The mapping, selecting and opening of data: the Records Management contribution to the Open Data project in Girona City Council

By Lluís-Esteve Casellas i Serra

The Open Data project of Girona City Council started at the beginning of 2013 in the framework of a wider project about the City's Open Government. The project is directly leaded by the Mayor's Office, and it is coordinated by the Municipal Unit of Territorial Analysis, with the collaboration of the IT Department, the Department of Records Management, Archives and Publications (SGDAP), and the Communications Office. The opening of the City's Open Government portal and the publishing of the first datasets are scheduled for the last trimester of this year. Nevertheless, it is an ongoing project and, consequently, it is necessary to consider it as a work in progress, beyond data maintenance and updating.

Why participate in the project?

It should be noted that our participation in the project was not assigned to us directly; we had to show our willingness to collaborate actively in it. At first, despite our decisive role in the Records Management at the City Council for the last 20 years, the vision of our task in the project was limited to unstructured information providers. That is, our role was limited by an absolutely traditional perspective of the archive function. However, we had two clear motivations: firstly, to be useful to the organization and secondly, to reinforce our position in it.

The conviction about our usefulness was based on our responsibility for the administration of the Council's Records Management System. This function allows us to have a much more precise knowledge of the roles and activities of the organization and, at the same time, a rather exhaustive control of the municipal information. Thus, from this perspective, the analysis of functions provided by the Records Classification Scheme and the basic data extracted from the General Register of Case Files can play a very relevant part.

On the other hand, and although our department enjoys a good position within the organization, records management still requires constant demonstration of its practical utility. Consequently, participation in innovative initiatives such as the Open Data project also makes us more visible.

Nevertheless, reinforcing our position is not only a matter of internal strategy; it is also a strengthening of the historical archives before society. Note that Open Data projects have a clear aspiration for the published data to be "permanent" and, also, to facilitate the enquiry and the reuse of information. Therefore, it seems logical to guess that such products can be much more attractive than the original source for most users, who basically seek information. The archives have to intervene in the preservation of these products and, at the same time, ensure their context and authenticity requirements, placing them in relation with the original sources.

Locating datasets

Open Data projects aim to achieve both administrative transparency and the reuse of public sector information. But there is often a tendency to prioritize the publication of information referring to transparency in decision-making, rather than to facilitate the reuse of data. This results in the inclusion of unstructured information in Open Data projects.

One of our goals was to avoid this approach, because we felt that the Open Data philosophy is to publish structured information, and also because it was indispensable to prevent the duplicity of the records available on the City Council's Electronic Portal. That is, we must distinguish between the publication of records and the publication of datasets.

Once the goal was clear, the main difficulty was identifying and locating which information datasets could be published. In this sense, Records Management allows a first global identification of the functions, activities, and producers, using the Records Classification Scheme and the General Register of Case Files. However, the information provided by the system is essentially quantitative, and to a lesser extent, qualitative. Therefore, these two tools provide a global perspective on the organization, difficult to achieve by other means, but we are aware that these tools are limited if we want to obtain greater detail.

Aware of these limitations, our intention was to work with two other instruments: the Application Catalogue and the Register of Personal Data Files, as Data Privacy is also one of our functions. From a records management perspective, the Application Catalogue focuses fundamentally on linking specific computer developments with functions and activities of the organization and, also, identifying the document or group of documents that they produce very precisely. It is interesting to notice the difference between the records management viewpoint of the catalogue and the technological viewpoint, which is more focused on the type of technological tool and management processes.

The main objective of the Application Catalogue is to identify data in information systems constituting records in order to facilitate their management and preservation. Therefore, it constitutes a very useful tool to compile a data mapping of the organization. The Register of Personal Data Files offers complementary information, since it details the structure of data within each file.

The result of crossing all these instruments has allowed us to identify 154 datasets that could be published. However, the extraction of data from these datasets is not currently possible. On the one hand, data extraction is not always easy from a technological point of view, because a dataset often resides in various database tables, and appropriate filters are required. Moreover, this could be more complicated if the dissociation of personal data is also needed. On the other hand, we realized that we would be more effective if we associate the Catalogue with the Register of Personal Data Files, and if the relevant data of records or group of records identified in the Catalogue were described more accurately.

Finally, then, this more exhaustive mapping constitutes one of our lines of development for the project. In the short term, we decided to select essential datasets identified in models of good practice of transparency in Catalonia and easy to obtain: the financial management data of the City Council and its associated organizations, the remuneration of politicians and trusted members of the government, population, planning permissions, economic activities, pollution and noise pollution data, energy consumption, municipal buildings and services, etc.

Data selection

The data selection processes differed according to the original data source. Thus, for example, one thing is selecting data from pollution detection devices, automatically generated in a structured way, and the other, selecting data derived from administrative management processes, which normally combine a manual, semiautomatic or automatic entry of data.

The Records Management Department has focused on the datasets derived from the administrative management of the organization, based on the following criteria:

- Defining the chronological scope of datasets.
- Identifying the datasets in a clear way.
- Selecting the data to be published from the dataset.

Determining the chronological scope is not just a matter of greater or lesser availability of the original sources, but also the consistency and sustainability of future updates. In this sense, we have taken two decisions:

- Only data in force from continuous administrative registers will be published; that is the case of economic activities, for instance.
- Only datasets from the current year, or the year immediately preceding it, will be published on the City's Open Government Portal. Previous years will be managed, preserved and published on the Municipal Archive website.

This second choice, we believe, is particularly relevant because the Municipal Archive may offer the Open Data datasets together with the corresponding original documents, in either a structured or unstructured information format or not. Moreover, the possibility to collaterally compare the Open Data product with the original will strengthen their authenticity requirements.

Nevertheless, it is not yet decided whether disposition rules of the originals will apply to Open Data products, especially when they refer to their disposal. Despite the fact that, in principle, this application should respond to the determination of the informational value, the weight of Big Data tendencies creates relative uncertainty in this regard.

Regarding the precise identification of datasets, the process is closely linked to the identification of specific activities within the organisation. For example, in the function of "population management" we have to distinguish what dataset exactly constitutes the updated register, in this case the Municipal Register of Inhabitants, from the dataset of variation data for a period of time, that is to say, annual modifications of the Register. This distinction is important because both datasets reside in the same information system and not discriminating them would hinder their understanding and, if necessary, its correct reuse.

As for the selection of data to be published, we have followed two different procedures, according to the documentary form of the original record. If the original record was in the form of a register we selected only the data included in the form of the original record, usually identified by comparison with their precedent on paper.

This is the case of the Municipal Register of Inhabitants, information derived from the municipal budget or its final settlement. In this sense, it is very useful to have a previous strategy of records preservation in databases. Firstly, because it contributes to identify the records, and secondly, because the extraction processes to preserve the records can also be used for the Open Data project. In fact, the Open Data project reinforces the development of these preservation strategies.

Furthermore, if the function or activity is reflected in a case file, the selection is based on the identification of key data in the administrative processing and, at the same time, on the qualitative data that differentiates them from a case file of the same nature. This has been the case, for instance, of planning licences, which have allowed us to obtain a non formal register more complete than the traditional case files catalogue.

The experience in both selection processes has reinforced our emerging strategy of records preservation in databases, and it has also led us to reach a consensus with the IT Department on two principles:

1. To promote the creation and formalization of specific registers based on administrative processing, as a complementary development of the General Register of Case Files.

2. To incorporate the Open Data perspective in the analysis prior to the computerization of administrative proceedings.

Data opening

To open data, it has been necessary to apply processes of dissociation of personal data in practically all cases. In fact, some datasets have not been opened because of the difficulty to assure personal privacy. These difficulties convinced us of the necessity to identify and analyze these cases specifically. The aim of this analysis was to introduce guidelines about information entries to facilitate automatic dissociation. The results allow us to show, for example, accounting operations at the most detailed level.

On the other hand, the quality of data is undoubtedly a key point in the publication of datasets of any Open Data project. This question can only be solved by focusing on the creation of information, at the time of its capture. Therefore, Records Management can have a prominent position because of the transversal function it develops.

Our transversal function in Records Management allows us to identify the appropriate interlocutor for making improvements to the quality of data any better than other department. This is not at all irrelevant, because the interlocutor is not only a qualified person who can provide us with information, but also the person with whom we already share experience, knowledge and, most importantly, vocabulary.

For this reason, our interlocutors from producer units have been essential to understand the meaning and relevance of specific data, and this is a key factor, because the data has to be easily understandable and valuable in each dataset. Thus, the description has to be as accurate, clear and precise as possible, which is why our professional specificity places us in a better and strategic position within the organisation.

What is more, the dataset description has to be understandable in relation to the content, and, at the same time, their production context. We are quite satisfied with our contribution in this domain, because descriptions have become substantially better. Dataset descriptions have integrated content information, original sources, the legal references that generate the records, and even limitations of data extractions to offer the possibility of corroborating their authenticity.

What have we learned?

1. The Records Classification Scheme provides valuable information to detect datasets susceptible to be published, especially if we identify the specific applications that produce records.

2. The Application Catalogue is very useful if it is linked to the functions and the activities of the Classification Scheme. In this case, including the structure of selected essential data and its description is very convenient.

3. Defining the chronological period of datasets to be published and taking care of Open Data Archives can reinforce the role of traditional historical archives.

4. The strategies of records preservation in databases make the data selection easier for Open Data projects. At the same time, Open Data projects can be an important incentive to develop or reinforce these strategies.

5. Including the Open Data perspective in the previous analysis to implement workflow solutions might facilitate data extraction in the future. Furthermore, it allows significant economic and resource savings and it avoids carrying out specific studies afterwards.

6. The appraisal of significant data from case file processing should be useful to strengthen the creation and formalization of administrative registers. These administrative registers could be integrated to the strategies of records preservation in databases.

7. Data quality is one of the most important problems in Open Data projects, and it is only possible to solve it when the information is created. In fact, the only viable way is to link data quality with the Records Management Systems. Even to promote data quality we believe that users should visualize on the screen which fields will be published and which of them have a special protection.

8. In some cases, data dissociation can be difficult, so Records Management can contribute to do specific analysis of these cases and will make it easier to introduce guidelines about information entries, thus facilitating automatic dissociation.

9. The transversal function of Records Management provides us with the important role of intermediation. This role allows us to identify adequate interlocutors of producer units and, at the same time, to be facilitators of solutions to specific problems.

10. An accurate description of the context of dataset creation allows users to understand them better, as well as contributing to the reinforcement of their

authenticity.

In conclusion, we are convinced that our contribution to Open Data projects is not irrelevant at all and, moreover, it could be essential in the future for the transformation of data from information systems into documentary heritage.