DELIVERABLE D2.2

Criteria for the selection of datasets

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	24th February 2014
Author(s)	Martin Nečaský, Jan Kučera, Dušan Chlapek, Peter Hanečák, Gabriel Lachmann, Anisa Rula, Peter Beňo
Responsible of the deliverable	Martin Nečaský
Email	necasky@ksi.mff.cuni.cz
Reviewed by	Miroslav Konečný
Status of the Document	Final version
Version	1.0
Dissemination level	PU (Public)

Table of contents

1 Deliverable context

1.1 Purpose of deliverable

This deliverable is the output of Task 2.3 – Criteria for the selection of datasets. The purpose of the deliverable is to define criteria for the selection of datasets that will be published by the consortium during the project. Criteria prevent the project from overrunning its budget and time constraints by focusing on too many datasets. On the other hand, the criteria enable to achieve full objectives of the project.

Deliverable objectives:

- Provide a definition of a dataset from the project point of view.
- Propose the set of criteria for selecting datasets for the project. Subset of the criteria must be criteria for selecting those datasets that will be published as Linked Open Data.
- Provide rules for evaluating the criteria.
- Demonstrate the evaluation on a sample of real-world datasets.

1.2 Related Documents

- List of related documents from project:
  - DOW, page 17
- Attachments:
  - Document “Dataset attributes”

2 Methodology used

2.1 Methodology

1. As the first step, we created the definition of a dataset from the project point of view. It was necessary to explicitly state what collections of data items will be considered as a single dataset. This is important since the project has to publish a certain number of datasets and it must be easily verifiable whether the number has been achieved or not. If a proper definition of a dataset would not be given then it could be questionable whether two published datasets are really two different datasets or whether they should be considered as a single dataset. This deliverable provides the definition that enables to answer such questions easily.
2. As the second step, we compiled a list of various attributes of datasets. We also identified a subset of the attributes that are mandatory for each dataset. The mandatory attributes are crucial for the selection criteria defined in this deliverable. The other attributes are important for further estimations, e.g., time necessary for publishing a dataset, etc. In Deliverable 3.1 we will compile the initial list of datasets. Each dataset will have values of all mandatory attributes assigned and also values of the optional ones when appropriate. The attributes are provided as an attachment to this deliverable.
3. As the third step, we proposed the selection criteria. Since the criteria must help us to quickly decide whether a particular dataset will be involved in the project or not, we defined a small number of *hard criteria*. If the dataset does not meet any of the hard criteria, it is not further considered. Next, we defined *soft criteria* that will enable us to sort all datasets that passed the hard criteria according to their relevancy to the COMSODE project. We also have a set of criteria for selecting those datasets that will be published as Linked Data.
4. As the fourth step, the attributes and criteria have been discussed with the User Board during the meeting in Prague on 28th of January 2014. Their comments were implemented to the attributes and criteria.
5. As the fifth step, we provided a guideline of how each criterion should be evaluated. This is necessary to obtain consistent evaluations from different evaluators of the candidate datasets.
6. As the last step, we evaluated the criteria on a few real-world examples.

2.2 Partner contributions

The responsible partner for the deliverable is DSE CUNI who is also the main author of the deliverable. DSE CUNI worked on the definition of a dataset, on a list of attributes, selection criteria and examples.

EEA and MoI SR contributed to the list of attributes and selection criteria. They also reviewed the definition of a dataset.

All other partners contributed to the deliverable mainly in a form of regular weekly discussions. Every partner also reviewed the definition of a dataset. Each of them also contributed to the list of attributes and reviewed the selection criteria.

It is important that every partner participated on detailed discussions about the definition of a dataset, the list of attributes and the selection criteria. Now, the consortium is synchronized on all details that are necessary for further deliverables, mainly Deliverable 3.1 – Final version of the selected datasets list.

3 Structure of the document

This document describes the overall approach of the COMSODE project to the selection of suitable datasets that will be published during the project.

3.1 Chapter 4

Chapter 4 introduces the definition of a dataset. It shows how COMSODE partners understand the notion of a dataset. It helps to decided whether a collection of data items is a single dataset, a part of another dataset, or whether it should be split to more different datasets.

3.2 Chapter 5

Chapter 5 briefly introduces the definition of Open Data. This is just a short review of existing definitions of Open Data.

3.3 Chapter 6

Chapter 6 introduces the proposed attributes of datasets. It explains in detail the mandatory attributes which must be filled for each considered dataset, to be later evaluated based on the criteria introduced in this deliverable.

3.4 Chapter 7

Chapter 7 introduces the dataset selection workflow. It refers to the the criteria proposed in the following chapter. It will serve as the base for the methodology for fulfilling the Deliverable 3.1.

3.5 Chapter 8

Chapter 8 is the core chapter of the deliverable. It introduces the selection criteria and rationale for their existence. Three kinds of criteria are proposed – hard, soft and Linked Open Data criteria. Assessment guideline for each criterion is provided as well.

3.6 Chapter 9

Chapter 9 introduces formulas for scoring datasets on the base of the criteria.

3.7 Chapter 10

Chapter 10 demonstrates the selection criteria, assessment guidelines and formulas on a set of real-world datasets.

3.8 Chapter 11

Chapter 11 concludes the deliverable and provides executive summary.

4 COMSODE Dataset Definition

This section provides basic definitions that will be used internally by the COMSODE project consortium. The definitions are intended for the purposes of the COMSODE project. We use them to clarify the basic terms among COMSODE partners. Therefore, each notion defined in this document has to be considered as COMSODE specific. The most important is the term of “COMSODE dataset” (Definition 6) that should not be considered as a definition of what is a dataset in general. The primary purpose of the definition is to enable COMSODE project members to reason about each proposed dataset candidate - whether a given candidate can be considered as a dataset for the purposes of the COMSODE project or not. Before the definition of COMSODE dataset, we introduce some auxiliary terms.

Definition 1: An ***ODN provider*** *is an entity that runs its own instance of the Open Data Node (ODN).*

Definition 2: An ***ODN user*** *is an entity on whose request we decided to work with a dataset during the COMSODE project. It can be an owner, a publisher or a consumer of the dataset or a combination of these.*

Definition 3: *An* ***owner*** *of a dataset is an entity that holds rights to the dataset or is legitimate to make decisions about the dataset.*

Definition 4: *A* ***publisher*** *of a dataset is a (legal) person or a (legal) entity responsible for making the dataset available as Open Data (see Chapter 6).*

Definition 5: *A* ***consumer*** *of a dataset is a person or an entity who uses or is willing to use the dataset for its purposes.*

It is important to identify an ODN user of a dataset before we decide to work with the dataset. However, it is not important and necessary (at least at the first stages of the project) to identify every possible user - an owner, a publisher and all possible consumers. We just need **some examples of typical ODN users**.

A dataset without an ODN user is not interesting for the purpose of COMSODE project. Ideally, the ODN user is an entity who requires our help to publish the dataset in some ODN instance of some ODN provider and who agrees with the publication. In most cases, it will be the **owner** of the dataset. If we are not able to get the request/agreement from the owner, it can be a **publisher** who has the agreement from the ODN owner and who requires our help. If we do not have this kind of users, **the ODN user can be a consumer - an individual, a group of individuals or an organization or group of organizations** who asks us to publish the dataset. For the purpose of the COMSODE dataset definition we do not distinguish each consumer as a separate ODN user - we will distinguish him or her if we need to manage the individual relationships to him or her. However, for the purpose of the COMSODE dataset definition, the group of all possible consumers represents a single identified ODN user of the dataset. It is enough if we find some consumers who represent this group. If there is neither the owner/publisher nor the consumer, we can act as an **internal ODN user** (someone from the COMSODE consortium who wants to work with a dataset for some reasons). However, we should always try to find some external user or, at least, a group of consumers.

Definition 6: *A **COMSODE dataset** (or shortly a **dataset**) is a collection of data items such that*

1. *all data items in the dataset share the same main **topic,***
2. *all data items in the dataset share the same common **data schema**,*
3. *there is one and only one **ODN user** who requires publishing the dataset in some ODN instance.*

The definition distinguishes a collection of data items from a dataset. A dataset is any collection of data items that meets the criteria 1-3 given by the definition. *A data item is a set of facts about an entity. It can be encoded, e.g., as a row in a table, an element in an XML document or a set of RDF triples having the same subject.* Usually, any collection of data items is considered as a dataset. However, this is too coarse-grained for the purposes of the COMSODE project. We need to work with a certain number of datasets in the COMSODE project and we need exact criteria that will allow us to count a certain collection of data items as a dataset. Definition 6 provides such criteria. Let's discuss them in more details:

(1) Two collections of data items can be distinguished as two separate datasets if they have different main topics. For example, if the main topic of one collection is "public contracts" and the topic of the other one is "public budgets" then each of them can be considered as a separate dataset. If both collections have the same topic, e.g. "public budgets", they will be considered as the same dataset on the base of the first criterion and the second criterion must be applied to distinguish them.

It is necessary to fix the vocabulary of possible topics. The topics can be neither too coarse-grained nor too fine-grained. A larger community should also accept the vocabulary and it should be translated into official EU languages. Therefore, we decided to use EUROVOC as the vocabulary for our topics.

(2) Two collections sharing the same topic are considered as two datasets **if data items in one collection have a different data schema than the items in the second collection.**

For example, consider two collections with the topic "public contracts" published by the same publisher. One collection consists of public contracts which exceed a given amount of money and is published according to law with a given data schema. The other collection, which consists of public contracts and is below the limit, is published only optionally and with a different, simpler schema. In that case both collections are considered as two different datasets.

Another example is the case when the publisher changes the data schema. The collection published before the change has a different schema than the collection published after the change and they are therefore distinguished as two separate datasets.

This is important because when the data schema is changed it may result in a nontrivial time required to update publication/consummation procedures of the collection. Therefore, we consider the above collections as two datasets because the work required to publish/consume the collections is doubled.

Note that we consider datasets with explicitly expressed schemas as well as datasets with only implicit schemas. A dataset has an explicitly expressed schema if there is an expression that describes the structure (and possibly semantics) of the data items in the dataset in some standard machine interpretable language.

For example, if we have an XML dataset then its explicitly expressed schema can be an XSD document which describes what XML elements and attributes can be used in the XML dataset and which structure the XML elements have. On the other hand, a dataset has an implicit schema if there is no explicitly expressed schema but the schema can be, at least partly, extracted from the dataset. For example, an XML dataset contains XML elements and attributes and we can see what these XML elements and attributes are and what structure the XML elements have. Therefore, the schema is encoded in the dataset implicitly and we are able to extract it, at least partly.

And, third, let us assume two collections with the same main topic and the same data schema. We count them as 2 datasets if there are 2 different ODN users of the collections. These ODN users can be different owners or publishers of the collection. In other words, if two ODN users ask us to publish two collections of data items with the same schema and the same topic we will count these collections, for the purposes of the COMSODE project, as two separate datasets.

Example: suppose that two different public bodies are willing to publish their public contracts as open data and they ask our help with the publication. This means that there are two collections of data items with the same topic ("public contracts") and they both want to use the same data schema for publication. Although, there are two publishers, the collections share the same data schema and thus the two collections are considered as two datasets. This is important because we have to work with each publisher separately and make a lot of negotiations and other steps before the publication itself. As a result, the work is doubled.

If we have no request for publication of a certain dataset from its owner or publisher but we have a request from one or more potential consumers then we can publish the required dataset using ODN by providing one COMSODE dataset. The number of consumers is not important regarding the resulting number of COMSODE datasets because we do not want to multiply the number of resulting datasets by the number of consumers. Consumers consume datasets. Datasets might exist even without the consumers. A new consumer of the dataset does not mean a new dataset. However in order to be able to prove the interest of potential consumers in some datasets we should record their requests. Suppose, from the previous example, the two public bodies publishing public contracts. Suppose they already publish their collections somehow on their own without asking our help. If we have consumers who seek to publish the collections in ODN because they want to exploit some features of ODN (e.g., APIs) then, for the purposes of the COMSODE project, we count both collections as one dataset.

In the project we can also publish datasets for our own purposes. We can publish them even if no one of the owners or publishers asks us to do so and if no consumer requires publication of the particular dataset. If we consider some datasets to be very important to us we can publish them using ODN. In this case we will play the role of the internal user.

Table 1 below summarizes basic rules for deciding whether there are one or more COMSODE datasets in situations that differ in who is the ODN user and many of the users there are.

Situation ID	**Same schema ?**	**Same topic?**	**# (external) users**	**# consumers**	**Internal user?**	**# COMSODE datasets**
S1	Yes	Yes	N	N/A	N/A	N
S2	Yes	Yes	1	N/A	N/A	1
S3	Yes	Yes	0	1-N	N/A	1
S4	Yes	Yes	0	0	Yes	1

Table 1: basic rules for deciding the resulting number of COMSODE datasets

Sometimes it might be possible to split a candidate dataset (after evaluating the topic, schema and user) to two or more datasets. Each candidate should be evaluated in order to decide whether it represents a single or more than one dataset. There are several possible reasons for splitting a dataset to more datasets which should be considered:

1. If data items in the dataset contain sub-items which can be
   a. meaningfully linked from different datasets;
   b. meaningfully reused for different purposes by various consumers;
2. If there is a subset of data items in the dataset which can be
   a. meaningfully linked from different datasets;
   b. meaningfully reused for different purposes by various consumers;

However when splitting a candidate into more fine grained datasets the first three conditions of Definition 6 still hold true. Hence, **all the data of a subset must have the same topic, the same schema and the same ODN user**. On the other hand, when comparing the derived subset they should differ from each other in schema or topic. This might sound like it implies that the candidate dataset from which the subsets are derived must have had the topic or schema wrongly identified in the first place. The point is that if it is possible to separate some candidate dataset then it might be also possible to identify subtopics of the main topic that unites all the data of the candidate.

For example, consider we have a general topic “administrative control” and that different public sector bodies in different domains like drugs, food and drinks, telecommunications etc might perform this topic. This would result in splitting the original dataset into more different datasets - one dataset for each of these domains.

Similarly, when it is possible to derive a subset that can be used on its own then this subset can contain only a part of the schema of the original candidate.

There are no exact rules for splitting candidate datasets into more detailed subsets. It is not even required to perform the splitting. Some common sense should be applied. In the ideal scenario the granularity of the dataset should be discussed with the owner or publisher because it will probably depend on the context and needs.

5 Open Data Definition

In order to achieve a common understanding of the term Open Data a definition of this term is provided in this section. This definition is based upon a definition of Open Data from the Open Data Cataloguing Strategy of the Czech Public Administration [1] which adapts the Open Government Data principles described in [4].

Based on the Open Data definition in [1], Open Data is any data published on the Internet that is:

1) **Complete** - all data of a dataset should be published if possible. Completeness might be defined by the legislation or by the owner of the datasets. Because the owner is authorized to make decisions about his or her datasets s/he can specify what data make up a complete dataset.

2) **Easily accessible** - data should be accessible and discoverable using common tools and methods.

3) **Machine-readable** - data should be available in machine-readable formats. According to [3] machine-readable format is "*structured so that software applications can easily identify, recognize and extract specific data, including individual statements of fact, and their internal structure.*"

4) **Using commonly owned (open) standards** - formats that are freely available to anyone without restrictions (open formats) should be used, other formats may be used if it is possible to easily transform such formats into some open formats using a freely available application.

5) **Available under explicitly stated terms of use (license) which allows its reuse with minimal restrictions** - terms of use (license) must be clearly specified and published so anyone can access them.

6) **Available to the potential users for minimal possible costs** - where charges are made for open datasets, general principle of the PSI re-use charging framework should be applied. Thus "*charges shall be limited to the marginal costs incurred for their reproduction, provision and dissemination*" [3]. Charges exceeding the marginal cost might be applied only on specific occasions where the responsible subject provides data upon request and where costs of fulfilling such request significantly exceeds the average costs of reproduction, provision and dissemination.

According to [1] Open Data should also be:

7) **Primary** - if possible published data should be the same as the data originally collected. Apart from this general rule the following types of datasets should be considered as primary datasets:
    a. basic data or data from the public registries, especially if it is treated as reference data by the legislation,
    b. aggregated data in case that it is not possible to publish the original data for some objective reason, e.g. due to the possible privacy violation,
    c. aggregation of other publicly available (open) data in case that the aggregation process or algorithm is described and links to the source data are provided.

8) **Timely** - publication of the data should not be unnecessarily delayed, it should be made available in reasonable time frame given by the tasks necessary to publish the dataset.

9) **Non-discriminating** - no individuals, group of individuals or other type of subject should be restricted from accessing or using the dataset.

10) **Permanent** - dataset should be available on-line at least for a period of time specified by its owner.

A dataset is considered to be compliant with the Open Data definition if it meets at least the attributes 1-6. Attributes 7-10 are optional. This approach allows statistics and aggregate data to be published as Open Data.

6 Dataset Attributes

In order to be able to assess datasets and select suitable set of datasets that will be published in the COMSODE project each of the datasets or candidate datasets should be described using a set of attributes. Therefore a rich set of metadata has been identified.

Because the proposed set of metadata contains quite large set of attributes it might be too time consuming to describe every datasets with all the attributes. Besides this, some of the attributes must not be applicable to every dataset. Therefore, we distinguish between mandatory and voluntary attributes. Mandatory attributes must be filled in for every dataset or dataset candidate. In contrary it is allowed to leave the voluntary attributes blank or fill them in only cases when the required information is available.

Complete description of the mandatory and voluntary metadata attributes is provided in the document “Dataset attributes” attached to this deliverable. In this section we will provide only description and rationale for the mandatory attributes. A description of the mandatory attributes is provided in Table 2:

Table 2: Description of the mandatory metadata attributes

ID	Attribute name	Attribute description
A1	Dataset ID	Unique ID of a dataset
A2	Dataset name	Name of a dataset in English
A72	Dataset name in local language	Name of the dataset in the local language, i.e. language of the country of origin of the dataset
A4	Owner	A person or an entity that owns a dataset, i.e. s/he holds rights to the dataset or s/he is legitimate to make decisions about the dataset
A10	Description	Description or documentation describing collection process of the data, intended use of the data, limitations of the data etc.
A11	Topic	Domain of the dataset (culture, finance, agriculture, health)
A16	Current formats	Current formats in which data of the dataset is available. It can be structured formats (CSV, XML, etc.), semi-structured formats (HTML, etc.), non-structured textual formats (text, machine-readable PDF, DOC, etc.) or non-structured images (scanned documents). All available formats should be listed here, even if data in some formats is not publicly available on the Internet.
A18	Target data formats	Data formats in which data is planned to be released
A22	Existence of a schema	Attribute that captures information whether the dataset has its schema defined or not. Schema exists in case that it is formally defined and described or in case that we are able to describe the schema even if the formal description is not available.
A24	Schema expressed	Schema of the data expressed for example in a form

		of a semi-formal (UML, ER) or formal model (ontology, XML schema).
A25	Data items have identifiers (keys)	Attribute which specifies whether data items (objects, entities or rows) have values which uniquely identify them in the data set. It also says whether the identifiers are natural or not (=artificial)
A26	Identifiers expressed in the schema	Attribute that captures information whether the data schema specifies identifiers (keys) of data items.
A41	Incremental data	Information whether periods contain only increments (only new data items) or there can be changes to data items from previous periods.
A42	Terms of use/licence	Terms of use or licence under which the data is released
A47	Potential consumers of the data	Individuals, entities or groups that might consume the published dataset in order to make use of it
A48	Secrets	Data contains personal information or business secrets
A49	Distributed dataset	Attribute that captures information whether the data is unique to one subject or it is produced/published by more subjects in a distributed manner (dataset specific to a particular institution vs. datasets published by each institution in the EU)
A55	Status	Current status of a dataset. Active dataset is actively curated, inactive datasets are no longer curated or no longer available. Even datasets released only once can be active as long as they are curated.
A62	Creation Date	Date of creation of the resource
A64	Valid Date	Date (often a range) of validity of a resource.
A68	ODN User	An entity that requires our help (help of the members of the COMSODE project) with publishing data on an ODN instance.
A69	ODN provider	An entity that runs the Open Data Node (ODN) for one or more customers.

Mandatory Attributes Rationale:

A1 - Dataset ID: Every dataset needs unique ID, it can be the URL of the dataset.

A2 - Dataset name: Every dataset needs a name so we can refer to.

A72 - Dataset name in local language: In order to be clear about to which dataset we are referring to while communicating with the responsible person (owner/publisher) of the country of origin of the dataset we should use the dataset name in local language (unless the communication is held in English).

A4 - Owner: We have to know who is authorized to make a decision to publish a certain dataset as Open Data

A10 - Description: Enough detailed description provides the necessary information about the dataset and helps consumers to understand how they should interpret the data.

A11 - Topic: Assigning one or more topics to datasets helps to classify them into domains. COMSODE project needs to work with datasets from various domains.

A16 - Current formats: It is necessary to know in which formats the data are available. We consider formats that enable processing by machines (i.e. are machine readable) and formats that are not machine-readable (scanned documents). Machine readable formats are those which are structured (e.g., CSV, XML or semi-structured (e.g., HTML). They also include textual formats (TXT, machine-readable PDF or DOC) which can be processed by machines in some way (full-text indexing, extracting data on the base of some structural patterns, etc.). There can be two or more files containing the same data in different formats, e.g. CSV, XLSX and XML.

A18 - Target data formats: This attribute captures important information about the future plans regarding some datasets.

A22 - Existence of a schema: For analysis and use of a dataset, it is important to know whether a schema is defined for the particular data because it determines how the data can be processed and manipulated.

A24 - Schema expressed: In order to understand the importance of the schema and its relationships to other datasets, we need to know the structure and semantics of that dataset. Therefore, it is important for us to know whether an explicitly expressed schema of the dataset already exists somewhere.

A25 - Data items have identifiers (keys): Information whether the items present in a dataset can be uniquely identified is important for linking datasets, i.e. to decide whether the dataset should be published according to the Linked Data principles. It is also important to know if the identifiers are natural (= real-world) or artificial (e.g., auto-generated in a sequence). Natural identifiers can be directly used to create their URL forms. It is possible to develop URL forms based on artificial keys as well. However, artificial keys are often dataset specific and, therefore, it requires knowledge of the particular dataset in order to properly identify some entity. If artificial keys are used in some dataset it should be always analyzed whether more suitable identifiers can be used instead.

A26 - Identifiers expressed in the schema: Knowing that the identifiers are explicitly expressed in a schema of the dataset might save us significant amount of time needed to analyze what the identifier are.

A41 - Incremental data: It is important to know whether updates to the data have impact to the previously published data or they are only new incremental data.

A42 - Terms of use/licence: Terms of use or license determines what operations with the data a user is authorized to do. License or terms of use determines the legal openness of Open Data.

A47 - Potential consumers of the data: COMSODE will be successful only when the datasets will find their consumers and when applications will be built on top of them. Therefore, it is crucial to identify potential consumers or their groups.

A48 - Secrets: Some datasets might include personal information or business secrets. Datasets with those kinds of secrets must be carefully anonymized and aggregated.

Therefore, this attribute provides us important information that is needed to properly assess the effort needed for COMSODE to publish such dataset.

A49 - Distributed dataset: We need to distinguish whether the dataset is published in a centralized or decentralized way. The decentralized way is interesting because of integrating distributed parts and republishing them in a common format.

A55 - Status: During the project, some datasets identified in the early stages might disappear or they might become obsolete. We need some attributes to flag obsolescence or no longer available datasets.

A62 - Creation Date: We must know when the dataset was created in order to know how old it is and to assess how often it changes.

A64 - Valid Date: We must know whether the dataset is still valid or what is the validity period of the dataset. We should focus on publication of the valid datasets.

A68 - ODN User: For the purpose of the project reporting according to the project success criteria we must know how many datasets were published using the Open Data Node. We propose to distinguish a dataset based on its topic, schema and the ODN user. Therefore we need the ODN user attribute.

A69 - ODN provider: Number of the ODN instances is one of the project's success criteria. Therefore, there should be several ODN providers. We need to know if and which ODN instance will be used to publish a particular dataset.

7 Dataset Selection Workflow

Dataset selection workflow consists of the following steps:

1. **Development of the initial list of potential collections of data items (initial candidate list)** - every COMSODE project member will propose some potential collections of data items that might become datasets later. Moreover, the broad public will be given an opportunity to propose a collection of data via an online tool on COMSODE.eu - the project website.

   It is the goal of the COMSODE project to validate usability of Open Data Node for publication of various types of datasets by using it to publish several datasets from different domains. Project members thus collect information about potential datasets in order to compile the initial list to choose from.

2. **Assessment of the dataset candidates using the hard criteria** - once the amount of proposed potential collections of data items reaches amount allowing sufficient freedom of choice, the first step of dataset selection can be performed. In this first step each candidate collection of data items will be assessed using a set of basic criteria called "*Hard criteria*" (see below). Every candidate that fails to satisfy some of the hard criteria is removed from the list. All remaining datasets are passed to the next step.

3. **Selection of datasets to be published during the COMSODE project** - in this step the proposed datasets (at this stage we can call the proposed collection of data items dataset because the proposed collections were checked for compliance with the COMSODE dataset definition in the previous step) that result from the previous step are assessed using the soft criteria. Based on the attributes of each of the proposed datasets a score is calculated using a formula described in Chapter 10. Alongside this score, an expert estimation of effort needed to publish the datasets will be performed. In the next step proposed datasets are sorted according to the achieved score. The list will be evaluated from the top (datasets with highest score) to bottom. Datasets at the top of the list are the most preferred. However datasets which are either too costly to publish (based on expert effort estimation) or belong to same topic as too many other datasets will be ruled out. Ruling out datasets with frequently repeated topics or domains might make room for dataset from some other topic/domain and thus increasing diversity of the final dataset list. Evaluation will continue until enough datasets to meet the project goals are selected (i.e. sufficient number of datasets, etc.).

4. **Selection of subset of datasets to be published as Linked Open Data during the project** - datasets selected in the previous step are assessed using Linked Open Data selection criteria. These criteria help us to identify datasets that are suitable for publication using the Linked Data principles. Based on the attributes of the dataset a score is calculated using a formula described in Chapter 9. Similarly to the previous step we will try to balance dataset scores, estimated effort and diversity to select sufficient amount of datasets that will be published as Linked Data to meet project goals.

Preliminary assessment of the effort needed to publish the proposed collections of data items, or datasets respectively, will be performed during the step 3. However, detailed analysis of the selected datasets might reveal that publication of some of the selected datasets is not feasible. In that case respective datasets will be removed and they will be replaced with other datasets based on the assessment of their score, estimated effort and type.

We will need to balance similarities and differences of the selected datasets. **A dataset that has the same topic and (optionally) schema with another already selected COMSODE dataset but has a different owner/publisher is interesting to us**. We want to have datasets that we can integrate together from various publishers (ideally from various countries and possibly also various ODN instances) and provide them in a unified way to consumers. We will not be able to demonstrate this if we do not have such kinds of datasets.

On the other hand, it is our goal that published datasets cover a (reasonably) broad range of different topics. We do not want to have datasets which all have the same or very similar main topics. With a broader range of topics covered we can target broader set of our potential ODN users and we will be able to better demonstrate various features of ODN.

During the public consultation to the revision of the PSI Directive 2003/98/EC a set of core dataset domains was proposed [2]. Therefore we are going to publish datasets from the domains described in Table 3 during the COMSODE project. However, we are not going to limit the final set of datasets to these domains only. If there is other interesting dataset in other domains they might be published as well.

Table 3: Possible dataset domains and example topics [2]

Domain	**Example topics**
Companies	company/business register
Crime and Justice	crime statistics
Earth observation	meteorology, /weather, agriculture, forestry, fishing
Environment	pollution levels, energy consumption
Geospatial	topography, postcodes, national maps, local maps
Education	list of schools, performance of schools
Finance and contracts	calls for tender, future tenders, local budget, national budget (planned and spent)
Government Accountability	government contact points, election results, legislation and statutes, salaries (pay scales), hospitality/gifts
Global Development	aid, food security, extractives, land
Health	prescription data, performance data
Statistics	national Statistics, Census, infrastructure, broadband penetration, wealth, skills
Transportation	public transport timetables, access points

However, if we prefer diversity of datasets too much it will lead to an unacceptable amount of work out of range of the COMSODE capacities. Note that this criterion does not consider the difference of schemas. We do not prefer the situation where datasets with the same topic (or close topics) have different schemas – this would lead to unacceptable amount of work.

Balancing similarities and diversities of datasets might seem difficult. On one side, we want datasets to share their main topics. On the other side we want to achieve a certain level of diversity of those topics. However, we do not see similarities and diversities as strictly contradictory. We want to achieve the “ideal” state where we work with an interesting amount of main topics and for each topic we have several datasets with homogeneous schemas. This maximizes usefulness (our datasets cover the interest of many consumers) and minimizes the amount of work we have to perform (less schemas means less work). We will achieve the desired balance in discussion within the project team and thus we will not set hard metrics for optimization of similarities/diversities of the selected datasets.

8 Dataset Selection Criteria

Datasets that are going to be published in the COMSODE project are selected using a set of criteria. There are 3 main types of criteria:

- **Hard criteria** - if a dataset candidate does not satisfy any of these criteria it means that this particular candidate will not be selected for publication during the COMSODE project;
- **Soft criteria** - every dataset candidate that satisfies all the hard criteria is going to be assessed using these criteria; based on the assessment a priority ranking will be given to each dataset;
- **Linked Open Data selection criteria** - datasets selected for publication in the COMSODE project will be assessed for suitability to be published as Linked Open Data; similarly to the soft criteria assessment priority ranks will be given to the candidate datasets based on the assessment of the Linked Open Data selection criteria.

8.1 Criteria Description

8.1.1 Hard Criteria

Description of the hard criteria is provided in Table 4.

Table 4: Description of the hard criteria

ID	Name	Description	Related Attributes	Values
CC1	Dataset definition compliance	Is it compliant with the definition of the COMSODE dataset?	A11, A22, A24, A68	Yes/No
CC2	Open Data publication feasibility	Is it possible to publish dataset as Open Data (see the Open Data definition above)?	A4, A5, A42, A48	Yes/No
CC3	Mandatory attributes compliance	Does it have all mandatory attributes filled in?	see table 2	Yes/No

Dataset candidate must meet all the Hard Criteria in order to be evaluated. Datasets that do not meet any of the Hard Criteria will not be evaluated and they will not be published during the COMSODE project.

Hard Criteria Rationale:

CC1 - Dataset definition compliance: Assessment of the COMSODE dataset definitions that allows us to properly and consistently specify and describe datasets, and to count them as well. By assessment of the compliance with the definition we are able to distinguish between one standalone dataset, group of datasets or subset of another COMSODE dataset.

CC2 - Open Data publication feasibility: If an identified dataset is a COMSODE dataset according to Definition A it must also be possible to publish it as open data. If the owner of the dataset does not allow publication or it already publishes the dataset but not as open data (e.g., there are restrictions given by a license under which the dataset is published) then we will not be able to publish the dataset as open data. We need to quickly rule out dataset that the project won't be able to publish. In case when owner or publisher is unknown and license is unknown or does not allow publication, COMSODE have to rule such dataset out as we are not permitted to publish the data as it is and we are also unable to negotiate the terms which would allow us to publish it. It may be the case that we will be able to negotiate with the owner or publisher of the dataset and find agreement on publishing the dataset as open data.

CC3 - Mandatory attributes compliance: We need to know values of the all mandatory attributes to be able to decide whether all hard, soft and linked open data selection criteria are met. We also need them to plan how we will work with the dataset if selected for publication by the project.

8.1.2 Soft Criteria

There are 2 groups of soft criteria:

1. **Technical criteria** - criteria related to the technical aspects of datasets, e.g. data formats;
2. **Assessment of subjects criteria** - criteria related to the subjects of the datasets, e.g. customer, published.

Description of the soft criteria is provided in the table 5.

Table 5: Description of the soft criteria

ID	Name	Group	Description	Related Attributes	Allowed values
TC1	Current formats	Technical	Assessment of the machine-readability of the format.	A16	0-5
TC2	Schema	Technical	Is the dataset schema expressed in formal or semi-formal form?	A24	0, 1, 3, 5
SC1	Availability of the ODN user	Subjects	What type of ODN users does the dataset have? A dataset with confirmed ODN user has higher priority.	A68	1, 3, 5
SC2	Willingness to publish data	Subjects	Is the owner (or potential publisher) cooperative and eager to publish the dataset with COMSODE (using ODN)?	A68, A4, A5, A6	0, 1, 3, 5
SC3	Willingness to run ODN	Subjects	Is the owner (or potential publisher) willing to run his/her own ODN instance?	A69	0, 1, 3, 5

Soft Criteria Rationale

TC1 - Current formats: We prefer datasets that already exist in some machine readable formats. The dataset does not necessarily has to be already published somewhere in that format. It is enough for us if it exists somewhere in the data storage of the owner who agrees to cooperate with us.

TC2 - Schema: It is harder for us to correctly understand a dataset which does not have a schema expressed. The schema should describe or at least imply the semantics of the dataset. It should be ideally complemented with some human readable documentation. Without the schema we risk to work with the dataset incorrectly and it will be a more time consuming work.

SC1 - Availability of the ODN user: We prefer datasets which have an ODN user associated. According to Definition 2, an ODN user is an owner, a publisher or a consumer of the dataset. Therefore, do not interchange the term ODN user with the consumer of the dataset. We prefer an owner or a publisher. If there is no such ODN user, it can be some consumer. And, in certain cases it can also be us, internal consumer, if there is a reason for it (e.g., we will need some datasets for linking).

SC2 - Willingness to publish data: An owner who hasn’t published the dataset yet but is willing to do so is much better for us than someone who does not. We do not have resources in the COMSODE project to convince owners. Another good situation is the case when a publisher who has already published the dataset wants us to republish it in a better way using an ODN instance. Another compromised solution is the case when the owner or some publishers have already published the dataset but not in a suitable way. If there are no objective obstacles, we can harvest the dataset without the cooperation of the owner/publisher and republish it in our ODN instance in a better way (with respect to SC1).

SC3 - Willingness to run ODN: COMSODE project aims not only at publishing datasets but also on implementing ODN instances. Therefore, a higher priority is given to datasets whose owner or publisher is willing to run his or her standalone ODN instance.

8.1.3 Linked Open Data Selection Criteria

Description of the Linked Open Data selection criteria is provided in Table 6.

Table 6: Description of the Linked Open Data selection criteria

ID	Name	Description	Related Attributes	Allowed values
LC1	Data identifiers	Score for data identifiers. Datasets with natural keys are preferred because they are easier to link to other data.	A25, A26	0, 3, 5
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).	A24, A25, A26	0-5

Linked Open Data Selection Criteria Rationale:

LC1 - Data identifiers: Keys/identifiers are crucial for linking datasets. We must assess what type of keys/identifiers are used in a dataset because these types of keys determine how easy or difficult would be to link datasets. Identifiers/keys are important for the assessment of the linking potential and for the estimation of the effort needed to publish data as Linked Data. Datasets with natural keys are the best candidates for linked datasets because the identification of such entity does not need arbitrary agreement between parties about how to identify that entity.

LC2 - Linking potential: In order to the power of the Linked Data technology we need to publish datasets that actually interlinks. Therefore, we must assess the potential of a dataset to be linked to other datasets. There are two types of good candidates for linked datasets. Datasets of the first type are those that will be frequently used by other datasets like addresses - address is frequently used among datasets. We expect datasets of this type to be linked from other datasets quite often and therefore they are good candidates to us.In the second type are those datasets which provide many links to other different datasets. E.g. datasets that contain data about business entities, addresses and that use standard vocabularies or taxonomies.

We do not expect many datasets to link towards datasets of this type but they are still good candidates for use because of the high number of links leading to other datasets - probably datasets of the first type. Basically we will need a balanced set of datasets of both types because they are complementary. It is not possible to develop interlinked data space only with datasets of a single type.

8.2 Criteria Assessment Guidelines

This section contains guidelines how the assessment of each of the criteria should be performed.

8.2.1 Hard Criteria

[CC1] Dataset definition compliance

Possible values and their meaning:

- **Yes** - dataset candidate is compliant with the COMSODE dataset definition (see Definition 6)
- **No** - dataset candidate is not compliant with the COMSODE dataset definition

Guidelines for the assessment of the CC1 criterion are provided in Table 7 below. The text below Definition 6 also provides some additional explanation with examples.

Table 7: CC1 assessment guidelines

Value	Guidelines/rules
Yes	● All data items in the dataset must share the same main topic. ● All data items in the dataset must share the same common schema. ● There must be one and only one ODN user who requires publishing the dataset on an ODN instance of some ODN provider (the ODN user does not necessarily have to be the provider). *(Note: If there are two different ODN users, then we have two datasets.)*
No	● Some data items in the dataset do not share the topic with the rest of the data items. ● Some data items in the dataset do not share the schema with the rest of the data items. ● There is none or multiple ODN users who requires publishing the dataset on an ODN instance. *(Note: Answer “No” probably leads to separating the collection to more datasets)*

[CC2] Open Data publication feasibility

Possible values and their meaning:

- **Yes** - it is possible to publish dataset as Open Data
- **No** - it is not possible to publish dataset as Open Data

Guidelines for the assessment of the CC2 criterion are provided in Table 8 below.

Table 8: CC2 assessment guidelines

Value	Guidelines/rules
Yes	● Dataset has already been published as Open Data (under some open license) OR owner of the dataset is willing to publish it as Open Data (under some open license). ○ For the definition of the term Open Data please see Chapter 6. ● Dataset does not contain any secrets (attribute A48) OR it is possible to remove the secrets (e.g. anonymization, aggregation)
No	● Dataset has not been published as Open Data AND the owner of the dataset is not willing to publish it as Open Data. ● Dataset does contain some secretes AND it is not possible to remove these secrets, i.e. anonymization is not possible or the dataset would be useless when anonymized.

[CC3] Mandatory attributes compliance

Possible values and their meaning:

- **Yes** - dataset has all mandatory attributes filled in
- **No** - one or more mandatory attributes of the dataset are left blank

Guidelines for the assessment of the CC3 criterion are provided in Table 9 below.

Table 9: CC3 assessment guidelines

Value	Guidelines/rules
Yes	● Each mandatory attribute is filled in. ● Value of each mandatory attribute is within the defined range. ● Value of each mandatory attribute is meaningful.
No	● Some of the mandatory attributes are left blank. ● Some of the mandatory attributes have values out of the allowed range. ● Some of the mandatory attributes has meaningless values, e.g. “ABC” instead of the dataset name.

8.2.2 Soft Criteria

[TC1] Current formats

Possible values and their meaning:

- 0 - format that is very difficult for machine processing, it is not structured
- 1 - format that is very difficult for machine processing but it is application independent
- 2 - format that allows partial structuring of the data
- 3 - structured format that partially allows description of the schema
- 4 - structured format that allows description of the schema and partial expression of the semantics and linking as well
- 5 - structured format that allows description of the schema, expression of the semantics and linking of the data

Guidelines for the assessment of the TC1 criterion are provided in Table 10 below.

Table 10: TC1 assessment guidelines

Value	Format	Application independence	Structured data	Existence of schema description language	Semantics	Linking
0	PDF	No	No	No	No	No
0	DOC(X) RTF	No	No	No	No	No
1	TXT	Yes	No	No	No	No
2	HTML	Yes	Partly	No	No	No*
2	XLS(X)	No	Partly	No	No	No
3	CSV	Yes	Yes	Partly	No	No
3	JSON	Yes	Yes	Partly	No	No

4	RDB[#]	Yes	Yes	Yes	No	Partly[##]
4	XML	Yes	Yes	Yes	No	Partly**
4	OData	Yes	Yes	Yes	Partly	Partly
5	RDF	Yes	Yes	Yes	Yes	Yes

* HTML enables linking documents. However, it is not possible to link data entities.
** XLink language can be used for linking. However, it does not support semantic linking as RDF does.
# RDB = Data exported from a relational database management system (i.e. dump of a relational database) in a form of SQL DDL expressions (= relational schema definition) and SQL DML (INSERT) expressions (= data).
## SQL DDL enables to express key-foreign key pairs within one database (= a set of relational tables)

If some dataset is available in more than one format then we use score to evaluate the best available format to use.

[TC2] Schema

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

Guidelines for the assessment of the TC2 criterion are provided in Table 11 below.

Table 11: TC2 assessment guidelines

Value	Guidelines/rules
0	● There is neither a logical schema of the dataset expressed in a machine readable notation (e.g. XSD for XML, SQL DDL for relational database, etc.) nor a textual description of the schema which can be read by a human.
1	● There is no logical schema of the dataset expressed in a machine readable notation. However, there is a textual description of the schema which can be read by a human.
3	● There is no logical schema of the dataset expressed in a machine readable notation. However, there is a conceptual schema expressed in some standardized graphical notation, e.g. UML class diagrams or ER diagram.
5	● A logical schema of the dataset is expressed in a machine readable notation. E.g., there is an XSD schema, listing of SQL DDL commands or an ontology expressed in RDF Schema/OWL, etc.

[SC1] Availability of the ODN user

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

Guidelines for the assessment of the SC1 criterion are provided in Table 12 below.

Table 12: SC1 assessment guidelines

Value	Guidelines/rules
1	● There are no potential consumers identified. ● None of the owners or publishers of the data asked our help to publish the dataset. ● But the dataset is important to at least one member of the COMSODE project consortium so there is an internal dataset user.
3	● None of the owners or publishers of the data asked our help to publish the dataset. ● But one or more consumers of the dataset requested publication of the dataset by the COMSODE project.
5	● Owner or publisher of the datasets asked the COMSODE project for help regarding publication of the dataset.

[SC2] Willingness to publish data

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

Guidelines for the assessment of the SC2 criterion are provided in Table 13 below.

Table 13: SC2 assessment guidelines

Value	Guidelines/rules
0	● Owner (or the potential publisher) is not willing to publish the dataset as Open Data and we are certain about it. ● Willingness of the owner (or the potential publisher) to publish the dataset as Open Data is unknown for us and it is not possible to clarify the attitude of the owner/potential publisher.
1	● Based on the information received from the owner/potential publisher or based on information from reliable sources it is clear that the owner/potential publisher has not made the decision yet, however it seems likely that the database will not be published. ● It is possible to make a qualified estimation that the dataset will probably not be published as Open Data.
3	● Based on the information received from the owner/potential publisher or based on information from reliable sources it is clear that the owner/potential publisher has not made the decision yet, however it seems likely that the database will be published. ● It is possible to make a qualified estimation that the dataset will probably be published as Open Data.
5	● Owner or the potential publisher is willing to publish the dataset and his or her attitude has been confirmed. ● Owner or the potential published published the dataset as Open Data.

[SC3] Willingness to run ODN

Possible values and their meaning:

- 0 - owner/publisher is not willing to run its 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 - ODN user is willing to run its own ODN instance

Guidelines for the assessment of the SC3 criterion are provided in Table 14 below.

Table 14: SC3 assessment guidelines

Value	Guidelines/rules
0	● Owner/publisher is not willing to run its ODN instance and his or her attitude was confirmed. ● We do not know or we are not able to judge the willingness of the owner/publisher to run its ODN instance.
1	● Owner/publisher has not decided yet, however it will probably not run his or her own ODN instance. Judgment is based on information obtained from it or on information from reliable sources.
3	● Owner/publisher has not decided yet, however it will probably run its own ODN instance. Judgment is based on information obtained from it or on information from reliable sources.
5	● ODN user is willing to run its ODN instance and his or her attitude was confirmed.

8.2.3 Linked Open Data Selection Criteria

[LC1] Identifiers of data

Possible values and their meaning:

- 0 - data has no identifiers
- 3 - artificial identifiers are used
- 5 - natural identifiers are used

Guidelines for the assessment of the LC1 criterion are provided in Table 15 below.

Table 15: LC1 assessment guidelines

Value	Guidelines/rules
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

If there are more items/entities in the dataset and the nature of identifiers vary between entities we use score to evaluate which was the most frequent type of identifier used. If the types of identifiers are evenly distributed in the dataset the highest possible score is used.

[LC2] Linking potential

Possible values and their meaning:

- 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

Guidelines for the assessment of the LC2 criterion are provided in Table 16 below.

Table 16: LC2 assessment guidelines

Value	Guidelines/rules
0	● Entities of the dataset do not exist in other datasets. ● Entities of the dataset do exist in other datasets, but it would be really difficult to interlink the datasets.
1	● Some entity or entities of the dataset also exists in 1-2 other datasets.
2	● Some entity or entities of the dataset also exists in 3-4 other datasets.
3	● Some entity or entities of the dataset also exists in 5-6 other datasets.
4	● Some entity or entities of the dataset also exists in 7-10 other datasets.
5	● Some entity or entities of the dataset also exists in more than 10 other datasets.

9 Dataset Selection Formula

We are going to use weighted sums to calculate overall score for the dataset using the soft criteria, or the Linked Open Data selection criteria.

9.1 Non-linked datasets selection criteria formula

Based on the soft criteria each of the dataset candidate is scored according to the following formula:

$$R_{score} = W_{TC1}*TC1 + W_{TC2}*TC2 + W_{SC1}*SC1 + W_{SC2}*SC2 + W_{SC2}*SC2$$

Where:

- R_{score} is the total score resulting from the assessment using the soft criteria
- W_{TC1} is the weight of the TC1 criterion
- TC1 is the value of the TC1 “*Current formats*” criterion
- W_{TC2} is the weight of the TC2 criterion
- TC2 is the value of the TC2 “*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

Weights of the criteria used in the formula are provided in Table 17. However, these are only initial weights. We suppose that we will need to augment them during evaluation of candidate datasets later.

Table 17: weights of the soft criteria

ID	Criterion	Weight	Weight ID
TC1	Current formats	0,1	W_{TC1}
TC2	Schema	0,1	W_{TC2}
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}

9.2 Linked datasets selection formula

Based on the Linked Open Data selection criteria each of the dataset candidate is scored according to the following formula:

$$L_{score} = W_{LC1}*LC1 + W_{LC2}*LC2$$

Where:

- L_{score} is the total score resulting from the assessment using the Linked Open Data selection criteria
- W_{LC1} is the weight of the LC1 criterion
- LC1 is the value of the LC1 "*Identifiers of data*" criterion
- W_{LC2} is the weight of the LC2 criterion
- LC2 is the value of the LC2 "*Linking potential*" criterion

Weights of the criteria used in the formula are provided in Table 18. Again, they are only initial weights. We suppose that we will need to augment them during evaluation of candidate datasets later.

Table 18: weights of the Linked Open Data selection criteria

ID	**Criterion**	**Weight**	**Weight ID**
LC1	Identifiers of data	0,6	W_{LC1}
LC2	Linking potential	0,4	W_{LC2}

10 Examples

Examples that demonstrate assessment of the dataset candidates using the hard criteria, soft criteria and the Linked Open Data selection criteria are presented in this section. Description of the example datasets is provided in Table 19.

Table 19: Description of the example datasets

Dataset ID	Dataset name	Country	Description	Topic	Secondary topics	ODN User
CTO_39	Register of providers of services of electronic communic a-tions	Czech Republic	Register of providers of services of electronic communications and of their fields of endeavour according to the general authorization	Registratio n of a company	Internet access provider, telecommuni cations	Czech Telecomm unication Office
CTIA_1	CTIA Inspection s	Czech Republic	Data about inspections performed by the Czech Trade Inspection Authority	Contract	Consumer protection	Czech Trade Inspection Authority
CTIA_2	CTIA Sanctions	Czech Republic	Data about sanctions imposed by the Czech Trade Inspection Authority	Contract	Consumer protection	Czech Trade Inspection Authority
CTIA_3	CTIA Bans	Czech Republic	Data about goods banned by the Czech Trade Inspection Authority	Contract	Consumer protection, product safety	Czech Trade Inspection Authority
MICR_1	List of public sector bodies	Czech Republic	List of public sector bodies in the Czech Republic and basic information about them	Public institution		Ministry of Interior of the Czech Republic

Datasets of the Czech Telecommunication Office (CTO), Czech Trade Inspection Authority (CTIA) and the Ministry of Interior of the Czech Republic (MICR) were selected as examples. Assessment of these datasets is provided in Table 20.

Table 20: Assessment of the example datasets

Dataset ID	Hard Criteria Assessment	TC1	TC2	SC1	SC2	SC3	LC1	LC2	Rscore	Lscore
CTO_39	Passed	4	5	1	5	0	5	5	2,6	5
CTIA_1	Passed	3	1	1	5	5	5	2	3,6	3,8
CTIA_2	Passed	3	1	1	5	5	3	1	3,6	2,2

CTIA_3	Passed	3	1	1	5	5	3	1	3,6	2,2
MICR_1	Passed	4	5	5	5	0	5	4	3,4	4,6

Datasets of the Czech Telecommunication Office and the Ministry of Interior of the Czech Republic are currently provided in better machine readable formats that the datasets of the Czech Trade Inspection Authority. However CTIA representative has become member of the COMSODE User Group and expressed interest in the Open Data Node. Therefore CTIA datasets received better SC3 score which contributes to better Rscore. We should prefer these datasets over the datasets owned by the Czech Telecommunication Office and the Ministry of Interior of the Czech Republic.

Dataset of the Czech Telecommunication Office received better linking potential scoring due to the larger number of datasets from the telecommunications domain which might be linked to this example dataset.

Public sector bodies (MICR_1) might be hypothetically linked to any datasets that contains data about some of the public sector bodies. Thanks to this a high linking potential score was awarded to the list of public sector bodies. According to these criteria CTO_39 and MICR_1 are more suitable candidates for publication in a form of Linked Open Data.

11 Summary

This document describes the overall approach of the COMSODE project to the selection of suitable datasets that will be published during the project. The definition of the term ***dataset*** is provided. This definition is by no means overarching definition of dataset. It is a COMSODE project specific definition because it reflects some of the project specific issues like implementation of the Open Data Node (ODN). In the document, a selection workflow and datasets assessment criteria are described. Examples and conclusions are provided at the end of this document.

Proposed approach to selection of the datasets suitable for publication can be summarized in following 4 steps (the selection workflow):

1) Development of the initial list of potential collections of data items (initial candidate list);

   A set of attributes was defined to describe candidate collections of data and candidate datasets. This set contains 22 mandatory attributes and other voluntary attributes that are suitable for capturing more detailed information about datasets.

2) Assessment of the dataset candidates using the hard criteria;

   In the second step a set datasets to be published during the project is selected. Priority rank for each of the candidate dataset is calculated using the soft criteria and the defined formula. It is an aim of the project to select datasets with high priority ranks that are feasible to publish within the scope of the project that also cover at least some agreed upon set of topics (e.g. finance and contracts or state inspection results) that will allow us to demonstrate various features of the Open Data Node. The project members will choose preferred domains. Public call to propose domains or collections of data will be used to gather data that will support the decision-making.

3) Selection of datasets to be published during by COMSODE project;

   Our aim is to select datasets with high priority ranks that are feasible to publish within the scope of the project that also cover at least some agreed set of topics (e.g. finance and contracts or state inspection results) that will allow us to demonstrate various features of the Open Data Node.

4) Selection of subset of datasets to be published as Linked Open Data during the project.

   Last step of the dataset selection workflow is dedicated to the selection of a subset of the selected datasets that is suitable for publication as Linked Open Data. Analogous approach to the previous step is applied. However different set of criteria is used because different aspects must be taken into account when selecting Linked Open datasets.

Three types of criteria are going to be used for the assessment of the candidate collections of data and the candidate datasets:

- ○ *Hard selection criteria* - if a dataset candidate does not satisfy any of these criteria it means that this particular candidate will not be selected for publication during the COMSODE project;

- *Soft selection criteria* - every dataset candidate that satisfies all the hard criteria is going to be assessed using these criteria; based on the assessment a priority ranking will be given to each dataset;

- *Linked Open Data selection criteria* - datasets selected for publication in the COMSODE project will be assessed for suitability to be published as Linked Open Data; similarly to the soft criteria assessment priority ranks will be given to the candidate datasets based on the assessment of the Linked Open Data selection criteria.

Qualified estimate of the effort needed to publish the datasets is going to be performed, with the goal to select a balanced set of datasets for publication. We are going to balance priority ranks, effort needed to publish the datasets and diversity of datasets in terms of problem domains and types of data of the selected datasets.

In order to help project members to perform the assessment of the candidate collections of data and candidate datasets a set of guidelines is also provided in this document.

Using the approach described in this document will lead to to selection of a set of datasets that covers enough topics, domains and types of data that will allow demonstration of various features of the Open Data Node and it applicability for different scenarios. Moreover, the project will also help to select datasets where effort needed to publish them does not exceed the scope of the project.

References

1. CHLAPEK, Dušan, KUČERA, Jan a NEČASKÝ, Martin. Koncepce katalogizace otevřených dat veřejné správy ČR [Open Data Cataloguing Strategy of the Czech Public Administration]. In: *Vláda ČR* [online]. September 2012. [cit. 2012-11-28]. Available from: http://www.korupce.cz/assets/partnerstvi-pro-otevrene-vladnuti/otevrena-data/Koncepce-katalogizace-otevrenych-dat-VS-CR---zkracena-verze.pdf

2. European Commission. Consultation on guidelines on recommended standard licences, datasets and charging for the re-use of public sector information. In: *Europa.eu* [online]. 2013 [cit. 2013-11-12]. Available from: http://ec.europa.eu/yourvoice/ipm/forms/formpdf/PSIguidelinesen.pdf

3. European Union. Directive 2013/37/EU of the European Parliament and of the Council of of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information. In: *EUR-Lex* [online]. 26 June 2013 [cit. 2013-08-23]. Available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:EN:PDF

4. Sunlight Foundation. Ten Principles for opening up government information. In: *Sunlight Foundation* [online]. 11. August 2010. [cit.: 2012-02-20]. Available from: http://sunlightfoundation.com/policy/documents/ten-open-data-principles/