DELIVERABLE D5.5

Contribution to international standards and best practices

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	22. 7. 2015
Author(s)	Jan Gondol, Lubor Illek, Ivan Hanzlik and other members of the COMSODE team
Responsible of the deliverable	Jan Gondol
Email	gondol@gondol.sk
Reviewed by	User Board Members
Status of the Document	Final
Version	1.0
Dissemination level	PU Public

Table of Contents

0) Meta Information

Deliverable 5.5 is an updated version of Deliverable 5.3. The major updates are included in the following sections:

- **New Chapter 5** ("Contribution to standards and best practices by COMSODE") -- policy activities, cooperation with other projects, work on the creation of **open data standards**.
- **New Chapter 6** (Open Data Ready standard description).
- **New Chapter 7** ("Case study: Open Government Partnership as platform for advancing Open Data") -- **identified best practices** for cooperation between the government and civil society actors.

Deliverable 5.3 was drafted in 2014 and **has now been updated and extended to reflect the situation in mid-2015**. Also, new activities that took place between Month 12 and Month 21 within the COMSODE project have been included to create this document. **Feedback** from COMSODE consortium members and COMSODE User Board members has been collected for Deliverable 5.3. This feedback has been fully implemented and numerous other small edits have also been made.

Description of the deliverables (according to the DoW) is as follows:

D5.3 – Contribution to international standards and best practises (Month: 12, Responsible: MoI SR)
Supplemental documentation for D5.2 giving publishers (mainly public bodies) broader context about international Open Data standards and best practices. – first draft version.

D5.5 – Contribution to international standards and best practises (Month: 21, Responsible: MoI SR)
Supplemental documentation for D5.2 giving publishers (mainly public bodies) broader context about international Open Data standards and best practices. Documentation will be published on project website.

Task 5.3 (M8-M21): **Contribution to international standards and best practices** (**MoI**; all)
The material developed within this task will supplement the above mentioned methodologies and put them into context with international Open Data standards and best practices. It will give the reader a quick crash course about an even broader context, describing briefly what Open Data is and why it is important. Material will be aimed at people with IT background; general parts are expected to be understandable also for the general public. This documentation will be released under a suitable open license too.

The beneficiaries commit to contribute to and participate in focused concerted actions, themed seminars or special interest groups, for example the European Data Forum, LOD2, OGP, OSIN, Open Data Initiative, Services for smart eGovernment, EU-standardisation initiatives, openforumeurope.org, etc. and to try to convince other public bodies and governments to adopt methodologies delivered by COMSODE as their own official recommendations.

In other words, **this is an introductory document, a quick-to-read text, a "crash course"** (to quote the DOW).

Intended audiences

- **Decision-maker** (data provider), not deeply informed (e.g. from local government) -- low IT education. People asked him/her about publishing data, legislation tells them to publish it, they heard some buzz words (read about the topic in a magazine) -- what should they do? We give them some pointers.
- **IT professional** (works for the Decision-maker) and was asked to publish LOD. What should they know? The following pages provide an overview.
- **OD User** -- general public or IT professionals who want to use the data, as well as professional reusers (journalists, lawyers, engineers, etc.).

1) Introduction

There has been a lot of public discussion about Open Data in the recent years. The subject was approached from the **perspective of policy** (by politicians), **technology** (by IT professionals), as well as practical **day-to-day utility** (by regular citizens who use the apps enabled by Open Data), among others.

When it comes to Open Data, different stakeholders often have very **different goals** (the consumer would prefer to have all the data, the public organization is thinking about the costs/benefits and potential risks of publishing). They also speak in **"different languages"** (the politician doesn't understand what the "IT guy" is saying, etc.) and therefore it's often hard for them to understand each other and work together. This material intends to provide an **introduction to all the major stakeholders** written in a way that's easy to understand. When it's understood what the main issues are from the perspective of everyone involved, it may be easier to work together.

Perhaps you, the reader, are in one of the roles. Hopefully this text will provide **a brief introduction, quick overview and a useful pointer** to where you can learn more.

1.1) COMSODE project introduction

This material is best understood within the **context of the COMSODE project**. COMSODE (Components Supporting Open Data Exploitation) is a EU-funded project which has three main outputs (as in many places throughout the document, we'll be simplifying): 1) creation of **software tools** to help organizations publish their data, 2) creation of **methodologies**, and 3) **publication of datasets** as well as example applications using them. The tool "Open Data Node" (ODN) created by COMSODE is free to use, free to modify and free to distribute. The second component, methodologies, teach organizations about Open Data best practices and show them how to use the Open Data Node (and partially other tools). And finally, COMSODE itself plans to publish 150+ datasets, using its own tools and its own methodologies (as well as several applications showing how to search through the data). The goal is to showcase that the technology works and how things may look when our software and best practices are both put into action.

The COMSODE project consists of several "work packages" which have **deliverables which could be useful**. This document is one of the deliverables but there are others, some of which are much more technical or detailed and could be of great help, especially to the technical users. We do not intend to introduce all the deliverables, only to provide a brief overview and choose a few documents that could be of particular interest.

Work Package 2, "Architecture and design", deals with collecting requirements for the creation of Open Data Node. Feedback from potential end users has been collected, so that the software fulfills all functional requirements. Also, criteria for the selection of datasets were compiled. Work Package 3, "Data analysis" focused primarily on the selection of the 150+ datasets published by the COMSODE project, as well as the technological details to make this happen (data transformation, cleansing, etc.). Work Package 4, "Development of software components and tools" is mostly about software development itself (the flagship product is called Open Data Node and can be found at http://opendatanode.org), but **Work Package 5 ("Development of methodologies")** is especially important to mention here.

Deliverable 5.1, "Methodology for publishing datasets as open data" is a methodology for publishers (mainly public bodies) about the steps and phases needed for publishing Open Data. It starts from the beginning of the publication activity ("what and why I must publish as open data?") to the result ("we have dataset suitable for publishing"). This is a **detailed catalog of steps** and activities that guide the organization throughout the publication process.

Deliverable 5.2 was called "Methodologies for deployment and usage of the COMSODE publication platform (ODN), tools and data" (draft version) and it's more specific than Deliverable 5.1. Think of it as a **manual for the tools** that COMSODE provides. The later, updated version, was published as **Deliverable 5.4**.

Finally, there is the document you're reading right now. The intention of **Deliverable 5.5** is to introduce the broader context of international Open Data standards and best practices. We wrote it to be an easy-to-read introduction for the various stakeholders. We want to **paint the "big picture"** (this in contrast to the previous documents, which are very specific) and point the reader to places where he or she can find more information. Deliverables 5.4 and 5.5 are final versions of D5.2 and D5.3, respectively.

There are other work packages in project COMSODE that we have not discussed for the sake of brevity. Just remember that the most important details are in Deliverables 5.1 and 5.2. **Technologically inclined readers** may find the deliverables from **work packages 2, 3 and 4** interesting as well.

1.2) Understanding Open Data

Open Data needs to be seen in a wider context. In a democracy, openness about what the government does and how it spends the public resources, is absolutely **crucial to the proper functioning of an open society**. Understanding and examination of government's activities is only possible when it's known what these activities are: when the public knows about the budgets and spending, when it knows about the plans and their implementation on **all government levels**, from local to national and international.

While in traditional democracies, access to such data has been possible through information legislation (such as Freedom of Information Acts / FOIAs), proper processing of such data is only possible when it's published in **machine-readable formats**, processable by computers. Imagine the difference between processing a table with a few thousand rows printed on paper versus a having a file importable to Microsoft Excel where it can be searched, filtered, and processed.

But availability of machine-readable formats is not enough: according to the **Open Definition** (http://opendefinition.org/), "A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike." In other words, **data needs to be freely available and re-usable**. There should be a permissive **license attached**. Having access to a file about government spending under a non-disclosure agreement (NDA) simply won't cut it. That kind of data isn't open data, even when it's machine readable and available in a digital form.

But Open Data is **not only about financial information**, it is much bigger. The government and various publicly-funded organizations produce vast amounts of other data, which could be used in new ways and bring about economic benefit, if it only was freely available. If the government created something for some primary purpose (e.g., mapping information or public transportation information), there could be **new and sometimes even unexpected positive new uses**. This is information re-use.

The line of thinking is this: if **the public has already paid** for the creation of the data (by paying the taxes, thus enabling the government and public organizations to fulfill their roles) then **it should be freely available**. Of course, there are limitations: releasing some data may not be desirable for the society as a whole (think about identities in the database of all ID cards: releasing them could enable massive identity fraud). While it's sometimes challenging to see all implications, the philosophy should be to **release all data unless it's somehow protected or sensitive**. Notice that we're not saying "all useful data": the reason is that some data may not be perceived as particularly useful by their "owner" but may be extremely useful for some individuals nevertheless (more on this later in the examples).

The astute readers may have noticed that the Open Definition doesn't speak specifically about governments or public bodies. And this is indeed true: Open Data is a concept that also applies outside the public sector. **Commercial organizations, non-profit universities**, as well as the entire **non-government sector or even individuals** can become not just consumers, but also publishers of Open Data. There are data catalogs specifically for the nonprofit sector and other entities outside of the government, they are easy to find on the Internet. The COMSODE project wants to make Open Data useful to businesses primarily as consumers, but we encourage the businesses to also publish their own data for the use by others.

In other words, Open Data is **NOT a "special category"** (or "different kind") of data -- it could be ANY machine-readable data in open formats distributed **under a permissive open license**. A computer file containing data could "magically" become Open Data simply by properly licensing it under a permissive license (such as Creative Commons Attribution). As long as it's "free to use, reuse and redistribute" (with no other strings attached, as we saw above), it is Open Data. By adhering to technical standards and by having the license attached, this data file has been given the attribute "openness".

The following chapters focus on the public sector, and argue that Open Data is an integral part of the Open Government. It should be also pointed out, however, that **Open Government includes more than Open Data**: in several countries including the Slovak Republic, **Open Education** has also started to become one of the topics discussed in Open Government activities (making educational resources open, providing open access to scientific publications). In the future, **Open Source** software may become another hot issue. The rationale is the same in all these areas: since the public financed the creation of the resources (data, educational materials, or software), it should be able to re-use them in new and even unexpected ways. And more importantly, while data, content and software are great places to foster openness, there are **other crucial areas, such as open governance**, open creation of policies with public participation, etc..

There are great **examples of what Open Data can do** for the society. For financial data, OpenSpending (https://openspending.org/) shows what public money is spent on. **Public transportation** applications are also based on publicly available open data, and so are many **map applications**. Some data found became unexpectedly popular, such as data about public toilets in Great Britain. Other examples from the **business generated with open data** are the open data incubators currently reusing open data with a business perspective. Examples can be found at http://www.finodex-project.eu/ (section: Results --> Start-ups).

2) For the Decision-maker

The following chapter is intended **primarily for politicians and executives** in public organizations. It discusses the benefits of opening up data from several perspectives and suggests what steps should be taken to get the data published.

2.1) Benefits of Open Data

Benefits of Open Data Can are manifold: a good summary is provided by the ePSIplatform in their Topic Report No. 2013/08 from August 2013[1]. Building on the work of Capgemini Group analysis, they recognize the three main areas of benefits.

[1] http://www.epsiplatform.eu/sites/default/files/2013-08-Open_Data_Impact.pdf

1) Benefit to government

- Increased tax revenues through increased economic activity
- Creation of jobs
- Reduction in data transaction costs
- Increased service efficiency (esp. through linked data)
- Increased GDP
- Encouraged entrepreneurship (economic growth)

2) Benefit to private sector

- New business opportunities for services / goods
- Reduced costs for data conversion (no need to convert into raw formats anymore)
- Better decision-making based on accurate information
- Better-skilled workforce

3) Benefit to NGOs / civil society

- Better informed monitoring
- New venues for project action: building tools/applications
- Increased sustainability potential through increased capacity

The ePSIplatform Topic report points out to **other benefits** mentioned by the Open Data Research Network, such as:

- Open data empowering transformation in specific sectors such as the financial one;
- Open data generating new kinds of Public-Private partnership models;
- Open data policies accelerating the process of private businesses releasing its own data;
- Open data disrupting traditional business models, lowering entry barriers and making the services industry more modular.

We can also point out to the benefits, such as **benefits for individual users**, not mentioned above. When data is available, mobile applications for smartphones that can make life easier can be created:

- Better navigation facilitated by mapping data, databases of points of interest, etc. (route planning, public transportation schedules).
- Easier interaction with the government (crime reporting, potholes / fixmystreet / odkazprestarostu)
- Etc.

Also, **more general benefits can be considered: better interaction** between governments and the citizens, **building of mutual trust** and improved public perception of those who publish the data pro-actively, and finally help with data cleanup from the users. If an organization publishes a dataset which contains errors, the users may notice them. When a feedback mechanism is provided, users can suggest corrections, which can be accepted or rejected by

the publisher. Cleaned data benefits both parties: the publisher (who receives corrections for free) and for the users as well (who offer corrections and receive cleaned up data in return).

There are other potential benefits that you may be aware of. This list is by no means exhaustive. It also highlights the fact that it may be very difficult to measure some of the outcomes, especially in financial terms.

The issue of benefits of Open Data has also been addressed by COMSODE Deliverable 5.1[2], which we recommend as supplemental reading (see the introductory chapter).

2.2) Economic Value of Open Data

Is it possible to quantify the impact of Open Data? We could see above the range of benefits that Open Data can provide and some of them are hard to quantify. How can one put an exact price tag making an individual's life easier or on improved decision making? Models that estimate the economic impact of Open Data **cannot provide accurate numbers but we believe that they are still useful**. Trying to quantify the effect of cost savings or economic growth can lead to deeper thinking about where the published data can provide most value. Is such value significantly greater than the costs related to publishing it? If the answer is known to be positive, it should be a priority to publish such data as soon as possible. This is an investment with great payoff.

How about quantifying the value of transparency? One could argue that lowering corruption by even very few percent would immediately pay off any costs related to financial transparency. This remains a speculation, which is possibly true but hard to quantify. But transparency can go beyond detecting possibly corrupt behavior, it helps us understand what is happening inside organizations. When we combine data from multiple sources, we can see **sources of inefficiencies** like overpaying for energy in an old building with improper insulation, inefficient scheduling of work by state employees revealed by data about their activities. Once people from the outside see what's happening on the inside, they can offer a fresh “outsider” perspective on how to improve the situation.

We could debate whether the well-known McKinsey study from October 2013 as well as similar studies are substantiated. The "liquid data" concept is interesting and the estimated economic value of opening up data (in hundreds of billions of dollars annually) is staggering. **Do these claims hold water? Can they be trusted?** We invite you, the reader, the prospective data publisher, to read the study personally (an executive summary is available), contrast it to other studies available, and come to your own conclusions.

[2] http://www.comsode.eu/index.php/deliverables/

A lot of data, once published in an open format, can appreciate economically and new value can arise from its re-use. Making a cost-benefit analysis in light of existing Open Data studies can be a good exercise. While we think that it may be **impossible to properly calculate all the economic impacts**, the potential for added value is there, even if it's hard to quantify. Further research in the domain of the economic impacts of open data is strongly encouraged.

For those interested in understanding the **value chain of open data**, we recommend the paper "Open data is fact now-when does the reuse start?" by Marc de Vries and Georg Hittmair[3].

2.3) Legal considerations

In a number of countries, access to information is a right guaranteed by the Constitution. **Constitutions often guarantee conflicting rights as well (such as the right to privacy)**, so these are sometimes in tension and have to be properly balanced.

When it comes to government information, countries typically have **Freedom of Information Acts** (FOIAs) which regulate in more detail who can request information, who has the legal obligation to provide it, under which circumstances, etcetera. In Slovakia, FOIA guarantees access to information to "everyone": the citizenship or legal age are not a limitation, and neither is personhood. A foreigner, a child, or even a company (legal person) can request information.

FOIAs may not always regulate the **electronic (digital) availability**: while in Slovakia it's possible to request the answer in a digital format, the law doesn't specify what kinds of formats must be provided. So if the public office prints out a document from Microsoft Office Excel on paper, scans it back to PDF and returns the result in this way, it's hard to challenge such behavior based on FOIA alone.

Specialized legislation may exist which deals with re-use of public sector information, most notably Directive 2013/37/EU on the re-use of public sector information (also known as the **PSI Directive**) and Directive 2007/2/EC known as the **INSPIRE Directive**. It is outside the scope of this document to discuss the INSPIRE directive (which is specific to geographic / geospatial data) but we will mention several key principles of the PSI Directive.

(Please be aware that Directive 2013/37/EU has amended the older Directive 2003/98/EC. All Member States are expected to transpose this directive into their national legislation by June 2015.) It is helpful to be aware that the PSI **directive is a so-called "minimum directive"**, introducing a minimal set of basic obligations common across Member States. In practice, this means that every Member State has to fulfill all obligations of the Directive but is **free to introduce legislation that goes beyond what the PSI directive requires**. This means that the

[3] https://www.w3.org/2013/share-psi/wiki/images/a/a3/Lisbon_deVries_hittmair.pdf

position of those who require data can be actually even stronger than what the PSI Directive says. If that is the case in a particular Member State, the Decision Maker / politician may have more obligations with regards to making data available. Such situation would favor the users of the data. If you are a politician, be aware that you can introduce stronger requirements and **push for more openness**. If you do so, keep the potential costs (financial, organizational) as well as benefits in mind as you try to find the right balance.

A very good overview of the PSI Directive was provided by the Open Knowledge Foundation at http://blog.okfn.org/2013/04/19/the-new-psi-directive-as-good-as-it-seems/ and is definitely worth reading.

So far we mentioned the **several legal frameworks** related to publishing the data: the very general national **constitutions**, more specific **national legislation**, as well as the overall **EU framework**. It is helpful to be acquainted with all of these legal acts to properly understand the legal context. Legal counsels of organizations (who understand the organizational context) as specialized information legislation consultants are definitely worth consulting.

There are **legal issues on the consumption side** as well (which relates to the actual use of the data). For example: to what extent is it possible to **rely on the data**? In other words, if an organization downloads information from a government data portal, can it be sure that the data is guaranteed to be correct? Here is a specific example: if a business uses a government-related web site to check a VAT number of their business partner abroad, can it be sure that it is OK to charge zero VAT for the cross-border business transaction? **What if the data is no longer updated?** What if a "man-in-the-middle attack" occurs and somebody modifies the data in transit?

Paying more attention to the above issue reveals several layer of problems: 1) guarantee that the data comes from the **correct entity** (was this data downloaded directly from the organization's official website of from a copy somewhere on the web?), 2) guarantee that the data hasn't been **modified in transit** (was this data transmitted through an encrypted connection or has its integrity been secured through other verifiable means?), 3) guarantee that the data is **fit for legal purposes** (which requires much more than simple technological measures and must be dealt with legislatively). All of these issues are complex and Open Data professionals are only beginning to tackle them properly. They become more complicated when data from multiple sources is combined. Steps are being taken to address these problems but for now, it's helpful to be even aware of the questions: "To what extent can I rely on this data? **Is this 'FYI' or can this be used in the court of law, if needed?** How can I verify this data and prove it is correct?" Decision makers should take steps to publish data in ways that gives their users legal guarantees (where appropriate) or at least a certain level of quality service.

2.4) Action Plan for publishing Open Data

As a decision maker, it is helpful to be aware that a **number of steps need to be undertaken** before the data is published. These include issues like:

- Mapping and understanding **organizational processes** (the reason: knowing what is going on, where data is produced, what the information flows are)
- Understanding the **technical requirements** (IT infrastructure)
- Understanding the related **costs** (either capital costs required for hardware, human resources costs, etc.)
- **Prioritization** (what data should be published first?)
- **Release schedule** (what will be published when?)

It is apparent that some of the questions can be **best answered by people other than the Decision Maker**. Forming a working group, or less formally, simply involving others in a **team effort**, can lead to the best results. From the experience of others, and similarly to other projects, inviting people with a "can do" mentality can be crucial.

There are several principles that might be helpful to follow:

- First publish data which is **most likely to be used soon**. (User-centric approach: publish what the potential users are already requesting.)
- Take it **one step at a time**: start with a few datasets and keep adding more.
- If the costs of publishing are very low (e.g. the data is already available for internal use, it just hasn't been published externally), don't worry too much how that data might be useful. There may be **unexpected uses**, especially when this data is re-combined with data from other organizations. People will figure things out. You can feel your burden lifted because you don't have to worry about all the uses. You don't have to worry about managing the process of apps creation (procurement, coding, hosting, GUI design),... -- all of these will be handled by the users / developers who decide that your data is useful and they want to work with it.

In this section, we pointed out only a few ideas which we consider very helpful. **For a much more thorough and systematic approach, see Deliverable 5.1** of project COMSODE which deals with the issues that will need to be discussed in much more detail. We recommend to work on the issues in a team because an understanding of both high-level issues known by the Decision Maker as well as understanding the technology background (known to the IT professionals) is crucial. COMSODE Deliverable 5.1 can be a basis for forming the **organization's own Action Plan**.

2.5) Common Concerns

A common worry is: **how much money** will need to be spent? And the good news is: typically the costs can be kept reasonably low. In many cases, it's possible to open up data without new infrastructure investments and with the use of freely available open source software. As a result, it's **possible to do a lot with very few resources**, with only time of IT personnel spent with bootstrapping the project (and modest server resources). COMSODE project in particular has made software and methodologies free, so organizations which are able to reuse them can save on procurement costs.

Services like **GitHub** (github.com) are extremely beneficial: they make collaboration on open source software and on published data (especially small- and medium-sized datasets) easy and free of charge for public repositories. There is even a **dedicated page for the government** at https://government.github.com/ which says: "Agencies use GitHub to engage developers and collaborate with the public on open source, open data and open government efforts. GitHub even renders common formats like text, CSV, and geospatial data." And indeed: data can be stored on Github and collaborated upon. If a mistake in data published on Github is discovered, **users can make a correction and offer their correction through a mechanism called "pull request"**. Anyone can suggest the corrections and the dataset owner can decide, whether to accept ("merge") the request, suggest a change to the correction before accepting it, or refusing it outright. Such actions are public, which can serve as a **record of the interactions / changes**, very helpful for the individual users as well as the government organization.

The switch to services like GitHub would mean opening up **unprecedented levels of transparency** and foster the spirit of collaboration. This often goes against the toxic culture festering in some public organizations but from the perspective of the Decision Maker, the target reader of this chapter, such change (at least in small degrees) is exactly what **should be encouraged and rewarded**. When the culture of collaboration gets its foothold, magic things with Open Data can happen. We recommend the article at http://readwrite.com/2014/08/14/github-government-ben-balter-open-source ("GitHub May Actually Be Dragging Government Into The 21st Century").

Finally, do not worry excessively about publishing datasets which contain errors. There is no need to hide the status quo. Users will be **fine with having the same data as you do, even if it's not perfect**. Imperfect data is usually better than no data, especially if the publisher is clear about the level of confidence and quality, explains what errors might be expected, and ideally provides a mechanism for reporting "bugs" and fixing them. Also, if there are legislative or other barriers to publish the full datasets: **even partial content may be useful**. While it's legally impossible to publish a database of all citizens because of privacy protection, aggregates with no personally identifiable information can also uncover very interesting insights: what are the

popularity trends of first names over time? In what regions is a particular last name / family name popular? What is the age distribution and how is it changing? How many people were born on a particular day? And so on. Organizations often flat out refuse to publish data because it's "protected" but while that may be true, **significant portions of such data can be made available without doing anything illegal**.

3) IT professional

This chapter is intended **primarily for workers – usually from IT department – tasked with technical execution of data publication** once the decision of doing so was done. But bottom-up approach also works, so while someone within the organisation wants to approach his boss asking for particular data publication it can really ease the process if he can show how means for „technical matters“ are already found.

While in most situations IT workers are viewed by managers as responsible and competent for solving all issues of data publication, we can only recommend involve various more roles within the organisation in this process, especially:

- **data owners** – they should know which data are processed, data value for organisation and for public, and should have detailed knowledge about data quality
- **legal department** – they can decide if data publication is possible and to what extent, they should produce the license for data use and usually deal with formal requests for data publication from outside the organisation
- **the organisation's management** – Open Data approach have several benefits for whole organisation, so it hopefully can be stated as global principle or strategy, and probably everyone knows how simple all things can be while having director standing on your side
- **IT department** – in optimal situation they must solve only particular problems relating to data extraction, preparation and making data publicly accessible

While there are many different ways how to solve technical tasks and manuals for all of them can be easily found, in this chapter we only briefly describe the most important of them. Many recommendations dealing with **data catalogue, data schemas, ontologies, identifiers and required software architecture** for Open Data publishing is covered by COMSODE Deliverable D5.1. If you intend to use Open Data Node (ODN) software created by COMSODE project, we can only recommend reading methodolgy of it's use in Deliverable D5.4 and User Manuals, available for download at http://www.comsode.eu/index.php/deliverables/

3.1) Data accessibility

While there are many interesting technical topics – eg. producing data linked to other data, automating dataset transformations, data quality assessment, data enrichment – for sure **the crucial point is to make content of published datasets easily accessible to users**.

In current practice there are two distinct methods how to facilitate access to the data:

- batch access implemented by **making available files** with contents of dataset
- query based access to selected parts of **data available through application interface** (API)

Basic difference between these methods lies in the **degree of interaction** between data provider and data user. Batch access through files is a service very easy to setup and maintain at the side of data provider, which does not need nor enables any active interaction with data while selecting and accessing them. At the other side API based approach requires usually substantial data processing at the data provider side for each user request, but its strength lies in minimising volumes of data transmitted to the data user and providing most actual data. There are hybrid approaches of course, for example ODN software supports both of these data access methods.

Batch access uses classical services for making files available online. Nowadays most common means are by use of HTTP, FTP or even BitTorrent protocol. But access can be also provided upon request, e.g. by e-mail or done completely offline, for example by using data storage medium (for extremely large datasets this can indeed be the most practical method of transfer) – but while designing your access procedures keep in mind that one of the benefits of Open Data approach is minimising complexity and needed resources to communication between data producer and data user.

The most important part concerning Open Data access through files is to **determine appropriate format of the file in which are data stored**. Here the key measure is to enable and simplify automated processing of stored data. Try to look at it from the point of view of user: how difficult (measured by needed software/algorithm, its computational complexity and accuracy of the result) it is to identify data in file? To extract data from it? To search for specific data? Not suitable are the data file formats which are proprietary as they impose unnecessary cost to the data user or sometimes their processing is not readily available at all.

Most data file formats can be divided into several categories:

- **formats not enabling automated data processing** or making this processing at great cost or inaccurate – for example, picture and media file formats, and formats to store unstructured data, like text documents, PDF, HTML – these are not suitable for Open Data publication

- **proprietary data formats** limiting their usage – for example DOC, XLS – not suitable for Open Data publication

- **open formats designed to hold structured data** – mostly CSV, JSON, XML – these are the main mean how to access Open Data

- **advanced formats** specifically designed for holding and processing large or semantically varied data – mainly RDF – these are standard for Linked Open Data and other high quality needs

Access to the data through API is more complex to set up. Firstly, the data provider must have data internally available in a structured form (database) and has to establish several infrastructure components:

- **Storage** for published data, both in terms of capacity and software handling – usually dedicated database instance. From the security perspective it is generally not a good idea to connect users directly to the production data stores.

- Application logic **processing user queries**. In some setups it is suitable to have a processor of some common query language – for access to the data stored in RDF, there SPARQL query language, which is derivative of SQL user for access to the data stored in relational databases. Otherwise it is possible to have defined special set of query constructs suitable to the domain model of data, which are usually translated into classical SQL queries.

- Publicly available **services where users specify queries and retrieve the results**. The most common ones are: Representational state transfer (and its implementation as RESTful API), SPARQL endpoints or custom designed Web Services.

- Maintain online **connection to retrieve data** from production data stores and enforce security. This can be rarely accomplished by a simple DB connect, so in this place data provider must use some ETL tools.

Various software packages integrate some or all these components to simplify process of building and maintenance of whole data access infrastructure for the data providers. Again, one example of such **integrated system** is ODN (Open Data Node) developed in COMSODE project, available at http://opendatanode.org/

Data providers can publish data by using their own infrastructure, or put them in the cloud (for example use a CDNN / content delivery network or application services such as DaaS / Data as a service). Some services are commonly provided to data producers at the central level in the country, particularly as part of national data catalog or OpenData portal.

For the data user one of the critically important tasks is the **ability to effectively obtain updates of data** (of course there are exceptions, particularly datasets containing historical data). The user must be able to answer the following:

- When will be more recent data available (as opposed to already received data)?
- How can it be determined which data are new or updated and which were deleted?
- How to reconstruct contents of the dataset at a particular time from the past?

Whatever method for data access is chosen, it is important for the users to have **reliable access** to the data. This can be achieved by adhering to several basic rules:

- **Invariability** - Points of access (for example URL to data file) should be kept constant. The same applies to the method of access, data structure and identification of individual objects within the dataset.

- **Capacity** – Data must be accessible to the user, in terms of both selected time and transmission capacity.

- **Rules** – Users must understand what they can do with the data (which is addressed by the license) and know the accuracy of the data in terms of their error rate and liability (for example if the data are legally binding)

3.2) Information security

There should be security considerations in the following areas:

- securing the IT environment of the data publisher organisation
- security (in the sense of „protection“) of the published data

Data publication and subsequent interaction with the users who work with these data almost always lies in **different security context from the production data processing** in the organization.

Therefore best practice is to put infrastructure necessary for data publication place in a **separate computing environment**. This measure is typically implemented at the network layer. Data publication infrastructure is usually placed into the network segment/environment for organization's servers directed toward Internet (as these provide public access to data).
One of the primary benefits of Open Data approach in terms of security is that published **data are not subject to any confidentiality requirements** – they are public. Non-public information should be excluded from data publishing. For example when Open Data Node (ODN) software is used and ODN servers are placed in the demilitarized zone (DMZ) or similar security zone, **all inputs should already be "cleaned"** from what data that has not to be disclosed - such cleaning should not be implemented in the ODN using its data transformations.

The question of data **integrity and authenticity** is the most interesting. There is a whole continuum of options in this field, but typically it is one of two situations: data are provided for information purposes only or data can be used as legally binding.

It is necessary to note that **even with "informative data" provider is responsible** to some extent for ensuring the accuracy of the data (as their integrity) – the provider is usually legally obliged to do so. A violation of data integrity can have many negative effects: from damage to the reputation of the organization to excessive loads on personnel tasked with solving the existing problem (especially communication with users). If data publication is for informational purposes only, normally it is sufficient to ensure the protection of integrity at the same level as web servers of the organisation. The authenticity of the data is achieved at the level of metadata, that means by declaration.

If the intention is to be **data useful for legally binding purposes**, there should be given high attention to ensuring the integrity and authenticity. It should be noted that in this area there is no generally accepted technical standard for machine-processed data. Best practice is to implement a mechanism ensuring data integrity and authenticity outside the infrastructure for data publishing, which means to implement it in the internal production environment of organization. There is mechanism of **electronic signature used in this process**, which is applied to the entire dataset (but, however, then the partial access to data through API is unsigned), or separate signing selected dataset entities (ie. if the data are in tabular form, each row in the table is signed separately). The authenticity of the data is then guaranteed through the signing certificate (the subject of certificate is the data provider) and relevant certification path to the trust anchor.

It is necessary to foresee and implement solution for the data availability, since if data is to be used seriously, there must be the a certain guarantee of availability for the users, or they must be at least rigorously informed about the service level parameters. Such information should include the **acceptable use policy** for all resources of the infrastructure (eg. capacity limits for data downloading, the allowed frequencies of API queries).

Indeed accessibility protection is usually the reason for **detailed monitoring of the use of data** publishing infrastructure and it's services. We recommend to monitor the current status and resources usage, as well as store historical data for possible subsequent analysis.

While enabling the access through the API the good practice is **not to create direct access to the database or application server of production environment** of the organisation, but to use a separate tool for processing API requests. Main reason is the security (protection of internal systems from unwanted external access) and protection of resources (intent is to guarantee that internal servers and infrastructure are not overloaded regardless of the amount of user requests). If there needs to be direct access to the production systems, there should be requirements for maintaining security included in the design from the initial phases of creating

publication infrastructure in this case. There are special tools created with the intent to simplify this task, for example ODN software produced by the COMSODE project is easy deployable, self-contained package, accessible and Open Source, yet maintained and supported.

3.3) Licensing

Processed data may be covered by some special **legal regime** defining and possibly limiting their use and reuse (for example personal data) – and this is especially likely while considering government data. Also there are usually some special rules applying for “dataset“, either as the copyright for compilations or a sui generis right for collections of data.

While form and scope of protection **varies in each jurisdiction**, in all cases it is necessary to reconcile with the legal terms for all parties can be clear which processing of data is allowed and which is not:

- data publisher has to be certain that publication of data is not prohibited be some special or general laws, for example there are special rules for publication of data about the safety of atomic power plants, and there are generic laws defining processing of personal data
- data user wants to know in advance what types of activities with the data he/she is allowed to do and have some protection to be not denied of these rights afterwards

Formal communication of the rules governing particular dataset publication and use are commonly named **license**. For any data publication to be called Open Data, these are the minimum condition that must be met (according to Open Definition, http://opendefinition.org/od/):

- **Use** - The license must allow free use of the licensed work.

- **Redistribution** - The license must allow redistribution of the licensed work, including sale, whether on its own or as part of a collection made from works from different sources.

- **Modification** - The license must allow the creation of derivatives of the licensed work and allow the distribution of such derivatives under the same terms of the original licensed work.

- **Separation** - The license must allow any part of the work to be freely used, distributed, or modified separately from any other part of the work or from any collection of works in which it was originally distributed. All parties who receive any distribution of any part of a work within the terms of the original license should have the same rights as those that are granted in conjunction with the original work.

- **Compilation** - The license must allow the licensed work to be distributed along with other distinct works without placing restrictions on these other works.

- **Non-discrimination** - The license must not discriminate against any person or group.

- **Propagation** - The rights attached to the work must apply to all to whom it is redistributed without the need to agree to any additional legal terms.

- **Application to Any Purpose** - The license must allow use, redistribution, modification, and compilation for any purpose. The license must not restrict anyone from making use of the work in a specific field of endeavor.

- **No Charge** - The license must not impose any fee arrangement, royalty, or other compensation or monetary remuneration as part of its conditions.

While with Open Data approach data publisher loses some control over data once it is published, it can be legitimate for him to state some **limitations or conditions for their reuse**. The most common are (according to Open Definition):

- **Attribution** - The license may require distributions of the work to include attribution of contributors, rights holders, sponsors and creators, as long as any such prescriptions are not onerous.

- **Integrity** - The license may require that modified versions of a licensed work carry a different name or version number from the original work or otherwise indicate what changes have been made.

- **Share-alike** - The license may require copies or derivatives of a licensed work to remain under a license the same as or similar to the original.

- **Notice** - The license may require retention of copyright notices and identification of the license.

- **Source** - The license may require modified works to be made available in a form preferred for further modification.

- **Technical Restriction Prohibition** - The license may prohibit distribution of the work in a manner where technical measures impose restrictions on the exercise of otherwise allowed rights.

- **Non-aggression** - The license may require modifiers to grant the public additional permissions (for example, patent licenses) as required for exercise of the rights

allowed by the license. The license may also condition permissions on not aggressing against licensees with respect to exercising any allowed right (again, for example, patent litigation).

For any data publication license to be **useful and practical** for data user, it should be:

- **Explicit** – The license should be fully expressed in written form.
- **Stable** – Rules stated in the license and the license agreement as a whole should not change in time, except for special situations (for example reflecting change in laws).
- **Legally valid** – The license should be legally valid and covering most possible jurisdictions, minimally data publisher jurisdiction.
- **Easy to use** – The license and its full meaning should be easily comprehended for the data user.

It is not easy to create a license adhering to all principles stated above. Furthermore, in most practical situations data users want to combine data from several datasets, possibly from different publishers of even countries. While doing this he must also **combine related licenses** and find out their intersection – and if defining the license is seen as difficult, combining several of them brings whole new level of challenge.

For these reasons it is strongly advised not to create new license but first to try reuse some existing and well known licenses in this field.

More detailed description of licensing process is covered by special section in COMSODE Deliverable D5.1, or here:

- http://opendatacommons.org/faq/licenses/
- http://opendefinition.org/guide/data/
- http://opendefinition.org/licenses/
- For compatibility of licenses, see COMSODE Deliverable 5.4, Section 3.1 -- Compatibility Checker tool Licentia -- available at http://www.comsode.eu/index.php/deliverables/

For further reading, we also recommend LAPSI Licensing Guidelines[4].

[4] http://www.lapsi-project.eu/sites/lapsi-project.eu/files/D5.2LicensingGuidelinesPO%20(1).pdf

4) Open Data User

Some of the content in the previous chapters was covered in Deliverable 5.1 of the COMSODE project in great details. That deliverable, however, did not mention one important stakeholder: the Open Data user. For the purpose of this chapter, we will think of the user as someone who **intends to use Open Data** for personal or other purposes, including business use[5].

One could argue that writing about Open Data users falls **out of scope of this deliverable** ("Contribution to international standards and best practices") but we consider Open Data users so important that we want to make sure that this chapter is included, so that both Decision Makers and IT professionals **do not forget about those at the end of the pipeline** for whom all the Open Data activities take place. This was famously illustrated by the scene in Monty Python's Meaning of Life where doctors (after filling up the entire hospital room with the latest medical technology) couldn't remember what was still missing -- after thinking very hard for some time, they finally figured it out: the patient.

There is a **wide spectrum of IT skills** and the scale ranges from the very novices to technological experts in various areas of IT. We do not pretend to write great "how to" manual for everyone. Instead, this chapter will point out some issues that the users may come across and for the sake of simplicity will divide all potential OD users in **two groups: "Citizens"** (with just basic computer skills) **and "IT professionals"**. Please note that the IT professionals discussed here are seen in a role different from those mentioned above: IT professionals in Chapter 3 were people responsible for publishing (those on the production side) and those discussed here in chapter 4 are on the consumption side of things. Their responsibility is not to provide data to others (so they need not worry about things like attaching the proper license for redistribution) and instead they will want to do something for themselves.

4.1) Citizen

Regular users typically **experience data indirectly**, in a processed form (as opposed to raw form). This often happens **through visualizations in newspapers**, magazines and even interactive visualizations on the internet. Not everyone realizes that there is typically raw data hiding behind the pie charts and graphs found in print media and that this raw data is often available to be reused, reanalyzed, revisualized and reinterpreted -- for anyone.

Where to find such data? It is helpful to realize that a lot of data is scattered all over the web. Just including a keyword and googling on the web can return that data. For example, searching for "**statistics filetype:xls site:uk**" (without quotation marks) in Google returns tens of thousands of Excel files from the ".uk" domain (such as www.gov.uk, but also www.shell.co.uk --

[5] http://theodi.org/blog/5-ways-better-open-data-reuse

these are all examples from the first search engine results page for the example query). Modifying the search query to include different keywords, different web addresses or different file types (such as "filetype:csv") will help change the search or narrow it down. Relatively **few people grasp the power of Google** and realize that they can search for PowerPoint presentations, PDF publications, or Excel files (among others) and find things they never knew existed. Investment in improving search skills pays many times over and there are quality tutorials available, such as Google's own course: http://www.powersearchingwithgoogle.com/

"Googling around", however, may not be the best way to find data (but it may be extremely helpful to know about nevertheless). In order to find local government data, it's sometimes **better to use a data portal**. These are typically in the form of "data.gov.DOMAIN", such as data.gov.uk, data.gov.sk, etc.. This is not a rule, however. In order to find a data portal for a given region or field of interest, use web search.

Some data portals provide filtering and visualization tools or other useful functionality, others do not. But the **data can almost always be downloaded** in formats that can be imported to Microsoft Office Excel, LibreOffice Calc, and so on. Once downloaded and imported, such data can be **filtered, sorted, aggregated, and even graphed / visualized**. So getting to know spreadsheet software is also a worthy time investment. Many people are amazed once they learn what their spreadsheet can do. While we think it that a great way to learn Excel is in a quality instructor-led course (with a highly skilled instructor teaching only a small group of students), there are also thousands of **YouTube videos** which can help if attending a course is not an option. The downside of these videos is varying quality, the upside is range of topics and target skill levels.

There are advantages of using Open Data portals instead of search engines. One is content curation: while some data portals are little more than dumps of junk data, others are much higher quality. It is **easier to rely on a dataset with proper meta data downloaded from a data portal** than on a random spreadsheet found on the "wild web". Another advantage is the service provided by the staff administering the data portals. Contact information is typically available and the staff can often answer questions about data, as well as even assist with **acquiring data that is not currently available**. This of course depends on the scope of the data portal (which is sometimes geographical, sometimes topical, etc.), as well as the difficulty level of obtaining the data, as well as time constraints of the administrator and other factors (some are friendly and go an extra mile while others do not). Last but not least, there is metadata: open data portals typically include useful information about the datasets, its authors, update schedules, etc., which helps make educated decisions about the use of the data.

Another way to get access to interesting data is to write a **FOIA request** (FOIA is Freedom of Information Act) to the party that has the information but but hasn't published it (this will typically be a branch of government or a public organization). It may be best to consult this step with an experienced person. While some public organizations are happy to cooperate, many aren't and

unless there is a specific legislation that explicitly compels them to publish data and the requester knows how to enforce it, some organizations will do all they can to avoid any extra work.

If any technical issues sound too complicated, it is best to **find an IT professional and ask for help**. It's the same as any other IT-related area: when encountering a problem, it is useful to find a human who can walk us through the issue. The same is true for legal issues: if the only way to obtain information is through a FOIA request, it is best to consult those who are experienced with this, such as administrators of public data portals or people in watchdog non-profits who often write FOIA requests themselves.

Many users will probably **experience Open Data through the applications** that use these data, such as public transportation schedules, mapping applications, and so on. Therefore it is not absolutely necessary to learn new IT skills or learn how to use FOIA legislation in order to appreciate the results of Open Data initiatives. Learning about semantic web technologies is definitely not necessary for non-interested people. But **acquiring a few new skills can be a fun exercise** and will help the interested individuals find their own answer to questions such as: "Why should I care about Open Data?", or "Why is Open Data good?" -- "What's in it for me?"

The next sub-chapter focuses on exactly the people who might *create* such applications (often called "apps" for short).

4.2) IT professional

We talked about "IT professionals" in the previous chapters. The IT professionals discussed previously were responsible for publishing the data in open formats and were typically employed in organizations that were producing data. The **IT professionals we talk about here** are different: these are **users of Open Data who also have more extensive IT skills** than most of their peers.

Most users (consumers) of Open Data typically don't know much about information technology. They can download files, do some basic processing and analysis in a spreadsheet but they do not know how to process the data in an advanced way. This can be a serious limit: for regular users, finding and downloading the file (using a search engine or a data catalogue) or filing a FOIA request can be the only way how to get access to what they would like to obtain. But what if they cannot obtain the data in a format that suits them? This is a deadlock.

Information is often published on the web in ways that make processing challenging: **tables in PDF files, web pages**, etcetera. If budget data or bus schedules happen to be scattered across hundreds of documents and their **publisher isn't very cooperative** in providing a better format, a regular user is stuck. He or she can copy-paste data manually (a time-consuming and more

importantly very error-prone process). This is far from ideal. A user with more advanced IT skills, however, has a much better variety of tools at hand to extract data, clean it up and post-process it.

We do not intend to provide a description of the whole process or go into much details. Instead, we will point out to a few concepts that will help readers understand what can be done. "IT beginners" will learn that their colleagues who are more experienced can **help them extract data even if the organization doesn't want to provide it**. Policy makers will see that failing to provide machine-readable formats doesn't mean that the data won't be extracted. Therefore playing games with delaying data publishing doesn't do much good. Anyone motivated enough can extract and transform data, today. And if a motivated individual can do this rather easily, how can an organization justify their "we don't have the resources to publish Open Data" rhetoric?

The process of data extraction is called **web scraping (or simply "scraping")**. It's typically done using a software tool which downloads web pages (or other documents, such as PDF files) one by one, extracts regions with interesting data (such as selected embedded tables, headings, or anything else), optionally extracts links to other documents (such as "next page") and saves results externally. Of course, **this process isn't always reliable**: documents aren't always consistently formatted, various forms of obfuscation can be used, but many of the problems can be worked around. Highly advanced scraping tools are available which can process even very complex web pages employing javascript, extract text from highly complex and messy PDF files, and so on. Some even use **learning algorithms to improve accuracy**, both unsupervised (where the tool learns by itself), supervised (where the user trains the tool) and semi-supervised. Others even employ computer vision and artificial intelligence. This is not always needed but it's good to know that when the need is there, options exist.

There are a number of scraping tools, from the simple **user-friendly ones using a GUI** (working in the web browser or as stand-alone apps), all the way to the "command line style" **scraping libraries**. IT professionals who process a lot of data will typically prefer using a scraping framework in their preferred programming language which can give them a great degree of control (such as Scrapy for the Python programming language). We also recommend the ScraperWiki platform and http://schoolofdata.org/ where individuals can learn scraping as well as other useful data processing techniques.

When scraping manually, it is nice to play fair. **Not all servers are built to withstand heavy downloading** with many requests per second, especially when a database back-end needs to be queried for each download. It is better to **throttle down the requests** and, if possible, do most of the downloading during off-peak hours where regular users won't be impacted (such as at night or during weekends). Some scrapers don't mind overloading servers when extracting large datasets (tens of thousands of documents), which sometimes results in the blacklisting of their IP address. While these "IP bans" may be easy to overcome (it's both easy and cheap to

simply "spin up" another virtual private server from Amazon AWS or from Digital Ocean), it doesn't mean that the scrapers should be reckless.

The servers often contain a file called **"robots.txt"**. This file can contain "disallow" statements listing what the server owner doesn't want to be visited by "robots" (i.e., automated tools, such as the web scrapers or "spiders"). These may or may not be enforceable in the court of law (this may depend on the context), but it's a **good practice to follow these instructions** whenever possible. If data hidden behind robots.txt needs to be obtained or if it's protected by some technological measures, it may be wise to consult the server owner first.

A **"full scrape" (download of all data)** may be needed for the initial download of data. If an update of the data is needed, it's good to analyze the data structure first to see if an incremental scrape will be enough. The data often contain identifiers containing temporal information (dates, version), so only a **partial download** is necessary. Whenever possible, partial (non-full) downloads should be preferred as they are best for both parties: a partial download puts less stress on the server (fewer data is transferred) and is also much faster (advantage for the person downloading data). So both sides are happy. If data accuracy is of great importance and data update policies are not well known (i.e., any data may change), it may be also wise to run a full scrape when needed to make sure that data stays fresh (or verify whether older data has been altered).

Once the data is downloaded, it can be **cleaned up and post-processed**, either using a tool such as Google's Open Refine (openrefine.org) or through a more manual process (such as the many Python libraries -- **Python is a language especially popular for data extraction and processing** and has a huge ecosystem of ready-made libraries which can be found at and installed from the Python Package Index and elsewhere).

There are a number of issues that could be discussed here: legal issues (what are you allowed to scrape? and how can you use it once you scrape it?), re-use of the data (linking of datasets is an especially tricky issue and can cause previously unrealized problems, such as emergence of personally identifiable information where there was none before), data quality (what should a person do when an error is discovered in data?), etc.. Even though many issues are out of scope, there are plenty of resources on the web to learn more. And as always, **COMSODE Deliverable 5.1 is a good overview** of potentially interesting topics, even though some will not apply to the “power user” and some may be missing (e.g., some issues related to application development).

For web scraping and data processing questions, there is a **huge number of tutorials** available on the web. **StackOverflow.com** is a good place to learn and find help. There are also courses, screencasts, and video tutorials available on **YouTube**. We aren't recommending specific tools because of the sheer volume of them, but we do suggest that people with IT skills proceed as follows: once you know about your favorite programming language (Python, Ruby, Java, etc.),

you can search for language + keyword + resource type (such as python scraping video, ruby pdf extraction gem, etc.). Once you find tools you'd like to be working with, search for documentation, tutorial videos and StackOverflow support. This should be more than enough to get you started.

In the previous chapters, we tried to show what the various interest groups may wish to achieve and what their needs often are. Their **objectives may conflict**, e.g. the politician may be worried about the costs (no money for new servers, limited time of IT personnel) and the user may want the data ASAP. It's probably best if these individuals started communicating and maybe even helping each other (does the extraction script work well? how about sending it to the organization, so they know about it and can point others to it? -- it's a band-aid solution and arguably an extremely poor one, but isn't it better than not being able to use any data at all, at least for the time being?). If we raised more questions than we provided answers, good. This could mean that deeper thinking may be going on. If it's followed by communication among those involved, even better.

The following pages point to other reading on the web that may be of interest, as well as other **activities, resources and references**.

5) Contribution to standards and best practices by COMSODE

In the previous chapters we talked about what you can do to publish and use open data successfully. Now we'd like to discuss some examples of what COMSODE members have done and recommend as best practice -- whether it is active involvement in the policy making process, cooperation with other organizations or standard bodies.

The goal of this section is to give an overview of some contributions to the best practices by COMSODE members. It doesn't aim to be exhaustive but rather tries to **point out some interesting activities** that have been taking place.

In **Slovakia, The Ministry of Interior** (COMSODE member) has participated in several policy-related activities. These are important because they try to introduce certain best practices (like the use of permissive licenses or correct use of metadata) and influencing the legal framework in such a way that these best practices can be widely used.

- **Working group on Copyright** -- proposals have been made to make Creative Commons Zero license compatible with the Copyright Act.
- Consultation on **PSI Directive transposition** -- Ministry of Interior has worked with the Ministry of Justice on issues like licensing, central cataloging of data requested in PSI regime, etc..

- **Open Government Partnership (OGP)** -- during COMSODE project, Slovakia prepared the new OGP Action Plan, which was approved by the government. It contains a strong section on open data, which doesn't just try to publish as much data as possible. It aims to publish the most widely requested data first. How is it done? One of the commitments Slovakia made was to run a public consultation, identify the highest-value datasets and work with government members to publish those datasets with priority. During OGP meeting in Georgia, this was recognized as a best practice to follow. Governments should be in close contact with stakeholders who use the data -- a two-side communication channel should exist and Slovakia tried to do just that.
- **Project eDemocracy and Open Government** -- Ministry of Interior has been involved in the project which has re-used both COMSODE methodologies as well as its software components to improve open data publication process in Slovakia. Why is this mentioned as best practice? Because due to the open source nature of COMSODE outputs, re-use could happen as the result, both for software as well as methodologies. These have been customized to fit the local context of the Slovak Republic.

In **Slovakia, EEA s.r.o.** (COMSODE member) has done important work on open data standardization. In the Slovak Republic, IT standards are under the auspices of the Ministry of Finance. This ministry has had a working group on standards where EEA (Gabriel Lachmann and Peter Hanečák, both members of COMSODE project) participated. This working group has been preparing the inputs for the Decree of Ministry of Finance No. 55/2014 on Standards[6], which applies to all information systems in both government and municipal organizations. Thanks to activity of COMSODE members (and of course also other participants) Decree now has a section on open data standards. The entire open data section has come into force on 15 March 2015. This decree defines open data quality standards (based on the five-star model), as well as open data itself (at least 3-star data, with explicit licensing, with free anonymous automatic access, non-discriminating usage rights, usable for any purpose, including commercial use and remixing). It also describes in detail file formats as well as cataloging requirements (minimum metadata that must be entered to data.gov.sk). As of 2015, work begun on standardizing referenceable identifiers (URL, thus going beyond "state of the art", i.e. URI) and code lists for Linked Data.

In the **Czech Republic, Charles University** (COMSODE member) has been working with the Czech government to adopt COMSODE methodologies. The Ministry of Interior of the Czech Republic links to the Methodology for publishing datasets as Open Data (COMSODE deliverable D5.1) from its web site related to Open Data[7]. This methodology has been used by the University of Economics, Prague in a project aimed at opening up data of The Supreme Audit Office of the Czech Republic. Currently, the Ministry of Interior is building Open Data

[6] Available in Slovak at http://www.zakonypreludi.sk/zz/2014-55 -- see especially § 51-53.
[7] See http://www.mvcr.cz/clanek/otevrena-data.aspx?q=Y2hudW09Mw%3d%3d (in Czech, for English please use Google Translator or similar)

publication standards for public bodies in the Czech Republic. The COMSODE methodologies served as the input for the definition of the standards.[8]

ADDSEN (Slovakia) has performed several activities to support the exploitation of open data in multiple domains and the acceptance of COMSODE results across Europe:

- Promotion of COMSODE methodological approach in EU-wide events (European Data Forum 2014, Share PSI workshop in December 2014)
- Collection of requirements and pilot opportunities (SEMIC - Semantic Interoperability conference, CONT_ACT RIGA 2015 - The impact of ICT on the design and delivery of public services)
- Testing of exploitation paths with various stakeholders (NetFutures 2015 conference, Open Innovation 2.0 conference, Collective Awareness Platforms 2015 conference)

Other activities by COMSODE members and COMSODE User Board members have also taken place, this is by no means an exhaustive list. COMSODE consortium members have also established active contact and cooperation with several FP7 Open Data projects found at https://twitter.com/dapaasproject/lists/ec-data-projects/members.

6) Introducing new standard: “Open Data Ready”

In this chapter we propose a new IT standard that could be adopted by organisations but it can also become part of national legislation. This is the draft that has to be improved, anyhow it gives enough information to formulate such regulation or strategy.

While cooperating with partners in real-world pilot publication projects we've identified (among other things) one important aspect: Decision to start publication of OD implies additional requirements on IT infrastructure, workflows and organizational structure. If an organisation wants to do it properly, such changes are necessary. Such changes can be expensive. COMSODE Methodology (D5.1, D5.4) can help these organizations to better focus their efforts and avoid (potentially expensive) mistakes.
It is a well known fact that it is much more costly to change systems that are already finished and in daily use, compared to changes being introduced to systems that are still in design or analytical phase.

So, to further aid the easier start of publication of Open Data, we propose that organizations (or even whole countries) that publish Open Data should adopt “Open Data Ready” standard.

[8] See COMSODE blogpost http://www.comsode.eu/index.php/2015/03/open-data-publication-standards-in-czech-republic/ for more details.

Fig.: Proposed Open Data Ready logo.
Once the standard is completed, this logo can be used as certification logo,
that may be used by organisations or software providers that comply with the standard.

6.1) Understanding “Open Data Ready”

“Open Data Ready” is a set of principles to follow when doing changes to existing information systems, processes, workflows and organizational structures aimed to prepare them for publication of Open Data later on.

The point is to take the advantage of “projects” that are designed to make changes (old hardware is nearing end of life and needs to be replaced, old software needs to be updated to add additional functionality due to changes in agenda, organizations is being restructured to optimize costs, etc.) and amend their scope with set of “Open Data Ready” requirements which will for little additional costs help execute steps, which will later greatly ease the start of publication of Open Data.

Such approach is deemed cost effective because it is much more efficient to introduce any changes (including those that are Open Data related) in analytical and design phases of almost any project compared to introduction of changes in later phases (implementation, testing, production).

And that's the main point of “Open Data Ready”: if an organization is doing any changes (for any reason), it can take the advantage of that painful process and use it as an opportunity to efficiently prepare also for the publication of Open Data by introducing a few additional requirements into analytical and design phases.

6.2) The levels of "Open Data Ready"

"Open Data Ready" is a set of principles/ requirements, which can be divided into three main categories differentiated on level where the requirements are applicable:

1. Dataset level
2. Information Systems level
3. Whole organization level (its processes and organizational structure)

Organization can decide at what level they want to apply "Open Data Ready" principles: If any changes are occurring at organization level, "Open Data Ready" requirements of that level can be chosen to be achieved. If only one particular information system (IS) is being upgraded, only subset of "Open Data Ready" principles applicable to IS can be chosen to be introduced into "requirement list" for this new IS. Etc.

Those principles are derived from generic COMSODE Methodology (D5.1). In very general terms, organization is "Open Data Ready" if it can easily execute all the steps as outlined in the methodology.

If organization is committed to follow the strategy, it is on a very good path to become "Open Data Ready" much sooner. And once "Ready", the actual start of Open Data publication is much easier.

"Open Data Ready" principles can also be adopted at national level, mandating or giving recommendations for all government organizations to follow when executing organizational changes or procuring/implementing new information systems.

To aid such activities even further "Open Data Ready" label can be used for:

1. Organizations, that successfully prepare themselves for Open Data
2. Information systems, that can readily provide not just intended functionality for primary users, but also provide Open Data to general public

6.3) Open Data Ready principles/requirements

Principle/ requirement	**Notes for Organisational level**	**Notes for IS level**	**Generic notes**
Ensure creation and	make sure relevant	make sure IS	this is mainly meant to ensure

maintenance of data sources	roles exists and are staffed staffed: OD Coordinator, Curator, etc.	documentation provides also this information	some iist is kept, details about its content are then the next requirement
Ensure creation and maintenance of necessary metadata about each data source, dataset	-"-		name, description, owner, to which IS it belongs + "address" where it is, main reasons/agendas it serves, mapping to processes (how it is created, quality, known limitations, etc.)
Ensure execution of analysis regarding publication of data as Open Data	-"-: Legal Expert		mainly: make sure that organization is owner of the data or at least have sufficient rights to act as if owner (mainly right to further assign usage rights to others for any purpose make sure data can be later licensed under public open license... or if it can't be achieved, document why not (also in relation to PSI or other applicable Open Data related legislation)
Ensure that data is actually in digital form, with sufficient structure and quality and well defined processes (creation, maintenance/curation, etc.)	-"-: Curator, Data Quality Manager and Expert, etc.		avoids later obstacles for Open Data publishing like "we have it only in paper form", "it contains lots of mistakes", "nobody understands the data", etc.
Ensure IS is able to Export-publish-provide data (or its	-"-: Owner, IT professional, etc.	make sure this functionality is there, is easy to use and	Make sure that open formats are used for datasets. Subset: that's mainly

subset) to general public		can be integrated with existing infrastructure (data catalog, web server/portal, etc.)	for data sources which contains data which is exception to OD (personal information, classified information) – existence of such data within data source should net prevent all publication, so sensitive information should be removed (anonymized) and the rest published, if feasible and sensible *Example: it is usually not possible to publish registry of all citizens as Open Data (including names, dates of birth, etc.), but at least few "reports" can be done: totals per year or month, male/female counts, given name counts, age distribution, region distribution, etc.*
ensure there are processes and roles created to publish Open Data	-"-: OD Catalog owner, etc. ensuring that OD is published, but also that organization is able to respond to queries and bug reports from general public	make sure that: - individual data items are not simply deleted but marked as deleted or invalid - general public/data users can easily determine changes in the data - data owner is able to fix the data (based on but not limited to request being received from general public) allowing him also to document the reasons for change	

		(linking the request from general public, law change, etc.)	

6.4) Open Data Ready and Open Data Node

As mentioned earlier, e.g. an Information System can achieve status “Open Data Ready” by providing certain additional functionalities.

To achieve that, “Open Data Ready” requirements have to be incorporated into IS at analysis and design phase. During that phase,supplier of this new IS can choose to implement this requirements on their own or re-use Open Data Node implementation for that.

Given that Open Data Node provides many functions required by “Open Data Ready” by default (e.g., data catalog functionality, feedback and social functions, ETL functionality, etc.), that it is Open Source and is modular, supplier/implementer can achieve even significant cost-saving by choosing to reuse Open Data Node or its parts (UnifiedViews, CKAN, Virtuoso, etc. as bigger modules or only certain CKAN plugins, DPUs or libraries as smaller building blocks).

Supplier may also choose whether they can re-use Open Data Node fully on their own (thanks to Open Source licensing) or may obtain consulting and integration services from ODN implementation partners (these are currently consortium partners plus some local companies).

This is an early version of “Open Data Ready” standard proposal, based on our experiences with national legislation, public institutions and organisation. It fills the gap between software, process and data quality standards. It can become a basis for such national or lower level legislation or regulations. Such standard can have direct and long-term impact of efficiency of adoption of open data principles on various levels.

We believe that such standard (if incorporated in legislation) may improve quality and lifespan of information systems in the EU public sector as requested repeatedly.

7) Case study: Open Government Partnership as platform for advancing Open Data

COMSODE has learned that there is one particularly effective avenue for supporting open data-related activities: Open Government Partnership. One of COMSODE consortium members,

Ministry of Interior of the Slovak Republic (or more particularly, Plenipotentiary of the Government for the Development of the Civil Society) is responsible for implementing Open Government Partnership (OGP) in Slovakia.

But what is OGP? "The Open Government Partnership is a multilateral initiative that aims to secure concrete commitments from governments to promote transparency, empower citizens, fight corruption, and harness new technologies to strengthen governance."[9] Since 2011, it has grown from 8 countries to 65 (as of July 2015)[10].

Each member country defines its own "openness" initiative and its own goals in the form of two-year action plans. The action plans are formulated in **cooperation between the government and the civil society** (that is the "partnership" word in "Open Government Partnership"). These two-year action-plans contain specific commitments which are then carried out by the respective countries. A large number of the commitments are related to open data[11].

During the COMSODE project, Slovakia's government approved the country's new OGP action plan[12]. Its very first chapter is dedicated to open data. It contains concrete commitments to publish data and it also introduces the **"pull" model**: whereas in the previous action plan, public organizations decided what data to publish (they "pushed" data out), in this new action plan there is an additional mechanism: a public consultation.

In cooperation with the National Agency for Network and Electronic Services (NASES), a public consultation was carried out: participants in the consultation were able to suggest which data should be published, which data could be improved, etc.. The results were processed, highest-priority datasets were identified and ministries and other responsible organizations in the Slovak Republic were notified of the results. Since the OGP action plan was approved by the government, the organizations which were in charge of the data now had a **government-given mandate to publish** this data in open formats.

During the June 2015 OGP meeting in Georgia where national contact points of countries involved in OGP gathered[13], it was stressed that governments and other public organizations should give priority to publishing data which is seen as highest-priority by its users. **Public consultation was therefore recommended as best practice** and the recent experience of Ministry of Interior in Slovakia (COMSODE member) was showcased.

[9] http://www.opengovpartnership.org/about
[10] http://www.opengovpartnership.org/countries
[11] http://www.opengovpartnership.org/explorer/landing
[12] http://www.opengovpartnership.org/country/slovakia/action-plan
[13] http://www.opengovpartnership.org/blog/alonso-cerdan/2015/06/02/fostering-peer-exchange-and-learning-european-region

Also, **best practices for cooperation with the civil society organizations** were presented in the Georgia meeting. These include: making the consultation timeline available, recommendation of advance notice (vs. ad-hoc consultation), awareness-raising activities (multiple stakeholders should be informed), online as well as in-person consultations (not everyone can participate personally, there should be at least several weeks of opportunity for people to participate online), ongoing regular dialogue in addition to one-off events.

Cooperation between the civil society and the government, understanding between policy makers and technology experts and the drive to make the society more open can help make the open data efforts successful. Hopefully this COMSODE deliverable will help in creating this understanding.