DELIVERABLE D3.1

Final version of the selected datasets list

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	30th May 2014
Author(s)	Martin Nečaský, Jan Kučera, Dušan Chlapek, Jakub Klímek, Peter Hanečák, Jan Gondol, Anisa Rula, Roberto Cornacchia, Veronika Belianska
Responsible of the deliverable	Martin Nečaský
Email	necasky@ksi.mff.cuni.cz
Reviewed by	Miroslav Konečný
Status of the Document	Final
Version	1.0
Dissemination level	CO Confidential, only for members of the consortium (including the Commission Services)

History

Version	Date	Description	Revised by
0.1	2014-04-30	The initial list of candidate datasets finalized.	Martin Nečaský
0.2	2014-05-05	The deliverable document created, outline of the deliverable prepared.	Martin Nečaský
0.3	2014-13-05	The process of compiling the initial list of candidate datasets described.	Martin Nečaský
0.4	2014-14-05	First version of 150 COMSODE datasets selected and presented to the partners	Martin Nečaský, Jan Kučera, Jakub Klímek, Dušan Chlapek
0.5	2014-14-05	Statistics of the initial candidate datasets + statistics of the candidates which passed hard criteria added and discussed.	Martin Nečaský
0.6	2014-16-05	Description of the deliverable contents added to the section 3. Minor corrections of the text.	Jan Kučera
0.7	2014-20-05	Statistics updated	Martin Nečaský
0.8	2014-21-05	Discussion of the selected datasets added	Martin Nečaský
0.9	2014-22-05	Final version of deliverable - first draft	Martin Nečaský
0.10	2014-26-05	Final version of deliverable - second draft	Martin Nečaský
0.11	2014-27-05	Deliverable finished	Martin Nečaský
1.0	2014-30-05	Deliverable reviewed	Miroslav Konečný

Table of Contents

1. Executive summary

The purpose of this deliverable is to choose 150 datasets that will be published by COMSODE project to demonstrate the operation of Open Data Node (a software solution for open data publication which will be developed by COMSODE project) and methodologies for open data publication (which will also be developed by COMSODE project). During the work on the deliverable, we identified 166 datasets, which is a bit more than the required 150 datasets.

This document describes how the final 166 datasets were selected. We present the initial list of candidates proposed by COMSODE consortium and associated partners. There were 387 candidates identified in total. We present the distribution of the candidates according to their country of origin and the domain they belong to.

We then describe the process of selecting the final list of 166 datasets. We proceed according to the selection criteria and the selection process defined in Deliverable 2.2. We show several statistics of candidate datasets that passed the first portion of criteria - so called hard criteria. 300 candidate datasets passed those hard criteria in total. Those candidates then proceeded to the evaluation against so called soft criteria. The evaluation resulted into a score for each candidate. The statistics of those scores are presented in the deliverable as well. In the end we present the final list of 166 datasets and how they were selected.

The final list of 166 selected datasets is fully presented in Attachment C of the deliverable.

The final list of 166 selected datasets is a base we will work with in the next steps of the COMSODE project. However, it may happen that some new datasets, which are better for the demonstration purposes of the COMSODE project, will appear. This may happen because the number of public institutions we cooperate with will probably increase during the project. In that case, we will replace some of the datasets on the final list with the new ones. For each case, a detailed explanation of the replacement will be provided.

2. Deliverable context

2.1. Purpose of deliverable

This deliverable represents the final version of the list of datasets that were selected for publication as Open Data in the COMSODE project. Development of the datasets list is described in this deliverable. It also explains how the criteria described in the deliverable *D2.2 Criteria for selection of datasets* were applied to select the final list of the datasets.

Deliverable objectives:

- Describe how the initial datasets candidate list was developed.
- Describe how the dataset selection criteria were applied to select the final list of datasets.
- Introduce the final list of selected datasets.
- Describe how changes to the final datasets list will be handled.

2.2. Related Documents

List of related documents from the project:

- DOW, version date 2013-08-06, page 11
- Deliverable D2.2 Criteria for selection of datasets

2.3. List of Attachments

The following documents are attached to this document

- **Attachment A - candidate datasets** : Shows the list of all datasets gathered by the partners of COMSODE project which were considered as initial candidates for the selection process.

- **Attachment B - evaluation of soft criteria** : Shows how the candidates were evaluated according to criteria specified in Deliverable 2.2

- **Attachment C - FINAL LIST OF SELECTED DATASETS** : Presents the output of this deliverable - the final list of datasets chosen by the COMSODE consortium for publication.

3. Methodology used

3.1. Methodology

We proceeded in the following steps:

1. Prepare a table where candidate datasets will be collected. Each candidate dataset will be filed as a row of the table. For each dataset we will fill in columns of the following kinds:
    - Columns for filling mandatory attributes introduced in D2.2.
    - Columns for filling values of evaluation criteria for selecting datasets from the candidates.
    - Columns for computing the final scores which will be used for selecting 150 candidates for the final list of datasets.
2. Collect the initial list of candidate datasets into the prepared table. A candidate dataset could have been proposed by any project partner, by broad public via www.youropendata.eu or by a member of the User Board through any of the partners. The candidate had to meet the definition of COMSODE dataset introduced in D2.2. For each candidate, mandatory criteria were filled in.
3. Divide the candidate datasets to those that passed the hard evaluation criteria defined by D2.2 from those that did not pass.
4. Fill in the columns for evaluating the datasets against the soft criteria defined by D2.2 for each candidate dataset that passed the hard criteria, including the effort estimation.
5. Compute the final scores as described by D2.2, sort the candidates according to the final scores and choose 150 candidates for the final list. The selection should proceed as described in D2.2.

3.2. Partner contributions

This deliverable is a joint effort of all partners of the COMSODE project. All partners contributed to the collection phase – the initial list of candidate datasets. The selection process was driven by WP3 leader - CUNI and intensively consulted by all partners during April and May 2014.

4. Structure of the document

This deliverable represents the final version of the list of datasets that were selected for publication as Open Data in the COMSODE project

Chapter 5 describes how the initial list of candidate datasets was developed and it presents various statistics about the collected datatsets.

Chapter 6 describes how the candidate datasets were assessed using the hard and the soft section criteria described in the deliverable D2.2. Results of the evaluation are also presented this chapter.

Chapter 7 introduces the final list of the selected datasets and it provides a short description of the selected datasets. The full list is available in Attachment C of this deliverable. The chapter also outlines how the list of the selected datasets might be further developed over the project. Although the list of the datasets presented in this deliverable should represent the final list of datasets published during the COMSODE project there should a process in place that will allow us to make changes to this list of datasets in case that there appear datasets which are more interesting for publication.

Chapter 8 concludes the deliverable and provides the executive summary.

5. Collecting the initial list of candidate datasets

In this chapter, we describe how we collected the initial list of candidate datasets. The purpose of this work was to collect all possible candidates we could work with. The only criterion was that the candidate had to be a valid dataset according to the definition of a COMSODE dataset introduced in Deliverable 2.2. These datasets were collected as a part of Task 3.1. Another datasets are those we will need later for interlinking. Those datasets were collected during the initial phases of Task 3.2.

We have collected the total number of 387 candidate datasets into the initial list. The full list is available in Attachment A of this deliverable.

5.1. Attributes of candidate datasets

For each candidate dataset, mandatory attributes were filled in according to Deliverable 2.2. We refer to Deliverable 2.2 for their full description. Here we repeat them for clarity.

- ID
- name in English and in the local language
- country of origin of the dataset
- description (in English)
- primary topic and secondary topics (EUROVOC)
- domain the dataset belongs to
- information whether the dataset is structured (i.e. has a schema) and whether the schema is expressed; information whether the schema also describes entity identifiers
- owner and publisher of the dataset (as defined by Deliverable 2.2)
- user and provider of the Open Data Node instance which will host the dataset (if unknown, COMSODE consortium is the user/provider as specified in Deliverable 2.2)
- contact person for the dataset if we have a contact to the owner/publisher of the dataset (usually, it is a public body) plus information whether the contact person is in the COMSODE user board
- information whether the specified user is also a potential Open Data Node instance provider - i.e. whether the user is willing to run its own instance for hosting the dataset
- homepage URL of the dataset if it exists
- COMSODE partner responsible for the dataset
- current data formats in which the dataset is available
- target data formats in which COMSODE project should publish the dataset
- information whether data items have some identifiers (even though they do not necessarily need to be explicitly expressed in the schema)
- information whether the dataset is incremental, i.e. whether new data items appear in time in the dataset
- information whether terms of use or licence for using the dataset exists and where we can find the full text of the terms of use/licence if it is publicly available
- potential consumers of the dataset

- information whether the dataset contains some personal or other secrets which can not be open by law
- information whether the dataset is distributed across different data sources
- information whether the dataset is active (maintained) or not (archived and not maintained/published yet)
- information whether there are some know limitations in publishing the dataset as opened

5.2. Statistics of collected candidate datasets

As was already mentioned, we have collected the total number of 387 candidate datasets. We provide some statistics of the collected candidates in this section.

5.2.1. Statistics by country

First, let us discuss the countries of origin of the candidate datasets. Each candidate is associated with a particular country. We concentrated mainly on datasets that come from the countries of the COMSODE project participants. However, we also collected several datasets from other countries. The following list shows countries for which we have more than 1 dataset collected:

- Slovak Republic: 151 candidate datasets
- Czech Republic: 124 candidate datasets
- Italy: 43 candidate datasets
- Netherlands: 25 candidate datasets
- Albania: 12 candidate datasets
- USA: 3 candidate datasets
- Croatia: 2 candidate datasets

We have also identified 3 candidate datasets that are published by the European Union as whole (by some EU institution). Therefore, they are not assigned to any particular country but to EU as a whole. There are also several countries for which we identified a single dataset. We do not list the countries in the text but refer to Figure 1 that displays a map with all covered countries.

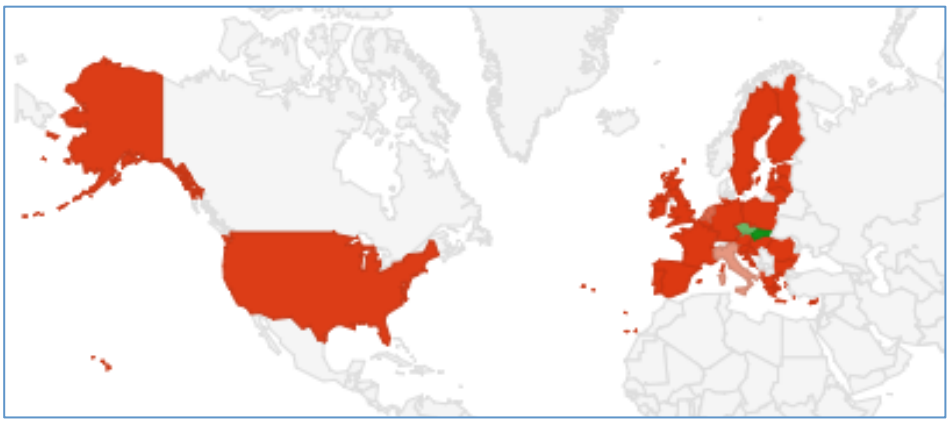

Figure 1: *Distribution of candidate datasets from the initial list on a world map excerpt*

5.2.2. Statistics by domain

Second, let us discuss the distribution of the identified candidates by their domain. We assigned one or two domains to each of the candidate dataset according to its topic. We identified 38 different domains. Figure 2 shows the numbers of identified candidate datasets for particular domains.

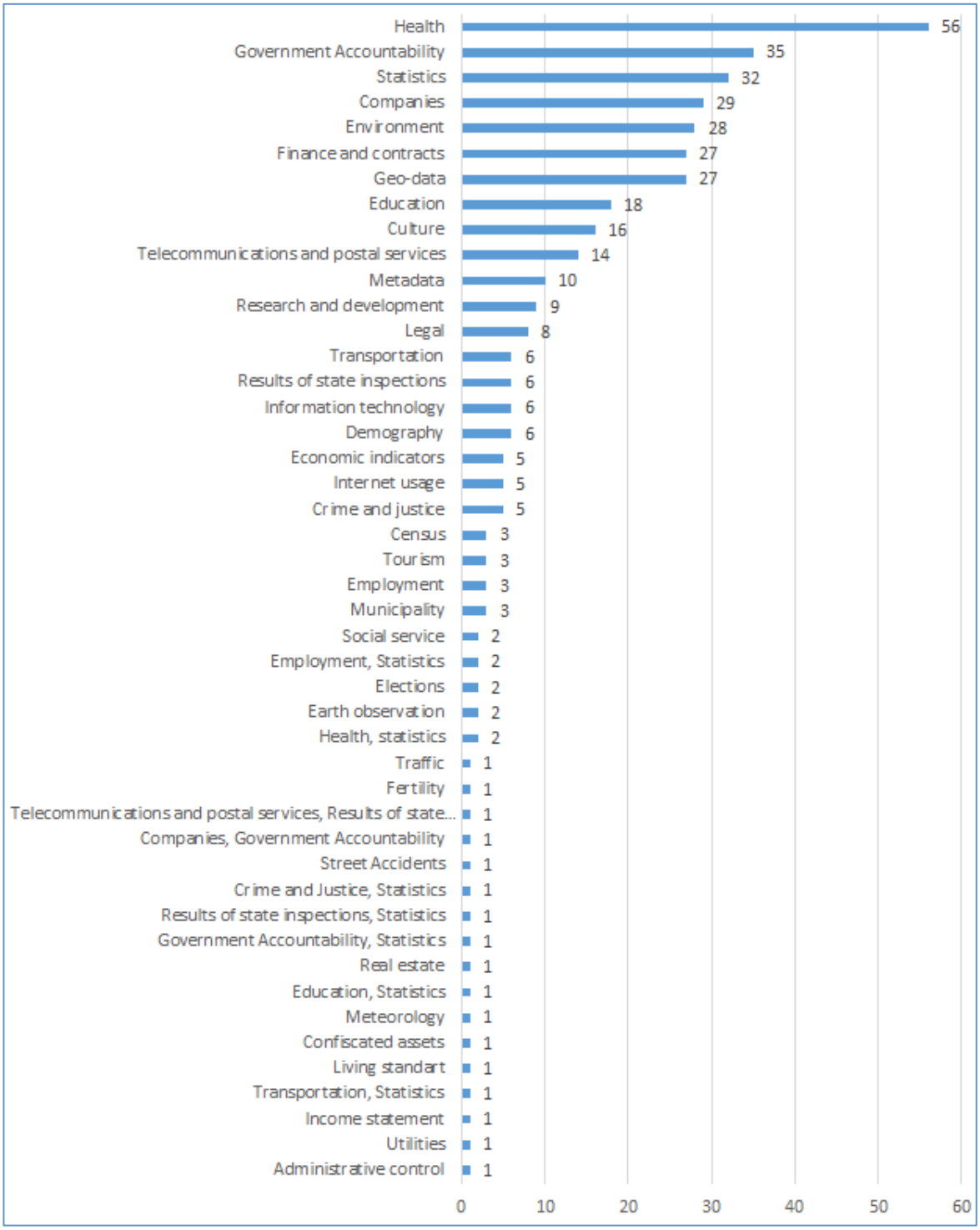

Figure 2: *Distribution of candidate datasets from the initial list according to identified domains*

6. Assessment of the datasets candidates

The next step of the process is to evaluate the candidates against hard and soft criteria specified by Deliverable 2.2.

6.1. Evaluation of hard criteria

First, each dataset is evaluated against the hard criteria. Evaluation of each of the hard criteria results to passed or not passed. A candidate dataset must pass all the hard criteria. If it does not pass any of the hard criteria, it is not further evaluated and is discarded. The following table repeats the hard criteria as introduced by D2.2. We refer to D2.2 for their detail explanation and rationale.

ID	Name	Description
CC1	Dataset definition compliance	Is it compliant with the definition of the COMSODE dataset?
CC2	Open Data publication feasibility	Is it possible to publish dataset as Open Data (see the Open Data definition above)?
CC3	Mandatory attributes compliance	Does it have all mandatory attributes filled in?

Table 1: *Hard criteria description*

From the 387 identified candidate datasets, 300 passed all 3 hard criteria. The remaining 87 candidates did not pass the criteria.

- All candidates passed the first criterion, CC1.

- 37 candidates did not pass the second criterion, CC2. Those candidates were proposed because they seemed to be interesting and nice to have but it was not possible to determine whether it is possible to publish them as open data. Usually it was not possible to contact the owner of the data to negotiate with him the publication.

- 67 candidates did not pass the third criterion, CC3. For those candidates it was not possible to collect values of all mandatory attributes. Usually we were not able to collect enough information about the dataset neither from public sources nor by negotiating with the owner of the dataset.

Attachment A shows for each candidate dataset whether it passed the hard criteria or not. The information is presented in the last column.

6.2. Evaluation of soft criteria

Those candidates that passed hard criteria were evaluated against the soft criteria. Soft criteria are described in detail in Deliverable 2.2. Here, we summarize them for clarity.

ID	Name	Group	Description
TC1	Current formats	Technical	Assessment of the machine-readability of the format.
TC2	Schema	Technical	Is the dataset schema expressed in formal or semi-formal form?
SC1	Availability of the ODN user	Subjects	What type of ODN users does the dataset have? A dataset with a confirmed ODN user has higher priority.
SC2	Willingness to publish data	Subjects	Is the owner (or potential publisher) cooperative and eager to publish the dataset with COMSODE (using ODN)?
SC3	Willingness to run ODN	Subjects	Is the owner (or potential publisher) willing to run his/her own ODN instance?

Table 2: *Soft criteria description*

The COMSODE project partners evaluated each of the soft criteria manually. The guidelines for evaluation were specified in Deliverable 2.2. Evaluating a particular dataset against a particular soft criterion means assigning a value between 0 and 5. Attachment B of this deliverable shows the values of the criteria in columns AO - AU. Let us note that 4 candidate datasets could not be evaluated because of insufficient information about them. Therefore, only 296 were evaluated against the soft criteria.

Let us discuss particular soft criteria and distribution of the datasets which passed the hard criteria. The presented statistics are computed on the base of columns AO - AU of Attachment B.

[TC1] Current formats

The purpose of this criterion is to asses the machine-readability of datasets. Possible values and their meaning are following:

- 0 - format that is very difficult for machine processing, it is not structured (e.g. PDF)
- 1 - format that is very difficult for machine processing but it is application independent (e.g. TXT)
- 2 - format that allows partial structuring of the data (e.g. HTML)
- 3 - structured format that partially allows description of the schema (e.g. CSV)
- 4 - structured format that allows description of the schema and partial expression of the semantics and linking as well (e.g. XML)
- 5 - structured format that allows description of the schema, expression of the semantics and linking of the data (e.g. RDF)

The result of evaluation of TC1 is summarized in Figure 3. It shows that most of the identified datasets that passed hard criteria are available for the COMSODE project in machine-readable formats (values 3-5). There is also a significant number of datasets that are available in HTML (value 2). These datasets are harder to process if we want them to be published as open data. Therefore, if these datasets are accepted to the final list of datasets we need to count with a higher effort to process them in the Open Data Node software.

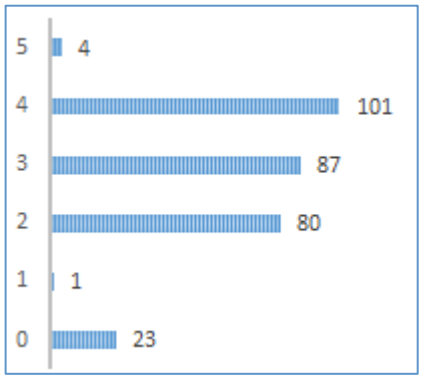

Figure 3: *Distribution of datasets according to TC1*

[TC2] Schema

The purpose of this criterion is to assess whether the schema of a dataset exists and how it is expressed. Possible values and their meaning:

- 0 - schema is not expressed
- 1 - schema is described in text
- 3 - schema is expressed in a semi-formal form
- 5 - schema is expressed using a formal model

The result of evaluation of TC2 is summarized in Figure 4. It shows that for most of the candidate datasets that passed the hard criteria there is no expressed schema available. However, there is still a significant number of datasets which have a schema expressed somehow. A formal expression (e.g., in a form of XML schema) exists only for 48 datasets.

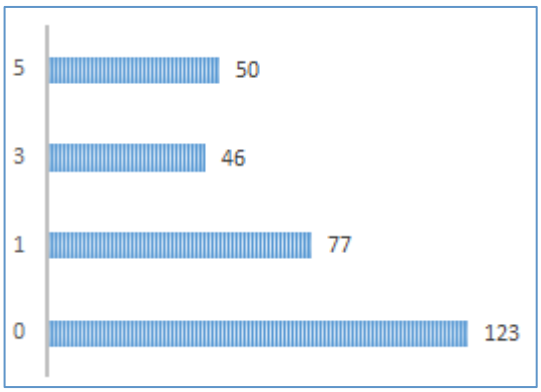

Figure 4: *Distribution of datasets according to TC2*

[SC1] Availability of the ODN user

The purpose of this criterion is to assess the type of ODN user for each dataset. ODN users are very important for the COMSODE project. Who is an ODN user is described in detail in Deliverable 2.2. Possible values and their meaning:

- 1 - internal user of the dataset
- 3 - 1-N known consumers as the ODN user
- 5 - owner or publisher of the dataset asked our help to publish the dataset

The result of evaluation of SC1 is summarized in Figure 5. It shows that we were able to identify real ODN users for 123 candidate datasets that passed the hard criteria (consumers of the dataset or owner/publisher of the dataset). This is a very good number since we are at the first year of the project. We expect this number to be increased during

the rest of the project as negotiations with the public bodies and software developers will continue. The goal is that for each dataset published by the COMSODE project we have some relevant and known consumers of the dataset or owner/publisher of the dataset who cooperate with us. Some datasets however can be legitimate for the project even if there is only the internal user (= COMSODE partners). Datasets that are important for linking with the other datasets (and were identified during our work on Task 3.2. of the COMSODE project) fall into this category.

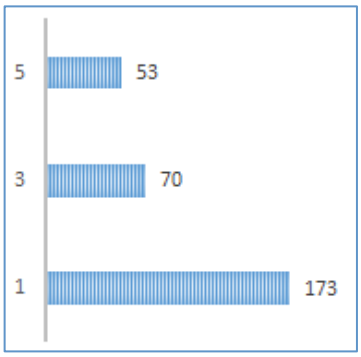

Figure 5: *Distribution of datasets according to SC1*

[SC2] Willingness to publish data

The purpose of this criterion is to assess whether the owner of a dataset is cooperative and eager to publish the dataset as open data. This criterion does not assess whether the owner will use the whole Open Data Node software. It is possible that the owner will just decide to put the dataset on the web without our software or only with some of its components or that s/he will just ask some of the COMSODE project partners to help him or her with the publication. What is important, however, is that the owner cooperates with the COMSODE project partners. Possible values and their meaning:

- 0 - owner (or the potential publisher) is not willing to publish the dataset or the possibility to publish the dataset is unsure
- 1 - owner (or the potential publisher) has not decided about the publication yet, but it seems likely it will not be published
- 3 - owner (or the potential publisher) has not decided about the publication yet, but it seems likely it will be published
- 5 - owner (or the potential publisher) is willing and ready to publish the dataset

The result of evaluation of SC2 is summarized in Figure 6. It shows that for most of the datasets it is possible to publish them as open data. Some of the datasets are already published as open data (i.e. they meet all criteria, including cataloguing and proper terms of use). However most of them require some improvement - it is to possible to download them from somewhere but they are not catalogued, terms of use could be specified more clearly and they could be provided in more suitable data formats.

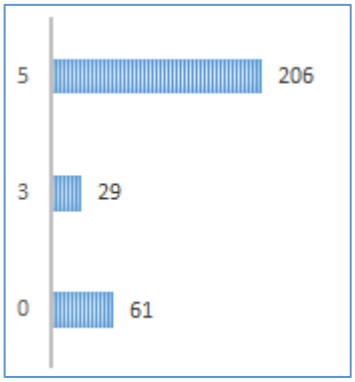

Figure 6: *Distribution of datasets according to SC2*

[SC3] Willingness to run ODN

The purpose of this criterion is to assess whether the owner or publisher of a dataset is willing the run its own Open Data Node instance for publication of the dataset. The previous criteria say whether we have established cooperation with the owner/publisher

(SC2) and if so, what form does this cooperation have. Possible values and their meaning:

- 0 - owner/publisher is not willing to run its own ODN instance or unknown willingness
- 1 - owner/publisher has not decided yet, however it will probably not run his or her own ODN instance
- 3 - owner/publisher has not decided yet, however it will probably run his or her own ODN instance
- 5 - owner/publisher is willing to run its own ODN instance

The result of evaluation of SC2 is summarized in Figure 7. SC1 showed that there are number of datasets whose owners/publishers and/or consumers want to use some ODN instance for publishing or consuming the datasets. SC3 shows the numbers of datasets that will be published by their owners/publishers using their own ODN instance. It shows that for 3 datasets we already have its owner/publisher who decided to use its own ODN instance. For 18 datasets, we are close to finalizing the negotiations with the owner/publisher of the dataset that he will use its own ODN instance. The rest of the datasets will be published on an ODN instance of some identified consumer of the dataset or on an ODN instance of some of the COMSODE project partners (e.g., Ministry of Interior of Slovak republic). However, we are at the beginning of negotiations with the owners/publishers. The number of datasets in category 3 and 5 will increase during the second year of the project.

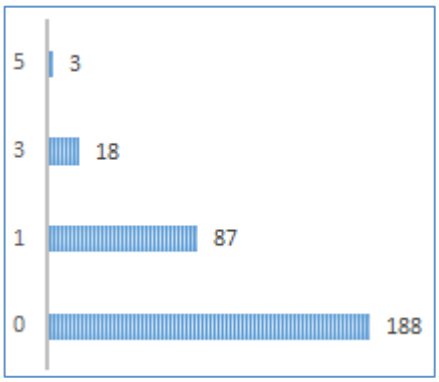

Figure 7: *Distribution of datasets according to SC3*

After values of soft criteria were assessed for all datasets that passed the hard criteria, we computed the final score for selecting 150 datasets from the identified candidates. The score was computed using the following formula:

$R_{score} = W_{TC2}*TC2 + W_{TC3}*TC3 + W_{SC1}*SC1 + W_{SC2}*SC2 + W_{SC2}*SC2$

Where:

- R_{score} is the total score resulting from the assessment using the soft criteria
- W_{TC2} is the weight of the TC2 criterion
- TC2 is the value of the TC2 "*Current formats*" criterion
- W_{TC3} is the weight of the TC3 criterion
- TC3 is the value of the TC3 "*Schema*" criterion
- W_{SC1} is the weight of the SC1 criterion
- SC1 is the value of the SC1 "*Availability of the ODN user*" criterion
- W_{SC2} is the weight of the SC2 criterion
- SC2 is the value of the SC2 "*Willingness to publish data*" criterion
- W_{SC2} is the weight of the SC3 criterion
- SC2 is the value of the SC3 "*Willingness to run ODN*" criterion

The weights enable us to prefer some soft criteria at the expense of the other. We set the weights shown in Table 1. The weights reflect the following preference: criteria related to ODN users have higher priority than the technical ones. The most important criteria are those which say whether the dataset can be published as open data and whether the owner/publisher is willing to use own ODN instance for dataset publication.

ID	Criterion	Weight	Weight ID
TC2	Current formats	0,1	W_{TC2}
TC3	Schema	0,1	W_{TC3}
SC1	Availability of the ODN user	0,2	W_{SC1}
SC2	Willingness to publish data	0,3	W_{SC2}
SC3	Willingness to run ODN	0,3	W_{SC3}

Table 3: *weights of soft criteria*

Figure 8 shows the numbers of candidate datasets that passed the hard criteria distributed by their final score. Most of the candidates have their score between 2 and 3. We sorted the candidates according to their final score and we used the resulting list as a base for selecting the 150 datasets we will work with in the COMSODE project.

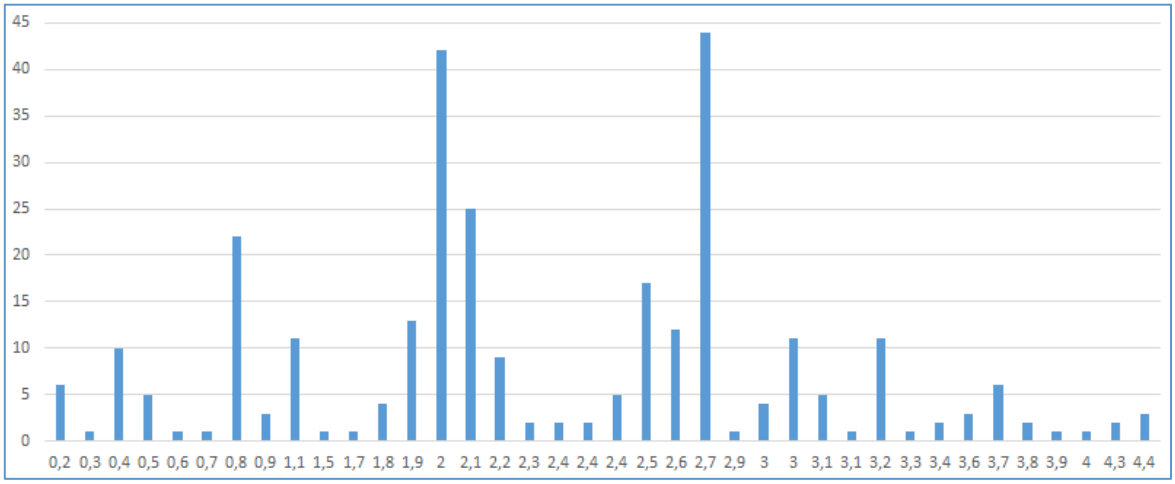

Figure 8: *Distribution of datasets according to their final score*

We marked the top 150 datasets in the sorted list as accepted. Then each partner could change the acceptance of any of its datasets to rejection and vice versa. The reason was that there can be some individual preferences of each partner which cannot be reflected by the numbered criteria. However, clear rationale had to be supplied by a partner when a dataset acceptance was changed manually. The goal was to achieve the following numbers of datasets each partner is responsible for:

- CUNI will be responsible for >= 50 datasets in the final list
- EEA+MoI will be responsible for >= 80 datasets in the final list
- SPINQUE+UNIMIB will be responsible for the rest of the datasets so that the total number of at least 150 datasets is achieved on the final list

After the selection by each partner was finished the following numbers of datasets accepted for the final list were achieved:

- CUNI : 53 datasets
- EEA+MoI : 90 datasets
- SPINQUE : 2 datasets
- UNIMIB : 21 datasets

In total, we have 166 datasets on the final list of datasets that is presented as the output of Deliverable 3.1. The accepted datasets are described in a more detail in Chapter 6 and their full list can be found in Attachment C of this deliverable.

6.3. Evaluation of linked open data criteria

We further evaluated the accepted datasets against the linked open data criteria. These criteria are also described in detail in Deliverable 2.2. Let us summarize them for clarity.

ID	Name	Description
LC1	Data identifiers	Score for data identifiers. Datasets with natural keys are preferred because they are easier to link to other data.
LC2	Linking potential	By this criterion we express the potential of a dataset to be linked to other datasets. It is a potential of a dataset to be either linked from other datasets (e.g. addresses are used in many datasets and therefore many datasets might provide links to address datasets) or a potential of a dataset to provide links to other datasets (e.g. datasets containing data about business entities might provide links to the business registry data, if it contains address of the entity as well it might also provide links to address dataset).

Table 4: *weights of soft criteria*

The COMSODE project partners evaluated each of the criteria manually. The guidelines for evaluation were specified in Deliverable 2.2.

[LC1] Data Identifiers

Possible values and their meaning are following:

- 0 - there are no identifiers used in the data
- 3 - artificial identifiers are used in the data
- 5 - natural identifiers are used in the data

The result of evaluation of LC1 is summarized in Figure 9. It shows that most of the datasets use natural or artificial identifiers. Those datasets are first candidates for publication in Linked Open Data format.

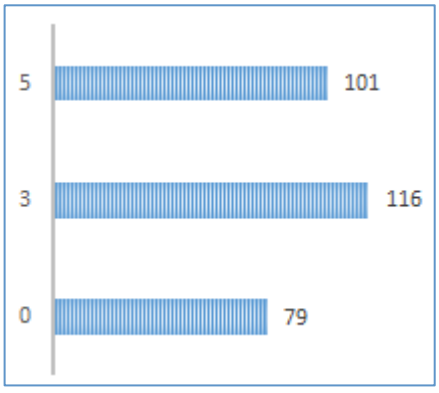

Figure 9: *Distribution of datasets according to LC1*

[LC2] Linking potential

Possible values and their meaning are following:

- 0 - dataset cannot be linked to other datasets and no other dataset can link to the it, or it would be very difficult to link
- 1 - dataset can be linked to 1-2 other datasets or 1-2 other datasets can link to it
- 2 - dataset can be linked to 3-4 other datasets or 3-4 other datasets can link to it
- 3 - dataset can be linked to 5-6 other datasets or 5-6 other datasets can link to it
- 4 - dataset can be linked to 7-10 other datasets or 7-10 other datasets can link to it
- 5 - dataset can be linked to more than 10 other datasets or more than 10 other datasets can link to it

The result of evaluation of LC2 is summarized in Figure 10. It shows that for most datasets we were able to find some other datasets for linking.

Even though we proposed a scoring function for Linked Open Data criteria in Deliverable 2.2, we decided that we will use those criteria only as a guideline and will choose the datasets we will publish as Linked Open Data manually during the project. We identified 144 datasets that were accepted after evaluating soft criteria and users who proposed them asked for publication in RDF format. At least 30 of them will be published as real Linked Open Data during the project (i.e., in RDF format and linked to some other datasets published by COMSODE project or someone else).

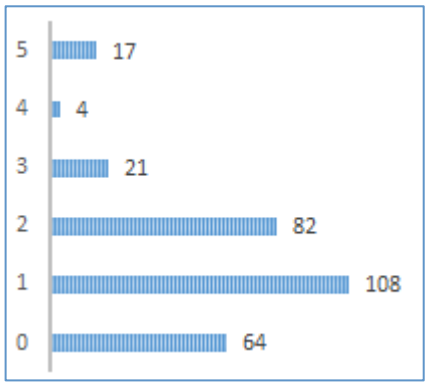

Figure 10: *Distribution of datasets according to LC2*

7. Final version of the selected datasets list

We compiled the list of 166 dataset that will be published by the COMSODE project. The list is provided in the Attachment C. The selected datasets are of various types.

7.1. Selected datasets by domain

The selected datasets cover 29 domains in total as shown in Figure 11.

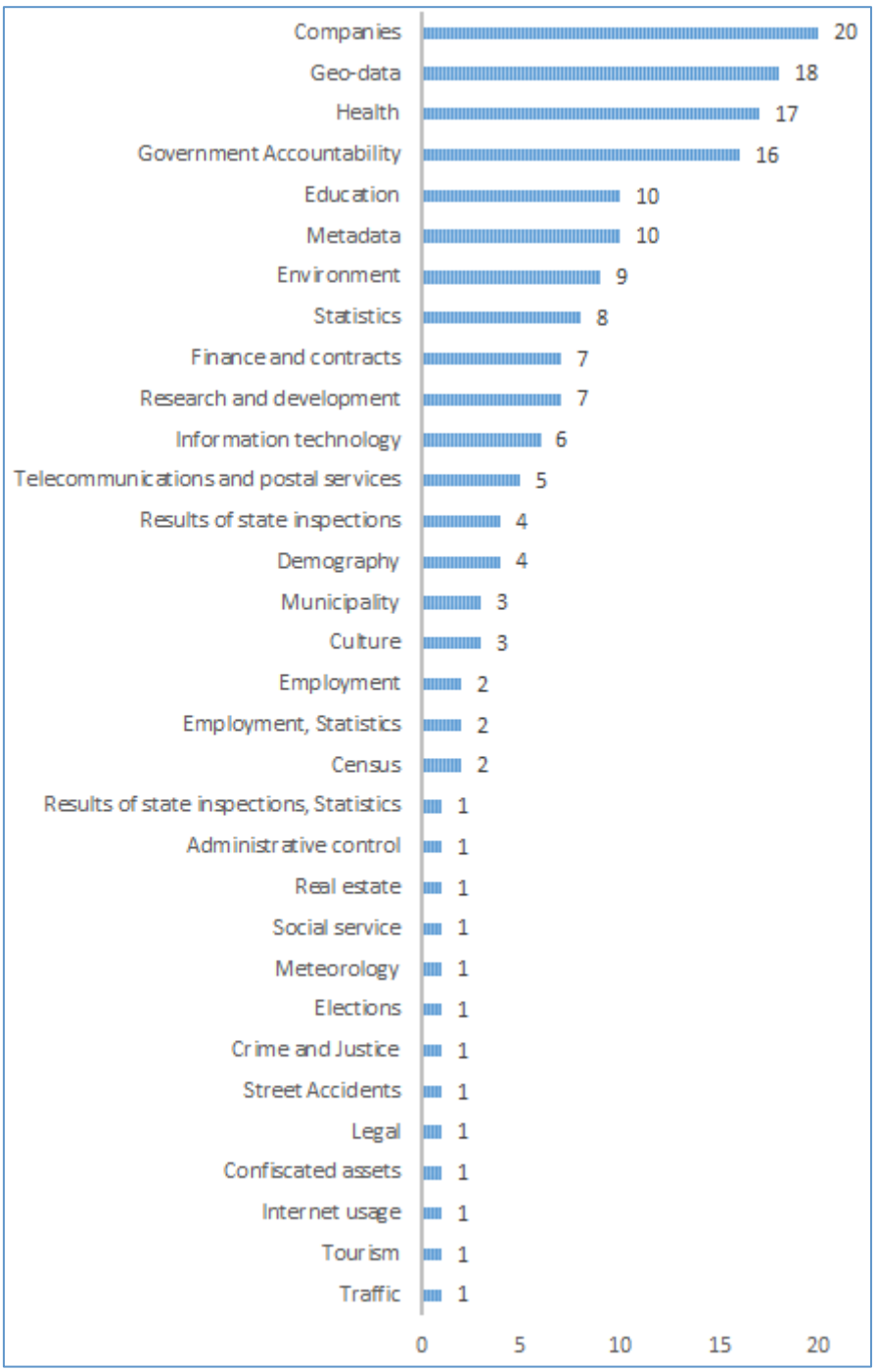

Figure 11: *Distribution of selected COMSODE datasets by domain*

7.2. Selected datasets by country

We will publish datasets from several countries, namely Slovak Republic, Czech Republic, Italy, Albania, Netherlands, and Spain. We also included several datasets from U.S. and datasets published by EU. It is depicted in Figure 12.

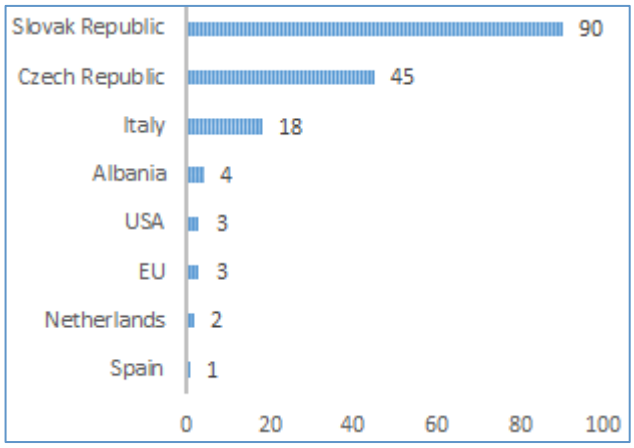

Figure 12: *Distribution of selected COMSODE datasets by their country of origin*

7.3. Selected datasets by estimated effort

We also estimated effort needed to publish each of the selected datasets. We tried to estimate effort in man-days needed to analyse, describe, transform, cleanse and publish the dataset on the web and list it in a chosen data catalogue. However, exact estimation is not possible. Therefore, we chose three levels of effort:

1. **Easy**: less than 1 man-day for a dataset
2. **Moderate**: less than 5 man-days for a dataset
3. **Difficult**: more than 5 man-days for a dataset

For example, most code lists were ranked as easy datasets. Datasets which are more complex, e.g. lists of business entities with several properties published, but they are well structured were usually ranked as moderate. Datasets that are not available in a nice structured format but harvesting from, e.g., HTML sites is first necessary, were ranked as difficult.

As Figure 13 shows, most of the datasets were ranked as easy or moderate. We also have several difficult datasets on the list. However, they are only few. This corresponds to the possibilities of the COMSODE project - it is not concentrated only to datasets publication but also to the development of the whole publication platform. On the other hand, thanks to having several difficult datasets, we will be able to demonstrate the possibilities of the publication platform on the whole range of different kinds of datasets.

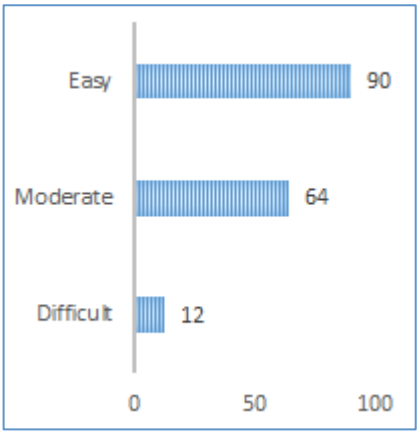

Figure 13: *Distribution of selected COMSODE datasets by their estimated effort*

7.4. Future changes in the selected datasets list

The presented list of 166 selected datasets represents the starting point of the publication efforts of the COMSODE project. The publication will proceed in the following months. It is possible that partners will identify other datasets which are better than those selected in this deliverable (according to the selection criteria given by Deliverable 2.2).

Therefore, it is possible that another ones that are more suitable for the purpose will replace some of the currently selected datasets. It will always be clearly explained why this replacement was made. The replacements and their explanation will be presented as a part of Deliverable 6.4 that will be prepared in the end of the project.

8. Conclusions

In this deliverable, we presented the list of 166 datasets selected for the publication by the COMSODE project. The list can be found in Attachment C of this deliverable. As we have shown, the list comprises datasets of various kinds. On one hand, there are large datasets that contain high number of complex objects. On the other, there are small datasets that consist of several simple values. From another point of view, the datasets cover several EU and non-EU countries and also several domains. Therefore, the list will serve as a good base for demonstrating all possibilities of the publication platform and methodologies developed by the COMSODE project.

During the selection process we proceeded in the steps specified by Deliverable 2.2. First, we identified initial candidate datasets. From these candidates we choose the final list by evaluating rigorous criteria. Therefore, this deliverable (together with Deliverable 2.2) can serve other projects as a methodology for selecting datasets for publication.