NTTS New Techniques and Technologies for Statistics 2017

Brussels, 14 to 16 March 2017

IDESCAT'S NEW STATISTICAL DISSEMINATION MODEL

Authors:

Josep Sort i Ticó, Deputy Director General of Information and Communication

Josep Jiménez Casanellas, Manager of the Department of Communication and Dissemination

Estela Tonzán Orio, Manager of the Department of Information Technology

ABSTRACT

INTRODUCTION

The Statistical Institute of Catalonia (hereinafter Idescat) was created in 1989. In its 26 years of existence, the Institute has seen how information systems have rapidly and radically changed and developed. Idescat's production and dissemination of statistics began in an environment still dominated by paper and, although current-day technologies, the data available and users' habits are light years ahead of those early days, it is necessary to evolve in order to address the new challenges of the information society.

In order to update its processes, in 2012 the Idescat management commissioned a master plan for its information systems in order to allow the introduction of a new model for the Institute's production and dissemination of statistics.

The existing dissemination model was based upon the generation of databases and applications organised by end products (derived from the stovepipe production model). This organization, which proved highly useful in the past, has become an unsustainable model which is unable to adapt to the challenges posed by the technological and information revolution.

For the users, the system offered statistical data which had been previously formatted in calculated and pre-defined tables. Information crossovers were limited to those the technicians had foreseen and users had to deal with different interfaces. Moreover, the Idescat website itself needed updating, not only in terms of usability, owing to the organization of its contents and the form of browsing, but also at a technological level, as users are tending more and more to consume statistical information via mobile devices, which means that the data have to automatically adapt to all kinds of formats.

The project which stemmed from the master plan launched in 2015 was named the **Cerdà Platform**, a technological platform for managing and hosting an integrated statistical information system which would change the traditional model of statistical production and overcome the challenges presented by the new information society of the 21st century.

This technological project, in partnership with the Qualitas Project, which adapted the GSBPM model to Idescat, represented a major challenge with regard to the modernization of the statistical production (a subject not covered by this paper), but also an excellent opportunity to renew the dissemination model.

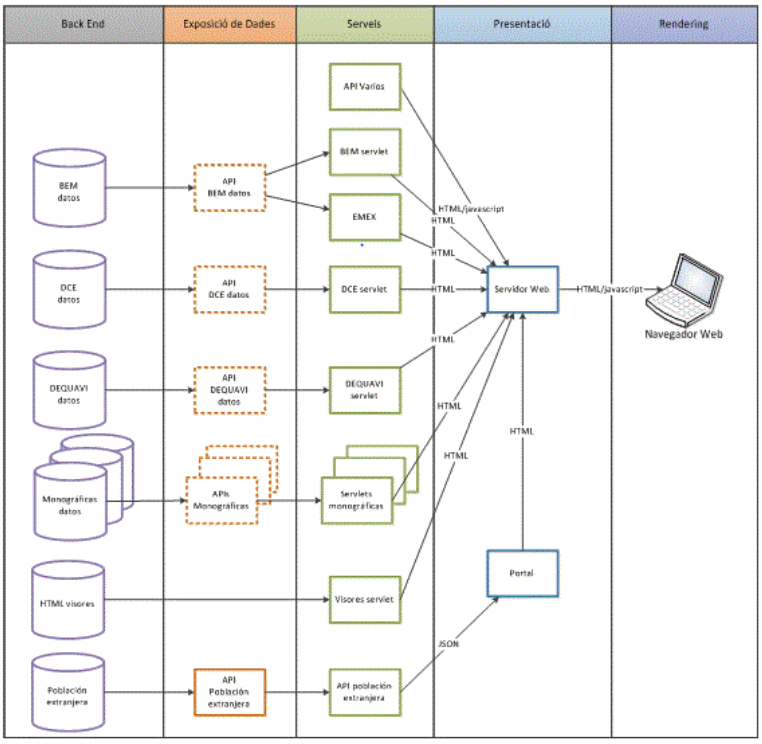
But migrating all the statistical information produced since the year Idescat was created to the new Cerdà Platform was conceived as a mid- and long-term project. Another challenge accompanied that of the change: how to align the new system with the existing one without this entailing additional difficulties for end users.

The solution finally chosen was not solely technological. The project has led to a major reorganization of the structure and organization of the statistical information produced and stored by Idescat, affecting the Institute's databases, applications and website.

The new dissemination model envisages a simplified data structure (a single dissemination database, a single viewer, a single thematic organization) more efficient dissemination management (centralised, flexible and automated) and more efficient data accessibility (different channels, depending on the user typologies, and a website adaptable to different devices and open data).

This paper indicates the starting point and the strategy followed and also lists the different projects Idescat has had to undertake in order to change its dissemination management model.

Idescat's information system prior to the Cerdà Platform



THE NEW MODEL

The point of reference taken for the new dissemination model to be implemented by Idescat were the conclusions of the master plan of the Cerdà Platform, in other words, for all Idescat's statistics it was necessary to have:

- A single production system.
- A single dissemination database.

This initial vision of the dissemination model, in accordance with the information collected during the analysis process, was complemented by other additional requirements:

- Structure (of the information): a single thematic organization.
- Management: centralised, flexible, automated and sustainable.
- Access: a single data visualization system, for different types of users, from any device, and the provision of an open data channel.

There was a starting point and the destination to be reached was known, but it was also understood from the outset that the results would not be obtained in the short term. Following the three years envisaged for the construction of the new system, a further x years would be necessary to migrate all the existing statistics onto the new Cerdà Platform. Therefore, it was vital to make decisions to cover the time which would go by between the points of departure and arrival.

The challenge consisted of changing the paradigm and simultaneously aligning the existing dissemination system with the newly-created one without affecting the day-to-day running of the Institute and, in addition, without hindering users' access to the statistics.

The answer to this question was clear: the introduction of improvements to the existing system could no longer be delayed. On the one hand, because users were already enduring some of the limitations of the system and, on the other, because there were too many external uncertainties: the Cerdà Platform construction project might be delayed due to budget constraints, the content migration might go more slowly than planned, etc. The decision not to wait meant that it would be necessary to act in parallel with the construction of the Cerdà Platform.

Upon the basis of this positioning, five strategic decisions were made to steer the project:

1. The existing dissemination system would be transformed and aligned with the one proposed for the Cerdà Platform, making the two models compatible in parallel. Any decision made with regard to either of them should apply to both.

2. The existing dissemination databases would be merged in order to obtain a definitive single repository which could be disseminated together with the one provided by the Cerdà Platform. This merger led to the creation of the multi-thematic database as a database for transition between the existing model and the new one.

3. The creation and use of metadata would be promoted in order to manage the publication of the statistics on the Idescat website and to automate the processes associated with the above publication as much as possible. This project resulted in the setting up of the QCOE system (the initials in Catalan correspond to control panel for statistical operations).

4. All the statistical data on the Idescat website would be offered as open data, whether they were on the Cerdà Platform or not. This project gave rise to the API services, one API for the Cerdà Platform and another for the multi-thematic database.

5. Priority would be given to the modernization of the Idescat website: a new architecture and form of browsing, pyramidal access to the information, a single viewer for all the statistics, visualization of the data on charts and maps and a website adaptable to any external device.

PROJECTS

The way of addressing and implementing the strategic decisions made was to launch a set of projects which, whenever possible, would move forward simultaneously and in the same direction.

Some of the projects would naturally come from within the Cerdà Platform project, while others would be born of the need to update the existing systems and would, if necessary, have an impact on the Platform. In both cases the projects had to be compatible, or, if appropriate, end up on the Idescat website in a similar way.

The projects undertaken, some still under execution, have been planned to move forward simultaneously and cover five different areas:

1. Cerdà Platform
2. Multi-thematic database
3. QCOE system
4. API services
5. Idescat website

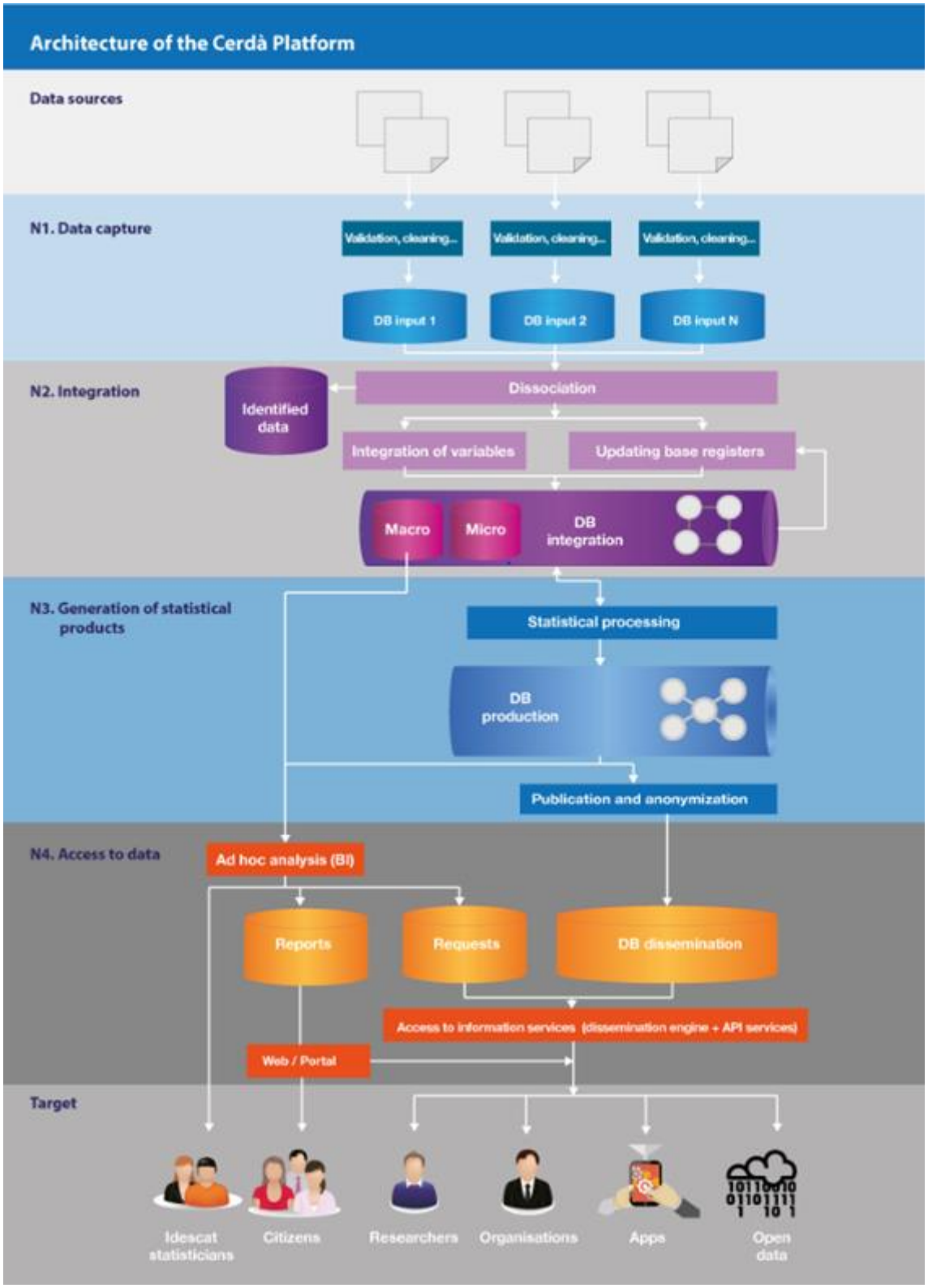
1- 1- Cerdà Platform

The construction project for the Cerdà Platform, initiated in 2015, basically consists of a technological platform for managing and hosting an integrated statistical information system which will allow the reuse and combination of the different data sources received and processed by Idescat. This new model will enable the Institute to do away with the traditional stovepipe production model, whereby each statistical operation is performed without any connection to any other.

This new approach to the integration of Idescat's statistical information is based upon the Swedish model created by Professors Wallgren and Wallgren.

At the core of this new model lie three basic statistical registers: population, entities and territory. This integrated system model seeks to make better use of the administrative registers for statistical purposes, with a consequent reduction in the burden on the reporting units and their costs, as the data allowing the above are linked to each other and can be reused.

Outline of the Cerdà Platform architecture



The Cerdà Platform divides the data into four levels and the communication between them is performed through a transversal layer of metadata and security protocols which guarantee the confidentiality and security of the data.

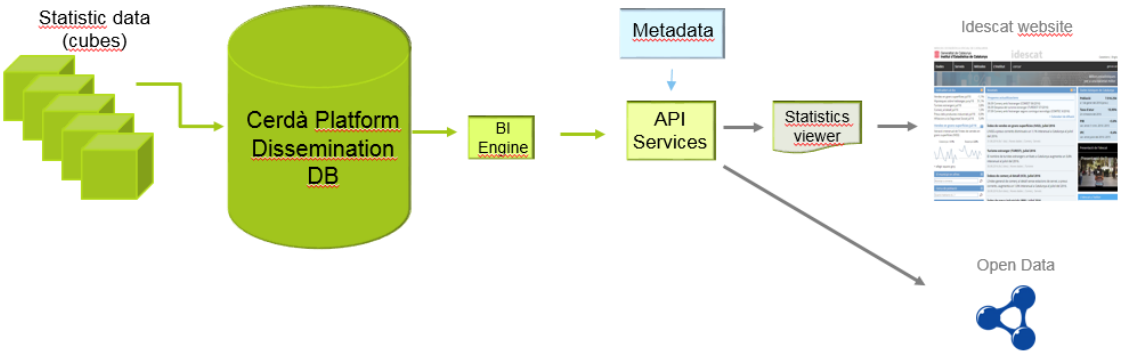
Level 1: Data collection. This level validates the completeness and quality of the incoming data, which may come from different sources (administrative records, surveys, private sources, etc.).

Level 2: Data integration. This level performs the pseudo-anonymization of the data (in other words, the substitution of any information which may directly identify individuals or companies with their respective codes) and the identification data are stored in a high-security location. Data quality processes are later applied when appropriate: merging or linking, detection of duplicates, allocation of missing data and other statistical methods.

Level 3: Generation of statistical products. This level prepares the statistical products for dissemination. We can interact with these products by using specialised software such as SAS, SPSS, R, JDemetra, etc. and the rules of statistical confidentiality appropriate in each case are applied.

Level 4. The Cerdà Platform dissemination system is integrated into the production process and is based upon four architectural elements: Dissemination database, BI (Business Intelligence) Engine, API and Display. All the statistical products are first entered into a single database (**dissemination database**), not in a traditional table format but rather in multi-dimensional cubes containing at least one variable, together with the dimensions of time and territory. The information in the database is extracted by a **BI Engine** which, with a metadata layer, can apply specific rules on statistical confidentiality and transform the data into XML format. These data are, in turn, read by an **API**, which translates them into the JSON-Stat dissemination format. This format will be that which enables us to disseminate data indistinctly via the website or via any other channel or device, according to the audience it is aimed at. Finally, the data are viewed by end users via a **Viewer** which can tabulate the results at their convenience.

Outline of the proposed Cerdà Platform dissemination architecture



Note: Idescat chose the JSON-Stat format, originally developed and promoted by a member of the Idescat team, as a standard for all its internal and external data services, for three main reasons: firstly, it is perfectly suited to Idescat's needs; secondly, it is a simpler format than others, such as the SDMX, promoted, among other organizations, by Eurostat; and thirdly, because it has already been successfully implemented in other European statistical institutes.

2- Multi-thematic dissemination database

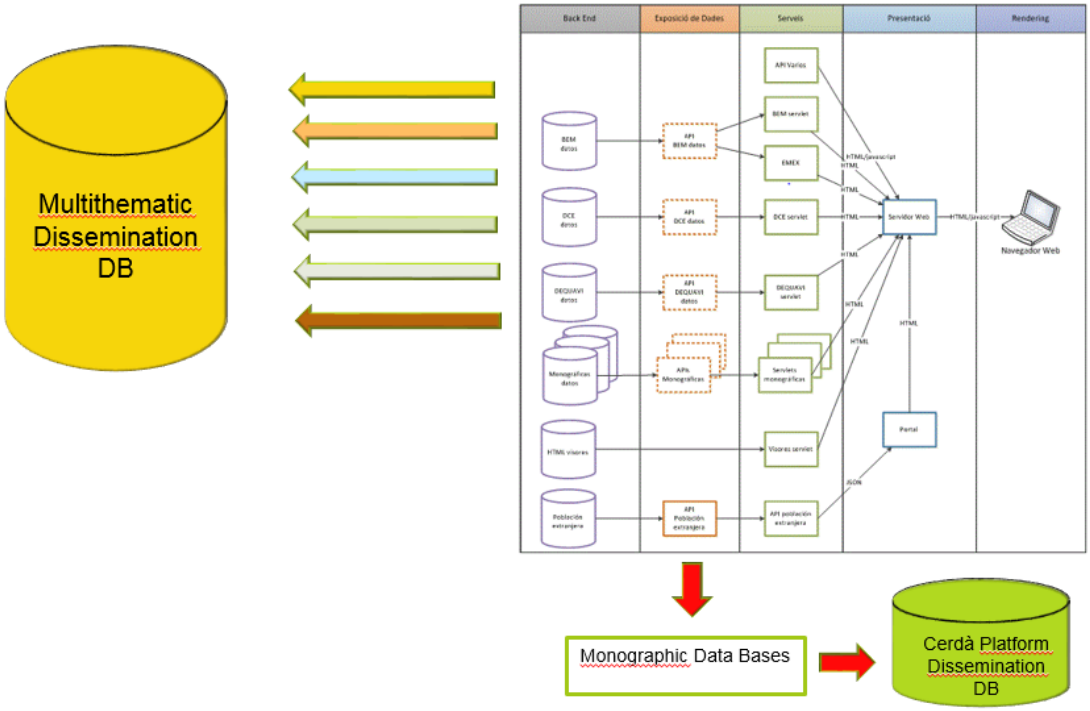
The decision on what to do with the existing databases while the Cerdà Platform was not yet completed and at full capacity was perhaps the most important challenge faced during the project.

The existing stovepipe model had led to the creation, at various times, of different databases programmed according to the requirements of the different types of statistical data to be disseminated. The existing system was the result of 26 years of statistical production. Many of these web applications were programmed without separating the data from their graphic presentation and were organised thematically, regardless of how they were to be disseminated. Many of them had different web applications and viewers, as a result of which the user experience differed according to the product viewed.

The existing dissemination databases could be grouped together as follows: the so-called monographic databases, which contained microdata and with which users could construct customised tabulations according to their needs, and the rest, the majority of them, which contained tables with calculated and pre-defined data - the most important of which was the database of municipal statistics (BEM).

In order to align the existing system with the model proposed by the Cerdà Platform (a single dissemination database and a single viewer), the decision was made to merge the databases containing tables with already calculated data into a single one to serve as a bridge until all the statistical data were available on the Cerdà Platform dissemination database. This transitional database was called the multi-thematic database.

The design of the new multi-thematic database began in 2013 and its aim was to offer the same services as the dissemination database on the Cerdà Platform. The requirements were a single dissemination database, implementation of all the features existing on the original databases (each database offered different features, according to its history), provision of the data separated from the graphic presentation layer, organization of the data by statistical operations, use of metadata to manage all the common aspects (territory, classifications, thematic arrangement, etc.), comprehensive territorial search for all the statistics, generation of a standard HTML and downloads of information in CSV format.



The merger ultimately affected three large databases (municipal, demographics and quality of life and other indicators) and a significant number of viewers and other historical publication formats used by Idescat. The nine existing monographic databases were excluded from this merger, as it was thought that they would be the first ones liable to be entered into the new dissemination database on the Cerdà Platform and it didn't make any sense to make the same journey twice.

The benefits for users, as yet not evaluated, are evident, as all the features have been extended to all the statistical data and, in addition, the desired objective on the project has been achieved: to have a single transitional database to which two basic services of the Cerdà Platform have been added: an API to offer open data and the Viewer of data and charts.

The multi-thematic database went into production and published the data on the website in June 2015. As of the date of the completion of this paper, both the API and the Viewer have been finished, but they are not yet available to end users. Their entry into operation is scheduled for 2017.

3- 3- QCOE System

The different levels of the Cerdà Platform and the communication of data performed through it are managed by a layer of metadata and security protocols which guarantee the confidentiality and security of the data. However, both the Cerdà Platform and, by extension, the new transitional multi-thematic database also had to have a consistent metadata system which would efficiently govern and automate the management of all the statistical data on the Idescat website.

For this purpose, a metadata management system called **QCOE** was designed and implemented. This system, whose basic unit of information is the **statistical operation**, contains all the metadata corresponding to the statistics and manages their publication on the website in three languages (Catalan, Spanish and English).

The QCOE system divides the metadata into four groups:

- **Description:** metadata which identify the statistical operation (code, name of the statistic, etc.).
- **Topics:** metadata which relate the statistical operation to the thematic classification and the order of visualization. Idescat has adopted its own classification of 25 topics which it applies both to the website and to all its applications and databases.
- **Statuses:** metadata which capture the life cycle of a statistical operation and its behaviour in each status.
- **Results:** metadata which include the type of result, where it is stored, the territory and links to the results.

Therefore, any changes registered in the QCOE have a direct impact on the visualization of the data on the Idescat website. The system also permits the linking of the statistical information with other Idescat applications, such as the results calendar, new features, press releases, the annual programme for statistical action, etc.

Idescat also uses **instrumental metadata** covering fundamental concepts for statistical production and dissemination. These metadata already existed prior to the project, but what has been done is to integrate them into the new model so that they can be used by both the Cerdà Platform and the multi-thematic database. Metadata are included here: the Territorial codes and identities database for the geographical dimension; the Time database for the time dimension; the Classifications database, the Variables database (under development) and the Concepts database (currently being created) for the classifications dimension; and, finally, the Methodologies module.

4- API services

The provision of statistics to users as open data was another of the project's essential requirements. In fact, Idescat has been working in this field since 2008, offering API (Application Programming Interface) tools in some of its high-impact statistical services.

Open data is a philosophy and a practice which aims to make some data freely accessible and reusable by third parties in an automated manner, without technical or legal limitations. Having free access to official statistics guarantees transparency, efficiency and equal opportunities when creating value.

The reuse of official statistics allows third parties (individuals, companies and organizations) to generate new products and services for other audiences and recipients. Public data must be reused by both citizens and companies, given that, in addition to provide transparency, they are a driving force for the development of the information and knowledge society. For this reason, there has been legislation in this regard, Law 37/2007 of 16 November on the reuse of public sector information, transposing Directive 2003/98/EC of the European Parliament and of the Council.

With this project, Idescat allows third parties to access to all the statistical data on the Cerdà Platform via a single API, using the JSON-Stat dissemination format.

In order to align the new system with the existing system, an additional transitional API has been implemented in order to download and reuse all the statistical data on the transitional multi-thematic database. Both APIs use the same interface, as a result of which users do not have to deal with two different tools.

With these two APIs Idescat guarantees free access to all the public statistics of Catalonia in an open format. In addition, the data are downloadable in CSV format (and also in open format) from any page on the Idescat website.

Note: As of the date of the completion of this paper, the two above-mentioned APIs have been developed and implemented but have not yet been made available to users, as Idescat is waiting for a corporate API manager to be set up to monitor their use.

5- The Idescat website

The corporate website is the ultimate and visible showcase of everything produced by Idescat and is also the portal of the public statistics of Catalonia. This powerful channel of communication has been and is the main element of the whole project, as it is the one which will be used by most users to access the statistical data which are made public. Therefore, it is one of the most vital elements.

The internal organization of the statistics using the stovepipe model had made it difficult, of course, to display all the statistical production on the Idescat website in a more usable and intuitive manner. Therefore, in this case too, it was essential to capitalise on the journey undertaken by the Cerdà Platform project to update the Idescat website.

The ultimate goal was to improve access to the statistical information, adapt the information to the different typology of users, reinforce understanding of the data and enable their adaptation to the different query devices.

The new model for the organization of the statistical data also meant some changes which clearly enriched the new website model: data organised as statistical operations, web applications separating the data from their graphic presentation, applications applying the same thematic organization, the QCOE, the new tool for managing the publication of the data on the website, etc.

The renovation of the website was not going to take place immediately. It could only be done in stages and as the other projects progressed. It was therefore proposed that throughout this period - perhaps three years - the website would undergo transformation and users would be informed of each of the important changes as they happened. Only at the end of the project would there be a complete renovation of the graphics (*look & feel*).

The website's renovation process is being conducted on five different fronts:

a) New architecture and browsing

A new structure and form of browsing have been created, based upon four pillars or sections around which all the information has been organised: data, services, methods and Institute. To make the usability of the website and access to the statistical data easier, the latter have been divided into 25 topics, each of which has its own page and organization. In addition, the search for statistics by territorial levels has been reinforced.

b) Pyramidal access to the statistical information

The website's thematic pages will display the information pyramidally (in other words, from less to more information), based upon four levels: basic data (with visible data), indicators (with visible data), statistics (with access to the tabulations of the statistical operations) and microdata (with access to pseudo-anonymised microdata, a level which is not yet operational). This new design has several advantages: on the one hand, it offers data to users from the first page, without the need to search for the different products, and, on the other, it enables users who browse via search engines and social networks to find the information they are looking for more easily.

c) Single viewer

This project, which comes from the Cerdà Platform, can also be used to access the multi-thematic database without users detecting any difference. With this viewer users can access the data and tabulate all Idescat's statistics in a single format and in a uniform manner. The system of publication in cubes leads to greater versatility, with only the limitations which have been specified in order to guarantee the statistical confidentiality appropriate in each case.

The viewer will be complemented in the first half of 2017 by a data visualization system using charts and maps. The model also envisages the publication of georeferenced data. This project already has a proven internal prototype which permits the generation of maps in *shape* format, and its integration into the Cerdà Platform is scheduled for the end of 2017.

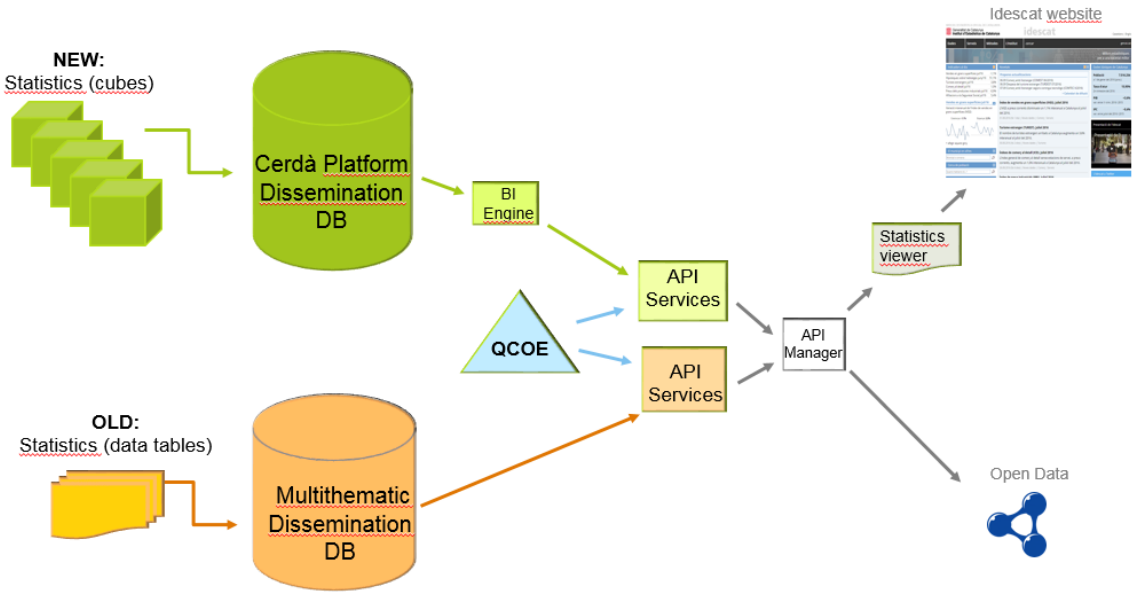
e) Adaptable website

At a technological level, the main innovation has been to transform the website into an adaptable (responsive) model, in other words, all the contents must automatically adapt to the device from which they are viewed, be it a personal computer, a mobile phone or a tablet.

The choice of such a model is largely related to the drastic change in the way users now access official statistics. The number of users who access the Idescat website using mobile devices is already approaching 35% and everything indicates that the figure will continue to increase exponentially in the coming years. The adaptable version also entails a change in the way in which the information is ordered and hierarchised. The structure of the homepage, with three vertical columns in its classical version, will become a portal made up of blocks when the final phase of the website renovation begins.

As of the date of the completion of this paper, the website is still undergoing transformation and many of the planned changes are not yet visible to end users. The first statistical data to appear on the Idescat website from the Cerdà Platform will be from the Tourism thematic area and will be published in the first quarter of 2017.

Outline resulting from the architecture of Idescat's new dissemination model



CONCLUSIONS

The project for the creation and implementation of the Cerdà Platform has been an excellent opportunity to update Idescat's statistical dissemination management model. The resulting system is better suited to the way data are consumed today and also enables Idescat's technicians to perform more efficient management.

Users can change the way they access the statistics. Now they have an interface enabling them to obtain the results in the amount and format which best suits them in each case. They can view the Idescat website in a much more intuitive and usable manner than before as a result of the better thematic organization of its contents and a much more user-friendly form of browsing. In addition, all the pages and data can be viewed on any device, mobile or otherwise, so the experience will be more satisfactory. Moreover, being able to guarantee that all the public statistics of Catalonia can be reused by third parties also places Idescat at the forefront of transparency and indicates the openness of the public Administration to its citizens.

In terms of internal management, the improvements have also been considerable. On the one hand, with the help of metadata, the QCOE system allows centralised management of the complete life cycle of all the statistical operations, including their final visualization on the website. Furthermore, the reunification of the databases, the standardization of the processes and the unification of the thematic criteria greatly simplify the management tasks and the maintenance of the information systems.

The decision to align the existing dissemination system with the newly-created one appears to have been a strategic success, allowing users to benefit immediately from the new model, without having to wait for an endless migration process to end.

With this new dissemination model, Idescat has reached a new scenario which will make it easier to face, in a better position, the challenges of the official statistics of the 21st century. There is still some way to go, but now it seems possible to address the new projects which are knocking on the door with fewer difficulties: big data…

Idescat's statistical dissemination management model is moving forward in parallel with other lines of activity, such as an active presence on social networks (Twitter, LinkedIn, etc.), the continuous improvement of the services for direct attention to end users and the media and, of course, quality management instruments such as the annual statistical results calendar and the rectifications register.

Josep Sort jsort@idescat.cat

Josep Jiménez jjimenez@idescat.cat

Estela Tonzán etonzan@idescat.cat