DELIVERABLE D3.3

Definition of ways to advertise the published datasets to Open Data catalogues

Project	Components Supporting the Open Data Exploitation
Acronym	COMSODE
Contract number	FP7-ICT-611358
Start date of the project	1st October 2013
Duration	24 months, until 31st September 2015

Date of preparation	August 2014
Author(s)	Oskár Štoffan, Gabriel Lachmann, Ivan Hanzlík, Tomáš Knap
Responsible for the deliverable	Oskár Štoffan EEA Communication Solutions
Email	oskar.stoffan@eea.sk
Reviewed by	Jan Kučera
Status of Document	Final
Version	1.0
Dissemination level	PU - Public

History

Version	Date	Description	Revised by
0.1	2014-07-20	The initial list of EU data catalogues	Oskár Štoffan, Jan Kučera
0.2	2014-07-21	The reduced list of candidate data catalogues created	Oskár Štoffan, Jan Kučera
0.3	2014-08-14	The deliverable document created, outline of the deliverable prepared	Oskár Štoffan, Jan Kučera
0.4	2014-08-20	The final list of selected data catalogues created	Oskár Štoffan, Jan Kučera
0.5	2014-08-20	The first version of deliverable presented and made available for review and comments to the partners	Oskár Štoffan, Jan Kučera
0.6	2014-08-25	DCAT-CKAN metadata fields mapping created, Metadata chapter created	Oskár Štoffan, Jan Kučera
0.7	2014-08-26	Methods of soliciting user feedback added	Oskár Štoffan, Jan Kučera
0.8	2014-08-28	Statistics added	Oskár Štoffan, Jan Kučera
0.9	2014-09-02	Conclusion added	Oskár Štoffan, Jan Kučera
0.10	2014-09-03	General findings added	Oskár Štoffan, Jan Kučera
0.11	2014-09-05	Selection criteria described	Oskár Štoffan, Jan Kučera

Table of Contents

Annexes:

- Attachment A - EU data catalogues.xlsx
- Attachment B - Candidate data catalogues.xlsx
- Attachment C - Final list of selected data catalogues.xlsx
- Attachment D - CKAN-DCAT metadata fields mapping.xlsx

1. Executive summary

The purpose of this deliverable is to define ways how to advertise each dataset from Deliverable 3.1 ‘Final version of the selected datasets list’ on various Open Data catalogues (country specific, EU wide, Global level).

The document provides an overview of methods and criterions applied in the process of selection of data catalogues. The final list of catalogues presented in Attachment C provides developers with detailed information required in the process of publishing including the possibilities to use automated form of publishing (via APIs). Mechanisms to capture public demand and user feedback from those catalogues as defined too.

Since the outcome of the delivery depends on responsiveness of data catalogue owners, there are still potential candidates for final selection due to missing information awaited in time of completion of this document. In case these information will be provided, the outcome of the deliverable can change (Attachments B, C) and the final list will be extended if appropriate.

2. Deliverable context

2.1 Purpose of deliverable

COMSODE will publish 150 datasets - see the deliverable D3.1 for the list of datasets selected for publication. The datasets are from various European countries, including Slovak Republic, Czech Republic, Italy, Netherlands, Spain and Albania. They cover more than 20 domains, e.g. business, e-health, environment, education or demographical statistics. In order to make them known to people (data users/consumers) we will (among other things) submit metadata about those datasets to various data catalogues. The goal of this deliverable is to identify all relevant data catalogues for this and map one or more relevant (in terms of location and domain) data catalogues to each dataset.

The actual submission of metadata will be done after each dataset is actually published. This submission is in the scope of Task 4.3 'Support work as required to meet the goals of the project'. Thus this deliverable also compiles all other relevant information about each data catalogue so that submission can be performed quickly and efficiently later on, including possible use of APIs to automate the whole process.

2.2 Related documents

- DOW_COMSODE_B_v16_full.doc, page 6, 19-20, 24-25
- Deliverable 3.1: Final version of the selected datasets list

2.3. List of attachments

The following documents are attached to this document:

- **Attachment A - EU data catalogues:** Shows the list of data catalogues as a result of research of existing Open Data catalogues across Europe.
- **Attachment B - Candidate data catalogues:** Shows the reduced list of candidate data catalogues after applying and evaluating the mapping of preselected datasets (D3.1). At this stage further criterions are evaluated for the final selection based on information collected from catalogue owners.
- **Attachment C - FINAL LIST OF SELECTED DATA CATALOGUES:** Presents the output of this deliverable - the final list of data catalogues chosen by the COMSODE consortium for submitting metadata information about selected datasets.
- **Attachment D - CKAN-DCAT metadata fields mapping:** Shows the list of metadata fields usable with each version of CKAN according to Attachment C - FINAL LIST OF SELECTED DATA CATALOGUES and the appropriate mapping between DCAT fields specified by W3C [9].

3. Methodology used

3.1 Methodology

These steps were made to achieve the final version of D3.3:

1. Table for the most significant data catalogues across Europe created (Attachment A).

- Columns for basic identification and description of catalogues created
- Columns for initial criteria and its evaluation created
    - criteria (C1, C2) for assessment of catalogues from geographic/domain perspective
- Information collected and criteria evaluated

2. Table for the reduced list of data catalogues based on evaluation of criteria in step 1 created (Attachment B).

- Columns for detailed description of preselected catalogues created
    - technology (cataloguing platform)
    - catalogue owners and appropriate contacts
    - existing consumer feedback gathering possibilities
- Columns for additional criterions (C3, C4.1 - C4.5) created
    - criteria considering publishing possibilities for COMSODE
    - criteria considering harvesting between catalogues
    - criteria considering manual / automated ways of publishing
- Information collected and criteria evaluated

3. Table for the final selection of data catalogues based on evaluation of criteria in step 2 created (Attachment C).

4. Table describing metadata fields for each CKAN versions and the appropriate mapping between the DCAT[1] fields specified by W3C created (Attachment D).

5. Research of existing methods of gathering data consumer feedback currently in use performed. Other possibilities extending current methods were proposed.

3.2 Partners contribution

This deliverable is a joint effort of most partners of the COMSODE project. The selection process was driven by EEA and intensively consulted by all partners during July and August 2014.

[1] http://www.w3.org/TR/vocab-dcat/

4. Structure of the document

This deliverable represents the final version of the list of data catalogues that were selected for publication of metadata describing the 150 datasets selected for publication in the deliverable D3.1.

- **Chapter 5 Initial list of catalogues**

This chapter describes the process of creation and collection of information about the initial list of Open data catalogues across EU (Attachment A). It also contains description of the methodology of evaluation of criterions defined.

- **Chapter 6 List of candidate data catalogues**

This chapter describes the next step of research performed on reduced list of Open Data catalogues, definition of the additional criteria for the final selection (Attach. B).

- **Chapter 7 Final list of data catalogues**

This chapter contains information about the selection process of final list of data catalogues (Attachment C) and the methodology of evaluation of criterions added in chapter *6. List of candidate data catalogues*.

- **Chapter 8 Contacting data catalogue owners**

This chapter describes the topics and the time schedule of communication with the data catalogue owners.

- **Chapter 9 Metadata**

This chapter describes the possibilities of publishing metadata for a dataset from a developer's perspective. The main focus is on CKAN cataloguing platform since the vast majority of data catalogues in the final selection are based on CKAN. The mapping between CKAN of different versions and DCAT is described as well as the methodology of publishing DCAT metadata fields missing in the CKAN specification, both for dataset and for distribution metadata.

- **Chapter 10 Methods of soliciting user feedback about datasets**

This chapter deals with existing ways how the user feedback is gathered by data catalogue owners. Different methods enabling users to add comments or rate datasets are described along with some new proposed techniques mentioned as well together with the potential benefits of using such techniques.

- **Chapter 11 General findings**

This chapter describes several interesting global activities of greater importance in the field of Open Data publication as a subsidiary outcome of the research of existing data catalogues.

- **Chapter 12 Conclusion**

This chapter concludes the deliverable.

5. Initial list of data catalogues

This chapter describes how the initial list of data catalogues was created. The purpose of this work was to collect as many existing data catalogues as possible regardless of country they belong to (within Europe). Data catalogues providing access to data at EU-wide or global level are listed as well. The outcome of this list is presented in Attachment A.

The initial list of catalogues was derived from existing lists, namely [1 - 6] and the basic catalogue functionality was verified (search for random datasets, preview of samples of a dataset, download of dataset) prior to adding a catalogue to the list.

The column 'Candidate for final selection' (if marked 'Yes') represents data catalogues proceeding to the list of candidate data catalogues (Attachment B). This is based on evaluation of initial criterions (C1 and C2) as follows:

ID	Name	Description
C1	Country	Is the country listed in the list of selected datasets[2]?
C2	Relevant dataset	Is there a dataset available that is relevant (from geographical, domain perspective) for the catalogue[3]?

The order in which the criterions are evaluated is from top to bottom.

Only catalogues passing criterion C1 are further evaluated against criterion C2.

5.1 Mandatory columns

For each data catalogue a list of mandatory columns were created:

- Data catalogue name / description
- URL (URL of the website where the catalogue resides)
- Country (the country a catalogue belongs to)

5.2 'Suitable candidate' column

Value contained in this column represents the result of evaluation C1 (based purely on the country to which the data catalogue belongs comparing to the list of countries selected datasets (D3.1) belong to).

[2] See chapter *5.2 'Suitable candidate'.*

[3] Multiple catalogues with the same group of relevant datasets which is a subset of other groups of datasets already assigned to other data catalogues are considered not meeting this criterion.

The list of countries the selected datasets (D3.1) belong to[4]:

- Slovak Republic
- Czech Republic
- Albania
- Spain
- Italy
- Netherlands
- EU-wide

5.3 'Suitable datasets for publishing' column

A column 'Suitable datasets for publishing' was created to show possible mappings between the catalogue and the available datasets (no harvesting between catalogues considered at this stage).

This column contains values:

- no relevant datasets available for this catalogue (domain specific or regional[5] catalogue not matching any of the selected datasets)
- Dataset ID (list of datasets suitable to publish on the catalogue)

5.4 'Candidate for final selection' column

A column 'Candidate for final selection' shows the result of evaluation of criterion C2 based on values in column 'Suitable datasets for publishing'. If the column contains value 'Yes', the data catalogue proceeds to the list of candidate catalogues (Attachment B).

5.5 Statistics of collected data catalogues

The total number of catalogues collected is 191. Figure 1 shows the distribution of collected data catalogues by countries they belong to. The affiliation of each data catalogue to a country is based purely on the geographical relation of datasets published.

[4] US datasets (Dataset ID: EH-2, EH-6, EH-7) will not be published within the pilot implementation
[5] Assuming that in general the regional (municipal, provincial, city, etc.) catalogues only contain data relevant to the specific area (catalogue of Florence doesn't contain Italian data on the national level).

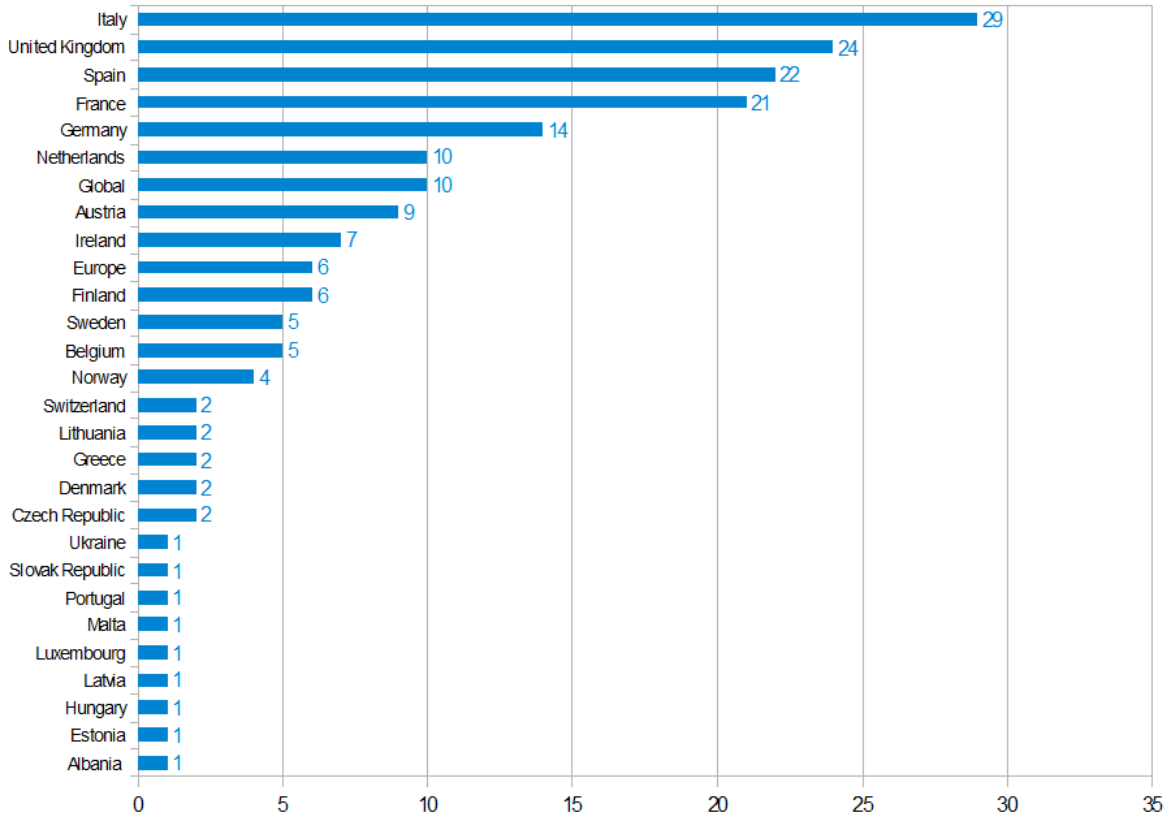

Figure 1: Distribution of data catalogues by countries

81 data catalogues out of 191 passed the first criterion defined in chapter *5.2 Suitable candidate*. Figure 2 shows the distribution of these catalogues by countries.

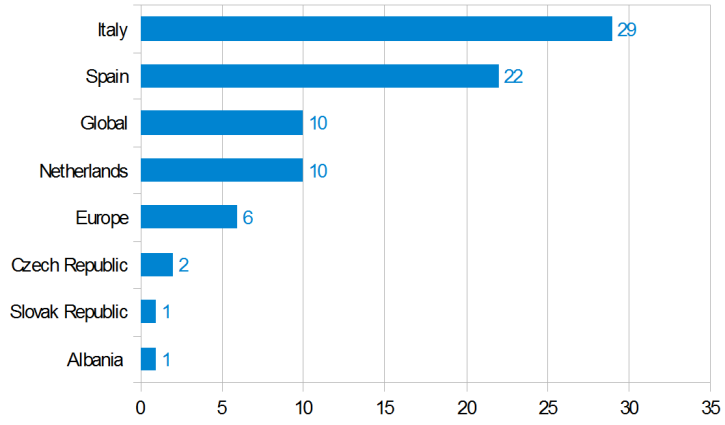

Figure 2: Distribution of catalogues passing the first criterion by countries

25 data catalogues out of 81 passed the second criterion defined in chapter *5.4 Candidate for final selection* and have proceeded to the Candidate data catalogues list (Attachment B). Figure 3 shows the distribution of these catalogues by countries.

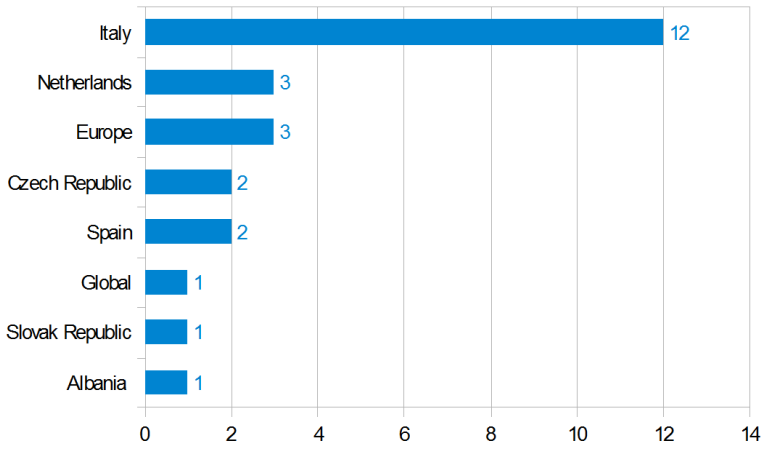

Figure 3: Distribution of candidate data catalogues for final selection by countries

6. List of candidate data catalogues

Applying criteria to the initial list of data catalogues (Attachment A) resulted in Candidate catalogues list (Attachment B). This chapter describes how this reduced list was created, what additional criteria have been added and evaluated. In order to collect additional information needed, in most cases, it was essential to contact the data catalogue owners directly. In many cases the evaluation process was highly dependent on the responses provided (colored cells are to be filled upon receiving responses from catalogue owners).

6.1 Mandatory columns

Some columns have been taken from previous list (Attachment A) and added to the rest of new mandatory columns. The mandatory columns in the list have been filled without the need to contact the catalogue owners:

- Country (taken from previous list)
- URL (taken from previous list)
- Owner
- Owner contact[6]
- Type of owner[7]

6.2 'Who can add entries manually?', 'Who can add entries via api?' columns

Not everyone whose intention is to publish a dataset is accepted to do so. It strongly depends on the policy of the catalogue owner. In general the national level data catalogues owned by governmental organisations do not allow 3rd parties to publish datasets even if the permission of dataset owner is given to them.

These two columns describe this policy both for manual and automated type of publishing (there are columns dedicated to automated ways of publishing in the list)

- only the data catalogue owner
- only the dataset owner
- anybody, given permission from the owner of dataset
- anybody, even 3rd party, must be registered/approved by the data catalogue owner
- anybody, even 3rd party, even anonymously
- not possible to add entries

The possibility to publish a dataset in a catalogue is the most important criterion for a data catalogue in order to be accepted for the final selection. The values presented are in most cases dependent on the responsiveness of the owner, however there can be additional

[6] Email contact was preferred, however in many cases a web based contact form was available only.
[7] GOV or NGO (governmental or non-governmental organisation)

information available that can be of some help to make it clear even without the owner's response (registration process description, public discussions, knowledge base, FAQ, etc.). In case such information was provided on the website of the catalogue, it was considered sufficient to accept the catalogue for publishing ('Notes about adding entries' column contains such information).

6.3 'Notes about adding entries' column

As mentioned above this column contains additional information about adding entries (catalogue records) in case they are provided by the catalogue owner or the website itself.

6.4 'Technology' column

This column was added to indicate the cataloguing platform used for each catalogue. Collection of this information was partially dependent on responsiveness of the owner, however in most cases it was quite straightforward to identify the platform based on information published on the website of the catalogue.

- CKAN
- DKAN
- Socrata SODA
- OData Open Data Protocol
- other (custom built platform)
- other (custom built platform with 3rd party API support)[8]
- unknown (no response)

6.5 'API URL', 'URL API Doc' and 'Info for developers' columns

The first two columns indicate the appropriate URL of API interface and the documentation for developers providing guidance on how to create the code that calls the API (if the dataset is intended to be published in automated way). The third column 'Info for developers' provides more information that can be of some further help for developers if provided by the site. See also chapter *9. Metadata* which is closely related to APIs.

6.6 'Dataset comments', 'Dataset ratings', 'Social networks', 'feedback gathering notes' column

In order to solicit feedback about the dataset from data consumers, some catalogues provide users with direct possibility of commenting and/or rating of the dataset. Along with the direct (custom built-in) features for comments/ratings consumers are provided with a possibility to

[8] Some custom built cataloguing platforms provide support also for 3rd party APIs that are commonly used (like CKAN).

comment (and share) a dataset on different types of social networks quite often. It is possible for the catalogue owner to collect this kind of feedback on different levels, see chapter *10.1 Methods of soliciting feedback currently in use*. In case some additional information was provided by the catalogue owner, it was recorded in 'feedback gathering notes' column.

6.7 'Harvesting of catalogue records' column

In theory the structure of data catalogues that belong to a geographic area should be hierarchical (EU-wide level catalogue harvests catalogue records from national catalogues harvested from municipal/regional data catalogues). In order to avoid any redundancies in publishing same datasets on catalogues of different or the same levels, we do check those harvesting settings with the owners. The 'Harvesting of catalogue records' column contains the information collected from catalogue owners.

- No harvesting (self explanatory)
- Irrelevant (publishing either not allowed or it is the only catalogue for the particular country in the list)
- links to catalogues or list of catalogues the catalogue harvests records from

6.8 Statistics of candidate data catalogues

6.8.1 Division of candidate data catalogues by the type of owner

The following figures (figure 4, figure 5) show the distribution of candidate data catalogues by the type of the owner. From this perspective the catalogues can be divided as:

- GOV - owned and maintained by organisation/s owned by the government (on national or EU level)
- NGO - owned and maintained by a group or community of people/enthusiasts interested in and passionate about Open Data in general.

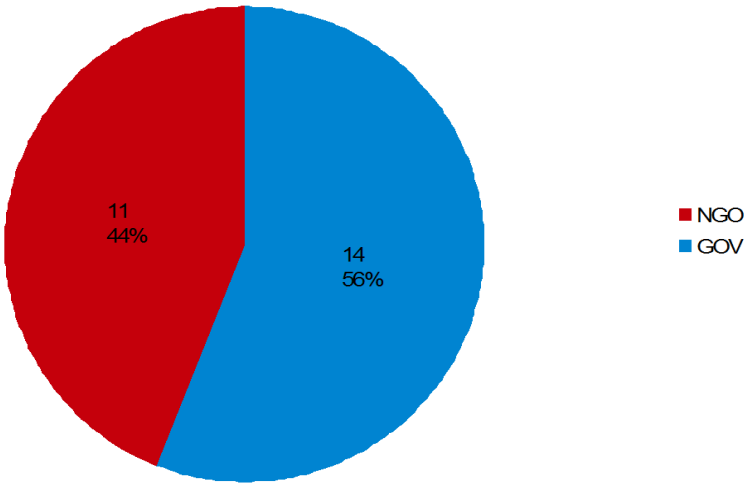

Figure 4: Division of candidate data catalogues by the type of owner

Some catalogues could also represent large projects as a collaboration of different types of organisations and contributors on different levels (NGO, GOV), in such case the catalogue was considered as GOV[9].

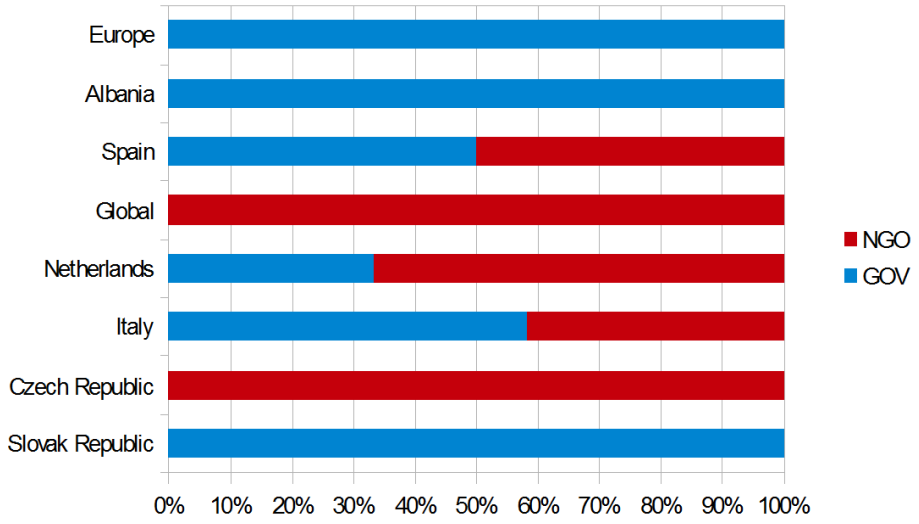

Figure 5: Division of candidate catalogues by the type of owner for each country

6.8.2 Division of candidate data catalogues by cataloguing technology

Figure 6 shows the utilization of different cataloguing platforms across the list of candidate data catalogues. A single catalogue is based on one cataloguing platform (CKAN, Socrata, custom built etc.), however it can be accessible by multiple 3rd party APIs of CKAN, DKAN, Socrata, SDMX, OData etc.[10]. Since there are catalogues in the list that we still have no response from the owner, the platform is marked as 'unknown' in such cases, however in many cases the platform can be identified without owner's help (the appropriate information and documentation is available directly on the website of the catalogue). The majority of catalogues uses CKAN cataloguing platform. Figure 7 shows the utilization of platforms separately for each country candidate catalogues belong to. Figures 6 and 7 also shows the subdivision of custom built catalogues marked as 'other' to those accessible using CKAN API and the rest.

[9] As an example http://publicdata.eu/ represents a catalogue of such type.
[10] As an example http://www.opendatahub.it/ represents a catalogue of such type. The cataloguing platform is custom built (Hammock) which conforms to W3C DCAT standards and provides support for multiple 3rd party interfaces (CKAN, Socrata, OData APIs) for accessing it.

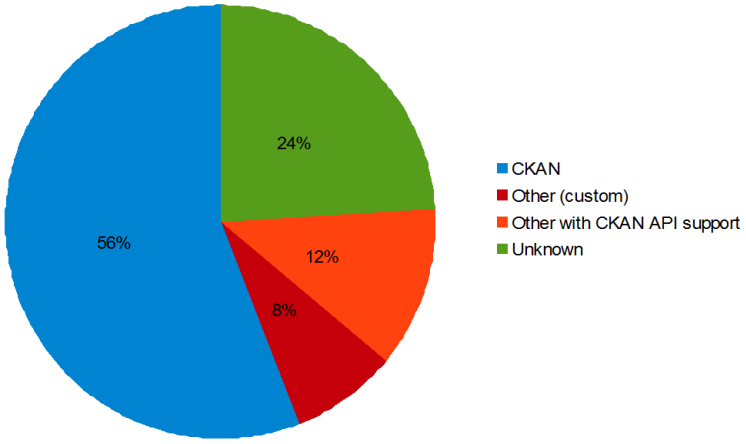

Figure 6: Utilization of cataloguing platforms across candidate catalogues

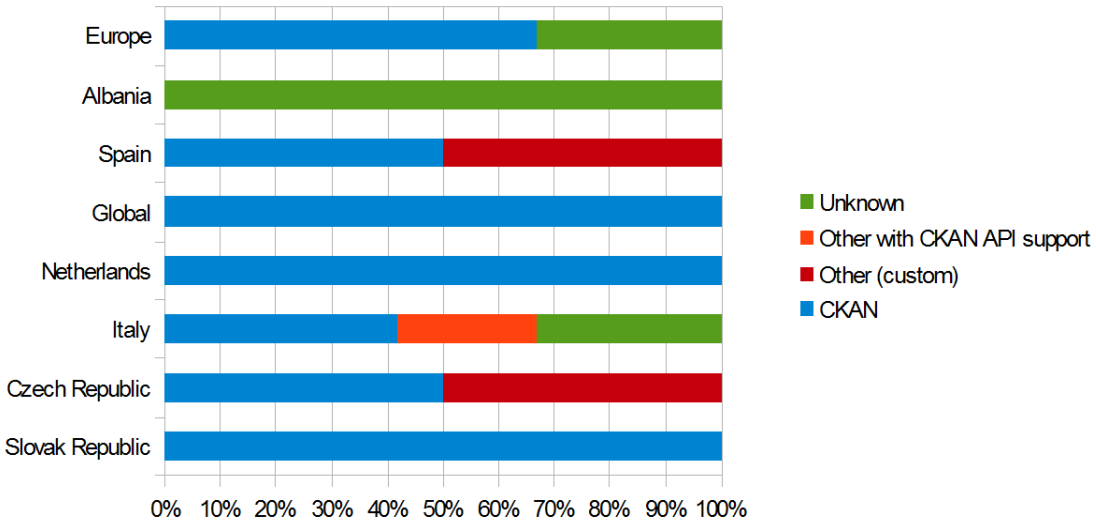

Figure 7: Utilization of cataloguing platforms for each country

7. Final list of selected data catalogues

The final list of selected data catalogues (Attachment C) represents the outcome of this deliverable. The list of candidate data catalogues (Attachment B) with 25 data catalogues represents a base for the final selection. The final list itself represents a reduced list of candidate data catalogues, the columns contained (described in chapter *6. List of candidate data catalogues)* are the same. The only column added is column 'List of datasets to publish (total number)' containing information about the datasets selected for each data catalogue.

The following criterions were required to be fulfilled for the candidate data catalogues to proceed to the final selection:

ID	Name	Description
C3	Publishing allowed	Is adding new catalogue records allowed by the owner for COMSODE[11]?

C3 criterion has to be met for every data catalogue proceeding to the final list of selected data catalogues (Attachment C).

In order to achieve the goal of deliverable i.e. to map one or more relevant data catalogues to each dataset and to be as efficient as possible at the same time, there are prioritization sub-criteria defined in case two (or more) data catalogues passing the C3 criterion are available for the same dataset.

The following sub-criterions should be evaluated for prioritization of catalogue for each dataset assigned to two or more data catalogues (column 'Suitable datasets for publishing' of Attachment A).

ID	Name	Description
C4.1	Ownership	one (or more) of the COMSODE consortium partners is the catalogue owner or is involved in operating the catalogue in some way
C4.2	Geographic hierarchy	the catalogue is on lower level of geographic hierarchy[12]
C4.3	Harvesting	records of the catalogue are harvested by another catalogue/s on the same level of geographic hierarchy

[11] Confirmed either by the owner directly (email) or the related information is available on the web site of the catalogue. There could be further conditions to be fulfilled i.e. registration of user/organisation and subsequent approval by the owner.

[12] The geographic hierarchy is defined (from highest to lowest level): Global, Europe, Country, municipal/regional.

C4.4	Automated adding of entries	automated adding of entries is available
C4.5	API for adding entries	CKAN API is available

If there are more than one catalogues left after evaluation of a sub-criterion or none of the evaluated catalogues satisfies the sub-criterion, the next sub-criterion should be evaluated. The order of evaluation is from top to bottom. If there are more than one data catalogues left after evaluating the C4.5 sub-criterion, one of them should be randomly selected for the final list.

The final list of selected data catalogues (Attachment C) comprises 7 data catalogues. The table below represents the shortened overview of selected catalogues:

Country	**Data Catalogue (URL)**	**List of datasets to publish**
Slovak Republic	http://data.gov.sk	all datasets for Slovak Republic
Czech Republic	http://cz.ckan.net	all datasets for Czech Republic not published as LOD[13]
Czech Republic	http://linked.opendata.cz	all datasets for Czech Republic published as LOD
Spain	http://opengov.es	all datasets for Spain
Italy	http://it.ckan.net	all datasets for Italy
Netherlands	http://www.rotterdamopendata.nl	all datasets for Netherlands
Global	http://datahub.io	all datasets for Albania, EU_EMA_PIL_01, EU_EMA_SPC_01

[13] Linked Open Data

8. Contacting data catalogue owners

All data catalogue owners from the Candidate data catalogues list (Attachment B) have been contacted already.

Areas discussed already:

- who can add entries to the catalogue
- publishing possibilities:
    - manual only
    - manual or automated (via API),
    - API URL and API documentation
- harvesting of records between catalogues (Italian, Dutch, European level data catalogues only)
- feedback gathering methods

Attachment B contains a sheet ‘Contacting owners’ where the detailed description of ongoing communication with catalogue owners can be found with appropriate dates and inquiries sent.

Optionally there are further areas to discuss:

- potential plans on how to improve the process of feedback gathering from data consumers
- steps towards more standardized support for metadata schemas and vocabularies

9. Metadata

According to D5.1 requirements, ODN should publish metadata about datasets in compliance with W3C DCAT (DCAT+VoID) scheme described in more detail here:

- http://www.w3.org/TR/vocab-dcat/ (DCAT)
- http://www.w3.org/TR/void/ (VoID) - in case of Linked Data

Properties of the classes presented in referenced documents (see links above) should be considered as mandatory metadata fields while publishing datasets in ODN data catalogue.

On public data catalogues side the support for such standardized vocabularies is not always ensured. Even if the catalogue uses CKAN platform, differences can be found in particular versions of CKAN. In order to ensure DCAT compliance, CKAN uses the Ckanext-dcat plugin to expose and consume metadata from other catalogues using documents serialized using DCAT[14].

For us to submit metadata to public catalogues, the metadata schema has to be determined. The schema should be properly described in the catalogue's documentation, ideally expressed in the description of service consuming a predefined source containing the schema (XML document, etc.). CKAN based catalogues expose the metadata schema directly via the API's function calls[15]. The 'URL API Doc' column of Attachment C contains link to appropriate API documentation of the platform used by the data catalogue. For further details see Attachment D 'CKAN-DCAT metadata fields mapping'.

9.1 CKAN-DCAT mapping

Since the final list of selected data catalogues contains exclusively catalogues based on CKAN (except linked.opendata.cz), this chapter provides developers with a proper methods how to submit DCAT compliant metadata to CKAN metadata scheme accepted by the target catalogue.

A separate file (Attachment D) was created to:

- name and describe the metadata fields available for use in each CKAN version[16]
- show the mapping of metadata fields between CKAN and DCAT specification

For each version of CKAN the following columns were created:

- DATASET (metadata fields dedicated for dataset description)
- Description (description containing also the data type)
- DISTRIBUTION (resource) (metadata fields dedicated for distribution description)
- Description (description containing also the data type)

For DCAT the following columns were created[17]:

[14] For more details see: https://github.com/ckan/ckanext-dcat

[15] The CKAN developer documentation http://docs.ckan.org/en/latest/ provides information about underlying metadata schema for each version of CKAN.

[16] CKAN versions from the final list of data catalogues (Attachment C) are included only.

- DATASET (RDF property) (properties of dcat:Dataset class as defined by W3C specification)
- DISTRIBUTION (RDF property) (properties of dcat:Distribution class as defined by W3C specification)

9.1.1 Missing CKAN metadata fields for dataset

The mapping is DCAT -> CKAN (we are looking for an appropriate dataset field in CKAN for each DCAT dataset field).

Metadata fields (for dataset) of DCAT not matching any available CKAN metadata field are in blue colored cells. For these fields (as well as for any additional field describing a dataset) 'extras' field can be used in CKAN (supported by all versions of CKAN). See chapter *9.1.3 Customizing dataset and resource metadata fields in CKAN* for details about 'extras' metadata field in CKAN.

9.1.2 Missing CKAN metadata fields for distribution

The mapping is DCAT -> CKAN (we are looking for an appropriate distribution field in CKAN for each DCAT distribution field).

Metadata fields (for distribution) of DCAT not matching any available CKAN metadata field for all CKAN versions in the list are in red colored cells.

For the following DCAT metadata fields (for distribution) an available CKAN field cannot be found:

- license (dct:license)
- rights (dct:rights)
- release date (dct:issued)

Since CKAN does not support an 'extras' filed for distribution, this cannot be resolved as described in case of dataset fields. In many cases the license should be valid for the dataset regardless of the distribution, however there are cases when a specific license needs to be mentioned for each distribution. If there is only one license to mention, we recommend to use CKAN's 'license_id' ('license' for CKAN v1.7.4) metadata field dedicated for dataset for the license description. If there are multiple different licenses to be mentioned for a dataset, we recommend to create separated CKAN dataset instances each with appropriate 'license_id' ('license' for CKAN v1.7.4) field for every subset of distributions with the same license. The latter recommendation is valid also for the 'rights' DCAT property, however there is no 'rights' field present for dataset in CKAN, so 'extras' filed should be used instead. The same is valid for DCAT property 'release date' (in case of CKAN v1.7.4). See chapter *9.1.3 Customizing dataset and resource metadata fields in CKAN* for details about 'extras' metadata field in CKAN.

[17] The detailed description of properties (property = metadata field) can be found here: http://www.w3.org/TR/vocab-dcat/

9.1.3 Customizing dataset and resource (distribution) metadata fields in CKAN

By default CKAN offers quite an extensive list of metadata describing a dataset and its distributions, however it is often required to extend this list by domain-specific, or other custom metadata.

"*Storing additional metadata for a dataset beyond the default metadata in CKAN is a common use case. CKAN provides a simple way to do this by allowing to store arbitrary key/value pairs against a dataset when creating or updating the dataset. These appear under the "Additional Information" section on the web interface and in 'extras' field of the JSON when accessed via the API. Default extras can only take strings for their keys and values, no validation is applied to the inputs and you cannot make them mandatory or restrict the possible values to a defined list*"[18] [7].

While creating 'extras' fields, follow the detailed naming convention in the separate sheet 'Extras fields for dataset' and 'Extras fields for distribution' respectively within Attachment D.

9.2 VoID support in CKAN

CKAN doesn't have a specific support for VoID as an RDF Schema vocabulary for expressing metadata about RDF datasets[19].

VoID covers four areas of metadata [8]:

- General metadata (following the Dublin Core model)
    - General metadata helps potential users of a dataset to decide whether the dataset is appropriate for their purposes. It includes information such as a title and description, the license of the dataset, and information about its subject.
- Access metadata (how RDF data can be accessed using various protocols)
    - Datasets in VoID are defined as sets of RDF triples. But the actual RDF triples are not part of the VoID description. Instead, *access metadata* is used to describe methods of accessing the actual RDF triples.
- Structural metadata (structure and schema of datasets, useful for querying and data integration)
    - This includes information such as the vocabularies used in the dataset, statistics about the size of the dataset, and examples of typical resources in the dataset.
- Description of links between datasets (helpful for understanding how multiple datasets are related and can be used together)

[18] By using CKAN's IDatasetForm plugin interface, a CKAN plugin can add custom, first-class metadata fields to CKAN datasets, and can do custom validation of these fields.
[19] For more information about RDF support in CKAN, read the CKAN Global User Group responses here:

https://groups.google.com/forum/#!topic/ckan-global-user-group/wB_UMvvH68s

10. Methods of soliciting user feedback about datasets

As described previously in chapter *6.6 'Dataset comments', 'Dataset ratings' and 'Social networks' column,* part of the research was focused also on collecting information about the methods currently in use by data catalogue owners to gather data consumer feedback.

10.1 Methods of soliciting feedback currently in use

During the research the information about different mechanisms was collected:

- by means of using (searching, browsing, previewing, downloading) of available catalogue records and manually searching for possibilities to provide some feedback
- by collecting the appropriate information about mechanisms in place from data catalogue owners

The following methods of soliciting feedback currently in use have been identified:

General level feedback[20]:

- discussion forums - dedicated for general level purposes, can (its dedicated sections) serve for collecting user's suggestions, ideas, data requests etc.
- dedicated email contacts

Dataset comments:

- adding a direct comment for selected dataset (discussion forum as built-in feature) - with possibility to reply to this comment similarly to discussion forum in general
    - anonymous user fills in the 'name' (nick), 'subject' and the 'comment' fields and confirms adding the comment to the forum
    - sometimes registration is required, so in order to add a comment, user has to log in first
    - users can subscribe to get email updates from a discussion
- adding a comment via social networks - along with the possibility to comment the dataset and further discuss it this method provides an additional feature to share the selected dataset record between social network users, connections, followers etc. The social network features can be integrated with the platform on different levels:
    - Fully integrated[21] - the user is required to log in to social network and use its features while the comments are shown also in the catalogue (or at least number of comments/shares/tweets etc. is shown)
    - Partly integrated - the user is required to log in to social network and use its features, the comments are not shown in the catalogue, however the catalogue

[20] As an example http://data.gov.uk/ is worth mentioning, especially the /data-request section dealing with the requests of registered users for new datasets. Users are provided with a discussion forum-like interface where others can reply to requests and further discuss them.
[21] As an example of full Disqus integration with the catalogue platform http://datahub.io/ could be mentioned.

owner is notified about the comments by email or by social network built-in features (the comment appears also on the Facebook page of the catalog etc.)

- No integration - the user is able to comment and share the dataset, however no feedback to the catalogue owner is in use

- sending a comment to catalogue owner/dataset owner/dataset maintainer by email (user has to search the website or metadata for an appropriate contact)

Dataset ratings:

- User can add ratings (like/dislike, X of Y) - log in locally or through social networks is usually required. There are two possible ways catalogues present these ratings:
  - custom built-in rating feature[22]
  - on social network - the same options for integration applies for it as in case of dataset comments (in case of full integration with the catalogue number of likes/dislikes etc. appears)

10.2 Proposed methods of soliciting user feedback

In general there are several ways how feedback can be provided by data consumers as well as gathered by data publishers. Even when a potential consumer of data searches, previews or downloads some data s/he provides some feedback unknowingly in terms of usage statistics that can be beneficial for publisher or catalogue owner in some way [10].

From this perspective (whether data consumer is involved in the process of providing feedback knowingly or not), we can divide the methods in two separate groups:

- methods involving users unknowingly (gathering usage statistics)
- methods involving users knowingly (gathering user feedback)

10.2.1 Gathering usage statistics

The process of gathering of usage statistics could be done on different levels:

- data storage level (gathering of usage statistics using technologies deployed on data storage - where the dataset is stored), these could include collection of the following statistics:
  - number of dataset downloads
  - number of API calls
  - number of unique IPs/sessions
- data catalogue level (gathering of usage statistics using technologies deployed on data catalogue - where the catalogue record is stored), these could include collection of the following statistics:
  - number of search results the dataset was included in (Spinque's Search usage logs could complement the statistics as well)

[22] Custom built-in solution providing users with dataset rating feature is used in case of http://www.datiopen.it/ catalogue (number of stars is used in this case).

- On-site activity (via Analytics) - number of clicks per dataset, etc.

For data owner the results of such statistics can be of some help to manage the performance of the data storage and its accessibility. Since the performance and accessibility is always limited, the statistics can reveal the real demand for a particular dataset that can lead to utilize mechanisms allowing to set limits of concurrent requests and amount of data transferred to preven one or few clients from degrading service for others. If the usage statistics show high demand for a particular dataset long-term, automated methods and rules can be set for moving such data to higher performance storages (and vice versa).

10.2.2 Gathering user feedback

As it was briefly mentioned in chapter *10.1 Methods of soliciting feedback currently in use,* these methods should focus on the following information to collect:

- user comments
- user ratings
- suggested corrections of data (errors found, etc.)

In order to collect this information cataloguing platforms could provide support for it. Some platforms like CKAN already do (partly). The methods in use could be:

- customized solution built in the cataloguing platform
  - own discussion forum, mailing list (comments, suggested corrections of data), survey
  - own rating system (ratings - like/dislike, X of Y, etc.)
  - own bug and issue tracking system (suggested corrections of data)
- 3rd party solution integrated with the cataloguing platform
  - public discussion forums, domain related etc. (comments, suggested corrections of data)
  - social networks (comments, ratings, suggested corrections of data)
  - bug and issue tracking systems (suggested corrections of data)
  - cataloguing platform to provide an API for 3rd party application developers that provides the application users with ability to comment/rate/suggest corrections[23] using the application
    - this option requires a discussion forum to be available in the cataloguing platform where the comments/ratings/suggested corrections will be published
- combination of the methods mentioned above
- no integration with the cataloguing platform (metadata schema extension)
  - author's/maintainer's contact listed in metadata

[23] In case a data catalogue contains a list of downloadable applications reusing a particular dataset/s published on the catalogue.

- discussion forum (public or belonging to data owner) link listed in metadata
- bug and issue tracking system link listed in metadata

In first three options mentioned above the feedback is provided to catalogue owners primarily (feedback information is collected on data catalogue level). In case the data owner is not the same as data catalogue owner, there could be an automated interface (API) provided by the catalogue ensuring the possibility for data owner to collect these data from catalogue owner and enable automated processing and evaluating of data if needed. In case the owner of data does not make use of such automated process, a simple access to the feedback information collected should be established for the representatives of the data owner.

The fourth option provides feedback only to the data owner.

11. General findings

During the research we recognized activities with ambitions to improve the Open Data publication, use and reuse, on different levels (locally, internationally, globally).

Some of these activities represent the outcome of projects (mostly as EU initiatives) with limited duration, however might represent the intermediate result as a base for future data publishing platform of a greater significance.

Opendata.eu

Publicdata.eu is one of the outcomes of a EU co-funded project, LOD2, on which more information can be found here: http://lod2.eu.

The portal was developed as a pilot of a future pan-European Open Data portal, the web site is still running, but the project will end its activities on August 31st 2014.

Dati.piemonte.it (HOMER project)

The dati.piemonte.it team is the leader of a federation of opendata portals that developed both through a cooperation with public administrations in Italy and also in the context of the HOMER project for data harmonization across regions in the Mediterranean area. The project is co financed by the European Regional Development Fund (ERDF).

The federation is open to anyone who wants to join. Currently there are partners from Italy, spain, Greece, Malta, France, Slovenia, Cyprus and Montenegro involved. Information about technical and practical requirements to join the federation can be found at http://homerproject.eu/project/homer-federation.

The overall goal of HOMER is to contribute to unlock the full potential of the Public Sector Information in the Mediterranean space, by contributing to make the all area a competitive territory, able to match global competition and to ensure a sustainable growth and employment for the next generations.

Datiopen.it

StatPortal OpenData (http://www.opendata.statportal.it), the open source framework used to implement DatiOpen is able to harvest all the data published in portals which expose CKAN API, Socrata Open Data API and SDMX (like I.STAT and EUROSTAT data catalogue). The harvesting process acquires all the meta information and is also able to analyze attachments and/or data links (CSV, SHP, XLS, etc.), import them into his PostgreSQL repository, analyze them to discover data content (e.g. statistical, geographical coordinates, administrative units, ecc.). The new updated release of this harvester is currently being prepared and will be available before December 2014.

The harvest comes from results achieved during a research and development project ODINet about harvesting, indexing and searching open data with semantic techniques in which datiopen.it is engaged (for more details see http://www.odinet.sister.it).

Openaire.eu

OpenAIRE aims to support the implementation of Open Access in Europe.

Open Access is the immediate, online, free availability of research outputs without restrictions on use commonly imposed by publisher copyright agreements. Open Access includes the outputs that scholars normally give away for free for publication; it includes peer-reviewed

journal articles, conference papers and datasets of various kinds.It provides the means to promote and realize the widespread adoption of the Open Access Policy, as set out by the ERC Scientific Council Guidelines for Open Access and the Open Access pilot launched by the European Commission.

OpenAIRE, a three-year project, will establish the infrastructure for researchers to support them in complying with the EC OA pilot and the ERC Guidelines on Open Access. It will provide an extensive European Helpdesk System, based on a distributed network of national and regional liaison offices in 27 countries, to ensure localized help to researchers within their own context. It will build an OpenAIRE portal and e-Infrastructure for the repository networks and explore scientific data management services together with 5 disciplinary communities. It will also provide a repository facility for researchers who do not have access to an institutional or discipline-specific repository.

OpenAIREplus (2nd Generation of Open Access Infrastructure for Research in Europe) is a 30 month project, funded by the EC 7th Framework Programme, will work in tandem with OpenAIRE, extending the mission further to facilitate access to the entire Open Access scientific production of the European Research Area, providing cross-links from publications to data and funding schemes. This large-scale project brings together 41 pan-European partners, including three cross-disciplinary research communities.The project will capitalise on the successful efforts of the OpenAIRE project which is rapidly moving from implementing the EU Open Access Pilot project into a service phase, enabling researchers to deposit their FP7 and ERA funded research publications into Open Access repositories. The current publication repository networks will be expanded to attract data providers from domain specific scientific areas. Innovative underlying technical structures will be deployed to support the management of and inter-linking between associated scientific data.

Access to and deposit of linked publications via the OpenAIRE portal will be supported by a Help Desk, and OpenAIRE's collaborative networking structure will be extended to promote the concept of open enhanced publications among user communities. Liaison offices in each of the project's 31 European countries work to support the needs of researchers in Europe. The project will also actively leverage its international connections to contribute to common standards, data issues and interoperability on a global level.

Datahub.io

DataHub is community run, and community owned Open Data catalogue and has a solid technical home within the Open Knowledge Foundation Labs, it provides free access to CKAN's core features.

Once upon time there was only one CKAN instance: ckan.net. It was renamed to datahub.io in 2011 to avoid confusion between CKAN (the software) and the website.

To meet the needs and expectations of those using it, Datahub created a team of administrators, curators, developers from the community of enthusiasts interested in Open Data domain. Today it provides a large number of datasets with global geographical coverage and brings together the LOD community.

12. Conclusion

In this deliverable, we presented the list of 7 Open Data Catalogues selected for submitting metadata of 150 datasets selected by the COMSODE project in Deliverable D3.1. The list can be found in Attachment C of this deliverable.

As we have shown, the list comprises a variety of data catalogues both government owned and non-government owned as well as catalogues with global geographical coverage and data catalogues collecting datasets on national and municipal level.

Information in Attachment C and Attachment D provide developers with all details needed for submitting metadata to the catalogues. In case some additional questions arise, contacts to data catalogue owners are provided.

Since all (except one) selected data catalogues are based on CKAN technology, the deliverable provides detailed description of mapping between DCAT and CKAN of different versions. Methods of soliciting data consumer feedback of selected catalogues are described along with proposed methods extending the current possibilities.

The outcome of this deliverable will be used in WP4 (Task 4.3) to perform actual advertisement.

13. References

[1] ŠEDIVEC, Tomáš. Nástroje pro katalogizaci otevřených dat. 2014 [cit. 2014-03-22]. Rozpracovaná Diplomová práce. Vysoká škola ekonomická v Praze. Vedoucí práce Dušan Chlapek.

[2] Datacatalogs.org
http://datacatalogs.org

[3] CKAN instances around the world
http://ckan.org/instances/#

[4] Open Data Websites (Portals and Catalogs)
http://data.okfn.org/data/okfn/opendatasites#data

[5] OpenGeoCode - Catalog of Open Data Portals
http://www.opengeocode.org/opendata/

[6] European Union Open Data Portal - Related open data websites
http://open-data.europa.eu/en/about

[7] Customizing dataset and resource metadata fields in CKAN
http://docs.ckan.org/en/latest/extensions/adding-custom-fields.html

[8] Describing Linked Datasets with the VoID Vocabulary
http://www.w3.org/TR/void/

[9] Data Catalogue Vocabulary (DCAT)
http://www.w3.org/TR/vocab-dcat/

[10] The 7 best ways to gather customer feedback
http://www.helpscout.net/blog/customer-feedback/