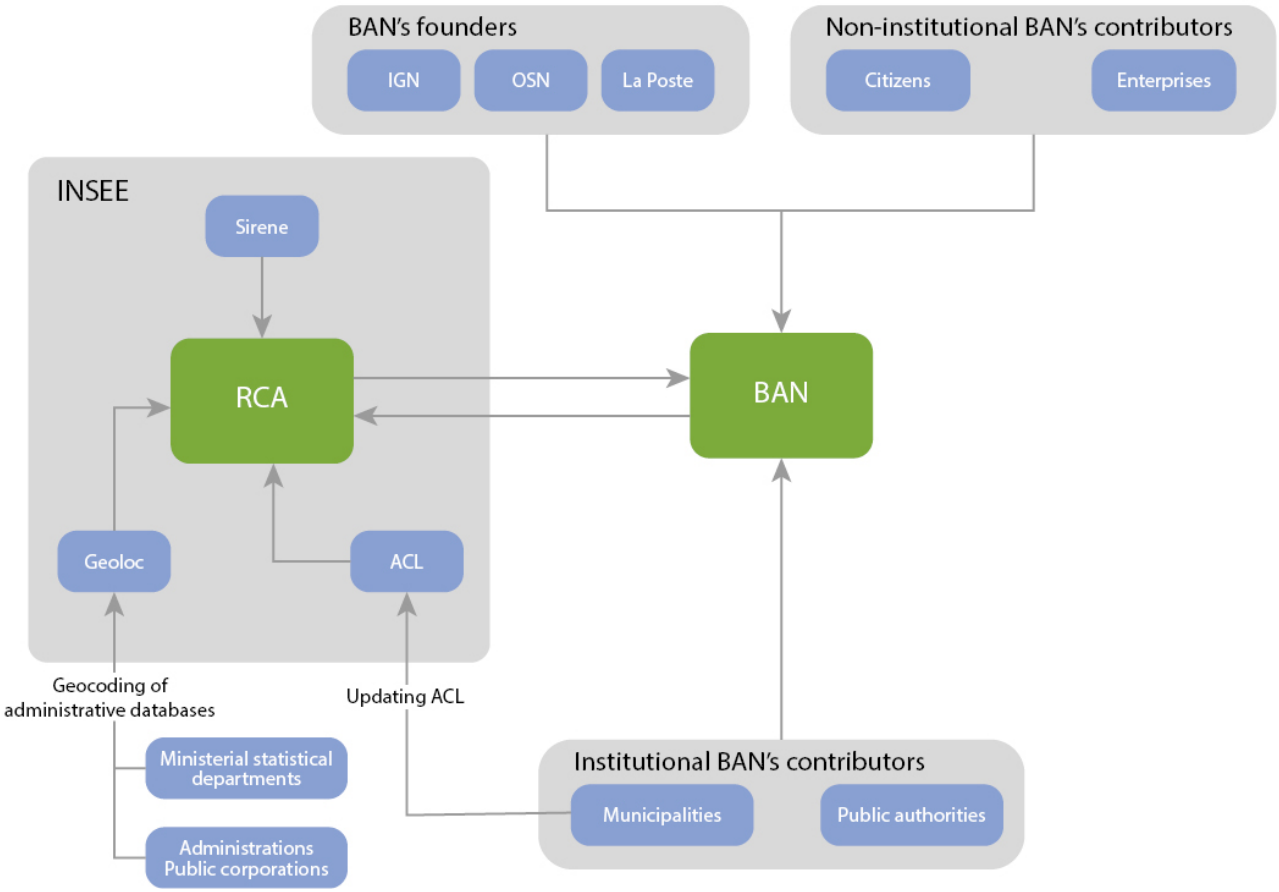

Figure 2: Proposed organisation for INSEE's common address register

INSEE initiated a questionnaire for statistical authorities within various ministries in order to identify their needs and expectations with respect to managing addresses and also to identify their involvement with the national address database (BAN) project.

Results

Action 1 : INSEE's ACL is a geocoded address database updated on an annual basis in cooperation with municipalities; however, it is limited to residential addresses in large municipalities with over 10 000 inhabitants.

The French tax authorities possess a dataset describing premises which, combined with the cadastral map, can be used to compile a geocoded address register. These addresses are positioned in the centre of parcels, which does not always allow for them to be identified by a precise location.

The national mapping agency provides free access, for administrations and institutions with a public service mission (including INSEE), to a geocoded address database whose data is mainly sourced from tax files enhanced with data provided by various other partners. Although there are regional variations in the quality of address positioning, which is especially poor in rural areas, the IGN carries out geometric processing which is capable of refining address positions while ensuring overall consistency with its other map resources.

OpenStreetMap (OSM) France launched a database (called BANO) in 2014. It includes addresses from contributors (based on a crowdsourcing principle) and from the tax authority's cadastral vector map, with a list of roads listed by tax authority and local authority in open data format. The association seeks help from its contributors to harmonise addresses and roads from these different sources. However, BANO remains an incomplete database and national coverage is uneven.

Many administrations and public services use address data on a daily basis to manage registers for their specific professional activities. These addresses are rarely geocoded at source and are more generally entered freely without any normative framework. As a consequence, substantial efforts would be required for geocoding this information.

The first comprehensive, nationwide address database in France (base adresse nationale; BAN) was officially launched in April 2015 and aims to ensure wide-ranging public involvement (of administrations, public corporations and citizens) in the creation and maintenance of a high-quality national address database. Discussions have taken place in workshops involving numerous users and potential contributors, with a particular focus on developing a data model to improve the database structure in response to different (user) needs.

Action 2 : the creation of an address register by merging various information sources requires investment in an efficient matching engine for the wording of addresses, which incorporates the common variations in address formulations.

There are discrepancies in the geographical positioning of addresses in the IGN, DGFiP and INSEE databases, which relate to the type of location method used. A non-standardised address may be difficult to geocode precisely on this basis alone if addressing efforts are not undertaken locally.

The development of an algorithm for the creation of an address register by merging different sources requires a form of calibration with a view to reducing the risks of omitting and duplicating addresses. In certain cases, monitoring the plausibility of matches among the addresses in several databases may require the inclusion of a manual validation stage. Depending on the desired level of precision, when none of the sources are able to provide a precise geographical location, positioning carried out on the ground may prove to be necessary. The initial ACL contained 5.7 million addresses, to which were added 687 thousand addresses for roads already existing in the ACL, 179 thousand addresses for new roads created in the ACL (in municipalities that already existed in the ACL) and 15.6 million addresses for small municipalities.

The initial results of a geocoding test for the statistical business register indicated that the processing carried out by IGN was more efficient, which is partly explained by the fact that its address register is more exhaustive in business sectors that are not so thoroughly referenced by the tax sources. However, even with the IGN system there remain a significant number of establishments for which precise geocoding does not exist.

Action 3 : almost all of the ministerial statistical authorities in France are interested in geolocation, mainly for the production of statistics or studies/analyses for local territories. Several authorities also expressed interest in indicators of proximity and access times (for questions of access to cultural establishments for example). In the short term, the BAN is not able to meet the main expectations of official statistics, but official statistics could consider feeding into the BAN and thus enable a greater degree of data sharing and better convergence of address systems.

Methodology

- Final report

External links

- INSEE
- Statistical office's maps
- Mapping portal

View this article online at https://ec.europa.eu/eurostat/statistics-explained/index.php/Merging_statistics_and_geospatial_information,_2014_projects_-_France