Consultant: Carlos Iglesias – contact@carlosiglesias.es
Reviewed by: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

Open Data management Guide

Precedents **2**

Open data principles **2**

Definitions **3**

Data treatment **4**
1 – DATA IDENTIFICATION 4
2 – DATA SELECTION 5
Restrictions on data publication 6
3 – DATA PREPARATION 7
Formats 7
Detail level 8
Quality and reliability 8
Data licenses 8
4 – DATA PUBLICATION 9
Metadata 10
ANNEX I – METADATA EXAMPLE 16
Dataset 16
Distribution 1 17
Distribution 2 17
ANNEX II – USE CASES 17
Datasets affected by intellectual and industrial property 17
Datasets with personal data 18
Datasets in proprietary formats 19
Datasets in unstructured formats 19
Datasets in image, audio or video formats 19
Datasets in formats that require previous treatment 20
Datasets with indirect access through directories, search engines and others 21
Datasets not digitized 21

Data is an essential resource with huge potential for the creation of social and economic benefits. Many valuable **datasets** are managed by Public Administrations, and the objective pursued by **Open Data** is the release of such data so that it could be *reused* in the benefit of everyone.

However, the amount and variety of data being managed within any Administration makes this a non-trivial task that will require the involvement and collaboration of all areas.

Consultant: Carlos Iglesias – contact@carlosiglesias.es
Reviewed by: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

Precedents

The Government of Aragon, by means of official agreement of 17 July 2012, is committed to the effective opening of public data following the recommendations of the 2003/98/EC European Directive on Public Sector Information Reuse, as well as its transposition in Spain through Law 37/2007. To that end the Aragón Open Data project was born, providing a complete platform that supports the publication of open data.

Moreover, to date March 4, 2013 the Spanish Technical Interoperability Standard for the reuse of Information Resources (hereinafter the NTI Standard) was published in the National Official Bulletin (BOE). This NTI Standard aims to facilitate and ensure data sharing and re-use in Public Administrations. For that it provides guidance on how to ensure the persistence of information, the selection of appropriate formats, as well as the right terms and conditions for re-use.

The publication of this standard is an important milestone, because it describes in detail the metadata used to ensure *findability* and interoperability of those information resources produced or held by the public sector. However, for this being effectively possible, it is crucial to **avoid any ambiguity** in the interpretation and final application of the standard. This fact is essential to ensure that data processing will be not only technically feasible, but also semantically reliable.

This document is a guide intended to serve as help and assistance in the process of data cataloguing, as well as to lead in the correct interpretation and application of metadata through the provision of clear indications and examples.

The purpose of this guide is to be distributed among the various agencies of the regional Government to provide them with the indications needed for the development, management and maintenance of all metadata consistently and in accordance with current standard. This way, data managers will be able to self-operate, taking also advantage of the tools that the project has previously made available through the Aragón Open Data platform.

Open data principles

Whenever we publish data, the Open Government Data principles[1] must be followed to ensure we do it in an effective way:

- **Complete**: all data is made available, with the only exception of that being affected by privacy, security or privilege special limitations.

[1] Open Government Data Principles: https://public.resource.org/8_principles.html

Consultant: Carlos Iglesias – contact@carlosiglesias.es
Reviewed by: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

- **Primary**: data is shared exactly as it was collected in the original source and with the same detail level, without further modifications, aggregations or intermediate operations.
- **Timely**: data is made available as quickly as necessary to preserve its value and avoid obsolescence.
- **Accessible**: data is equally available to the widest range of users and for the widest range of purposes.
- **Machine readable**: data is structured in a way that enables automated processing.
- **No discriminatory**: data will be available for anyone, with no additional identification or registration requirements.
- **Non-proprietary**: data is available in formats over which no entity has exclusive control to maximize its reuse potential.
- **License-free**: data is not subject to any copyright, patent, trademark or trade secret regulation, allowing anyone to use, reuse, share or redistribute it without further restrictions.

Throughout this document a set of useful guidelines to assist while trying to meet these principles are explained.

Definitions

The definitions of some key open data related concepts used throughout this document are included below.

Data Catalog: electronic repository where data and documents are stored along with the correspondent metadata.

Dataset: a series of data grouped around a common theme or subject that characterizes them. It may be composed of one or more distributions in different formats.

Data: minimal formal representation of facts, concepts or instructions that is suitable for communication, interpretation or processing by individuals or automatic means.

Open Data: data opening and sharing initiative in any form that facilitates its *reuse* by others.

Personal data: any information in any form which may enable the identification of an individual, either directly or indirectly.

Consultant: Carlos Iglesias – contact@carlosiglesias.es
Reviewed by: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

Distribution: data available in a particular presentation format: a file type, a web service or any other. The same dataset could have multiple distributions in different formats to facilitate reuse with a wider range of programs and tools.

Infomediary: any agent that collects, organizes and adds value to large amounts of data coming from different sources, thus acting as an intermediary between data suppliers and end users.

Metadata: data that defines and describes other data. There are different types of metadata depending on its application.

Reuser: natural person or legal entity that reuses public sector information, whether for commercial or non-commercial purposes, provided that such usage does not constitute a public administrative activity.

Other definitions and acronyms related to the reuse of public sector information are available in the guide published on the subject by the Spanish National Interoperability Framework[2].

The data process

It is necessary to establish an initial collection all data at a single access point for an optimal management from the reuse perspective. This is usually made through a **data catalog** that will facilitate access for potential re-users.

The datasets catalog is generated on the basis of information and metadata provided by the different data managers, which will be eventually published through the online platform developed by the Government of Aragón http://opendata.aragon.es/, enabling a complete inventory of all reusable public datasets in the region.

The ultimate goal is to catalog the information about the data currently available in all agencies of the Government of Aragón and other related entities. For that, it will be necessary to follow a series of steps outlined in the following sections:

1 – DATA IDENTIFICATION

A data inventory process should be conducted to be able to know what the scope and extent of existing data records is. That inventory should be considered only as

[2] Reuse of public sector information – Definitions and Acronyms: http://datos.gob.es/sites/default/files/definiciones-acronimos_risp.docx

the starting point of a process that will be **progressively expanding to include all areas and agencies**. In addition, since the production and update of data is continuous, this should also be seen as a **cyclical and constant process** within each area and body that need to be repeated and made lasting over time.

In general, data managers must assume the initial identification of existing datasets and the subsequent detection of new datasets as their own responsibility, as well as the control of changes and updates as they occur.

The following are some usual milestones where new datasets may be emerging during the course of daily activity:

- When a **new activity or project** starts;
- When significant changes affecting the technology and/or **information systems** in use are introduced;
- When additional information is needed to inform **decision-making**;
- When **new information needs** are requested from other areas or bodies;
- When **legislative changes** affecting the scope and purpose of action occur;
- When new **data and information demands** are received from the public;

Apart from the above, there may be many other factors that can lead to the massive creation or updating of data, depending on the specific procedures in each area or body.

On the other hand, data may only available through non-electronic means - such as in the case of printed material - or just on *physical* media - such as CDs; DVDs; or others. Then, we will need to locate the original data source at the internal databases and systems in order to be able to extract the corresponding dataset in electronic form before its final publication.

Digitalization of any content that is not yet available in an appropriate electronic format is also a reasonable long-term objective that should be considered to enable reuse of historical data.

2 – DATA SELECTION

The final goal is to catalog **all available data**. However, as time and resources are always limited, each area or body should determine what information is the most valuable from the *reusability* point of view to be able to prioritize publication.

Due to the different nature of the agencies and the information they manage, prioritization criteria must be adapted to the particularities of each case knowing

Consultant: Carlos Iglesias – contact@carlosiglesias.es
Reviewed by: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

the type of information they hold and the potential audience. However, there are some general criteria that should be always evaluated:

- **Regulations and laws**, as it is increasingly frequent the existence of rules and recommendations at international, national or regional level advising, or even laying down, an obligation to publish certain information.
- **Relevance to the public**, assessing the impact and potential usefulness of data for the general public, especially in those cases where specific data requests have already been received.
- **Relevance for companies**, assessing the impact and potential usefulness of the data for companies that will be making use of it, especially in those cases where previous public-private partnerships may already exist.
- **Relevance for Public Administration**, assessing the strategic value and potential usefulness of data for internal Governmental use or when sharing with other public bodies.
- **Update frequency**, given that the more dynamic data generally offers the best reuse potential, although this ultimately depends on the information nature.
- **Data volume**, taking into account both, current volume and future growth forecast. As a general rule, the higher data volume the best reuse potential.

Each area or body could also adapt and complete its own criteria list accordingly to the scope of activities and services, as well as any other particular rules.

Restrictions on data publication

While the general rule is just to **publish everything**, there are certain limitations when publishing datasets that must be respected, as in the following cases:

- Those restricted by **personal data protection**;
- Those subject to **statistical or commercial confidentiality**;
- Those affecting **security or public safety**;
- Those affected by third-party **intellectual property rights**;
- Those where a **prohibition or limitation on the right of access** may exists, or where legitimate rights are required to grant access;

For all the aforementioned cases any sensitive information needs to be filtered and removed from the datasets **prior to final publication**.

Consultant: Carlos Iglesias – contact@carlosiglesias.es
Reviewed by: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

3 – DATA PREPARATION

One of the main objectives when exposing public datasets is to enable and facilitate **automatic data processing**, since that is the only way to handle these large amounts of information efficiently. For such software-enabled processing being possible it is necessary first to use the **appropriate formats**.

Formats

Some criteria to follow when selecting formats are:

1. **Structured** formats that are directly **machine-readable**.
2. The most popular formats among our potential *reusers* or, alternatively, those **formats in widespread use by major consuming sectors** of such information, and therefore acting as standards, either official or *de facto*.
3. **Open** formats that allow **unrestricted** use, equal opportunities and possibly cost savings for all involved parties. Otherwise, although the data was published in the open, we would be forced to use some technologies tied to a particular vendor.
4. Those compatible with the provisions of the **National Technical Interoperability Standard**[3], particularly regarding file formats, document management and semantics.
5. Those that also provide embedded meta-information about the schemas or vocabularies that have been used to represent the information, such as XML or RDF[4].
6. Offer **different and complementary formats simultaneously**, provided they are appropriate for the specific use case. This will make it easier to meet the needs of a larger number of *re-users* with different preferences and expertise (businessmen, entrepreneurs, researchers, journalists, etc.)

The ideal target with regards to formats would be to get every dataset **at least with one distribution** using an **open, standard, structured and machine-readable** format. Common examples of such formats are: CSV (or TSV) for tabular data, WMS for maps or XML for structured data.

In any case, short-term efforts need to focus on providing direct and immediate *reuse* of the greatest amount of data whatever the existing formats are. When deciding on the final formats there should always be a balance between the reuse

[3] Resolution of 3 October 2012 from the Ministry of Public Administration, approving the Technical Interoperability Standard Catalog: http://www.boe.es/boe/dias/2012/10/31/pdfs/BOE-A-2012-13501.pdf

[4] It will be necessary to consult with the technical assistance of the area concerned or the global open data project management.

Consultant: Carlos Iglesias – contact@carlosiglesias.es
Reviewed by: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

potential of the current format and the efforts that may be needed for the transformation of such data into more suitable formats.

Detail level

Data must be provided with the **highest possible level of detail**, so that each re-user will be able to perform the necessary data treatment for their specific purposes after. It is therefore important to **respect the original format and details of data**, avoiding any modification or alteration prior to publication, even when the aim was to facilitate data readability. The only exception here is any filtering that may be necessary for **privacy**, **confidentiality** or **security** reasons as discussed above.

Different data *aggregations* may also be provided in order to facilitate interpretation as a **supplement**, but never as a substitute for granular data, since otherwise we could be preventing access to certain relevant information to the *infomediearies'* needs.

Quality and reliability

Published datasets should also have **reliable** content that is appropriate to be directly exploited, so it should not present **quality** problems beyond what is reasonable. If that is not the case, it will be necessary to clean the data before publication, determining the measures required to carry out the necessary repairs.

Data licenses

It is recommended that the publication of any dataset is performed according to the basic provisions of the R.D. 1495/2011 by default, including:

- **General conditions** for data reuse, such as to forbid meaning distortion or the obligation to mention the original data source among others.
- The **exclusion of liability** for the publisher in any subsequent data usage.
- **The responsibility of the *reuser* agent** with regards to data usage.
- Information on how to proper reflect the original source **attribution**.

When, exceptionally, a given area or body opts for the application of an alternative licensing regime in the provision of data, that should be only done through a license compatible with the open data principles.

In general, the applicable license for any dataset within the Aragón Open Data initiative will be:

Consultant: Carlos Iglesias – contact@carlosiglesias.es
Reviewed by: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

- Creative Commons-Attribution 4.0[5] (CC-BY 4.0)

A different license could be applied exceptionally[6] when it also fulfils the open data principles, such as one of the *Open Data Commons*[7] licenses:

- Public Domain Dedication and License (PDDL).
- Attribution License (ODC-By).
- Open Database License (ODC-ODbL).

4 – DATA PUBLICATION

Last step will be to publish not only datasets, but also any other supporting **metadata** that could be useful for better localization, classification and reuse.

Data and metadata will be collected by the designated coordinators in different areas and bodies, to be later shared by means of any of the available options:

- Metadata templates.
- Tables for bulk data upload.
- Directly through the online catalog[8] at http://opendata.aragon.es/ (coming soon).

In case of any doubt regarding procedures, it is recommended to contact directly with the global project management at opendata@aragon.es.

For enabling publication data should be available either through a direct web link to the original data source or through some type of file in any format to be manually uploaded to the platform. In this last case, data will be updated by uploading a new file to the platform with the previously established frequency (this way the historic of data will be available as well).

Metadata

Metadata is an **essential element** when cataloguing datasets because it is the element that allows us to sort and find information effectively. Therefore, it is very important not only to provide the biggest amount of metadata, but also to do it the right way by carefully following the instructions given in this section.

[5] CC-BY 4.0: http://creativecommons.org/licenses/by/4.0/
[6] See a use case for datasets being affected by intellectual or industrial property in Annex I.
[7] Open Data Commons data/databases license: http://opendatacommons.org/licenses/
[8] The catalog is based on the Open Source platform CKAN http://ckan.org/ developed by Open Knowledge, with several modifications and adaptations made by the Government of Aragón and published through: https://github.com/aragonopendata/Aragon-Open-data-Website.

Consultant: Carlos Iglesias – contact@carlosiglesias.es
Reviewed by: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

In case any doubt with regards to the meaning or format of metadata remains, it is preferable to consult directly with the global project management[9] to avoid any problem that may affect metadata quality.

In this section metadata is divided into:

- **Required:** metadata that must be provided given its special relevance and also for compliance with the applicable law.
- **Recommended**: metadata that, although not being required to comply with current legislation, is also highly recommended given its relevance for an adequate data classification.
- **Optional**: metadata that, although being equally recommended, may not always be available.

Furthermore, sometimes metadata describes datasets and others to their different compounding distributions[10].

SUMMARY METADATA TABLE		
DATASETS		
REQUIRED	RECOMMENDED	OPTIONAL
Title	Tag(s)	Last update
Description	Terms of use	Update frequency
Theme(s)	Contact email	Geographical coverage
Publishing body		Temporal coverage
		Validity
		Related resources
		Regulations
DISTRIBUTIONS		
REQUIRED	RECOMMENDED	OPTIONAL
Distribution	Name	Additional format information
Format	Description	

[9] Support available at opendata@aragon.es.
[10] See the definitions section at the beginning of the document.

Consultant: Carlos Iglesias – contact@carlosiglesias.es
Reviewed by: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

REQUIRED METADATA

This section describes the minimum metadata to be provided, both for each of the datasets that will be published as well as for every single distribution format.

DATASETS

Title: The representative name of the dataset.

A short text is recommended with no more than 10-12 words.

Description: Text description for the dataset content and main characteristics.

It is recommended to include information about the data types, origins, potential usefulness, limitations, etc. while avoiding replicating any information that may be directly available through other metadata fields.

The description should be as complete as possible while maintaining a reasonable extension of no more than 3-4 paragraphs. For those datasets involving certain complexity that may require further description we recommend using additional information metadata to include a reference to some available guide or manual.

Theme(s): Main category or theme of the dataset.

It will correspond to one or more values (two or three at most) selected among those available in the official classification being applied to the data catalog of the Government of Aragón:

Science and technology	Energy	Safety
Commerce	Finance	Society and welfare
Culture and leisure	Industry	Transport
Population	Law and justice	Tourism
Sports	Environment	Urban planning and infrastructure
Economics	Rural areas	Housing
Education	Health	
Employment	Public sector	

It is important to select the most appropriate theme(s), which in principle do not need to be directly related to the organizational structure of the Government of Aragón.

In the document where the taxonomy of the National Catalog[11] is defined (the one being used as reference), a list of the most frequent topics or possible subcategories

[11] Taxonomy of the Spanish National Catalog:
http://datos.gob.es/sites/default/files/files/12_tax_02.pdf

that correspond to each major theme is included as guidance while selecting the most appropriate one(s).

Publishing body: the body in charge of the dataset.

The body responsible for the data (a particular Service, Unit, Secretariat, Cabinet, Offices, Branch, Department, Institution ...) within the official organization chart of the Government of Aragón[12].

The body must always be indicated with the greatest possible specificity (service level or similar). When the responsibilities are not clear, the Government of Aragón will be indicated as the body in charge.

DISTRIBUTIONS

Distribution: will indicate how to locate a specific dataset distribution.

The file or resource will be indicated in one of the following ways:

- *providing a web address (URL such as http://www.example.org) where you can find the dataset in question (file, service, etc.)*
  or
- *selecting a file from your own computer that directly contains the desired data and uploading it to the platform.*

Format: indicates the format in which the dataset is being represented.

The format type is usually defined by a 3 or 4 letters code (CSV, XLS, HTML or JSON for example) and it usually corresponds to the extension of the file where the data is stored. Examples of frequent reusable formats are:

***CSV**: tabular data.*
***KML**: geographical information.*
***JSON**: data exchange.*
***ODS**: spreadsheets.*
***PX**: statistical data.*
***RDF**: semantic resources.*
***RSS**: data feeds.*
***SHP**: spatial data.*
***WMS**: georeferenced data.*
***XLS**: spreadsheets.*
***XML**: personalized data vocabularies.*

It is advisable to use always uppercase for improved readability, and to be sure about specifying always the final distribution format not an intermediate one[13] (e.g. if we need to access a CSV dataset distribution through an intermediate HTML page

[12] Organization chart of the Government of Aragón:
http://servicios.aragon.es/organigrama_publico/PublicoServlet?accion=1
[13] See use case for datasets with formats that require prior processing in Annex I.

because there is no other direct access, then the CSV format should be indicated, not the HTML one).

There is also a public record with all existing types of formats[14] that could be helpful, but if in any doubt it is also recommended to contact global project management[15].

RECOMMENDED METADATA

This section describes the recommended additional metadata for datasets and distributions that should be provided whenever it is available.

DATASETS

Tag(s): one or more specific textual labels for dataset classification.

It is preferable to use single word tags (or two-words at most) separated by commas. It is advisable also to limit ourselves to a reasonable number of tags when describing each dataset. We could say that generally 3 or 4 is a reasonable limit for example, prioritizing those we consider the most related ones and trying also to supplement the previously selected main themes.

Terms of use: designation of the terms of use or the applicable license.

The generalized option will be to use the "Terms and conditions for the reuse of public sector information" that have been previously described in the licenses section of this guide.

Generally speaking, the applicable license for any dataset within the Aragón Open Data project will always be:

- *Creative Commons-Recognition 4.0[16] (CC-BY 4.0)*

If in an exceptional case[17] it may be necessary the adoption of any other license type, this should be consulted first with the global project management through opendata@aragon.es.

Contact email: contact information.

A general contact email for the area or body publishing the data, or the personal email of the person who is directly in charge of the data in question.

DISTRIBUTIONS

[14] IANA Media Types: http://www.iana.org/assignments/media-types/media-types.xhtml
[15] Support available at opendata@aragon.es
[16] CC-BY 4.0: http://creativecommons.org/licenses/by/4.0/
[17] See for example the use case about datasets being affected by intellectual or industrial property in the annex to this document.

Name: It will be the representative title for a dataset distribution.

A short text with no more than 10-12 words is recommended. The name should focus on the differences between this distribution and any other.

Description: Descriptive text about the distribution content.

Descriptions should be provided with a reasonable extension of no more than 2-3 paragraphs. It is recommended to focus on any information that is particularly distinctive for the distribution in question.

OPTIONAL METADATA

This section describes the additional metadata that can be provided for a specific dataset. Although the fields requested in the section are not strictly required, it is still highly recommended to facilitate such information when it is known to improve quality and usefulness of available metadata.

DATASETS

Update date: last update date for the dataset.

The date of the last update on the content (not on the metadata) of the dataset will be provided following the YYYY-MM-DD format, i.e. the year with a four-digit number, the month with a two-digit number (01 to 12) and the day also with a two-digit number (01 to 31).

For example: 2011-11-14 for November 14, 2011.

Update frequency: represents the estimated period of time between two consecutive dataset updates.

It will be represented by the closest value from the following predefined ones (lowest to highest update):

Triennial, biennial, annual, biannual, four-monthly, quarterly, bi-monthly, monthly, fortnightly, tri-monthly, weekly, bi-weekly, tri-weekly, daily, hourly and real-time

For example:

- *Update every year:* ***annual***
- *Update every two months:* ***bi-montly***
- *Update two times a month:* ***fortnightly***
  Update two times a week: ***bi-weekly***

Geographical coverage: represents the geographical area covered by the data.

If it is the same as the general coverage of the Catalog (Autonomy of Aragón) it is not necessary to specify any, but that is not the case when the coverage differs from the general one (being a province, municipality, county, etc.).

The geographic scope should always be stated with as much specificity as it is applicable in each case[18] (either at the provincial, county or municipality level). When there is no other clear scope the Autonomy of Aragón should be used.

Temporal coverage: identifies the start and end dates of the period of time covered by the dataset.

Two dates corresponding to the start and end of the period of time covered by the data following the YYYY-MM-DD format, i.e. the year with a four-digit number, the month with a two-digit number (01 to 12) and the day also with a two-digit number (01 to 31).

For example: start 2013-01-01 and end 2013-12-31 represents be the period from January 1, 2013 until December 31, 2013.

Validity: identifies the final validity or "expiration" date for the data from which it could be losing its relevance for some reason (modification, update, etc.)

The expiration date will be indicated following the YYYY-MM-DD format, i.e. the year with a four-digit number, the month with a two-digit number (01 to 12) and the day also with a two-digit number (01 to 31).

For example: 2014-01-31 indicates that the data is valid until December 31, 2014.

Related resources: Web addresses (URLs such as http://www.example.org) pointing to pages, documents or other resources that contain extensive information on the dataset.

For example, this could be a link to a PDF document where the methodology used while collecting the data is explained, or a link to an interactive application through which the data is explained in a visual form.

Regulations: web addresses (URLs such as http://www.example.org) pointing to pages, documents or other resources that contain rules or regulations affecting the dataset.

For example, this could be a link to a web page with a law or ordinance regulating the data that is part of the dataset.

DISTRIBUTIONS

[18] A complete list of counties and municipalities can be found at: http://idearagon.aragon.es/toponimia/comarca.htm

Additional format information: web addresses (URLs such as http://www.example.org) pointing to pages, documents or other resources that contain additional information, usually of a technical nature, on the distribution format.

For example, this could be a link to a website where you can find more information on the format, or a link to an XSD schema associated with a XML dataset.

ANNEX I – METADATA EXAMPLE

Below is a complete example of what could be the appropriate metadata for a *hypothetical* dataset of public examinations and employment with two distributions in different formats.

Dataset

Title: Health service public examinations call.

Description: Competition calls for public examinations and employment for the Aragonese Health Service that have been published in the BOA from 01/01/2014, including the title of the call, official syllabus and time limits for procedures.

Theme: Employment; Public Sector

Tag(s): public examinations, call

Publishing body: Aragonese Health Service

Contact email: salud@aragon.es

Terms of use: Creative Commons – By 4.0

Temporal coverage: from 2014-01-01 to 2014-12-31

Update frequency: weekly

Update date: 2014-06-28

Validity: 2014-12-31

Geographical coverage: Autonomy of Aragón

Related resources:
http://aragon.es/Salud/Oposiciones/ListaAprobadosEnfermeria.pdf

Regulations: http://boa.aragon.es/oposiciones/bases-convocatoria=271213

Distribution 1

Name: Last calls in JSON.

Description: Real time access to the details of the last three published calls in the current year through JSON format.

Distribution:
http://boa.aragon.es/oposiciones/convocatoria=271213&format=JSON

Format: JSON

Additional format information: http://json.org/

Distribution 2

Name: Calls record in XML.

Description: Access to detailed record of all calls in XML format.

Distribution:
http://boa.aragon.es/oposiciones/convocatoria=271213&format=XML

Format: XML

Additional format information: http://boa.aragon.es/oposiciones/esquema.xml

ANNEX II – USE CASES

In this section guidelines to follow in certain frequent use cases when publishing datasets are provided.

Datasets affected by intellectual or industrial property

When external agencies or companies are involved in data generation and management due to service contracts or agreements, we should check whether they have effective intellectual and industrial property rights that are required to enable effective reuse. Otherwise, we could either:

- Get the required rights assignments from the rightful owner.

  or

- Do not allow the re-use of any affected datasets.

If you choose to apply for the transfer of rights and these are not in accordance with the general terms of use previously established by the Government of Aragón, you could opt for setting specific reuse conditions for the affected datasets that will be specifically tailored to the rights assignment finally obtained.

For example, if a given body carries out an exhibition of artistic works from various authors and would like to publish a dataset with photos and other data for all these works, a permission of their respective owners should be obtained first.

Datasets with personal data

Law 11/2007 of 22 June, on electronic access of citizens to public services, provides in article 4a) that the use of information technologies should include **respect for the rules on personal data protection** as well as **compliance with the right to honour and privacy**.

Therefore, to support the publication of datasets observing the established rules a process known as *irreversible dissociation* must be applied in order to **avoid possible identification of people**. It is recommended to proceed as follows for each of the datasets concerned:

1. Determine what are the conflictive elements in the dataset that could make it possible to identify a person: for example *name*, *postal address*, *email*, *phone or fax number*, *personal ID*, *social security number*, *signature*, *IP address*, etc.
2. Remove any identifying information prior to the dataset publication.

In the event that such **dissociation may not be possible** due to technical difficulties, cost or any other reason, then **the dataset must not be published**. It is also important to note that legal entities and individual entrepreneurs are excluded from data protection[19], as well as those files where only professional data of individuals is incorporated,.

For example the population census must necessarily contain at least the following information:

- *complete name;*
- *gender;*
- *residence address;*
- *nationality;*
- *place and date of birth;*
- *ID number;*
- *education level;*

Among all this data, some fields such as the complete name, the ID number or the residence address may enable the identification of a person, and therefore this is data that should be filtered or aggregated for anonymization before publication.

Datasets in proprietary formats

Although you should always opt for open formats as the first choice when publishing datasets, other proprietary one may also be used, provided that they are at least structured formats.

Examples of quite popular structured and proprietary formats are XLS - spreadsheet format developed by Microsoft - or SHP - spatial data format developed by ESRI.

For example, if we have a dataset available in XLS format, we could also easily publish an additional distribution in CSV format to make the dataset also available in an open format.

[19] According to Royal Decree 1720/2007 for the implementation of the personal data protection Act.

Datasets in unstructured formats

In these cases, as for example with traditional text documents (TXT, PDF, DOC, ODT), the only possibility to enable automated reuse is usually via natural language processing. Unfortunately, this process may result in final documents not containing any useful structured information in most cases, making the vast resources invested for automated processing useless.

Therefore, such documents **should not be considered a priority while selecting datasets for publication**. If the documents were based on previously collected data, as frequently occurs with all kind of reports, you should also publish this information separately in the most appropriate structured and reusable format in each case, and the report as companion documentation.

For example, we may have a PDF report on a mobility study in the region in which numerous statistics on traffic, transport, inter-urban mobility and the like are discussed. Then, if we want to publish them for reuse such statistics should be published in a structured format suitable for reuse, as CSV for example, rather than simply publish the report in its current format. Finally, we could also include the report as additional documentation.

Datasets in image, audio or video formats

This is what usually happens when the data we want to publish is a set of pictures, videos or sounds. Some examples are a bitmap with cartographic information or a photoset of touristic areas in a city, whose natural format is usually an image (such as JPG, GIF, TIFF, etc.) or in the form of audio and/or video (MP3, WMV, MP4, etc.). In these cases we must proceed as with any other normal dataset, **providing also the metadata required at the time of cataloguing**.

This situation should not to be confused with other cases in which such images are used as data screenshots, data image exportation or simply printed documents that have previously been scanned. In these cases, the **data cannot be catalogued as is** because it would be impossible to reuse. For a proper cataloguing, **datasets need to be published in the original structured format**, be it proprietary or not.

For example, we may want to publish the photographic archive of a region. In this case data will be directly published in any graphic format, such as JPG or similar, as that is its "natural" format.

However, we may also want to publish the budget of the Government of Aragon for example. In this new case it would not be appropriate just to scan the relevant documents and publish them as a JPG document or any other graphic format. Instead, we should access the original data source to be able to publish this same data in a more appropriated structured format, such as a CSV document or a spreadsheet.

Datasets in formats that require previous treatment

Using formats that **require a previous treatment to access datasets** should be avoided because that could be limiting direct reuse, such as those that require a preliminary ***decompression*** process or to provide a **password** for content access.

Examples of formats that can be commonly affected by this sort of limitations include ZIP, RAR and other *compression* formats - which are not really final representation formats and can also easily be protected by *passwords*.

We should avoid this kind of compressed formats whenever possible by following some simple rules:

- If we only publish a single distribution file it is better not to use any compression format, but the final format directly.
- If we have a dataset divided across multiple files, it would be convenient to group all data in a single file before publication (e.g. using multiple sheets in a single spreadsheet instead of several individual spreadsheets separately).
- If we are going to publish the same dataset in several different formats (for example XLS and JSON), each of that formats corresponds to a separate distribution and therefore should be published individually, not in a single compressed file.

Still, there may be some cases where you will need to publish multiple files corresponding to a single distribution. In these cases, you may use a compression format for that, but the format to be indicated in the distribution metadata must be that corresponding to the final representation, not the one that has been used just for compressed encapsulation.

For example, we publish information on public transportation in GTFS format that, according to the official specification, consist of a series of separate CSV files containing various data on stops, routes, schedules, etc.

According to the specification, these files must be collected into a single ZIP file, but when providing the corresponding format metadata for this file we must indicate GTFS (the final representation format) rather than ZIP (the encapsulation format).

Datasets with indirect access through directories, search engines and others

Sometimes it may happen that there is no way of providing direct access to a dataset, being necessary in such cases some form of **interaction with a search or selection tool** (usually through a web application) that returns the desired data (usually in HTML format).

When such tool also allows **data exportation in some other structured format** (e.g. CSV, XLS, etc.), the priority should be cataloguing such structured exportation formats that are being offered, with the recommendation of including also a reference to the original tool as a kind of additional information.

For example, if we have an online application of the Government of Aragon that allows us to access a directory of health centres (by location, province, municipality, town, etc.), we should not be cataloguing such application page directly. Instead we should provide the same information through any structured format, such as a CSV file, and be cataloguing that dataset, including a reference to the application as an additional documentation resource.

Datasets not digitized

When we meet datasets that have been published only by **non-electronic means**, usually in the form of printed publications, we will need to locate the original data source in databases or internal systems in order to extract the equivalent dataset in electronic format.

If the source does not exist or is not reachable, the only alternative possibility to be assessed in the medium term is to **digitize** the contents of these publications in **an appropriate electronic format for easy reuse**.

For example, we may have suggestion boxes in the various offices associated with the Government of Aragon where citizens can enter their suggestions written on paper. Then, in addition to maintain an archive of these suggestions, we should also store them in some database or spreadsheet so that they could be easily reused.