From field notebooks to automatic mapping: the ‘Atlas Lingüístico Galego’ database

Xulio Sousa

Abstract

The use of computerized systems for geolinguistic data processing began in the 1970s with the production of maps for linguistic atlas projects. From that moment, dialect studies have continuously benefitted from the innovations that took place in the field of digital technologies. At present, linguistic geography projects are fully integrated within the Digital Humanities and are governed by the same principles that guide studies in this discipline: the interoperability of applications, free reuse of data and interdisciplinarity. This paper provides a brief outline of the structure and design of the database that currently houses the Atlas Lingüístico Galego materials, a linguistic atlas of Galician begun as a project in the 1970s. This database is being used as an information source for the publication of the project’s volumes, for research into variation and change of Galician varieties, and to contribute to the elaboration of geolinguistic projects that are greater in scope.

Keywords

Digital Humanities, Dialectology, Geolinguistics, Linguistic Atlas, Linguistic Database, automatic mapping, Geographic Information Systems

1 Background

In monographs focussing on the Digital Humanities, geolinguistics and dialectal studies in general tends to merit only some brief lines in those chapters that address the employment of geographical information systems or in which database design for linguistic documentation is discussed. The latest edition of Companion to Digital Humanities, recently released (Schreibman et al. 2016), features an article by Todd Presner and David Sheppard, which discusses the relevance that the use of geographical information systems is providing for studies in the humanities (visualization, interpretation and linguistic data analysis). In this chapter, reference is made to how geolinguistics has benefitted from these new applications. The authors establish the first use of digital technologies as being in a geolinguistics project in 1983, with the computerised printing of some of the maps of the Atlas Linguarum Europae (ALE) under the supervision of Mario Alinei (2008). In a work that revises the methodology of traditional and modern

If this last reference were considered correct, it becomes clear that dialectology, as a discipline focused on the study of linguistic variation, included the use of computers in the analysis and processing of data shortly after the rest of the disciplines that are now part of the area of study that is usually referred to as Digital Humanities. Geolinguistics, as sub-discipline of dialect studies tasked with representing and analysing geolinguistic data, was directly favoured by the revival embodied by the use of computer systems for dialect data processing and its significance for philological and linguistic studies. The application of quantitative analysis methods in dialect studies acted as a new incentive for the incorporation of computerized data processing systems (Séguy 1971; Goebl 1989; Nerbonne & Heeringa 2010; Dubert & Sousa 2016). From the 1970s to the present, progress in the application of computerized systems to geolinguistics has profoundly changed the interpretative and display methods of dialectal information (Hoch & Hayes 2010).

The changes that have taken place in the field of linguistic studies with the incorporation of information technology and methodological innovations have prompted the modification of geolinguistic information usage and preservation techniques (Thieberger & Berez 2012). Linguistic atlases are projects that require data storage and management that is extremely heterogeneous in nature (text documents, drawings, photos, sound files, video files, etc.). This information should be structured so as to ensure its preservation and facilitate consultation and use. The data must be made available to the scientific community in formats that expedite processing and analysis.

In this article, I intend to give an account of the main features of the *Atlas Lingüístico Galego*, computerization project, and primarily of those aspects that refer to the database design and the process of automatic generation of cartographic representations. I shall focus on the description of the structure of the current database, and more modernized adaptation of the complete original database (Sousa 2004), as well as the presentation of the direction to be taken by the project in forthcoming years. Section 2 provides information about the history and current status of the project. Section 3 presents the materials that comprise the ALGa documentary archive. The structure and characteristics of the applications currently used today for the production of printed volumes are set out in 4 and 5. Finally, the sixth section proposes lines of work that will be developed in the future.
2 The ‘Atlas Lingüístico Galego’

In the same decade in which computers began to be used in geolinguistics projects, fieldwork began for the development of a linguistic atlas for Galician. In 1974 the Instituto da Lingua Galega (ILG), a research centre at the University of Santiago de Compostela, launched the *Atlas Lingüístico Galego* project with the aim of documenting the linguistic variation of the Galician language. The legacy of this project are the materials that are stored today at the ILG, and which include survey notebooks, recordings and a good number of written material with ethnographic and linguistic information concerning 167 locations in the Galician linguistic domain. All the information collected by the researchers who participated in the survey is the primary source for much of the knowledge we possess today regarding rural Galician over the last quarter of a century. The project began during a period in which the study of linguistic variation was already in crisis because of the revolution in methodology and principles that began with the development of Variationist Sociolinguistics. Despite all the objections towards one-dimensional linguistic atlases, the materials provided by such projects continue to be a fundamental source for the knowledge of linguistic variation today.

ALGa is the first linguistic atlas dedicated to the exclusive study of the Galician linguistic domain through the application of geolinguistic research methods. Galician geolinguistic varieties have been the subject of study in the *Atlas Lingüístico de la Península Ibérica* (ALPI), an atlas project focussing on peninsular Romances varieties which began in the first third of the twentieth century under the tutelage of Ramón Menéndez Pidal and the supervision of Tomas Navarro Tomas (García Mouton 2012). The ALPI was envisaged as continuing in the wake of the first projects of European geolinguistics and in particular the *Atlas Linguistique de la France* (ALF), by Gilliéron, and the *Sprach- und Sachatlas Italiens und der Südschweiz* (AIS), by Jaberg and Jud. For the atlas of Peninsular Romance varieties, information was collected in 53 Galician locations between 1934 and 1935 (Sousa 2008). Although some Galician points were also included in subsequent geolinguistic projects (Alvar 1974), it was not until seventy years after the start of the ALPI surveys that fieldwork began for the preparation of an atlas which would was going to take up the study of spatial variation across the territory regarded as being Galician-speaking. The area investigated for the ALGa comprises localities in Galicia and in the administrative territories of neighbouring Asturias, León and Zamora. The number of points is considered far superior in the ALPI for the same territory, therefore the results provide much more detail regarding Galician linguistic variation.

According to the researchers on the project supervised by Constantino García and Antón Santamarina, a linguistic atlas of Galician varieties was undertaken for various purposes: i) to study diatopic variation at different levels of analysis
(phonetics, morphology, syntax, vocabulary); ii) to develop a dictionary of spoken Galician; iii) to bear witness to the influence of social factors on variation (age, class, occupation, academic training, etc.); and iv) to identify the direction of language changes (Garcia et al. 1977: 9–10).

Despite these ambitious claims and that the traditional methods of dialectology had already found themselves affected by the crisis initiated with the first sociolinguistic studies, the ALGa is a linguistic atlas conceived according to the principles of what Britain calls the “pre-sociolinguistic era” (Britain 2014). The variety documented in the project is that pertaining to speakers of rural areas, adults, with stable residence in the survey location and male. In addition, the selection of places surveyed excluded the most populated cities and towns. The information obtained would serve to draw a picture of the spatial variation in Galician linguistic territory, but not to be the starting point for research that took into account other variables associated with social factors (gender, age, occupation, etc.), as the researchers initially claimed. In the introduction to the project’s final two published volumes, the researchers are much more cautious and honest when pointing out the project objectives and the output that can be extracted from the analysis of the materials compiled: “the ALGa seeks to take a close look at the study of the characteristics of modern Galician, and also to offer a vision of reality that allows comparison with other Romance and even non-Romance domains” (ALGa 2015: 11). Furthermore, the information contained in the ALGa notebooks would contribute to our knowledge about the spoken Galician that would be the basis of the standard variety that became widespread from the 1980s. The materials corresponding to the sections of the questionnaire dedicated to grammatical variables were essential for understanding the vitality, distribution, spread and use of the forms proposed for the standard variety (Santamarina 1988; Ramallo & Rei-Doval 2015).

Therefore, the ALGa was developed as a small domain atlas that sought to deepen linguistic knowledge regarding rural Galician in the mid-1970s. The questionnaire designed for this task was modified for the context that would be studied, both in aspects directly related to the structure and characteristics of Galician and linked to Galician cultural reality. The main innovations regarding the foundational works of dialect studies have occurred in the use of the tape recorder for recording oral witnesses and the use of a camera to record images of objects of ethnographic interest.

Until now, the ALGa has been the basic source for Galician dialectology research. Analyses of Galician linguistic variation and proposals for dialectal division have so far been based on the materials collected as part of the ALGa project and therefore reflect rural Galician spoken in Galicia in the 1970s. These materials have been the basis for most studies of Galician dialectology conducted over
the past three decades, whether these have been the interpretation of simple answer maps like those of Álvarez Blanco (1983), González González (1984) and Dubert & Sousa (2002), or those based on aggregate data analysis (the most classic examples being Santamarina 1982 and Fernández Rei 1985; 1990), and those applying more sophisticated quantitative procedures, such as Álvarez et al. (2006), Dubert (2011) and Sousa (2006). In recent years, ALGa has also served as the basis for conducting analytical studies of linguistic change in real time, such as those carried out by González et al. (2002), Louredo Rodríguez (2014), Rodríguez Lorenzo (2012) and Sousa (2010; 2012).

For both the undertaking of some of that research especially that developed in the 1980s and 90s, and the editing of the first volumes published by ALGa researchers, data was collected individually by obtaining access to books used in the original field research or printed maps from the volumes that were being published. Data collection for this procedure is very laborious as it requires the investment of considerable time and extreme care in handling the original documents. The impossibility to access data in a simpler manner and the lack of files in digital format restricted the ability to perform a quantitative analysis of ALGa for decades and even hindered the edition of the first volumes.

The first ALGa volume, published in 1990, was produced in an almost handmade way, passing information manually from notebooks to maps and pages of notes (ALGa 1990; Sousa 2004). The second volume was prepared using a graphic design program for the cartographic representation of the results, thus improving the appearance of the maps and lessening the correction work. During the drafting of the third volume (ALGa 1999), the project team had the opportunity to meet other groups working on linguistic atlases and began to reflect on the need to computerize the development process of the volumes. On the one hand, it was necessary to automate the process of map making in order to make it manageable, but an urgent task was also to seek a new medium to all project materials in order to ensure their conservation and enable a better use of the dialectal information.

3 From field notebooks to the database

The first attempt to carry out the task of computerizing the ALGa materials took place in the late 1990s. At that time, several linguistic atlas computerization projects had already been launched and therefore there were models available for dialectal database design and for setting up a computer-aided mapping system (Bauer & Goebl 2000; Kretzschmar 2001; Aurrekoetxea 2008; Olariu & Olariu 2014; Kumagai 2016). First it was necessary to transfer the material from ALGa in a format that would guarantee conservation and enable usage through different types of applications (geographic information systems, web services, networking applications quantitative analysis, etc.). The materials collected during dialectal fieldwork are usually heterogeneous in nature and require a different approach.
The documentation registered for ALGa consists of three types of materials: field notebooks, sound files and written documentation.

3.1 Field notebooks

The field notebooks are the paper questionnaires in which the answers of the respondents in each of the locations researched were written down. For each of the 167 points that form the ALGa network, there is a 93-page book with 2712 questions and the phonetic transcription of the corresponding answers (Fig. 2). The questionnaire is divided into sections according to the type of linguistic information sought: phonetics (1–148), morphology (149–386), syntax (387–526) and lexicon (527–2712). The answers were transcribed using a phonetic alphabet originally developed for peninsular varieties by Tomas Navarro Tomas (1966) and used in most Iberian geolinguistic developed during the twentieth century. On the even pages there is complementary information regarding the answers noted on odd pages: explanations, sayings, indications as to the use and the validity of the answers, drawings etc.

The first task to be performed in order to ensure the conservation of the notebooks and ease of reference was scanning. A high quality PDF file was created for each of the field notebooks. The files were identified by the code assigned to each of the locations surveyed (a letter identifying province followed by a number).
1870  
cuna, berco  
mece, arrojar  
andaderas de los niños, andadeiras  
cama, cama  
partes da cama  
colchón, colchón  
xargón (colchón de foliado)  
almohada, cabozal  
almohadón (funda de la almohada)  
sábana, saba  
1880  
manta, mantas, roceador  
colcha  
despertarse, acordar  
ventana pequeña, ventá pequena  
contraventana, contra  
1885  
encalar, blanquear  
hogar, fogar (lugar donde se aldea o lume)  
lumbre, lume  
cerilla, misto  
encender el fuego, prendé lo lume  
1890  
enecer la luz, prendé la luz  
alar, alumear  
arder, ardér  
aceite, fumo  
apagar, matar  
soplillo, abanico  
chaspa, mexica  
pavesa  
rescoldo  
ceniza, cinza  
1900  
cinza pra abonar, borralla

Fig. 2. Page 68 of the notebook for A Gudiña (O-21).
3.2 Audio files

The ALGa project basic information was collected on paper questionnaires. Additionally, the researchers conducted audio recordings in order to take samples of contemporary speech. Most of these recordings do not contain answers to the questionnaire, but popular narratives, conversations, stories and even some example of popular music being sung. The original intention of the project researchers was to carry out recordings in all locations selected. The political situation that Spain was undergoing at the time (the end of a dictatorship) created distrust in many informants, therefore recordings could only be made in a small number of the localities surveyed. The audio files of the project were scanned in the middle of 2000 and constituted the first collection that was part of the Arquivo do Galego Oral project (Fernández Rei 2008).

3.3 Visual data

In addition to the field notebooks and sound recordings, and to accompany the questions of an ethnographical nature, the ALGa project also has an important collection of drawings and photos. The drawings are contained in field notebooks and were made by the same researchers (Fig. 3). A collection of photographs contains a total of almost a thousand images, most of them from farm implements and objects associated with traditional professions. This graphic material is scanned and each file is identified by reference to a location and the corresponding question on the questionnaire.

Fig. 3. Drawing linked to question 1003 angarela ‘stretcher’, Ribadulla notebook.

3.4 Computerisation of materials

As previously mentioned, the process of computerization and digitization for the ALGa project was first undertaken with the intention of preserving the original materials sourced from research and facilitating the editing of the expected volumes. The digitalisation of the survey notebooks and supplementary material (pictures and sound recordings) guarantees their preservation and also enables faster and easier access to materials. However, for the employment of the information contained in the notebooks, the development of a database that facilitated
the consultation of the materials, the preparation of the map volumes and other
types of research to be permitted (quantitative analysis, comparisons with mate-
rials collected at other times, advanced spatial analysis, etc.) was crucial. Fur-
thermore, the database had to be designed so that it would simplify the task of
transcribing all the information contained in the notebooks.

The computerization of the ALGa was a task with certain special features,
given that it did not have the characteristics of other computerized geolinguistic
projects undertaken in the same period. On the one hand, it was not a project
designed from its inception as a digital atlas; on the other, neither was it a finished
project whose materials had been edited and for which a new way of consultation,
dissemination and exploitation was sought. The fundamental research materials
had already been collected, the atlas edition had already been started and there-
fore the first intention during database design had to be that of facilitating the
publication of the planned volumes. In 1996, I presented the structure and basic
features of the first version of the ALGa database (Sousa 2004). Later technolog-
cal advances, the increase in size of the database (currently with nearly a million
records), the completion of transcription work, changes in the software used for
cartographic visualisation and new objectives established for the project led to a
series of upgrades and modifications in the design and structure of the database
being carried out. This new version of the database is presented below.

4 The ALGa database

In order to enable the employment of dialect data from different perspectives
(searches, index compilation, variation analysis, results visualization, diachronic
data analysis, etc.), the first step was to organize and structure the information so
that each linguistically registered form could be characterized with respect to all
the parameters considered in the original investigation. Despite attempts made in
recent years to establish a catalogue of classes and properties for data from geo-
linguistic projects, we still do not have a standardized ontology that has been put
into practice and evaluated. The proposal submitted for *Atlante Sintattico d’Italia*
(ASIt) by Bucci is one of the most detailed and complete, but has still not even
been implemented for the project for which it was designed (Bucci et al. 2014).
This shortcoming and the heterogeneity of these same geolinguistic projects have
been used to justify the difficulty of reusing data and the limited presence of this
type of projects in open repositories of linguistic data (LLOD).

Using the aforementioned proposal for classifications of geolinguistic mate-
rials (Buccio et al. 2014) as a guide, it is possible to suggest a basic organization
of information in three main areas: i) geographic location data; ii) data on the
informer subjects; and iii) linguistic data documentation. In a way, this distinction
was already the foundation of the ALGa database’s first structure (Sousa 2004).
Class i) corresponds to the table named *Puntos*, and class iii) consists of the contents of the *Tipos*, *Respostas* and *Cuestionario* (Fig. 4). The content of the current database is already organized on the basis of a distribution of information in these three areas: towns, locations and documentation (Fig. 5).

4.1  Locations

The *Location* class consists of information about localities that make up the ALGa network: post code, place name, parish, county, region, diocese, province, autonomous region and geographical coordinates of the location. This is the data that identifies each of the questionnaires, since each corresponds to one of the localities surveyed. This information allows each of the database answers to be geographically identified and also all the existing linguistic information for each of the survey points to be known.

4.2  Speakers

In the *Speakers* class, information on each of the subjects who provided the answers is provided. For each of the informants, name, age, sex, place of birth, place of habitual residence and other information of socio-economic interest (occupation, studies, other places of residence, etc.) is indicated. Although all informants
are associated with the survey location, the total number of respondents is more than 167, since in some points several people were interviewed.

4.3 Linguistic documentation

Strictly linguistic database information is organized in the class called **Documentation**. Here the information from field notebooks and also the non-textual content to which I have referred earlier (drawings, photos and sound files) is grouped together. Given its nature, it was considered convenient to distribute this data in two tables: a) questionnaire; and b) answers.

4.3.1 Questionnaire

The basic information contained in field notebooks: question statements, question number, sections (phonetics, morphology, syntax and vocabulary), page number for the question and the code that corresponds to the questionnaire, is registered in the Questionnaire table. Information is included that is not contained in the original questionnaires in the lexical category of questions (provided that these were of a simple format) and the semantic field to which the concept queried pertains. This table has a number of registers greater than the number of original questionnaire questions. The reason behind this is that a space had to be created for answers associated with questions of a complex character. For example, on the basis of data taken from question 2711, **Cubic measures**, new question registers were created for the answers concerning the various liquids referenced in the answers (2711.1 for water, 2711.2 for wine, 2711.3 for oil, etc.).

4.3.2 Answers

The table that stores all the documentation offered by informants is called **Answers**. This is the main database table and is linked to the rest of the tables that comprise it (Locations, Speakers and Questionnaire). In **Answers**, the transcript of the answers offered by informants in each of the surveyed points is stored. At present, each of the answers are transcribed with the symbols of the International Phonetic Alphabet (IPA) and a transcript with conventional spelling representing the most relevant phonetic and morphological features. For questions from the “Lexicon” section of the questionnaire, a transcript in conventional orthography is provided and is used to create the legend that accompanies each map in the published volumes. In this transcript, the representation of phonetic and morphological variations considered irrelevant is avoided. This table has several fields in which additional information recorded by investigators in the field notebooks (notes on the meaning, notes on usage, related information, etc.) can be transcribed. The table also contains references to the existence of non-textual information (images and recordings). Each one of the registers in the **Answers** refers
to a question from the questionnaire and is associated with the rest of the database tables.

As often occurs in geolinguistic projects based on the use of questionnaires, the ALGa researchers collected at certain times more than one answer to questions from the notebook. Multiple answers are quite frequent in the section dedicated to lexicon. Sometimes the answers are really lexical synonyms, but others are forms of related meaning or relate to different levels of employment (disused words, unusual forms, etc.). In order that this information could be represented on maps, a supplementary table was created in which each of the different questions are identified with a letter. It was believed that the map could represent a maximum of five answers associated with each point (A, B, C and D); in the event that more answers were collected, these are recorded in the “Notes” field.

<table>
<thead>
<tr>
<th>Location</th>
<th>Speakers</th>
<th>Documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locality</td>
<td>First name and surnames</td>
<td>Audio files</td>
</tr>
<tr>
<td>Parroquia</td>
<td>Sex</td>
<td>Graphic files</td>
</tr>
<tr>
<td>Municipality</td>
<td>Date of birth</td>
<td>Answers</td>
</tr>
<tr>
<td>Comarca</td>
<td>Professional category</td>
<td>Questionnaire</td>
</tr>
<tr>
<td>Province</td>
<td>Place of birth</td>
<td></td>
</tr>
<tr>
<td>Locality Code</td>
<td>Locality Code</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 5. Schematic structure of the current ALGa database.

4.4 Using the database

The ALGa project’s current database and the mapping display system are designed with commercial software. The database uses Microsoft SQL Server 2000® and its access interfaces were designed using the Access Data Project (ADP). ADP is a file format for Microsoft Access that contains information about the project, and enables database objects to be connected and consultation and editing forms to be designed. Fig. 5 shows the components and the fundamental relationship of the tables and objects that constitute the database. The index used to relate the different tables is the code assigned to each of the locations surveyed.
Access to the database is realized by means of an interface (ADP) that allows the three basic tasks that an operator can perform to be selected: editing of materials, querying and mapping.

Fig. 6. Editing application; question 1882 *espertar* ‘to awake’, point O-21 (A Gudiña).

The appearance of the editing application is presented in Fig. 6. This interface is used to enter the information on the basis of notebooks and to make corrections regarding the information transcribed. With the selection of a question from the questionnaire, the operator has access to all information related to the selected answer and the image of the questionnaire page on which the information is found. At the top of the window, data appears identifying the edited question (question number, the text of the question on the questionnaire and the identifying name of the question). In the central part is the page image from the original questionnaire where the question is found. This image always pertains to the question and location selected. At the bottom of the window appears the list of survey points (the left box), where the location on which to work is selected. In the central square, larger in size, fields are located that collect the different ways of transcribing the answers: an alphabetic code that identifies the answers (type), phonetic transcription (IPA Answer) transcription in conventional orthography (Orto Answer), transcript to be included in the printed map (Legend) and notes made by researchers. To the left of the window are two boxes that provide information about forms registered as answers: number of phonetic and orthographic variants, and variants of these relationships. For the question used as an example (Fig. 6) 28 phonetic variants were collected and 8 in conventional orthography.
The information contained in the database can also be accessed through a search interface. Through consultations, information can be obtained on all answers and additional information recorded in notebooks. It is possible to perform a search on the basis of the question number, locality code, a given semantic field, a string of sounds or graphemes, etc. Lists of variants associated with the same question indicating the frequency with which they appear in the database and calculations on the occurrence of certain linguistic forms (sounds, strings of sounds, morphemes, words and complex constructions) can also be obtained in a straightforward manner. The connections between the different tables that constitute the database enable the relationship shared by the points where form or forms sought were recorded may be established in all cases. Fig. 7 presents as an example the results table for the form *espertar* ‘to awake’: locations where it was registered (Locality code), questionnaire questions where the answer appears (Question number) and associated transcripts (IPA, Orto and Legend).

Fig. 7. Table of search results for variant *espertar* ‘to awake’.

The design of the current ALGa database was used as a model to develop the web application project for the preparation and publishing of the ALPI materials. This application enables online editing work to be performed and in the future allow consultations to be made regarding the materials (García Mouton 2016).
5 From database to mapping

The design and features of the ALGa database enable developed cartographic visualizations of the results to be produced in a simple manner. The database is connected with a GIS programme that immediately creates thematic maps. The connection between the data base and GIS enables, through the selection of a question number from the program outline, a map and legend with relevant symbols to be generated immediately (Fig. 8; the selection square with all questionnaire questions appears at the top left of the image).

For the production of maps for the ALGa volumes that are currently being edited, the ArcMap tool is used from the ArcGIS Desktop package, created by ESRI®. A first map is created automatically by selecting a question from the questionnaire and via the same application site. On this map, symbols are arbitrarily assigned (one colour for each different variant; Fig. 9). The editor must choose the shape and colour of the symbols that allow a better understanding of the geographical distribution of the variants associated with the variable that is represented (phonetic, morphological, lexical, syntactic, etc.; Fig. 10 shows an example of a map...
for a lexical variable that has been edited and ready for publication). The link between the database and the GIS allows more complex displays to be obtained: data representation from different sources, frequency analyses of segments, combination of diverse types of data (linguistic, economic, geographical, historical etc.).

The application for using the database and the mapping software are employed solely for internal use within the ALGa research team for the development and editing of hard copy volumes. Some of geolinguistic projects implemented in recent decades used cartographic display applications designed specifically or for use with web applications that allow simple format representations.

The use of GIS software, as used for the ALGa project, offers several advantages with respect to any of these other options. The GIS programs use standard formats which guarantee interoperability of the information processed and the continuous renewal of applications. Moreover, this type of software allows maps to be exported in formats that facilitate the work of editing the final results.
In addition, currently there are free code applications that have the utilities required in order to allow a not very experienced user to produce linguistic displays without difficulty and for free.

![Map](image)

**Fig. 10.** Map edited for publication with the results of question 1445 *teixugo, porco teixo* ‘badger’.

## 6 Conclusion and future work

In 1974, the first fieldwork began for the collection of the materials that would allow the first rigorous and reliable description of modern Galician’s geolinguistic varieties to be drawn up. The work of documenting, analysing and editing these valuable materials began at the same time as computer technologies began to be applied to dialectal studies. These innovations are now being used for the edition of the 12 volumes that will comprise the complete work.

The development of the ALGa database seeks to preserve the original materials from the research and advances in the employment of this valuable infor-
mation. The project co-ordinators continue to work in order to enlarge and com-
plete the database with additions that perfection and help disseminate the results:
the inclusion of visual fields for drawings and photos, the introduction of new
fields with information from the analysis of the answers (morphological segment-
tation, detailed semantic characterization, the linking of related answers, etc.),
incorporating information from other sources (etymology, synonyms, bibli-
ographical references, etc.), etc.

Amongst the tasks that are intended to be undertake in the future is also the
updating of the design and database components. This work is guided by two fun-
damental purposes: i) the use of free code applications, both in the database
(MySQL, MariaDB, PostgreSQL) and in the mapping application (QGIS and
OpenLayers); and ii) the open availability of all information related to the project.
For the latter objective we are designing a web application that allows the consult-
tation of information, its cartographic representation and also free data access for
the user who is interested in studying Galician geolinguistic variation. In the aca-
demic field, there are numerous geolinguistic projects that allow information to be
consulted and to download data in formats compatible with several analytical ap-
plications. In this area, a remarkable project, both for its design and ease of access
to all information, is the World Atlas of Language Structures (Dryer & Haspel-
math 2013). The ALGa project must participate in the current trend in the field of
academic research that promotes the development of infrastructures that ensure
the conservation of data, permit free access to information and make resources
available that enable the straightforward employment of materials (Sousa 2016).

Linguistic atlases, and geolinguistic projects in general, are projects that are
financed almost exclusively with public funds in all the stages of their develop-
ment. Therefore, it is the obligation of researchers to make all the information
collected in these projects available to the scientific community. Furthermore, all
those of us who are involved in geolinguistic research recognize that the results
of the linguistic atlases development projects rarely achieve the dissemination
deserved. At best the volumes of atlases published in hard copy eventually end
up being difficult to access in a few specialized university libraries. Indeed, lin-
guistic atlases published in hard copy are works that are difficult to handle and
in which is not always easy to find information.

Geolinguistic studies should avail of all means made available by technology
today (Kretzschmar 2013). Disseminating knowledge regarding variation in lan-
guages and facilitating access to information so that it can be explored from var-
ious perspectives can only redound to the benefit of the projects in which we are
involved.
Acknowledgements

This work was funded in part with the help of the Ministry of Economy and Competitiveness of the Spanish Government (project FFI2015-65208-P), the Galician Government and the European Union (GRC2013/40). I would like to express my gratitude to César Osorio for his crucial assistance in the design and development of the database and the automatic map preparation system. The Atlas Lingüístico Galego was developed at the Instituto da Linguaxe Galega at the Universidade de Santiago de Compostela thanks to the continuing contribution of the Galician Government (Dirección Xeral de Política Lingüística da Consellería de Cultura, Educación e Ordenación Universitaria) and the Fundación Barrié.

References


LLOD. *Linguistic Linked Open Data*. http://linguistic–lod.org/


