

GENERATION OF A TOPONIMICS GEODATABASE FOR ANDALUSIA IN THE ANDALUSIAN INSTITUTE OF CARTOGRAPHY (SOUTHERN SPAIN).

Torrecillas, C.⁽²⁾; Garrido, M^a T.⁽¹⁾; Risquete, B.⁽¹⁾;

(1) Instituto de Cartografía de Andalucía ICA (SPAIN)

mariat.garrido@juntadeandalucia.es; begona.risquete.ext@juntadeandalucia.es

(2) Departamento de Ingeniería Gráfica, University of Seville (SPAIN) torrecillas@us.es

Abstract.

Generating a place-name database is a difficult task. This communication describes the process used to generate a database for toponimics in Andalusia. The source has been an ARC/INFO GIS cartographic data model with 2,745 cartographic sheets at a 1:10,000 scale. We transformed this information into a place-names database with a few macros and a large number of non-automated processes. Furthermore, the Geographic Information Systems helped us to revise errors and in the final design of the geodatabase. The result is a geodatabase with more than 190,000 geo-referenced entries. These entries are going to be the core of the Andalusian Geographic Nomenclature – the next new ICA project.

1. Precedents: the main databases and toponymic works undertaken by the Andalusian Regional Government's Ministry of Public Works (Spain).

The toponymic actions undertaken by the Andalusian Regional Administration over the last twenty years (1985-2005) have been under the aegis of the Andalusian Regional Government's Ministry of Public Works (C.O.P.T.) through the different Directorates General that have held mapmaking responsibilities via the Andalusian Cartographical Institute since this latter's creation in 1993. Table 1 summarises the works undertaken and figure 1 shows the covers of some of the publications in this field.



Figure 1: C.O.P.T. Toponymic Publications.

Year(s)	Work
1985-1990	<ul style="list-style-type: none"> ▪ Creating and publishing the Andalusian Toponymic Inventory (I.T.A.) (1), drawing on 1:50,000, L-series military cartographic sources from the S.G.E.¹, registry entries of rural property at the C.G.C.C.T.², and cattle droving routes at I.C.O.N.A.³. It was published in 9 volumes and as a database (B.I.T.A.) containing over 120,000 toponimics. ▪ Creating the B.T.A.50: This was extracted from the B.I.T.A. and contained some 50,000 entries exclusively from 1:50,000 Military Cartographic sources to be used in the G.I.S⁴.
1989	<p>Publication of the Toponymic Inventory of the 1:300,000 Map of (B.T.A.300) (2).</p>
1994-1998	<ul style="list-style-type: none"> ▪ Revision work on the toponymy of the 1:10,000 Topographical Map of Andalusia (M.T.A.10) covering 476 sheets with 22,848 toponimics in the province of Cordoba and 276 sheets in the province of Cadiz. The revision was undertaken by contrasting the Map with other sources (the S.G.E.'s 1:50,000; the I.G.N.⁵'s 1:50,000 and the Land Registry's toponimics included in the B.I.T.A.). ▪ Field surveys undertaken in 16 boroughs in the province of Cordoba (1,862 toponimics) and in all of the boroughs in the province of

¹ S.G.E.: Army Cartographic Service.

² C.G.C.C.T.: Official Land Registry and Revenue Collaboration Management Centre, now known as the Official Land Registry directorate General.

³ I.C.O.N.A.: National Nature Conservation Institute.

⁴ G.I.S. : Geographic Information Systems.

⁵ I.G.N National Geographical Institute.

	Cádiz. Both the revision and field surveys of this latter were undertaken via an agreement with the University of Cádiz through its Philology Department, Faculty of Philosophy and Letters.
19 98-1999	<ul style="list-style-type: none"> ▪ Creation of the 1:100,000 Map of Andalusia (B.T.A.100)'s Toponymic Index. This was included in the publications: “Atlas de Andalusia”, Tome I (3) and the “Mapa Digital de Andalusia 1:100,000” (4). ▪ Creation of a Methodology for the geographical codification of the M.T.A.10 toponymy and the designing of a Database.
20 00	Creation of the urban toponymy of the publication: “Atlas de Andalusia”, Tome IV (5).
20 00-2003	Linguistic and geographical revision of the M.T.A.10 (Digital Vectorial v.1) toponymy for the eastern area of the region plus some Urban Conglomerations (Cordoba, Huelva, Malaga and Bay of Algeciras), obtaining the Toponymic Databases for the above Urban Conglomerations.
20 04-2006	Creating the Toponymics Database for Andalusia 1:10,000 (B.T.A.10).

Table 1: Work undertaken within the COPT with regard to toponymy.

Project Title	1:10,000 (B.T.A.10) Toponymics Database
Dates	From July 2004 – December 2006
Body Responsible for the Project	Andalusian Cartographic Institute, C.O.P.T., Andalusian Regional Government
Ambit	Andalusia
Language	Spanish
Source	1:10,000 Digital Vectorial (M.T.A.10 DV-v1) Topographical Map of Andalusia upon an S.I.G. ArcInfo Platform.

M.T.A.10 Sheets	2,745
Technology Employed	ArcGis v.9, aid programs generation in VBA.
Supports	CD-ROM
Format	Microsoft Access v. 2003
Contents	Major and minor toponymy in Andalusia classified by theme and geo-referenced by EPGS 23030, together with the province, borough and M.T.A. 10 sheet.
Objective	To create a standardised toponymic Database for Andalusian toponymy.

Table 2: Summary of B.T.A.10 Project.

2. - Generation of a Toponimics Geodatabase for Andalusia 1:10,000 (B.T.A.10).

This project began in mid-2004 and had one single objective: to generate a complete, geo-referenced and standardised database of toponimics for the whole of Andalusia using a scale of detail that would cover the whole of Andalusia – that of 1:10,000. To do this, a process was used that would be as automated as possible in order to have the database available in a short timeframe while using limited human resources. Table 2 summarises the project's characteristics.

2.1. Definition and Content.

To define the toponym database's structure one that had already been executed in 1999 was used as the point of departure, adapting it to the new requirements. The B.T.A.10 is defined by the fields and rules illustrated in table 3.

2.2. Methodology.

2.2.1. Definition of the Unit of Work and the Selection of Themes for the M.T.A.10 (D.V. v.1) Data Model.

The large number of 1:10,000-scale sheets (2,745) and the amount of data to be treated obliged us to simplify the process by working with sheets covering a greater ambit, later unifying the resulting partial databases and finally studying and revising the identifiers of each geographical entity that had a geographical location greater than the working ambit. Taking advantage of the fact that until recently the storage unit of 1:10,000 vectorial sheets in the ICA has been the ambit of a 1:50,000-scale sheet, this unit was chosen methodologically to create partial databases which would later be unified to generate provincial databases and finally a database for the whole of Andalusia.

The M.T.A.10 (D.V.v.1) has an extremely complete toponymy at all cartographic levels and thus the first methodological step was to select the information. Its Data Model comprises eleven thematic areas that contain different layers describing the entities represented, according to the vector model, by points, lines, polygons, and derived elements such as routes or regions. Independently of their graphic representation (point, line or polygon). Elements that helped to contribute information to the B.T.A.10 were selected. As the geo-referencing used is always by points, bi-dimensional elements were converted into a significant point; half a point for lines and a centroid point for polygons.

Field name	Type	Domain	Description	Rules
Identifier Fields				
id			Identifier (Ident)	Enumerate all of the register entries originating from the emptying of the MTA10 (DV v1)
				Geographical entities from MTA10 with: <ul style="list-style-type: none"> - <u>Point topology</u>: 1 geo-referenced register entry per MTA10 sheet, except in surface entities treated as point elements in MTA10, (such as places or natural areas)). - <u>Linear topology</u>: 1 geo-referenced register entry (half a point) per borough and MTA10 sheet for the same entity. - <u>Polygonal topology</u>: 1 geo-referenced register entry (centroid) per borough and MTA10 sheet for the same entity.
id_un		..∞	Identifier	Identify numerically and univocally each geographical entity
Indexing Fields				

<p>ame</p>	<p>N</p>	<p>..∞</p>	<p>Stri ng (80)</p>		<p>Geographical place name or identifier of each geographical entity</p>	<p>1. Exclusion of invalid register entries due to:</p> <ul style="list-style-type: none"> - repetition in two layers of the source's data model - their not having a proper name except for those where it has been omitted and which must be completed at a later stage (e.g. trunk roads, services or urban equipment). These are not included in the index field - their not fulfilling the "Id" field's rules. <p>2. Linguistic rules:</p> <ul style="list-style-type: none"> - nouns, adjectives, adverbs, cardinal and ordinal numbers: the first letter in capitals in initial and intermediate positions. - articles. Prepositions and conjunctions: the first letter in capitals and always in lower case when in an intermediate position. - accents according to spelling rules - variation or alternation of spellings; the name appearing on the MTA10 (Digital raster) except when it is evidently incorrect or when the alternative name/variant is documented in official sources such as the Spanish Ministry of Public Administrations' Registry of Local Entities.
-------------------	-----------------	------------	-------------------------	--	--	---

I Index		..∞	String (80)		Indexed place name	<ul style="list-style-type: none"> - The article or generic is placed after the proper or specific noun, separated by a comma (,). - Generics without a specific name are excluded. - Indexing is subordinate to geographical coding.
Thematic Field						
G Geo.-Code		..n	Ident	Geographical Code	5-digit geographical code indicating the type of entity. External key.	<ul style="list-style-type: none"> - Code representing 3 hierarchical levels with subordinate relationships. The first digit represents the first level, and two digits for both the second and third levels respectively, according to the detail established by the geographical classification employed. - Codes assigned with regard to their location on the map, the codification on the MTA10 (DV v.1) and the semantic value of the toponym's generic form.
Locating Fields						
S Sheet		..n	Ident	sheet	MTA10 sheet number	Format with a 6-digit maximum.
T Empal		..n	Ident	INE Code	Borough and province	Format with a 5-digit maximum corresponding to the Spanish National Statistics Office (I.N.E.)'s 2-digit province and 3-digit borough.

<p>X _coord</p> <p>Y _coord</p>		<p>..∞</p>	<p>l</p> <p>Rea</p>		<p>X and Y Coordinates on a UTM projection upon the ED50 Reference System , area 30 (EPSG: 23030)</p>	<p>The X and Y coordinates can, depending on the topology originating from the el MTA10 (DV v1), indicate:</p> <ul style="list-style-type: none"> - the points layer TP1's names for the toponimics (e.g. natural areas). - the entity for the rest of the toponimics originating in the points layers. - the entity in the case of linear entities. One point per route per borough and MTA10 sheet. - the entity in the case of polygonal entities. One half or centroid point per borough and MTA10 sheet.
Dating Fields						
<p>En try Date</p>		<p>..∞</p>	<p>e</p> <p>Dat</p>		<p>Date of toponym's entry into Database</p>	<p>Date format: day/month/year</p>
<p>Bt1 0 Rev. Date</p>		<p>..∞</p>	<p>e</p> <p>Dat</p>		<p>Date of BTA10 Revision or Correction</p>	<p>Date format: day/month/year</p>

Mda10		..∞	e	Dat	Date of MTA10 (DV) Revision or Correction	Date format: day/month/year
Notes Fields						
Notes	N	..∞	ng (60)	Stri	Diverse notes such as classification of corrections or modifications, etc.	Free text or standardised notes

Table 3: Fields and Rules Applied in the B.T.A. 10 (Note: C= Cardinality, O: Obligatoriness)

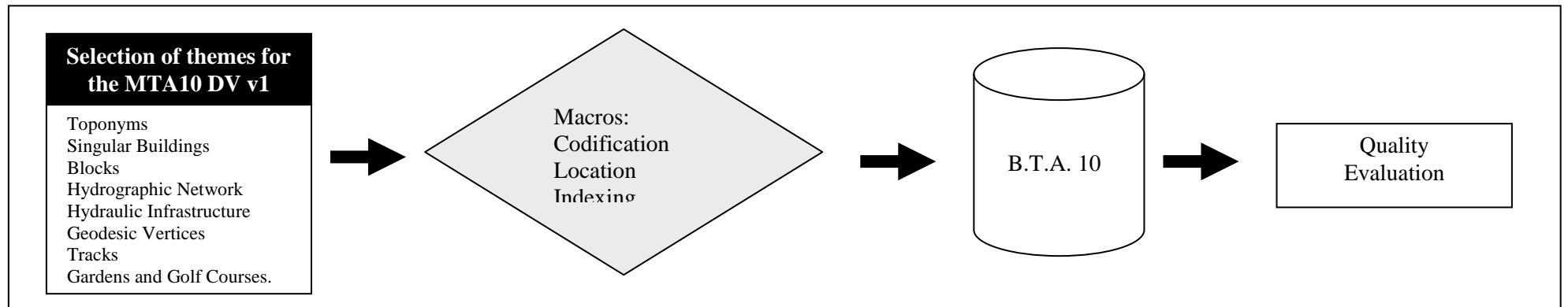


Figure n°2: Methodology used

2.2.2. The B.T.A.10's Semi-Automated Generation Process and its Subsequent Manual Revision.

An IT application has been generated in Arcgis VBA called “*Crear B.T.A.10*” [Create B.T.A.10] whose aim is to unify, to as great a degree as possible, all of the entered data to the definition in the database adopted. In order to achieve this, the programme executes the following tasks:

- Selecting the layers and, if they were originally linear elements, routes, polygons or regions – ARCINFO topologies – converting them to point topology. Converting them to a significant point consists of determining the mid-point for lines and the centroid for polygons. This conversion required individual treatments due to the peculiarities of some linear layers, such as the hydrographical layer which had previously been intersected with the boroughs layer in order to obtain a point for every river per borough.
- Unifying all of the layers of points and defining the fields that are part of the B.T.A.10.
- Eliminating selected duplicated register entries since some elements are included in the source model as polygons and lines.
- Codifying geographically or assigning the type of entity: by having recourse to the element or entity's values within the MTA 10 (DV v.1)'s field and to a list of key words created in a text file, each register/toponym is codified.
 - Indexing the toponym.
 - Entering its geo-location: X and Y coordinates and the I.N.E code⁶. Of the borough where it is located.
 - Entering the data of the register's creation.
 - Eliminating fields that are surplus to requirements and which have been used previously.

⁶ Spanish National Statistics Office.

This procedure is undertaken for each 1:10,000 sheet and then subsequently a first unification of is carried out with 16 sheets comprising the 1:50,000 ambit. This in turn is followed by a manual revision that involves verifying whether all of the names have been collected by the program and whether any names have not been codified or indexed, as well as checking to ensure that the indexing process has been performed correctly.

2.2.3. Identification of Geographical Entities.

Another application in ArcGIS has been generated in order to assign a unique numerical identifier (Id_un Field) to each geographical entity, something that is especially important in the case of linear entities (river courses) and surface entities (sierras, natural areas) with more than one register entry due to the fact that they cross more than one 1:10,000 sheet or borough. This process was undertaken first at province 1 level before requiring a short manual revision to ensure that the program was working correctly and to establish improvements in the program by including rules that had not been considered previously.

2.2.4. Unifying the Databases.

The last step is to unify all of the databases generated at province 1 level in order to generate the database at an automatic level. This unification requires the re-numbering of register entry identifiers and entity, something that is especially necessary for specific types of entity such as long river courses, large sierras and natural areas.

2.3. Evaluating B.T.A.10's Quality.

First it can be stated that the methodology is itself a form of quality control. Groups processing requires that in order to standardise, codify and index them, the data must be accurate. Should this not be the case, the entries are erroneous and they are corrected manually. Furthermore, the I.C.A. commissioned the Department of Cartographic, Geodesic and Photogrametric Engineering in the Higher Polytechnic School of the University of Jaén to perform an evaluation of the quality of its cartographic series, including the B.T.A. 10, whose results are to be found in table 4.

Internal Coherence with respect to the Data Model	Over 94%.
Process of obtaining the M.T.A.10 (D.V.)	5% of omissions and 3% of commissions.
Comparison with the Land Registry Database (B.D.C.)	Low level of coincidence (21.7%). The B.T.A. 10 is richer in toponimics, as well as being more rigorous, balanced and thorough. The B.D.C. could make a contribution of toponimics equivalent to 27% of the total.
Comparison with the M.T.N. 1:50,000	Low level of coincidence (16.4%). 15.8% of toponimics do not appear in the B.T.A.10.
Fieldwork	The evaluation had 36 informers, obtaining 271 answers (with a number of coincidences of three or more informers.) who confirmed the B.T.A. 10 toponimics. Only 19 indicated that the toponimics were incorrect and therefore the level of correctness is estimated to be about 93.8%.

Table 4: Summarised description of the BTA10's quality control.

3. Conclusions and the Next Lines of Research.

This project has suffered a large number of setbacks and to complete it, it has needed a large input of improvisation, initiative and creative problem solving as well as a great effort and a lot of patience when developing the applications. It is an unavoidable fact that "... we acquire experience over time and above all by the accumulation of analyses undertaken as the result of both successes and failures we have obtained when performing experiments and repeating processes" (6).

Some of the positive aspects of the B.T.A.10 are its economical nature, in both technical and human resources. As a result of this economy, the automatic conversion processes are rapid and they have made its execution more flexible and less time-consuming. We can also mention the simplicity of its structure – in which both the linguistic and geographical components are taken into account – making it possible to add and remove attributes in order to complete it with other documentary and oral sources. Finally, its content enables it to be adapted easily to the requirements of the Spanish Nomenclature Model (M.N.E.) (7) and it can thus be put into service for the Andalusian Spatial Data Infrastructure (IDEAndalusia) (8).

By generating the B.T.A.10, which contains some 149,500 toponimics with more than 190,000 point locations, a standardised process for the systematic collection and homogenous, geo-referenced toponymy of Andalusia – a toponymy that in many cases is in danger of being lost, has been set in motion. As the United Nations' Group of Experts in Geographical Names has observed, in an increasingly globalised world: "The benefits of such standardised names are apparent in communication... providing clarity and cost savings in commerce, planning, search and rescue, emergency preparedness, censuses, property rights, environmental management, etc." as well as in mapmaking; automatic navigation; tourism and communications.

Finally, it is an unavoidable fact that, as the new generations progressively break their ties with territory; many place names will die with the older generations that have kept them alive. Therein lies the challenge of conserving such an important cultural heritage, of disseminating it and encouraging citizens to use it by offering them the fruits of such works as the B.T.A. 10. Modest in its initial state, the B.T.A.10 had an ambitious goal: to compile, to give a new value to, and progressively to standardise Andalusia's toponymy so that it becomes widely known used through the Andalusian Autonomous Community's future geographical nomenclature.

Having concluded the project, new lines of action have been proposed, with most already in motion. Among others, such actions include:

- adapting the B.T.A.10's structure to the M.N.E.;
- performing a quality control of the toponymy obtained by contrasting it with standardised sources by entity type;
- expanding its content both in terms of sources and attributes;

- standardising progressively Andalusian toponymy with a view to creating the Andalusian Autonomous Community's future geographical nomenclature;
- raising awareness via the Internet by creating an interactive geographical names service where users could take part in correcting and standardising Andalusian place names.

4. Bibliography.

1. *Inventario de Toponimia Andaluza, Sevilla: Consejería de Obras Públicas and Transportes, 1990 (9 volúmenes)* [Inventory of Andalusian Toponymy, Seville: COPT 9 volumes].
2. *Inventario Toponímico del Mapa de Andalusia 1:300,000, Sevilla: Consejería de Obras Públicas and Transportes, 1989.* [Toponymic Inventory of the Andalusia Map, 1:30,000, Seville, C.O.P.T., 1989].
3. *Atlas de Andalusia. Tomo I. Sevilla: Consejería de Obras Públicas and Transportes, 1998.*[Atlas of Andalusia. Tome I, Seville, C.O.P.T.1998].
4. *Mapa Digital de Andalusia 1:100,000. Sevilla: Consejería de Obras Públicas and Transportes, 1999.* [Digital Map of Andalusia 1:100,000, Seville, C.O.P.T.]
5. *Atlas de Andalusia. Tomo IV. Sevilla: Consejería de Obras Públicas and Transportes, 2000.* [Atlas of Andalusia. Tome IV, Seville, C.O.P.T.2000].
6. Sebastián Más Mayoral: “*Experiencias y proyectos del I.G.N./C.N.I.G. en el campo de S.I.G.*” [Spanish National Geographical Institute / C.G.I.N Experiences and Projects in the S.I.G Field]March 2001. A presentation given in TERRITORIAL 2000, held in Pamplona from November 8th to 10th 2000.
7. *Modelo Español de Nomenclator v.1.2. (2006/10/16). SGT de M.N.E. de la Infraestructura de Datos Espaciales de España. Comisión de Geomatica del Consejo Superior Geográfico.* [Spanish Nomenclature Model (M.N.E.) v.1.2 (16/10/2006 M.N.E SGT belonging to the Spanish Spatial Data Infrastructure.

Spanish Superior Geographical Council's Geomatic Commission]
<http://www.idee.es>.

8. *Infraestructura de datos espaciales de Andalusia* [Andalusian Spatial Data Infrastructure]: <http://www.andaluciajunta.es/IDEAndalucia/>.

9. United Nations Organisation:
<http://unstats.un.org/unsd/geoinfo/default.htm>. Publicity leaflet by the United Nations' Group of Experts in Geographical Names.