DELIVERABLE D5.3

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	29th September 2014
Author(s)	Ján Gondol, Ľubor Illek
Responsible of the deliverable	Ľubor Illek
Email	Lubor.Illek@soit.sk
Reviewed by	Miroslav Konečný
Status of the Document	Final
Version	1.0
Dissemination level	PU - Public

Table of Contents

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 of this domain have often 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 supposed to be 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 expected roles. Hopefully this text will provide you with **a brief introduction, quick overview and a useful pointer** to sources where you can learn more.

1.1 COMSODE project introduction

This material can be 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 that 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, but **Work Package 5 ("Development of methodologies") is especially interesting** to mention right 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 is 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.

Finally, there is the document you're reading right now. The intention of **Deliverable 5.3** is to introduce the broader context of international Open Data standards and best practices. We wrote it this 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 haven't 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 didn't fall out of the blue sky; it is part of 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 isn't enough: according to the **Open Definition** (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 way **more than just financial information**. The government and various publicly funded organizations produce vast amounts of other data, which could be used in novel ways and bring 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 fulfil 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 much bigger than just 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 catalogues specifically for the non-profit 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 them to also publish 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 data that is 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, re-use and

redistribute" (with no other strings attached, as we saw above), it is Open Data. 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 the Slovak Republic, **Open Education** is also one of the main 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 (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. The list goes on.

2 For the Decision-maker

This 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. Building on the work of Capgemini Group analysis, they recognize the three main areas of benefits:

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 their 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 (e.g. crime reporting, potholes reporting, fix-my-street).

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 that 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, 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 is 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.

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.

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) that 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 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 formats must be available. 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 famous of which are 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 beyond 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 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).

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 web site 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 only become more tangled 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?"

2.4 Action Plan

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 these questions should be **better answered by people outside of Decision Maker reach**. Forming a working group, or less formally, simply involving others in a **teamwork**, 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 that is **most likely to be used by anyone very soon**.
- Take it **one step at a time**: start with a few datasets and keep on adding more.
- 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, etc.) -- 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 that we consider very helpful. **For a much more thorough and systematic approach, see Deliverable 5.1** of project COMSODE that 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 **organizations’ own Action Plan**.

2.5 Common Concerns

A common worry is: **how much money** will need to be spent for Open Data management? And the good news is: typically the costs can be kept 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.

Services like **Github** (found at 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 about publishing datasets that 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 IT professionals – 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 the 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

Whereas 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 methodology of it's use in Deliverable D5.2 and User Manuals.

3.1 Data accessibility

While there are many interesting technical topics – eg. producing data linked to other data, automating datasets 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 more sophistically by BitTorrent protocol. But if there is specific reason, transfer can be permitted only after request, possibly 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.

To put it summarily, most data file formats can be divided into categories:

- **Formats not enabling automated data processing** or making this processing at great cost or inaccurate – in this category are all 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 DOS, 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 used is RDF – these are standard for Linked Open Data and other high quality needs

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

- **Storage** for published data, both in terms of capacity and software handling – usually dedicated database instance. As we point out in the chapter dealing with security, 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 only processor of some common query language – for access to the data stored in RDF form there is special query language defined – SPARQL, 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 user specifies query and retrieve the results**. At this brief description we only want to mention most common ones: Representational state transfer (and its implementation as RESTful API), SPARQL endpoints or custom designed Web Services.

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

There are available various software packages integrating some or all mentioned 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 developed in COMSODE project.

Data provider can publish data by using their own infrastructure, or put them in the cloud (for example use CDN or application services such as DaaS). It is a common situation to provide some services to the data producers at the central level in the country, particularly as part of national data catalogue or OpenData portal.

For the data user one of the critically important tasks is **ability to effective update data** (of course there are exceptions, particularly datasets containing historical data). In this field, the user must address following questions:

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

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

- **Invariability** - Keep point of access (for example URL to data file) constant, and also 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** – User must understand what can he do with the data (see next chapter about 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

The security considerations should be given to two aspects:

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

Basis for security evaluation is finding that the data publication and subsequent interaction with the users who work with these data almost always lies in **different security context than production data processing** in the organization.

Therefore best practice is to put infrastructure necessary for data publication place in a s**eparate computing environment**. This measure is typically implemented at the network layer. Data publication infrastructure is mostly placed into the network segment/environment for organization's servers directed toward Internet (as these did well 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 all the time public. If you want this assumption to be effectively used to simplify data publishing, beware of the nonpublic information not be included in any form in the infrastructure for data publishing. For example if you use the ODN software, and have it placed in the 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.

More interesting is the question of data **integrity and authenticity**. Solution depends mostly on how much liability publisher wants to impose to the data. 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) – he 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 for web servers or sites of the organisation. The authenticity of the data is achieved at the level of metadata, which 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 most often**, which is applied to the entire dataset (but, however, then the partial access to data through API is unable), 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 trust anchor.

It would seem that if there is no need to address confidentiality, which means lack of necessity for **access control** also. However, access control is often important for the efficient management of resources used by data users - especially network capacity and utilization of servers making data available. Overall, **it is necessary to foresee and implement solution for the data availability**, since if data are to be used seriously, there must be 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 (e.g. 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 its services. We recommend monitoring the current status and resources usage, as well as storing 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 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 is allowed to do and have some protection to be not denied of this 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, minimally these conditions 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 endeavour.

- **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 come 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).

- **Legaly valid** – The license should be legaly 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 license adhering to all principles stated above. Furthermore, in most practical situations data users want to combine data form several datasets; possibly form different publishers of even countries. While doing this he must also **combine related licenses** and find out their intersection – and if you find defining the license difficult, combining several of them brings whole new level of challenge.

For this reasons it is strongly advised not to create new license but first try to reuse some existing and well known licenses in this field. You can find several of the most common licenses used for Open Data publication listed in chapter with resources of this document.

More detailed description of licensing process is covered by special section in COMSODE Deliverable D5.1, or further informations you can find in:

- http://opendatacommons.org/faq/licenses/
- http://opendefinition.org/guide/data/
- http://opendefinition.org/licenses/

4 Open Data users

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.

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 re-distribution) 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 re-used, re-analyzed, re-visualized and re-interpreted -- 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 -- 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, Libre Office 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 curating: 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).

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 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: if you encounter a problem, find a human who can walk you 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.

Many users will probably **experience Open Data through the applications** that use these data, such as public transportation schedules, mapping applications, and so on. 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, it may be just a bonus. 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 then 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 last detail. 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"?

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).

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 (depending heavily 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 are 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 favourite 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 Activities and resources

5.1 Open Data projects

In next table we summarise list of main projects trying to find solution for particular problems of Open Data publishing and reuse with which project COMSODE members established active contact and cooperation:

Project name	About the project	Online resources
DAPAAS Data Publishing through the Cloud: A Data- and Platform-as-a-Service Approach for Efficient Data Publication and Consumption	Developing a software infrastructure combining Data-as-a-Service (DaaS) and Platform-as-a-Service (PaaS) for open data, with the aim of optimizing publication of Open Data and development of data applications. Addressing the data consumption aspect by developing novel cross-platform interfaces to data applications, DaPaaS extensively covers the life cycle of cost-efficient data publishing and consumption. Backed by the development of a methodology for data use in the DaPaaS infrastructure, the project will deliver an intuitive platform that simplifies data publication, as well as cross-platform data consumption, thus enabling a sustainable infrastructure for efficient and simplified reuse of open data. Core innovations include: an open DaaS and PaaS, unified Linked Data access, integrated DaaS and PaaS for open data, lowering the complexity of open data publishing and consumption for non-experts.	http://project.dapaas.eu/ https://twitter.com/dapaasproject
ENGAGE project - 52,011 DS catalogue	An Infrastructure for Open, Linked Governmental Data Provision towards Research Communities and Citizens 52,011 Datasets in their catalogue	http://www.engagedata.eu/ https://twitter.com/engage_eu https://www.linkedin.com/groups/ENGAGE-eInfrastructures-

		Project-3867416
OPEN DATA MONITOR - Monitoring, Analysis and Visualisation of Open Data Catalogues, Hubs and Repositories	OpenDataMonitor provides the possibility to gain an overview of available open data resources and undertake analysis and visualisation of existing data catalogues using innovative technologies. By creating a highly extensible and customizable harvesting framework, metadata from diverse open data sources will be collected. Through harmonization of the harvested metadata, the gathered information can be structured and processed. Scalable analytical and visualisation methods will allow the end users to learn more about the composition of regional, national or pan-European open data repositories. during the OpenDataMonitor project, established open source software like CKAN will be adopted and extended. The research outcomes and technical developments will be combined in a demonstration platform, integrated in third-party sites	http://project.opendatamonitor.eu/ https://twitter.com/opendatamonitor
OPENCUBE - Publishing and Enriching Linked Open Statistical Data for the Development of Data Analytics and Enhanced Visualization Services	The ultimate goal of OpenCube project is to facilitate (a) publishing of high-quality linked statistical data, and (b) reusing distributed linked statistical datasets to perform advanced data analytics and visualizations. Towards this end, OpenCube project will develop open source software tools in terms of (a) standalone applications and (b) extensions of two linked data management platforms, namely Swirrl's PublishMyData and fluidOps' Information Workbench, for publishing and reusing high-quality linked statistical data.	http://opencube-project.eu/

	Results: • The OpenCube toolkit comprising standalone tools for publishing, visualizing and performing data analytics on top of linked statistical data. • The OpenCube extension of the PublishMyData platform. • The OpenCube extension of the Information Workbench platform. • The OpenCube Pilots in three public authorities and businesses to validate and prove the usability and effectiveness of the developed tools and guidelines.	
EUCASES- EUropean and National CASE Law and Legislation Linked in Open Data Stack	EUCases will develop a unique pan-European law and case law Linking Platform transforming multilingual legal open data into linked open data after semantic and structural analysis. It will reuse the millions of legal documents from EU and national legislative and case law portals. The developed components cover the entire publication stack: we collect the data from institutional portals, enrich them using partner's or open source language technologies and ontologies, and publish the documents as linked open data in XML, with metadata and legal ontologies in RDF, to facilitate access, navigation, multilingual search and reuse. EULinksChecker will interactively assist legal professionals while editing or browsing documents by identifying and establishing connections with regulations and legal ontologies.	http://www.eucases.eu/start/
LinDA - Linked Data and Analytic tools for SMEs	Linked Data Analytics- the renovation and conversion of existing data formats into structures that support the semantic enrichment and interlinking of data. To create cross-platform, extensible software	http://www.linda-project.eu/

	framework that provides a rule-based system for renovating and converting a wide range of supported data containers, structures and formats into arbitrary RDF graphs. The framework can be used to develop custom solutions for SMEs and public sector organisations or be integrated into existing open data applications	
MELODIES - Maximizing the Exploitation of Linked Open Data In Enterprise and Science	The MELODIES project will apply the latest technologies in cloud computing and data-handling to exploit these data to their best advantage. The outputs of the MELODIES project will be eight new environmental services built using Open Data. These are products and/or applications that span a range of domains including agriculture, urban ecosystems, land use management, marine information, desertification, crisis management and hydrology. A different organisation leads the development of each of these applications, building on previous experiences in their sector.	http://www.melodiesproject.eu/ https://twitter.com/MelodiesProject
Smart Open Data - EU national parks and protected areas	Decsription in Czech: http://www.geobusiness.cz/2013/11/otevrena-datova-infrastruktura-projekt-smartopendata/ SmartOpenData will create a Linked Open Data infrastructure (including software tools and data) fed by public and freely available data resources, existing sources for biodiversity and environment protection and research in rural and European protected areas and its National Parks	http://www.smartopendata.eu/ https://twitter.com/SmartOpenData
	Big Data Public Private Forum (BIG) is	http://www.big-project.eu/

BIG project	working towards the definition and implementation of a clear strategy that tackles the necessary efforts in terms of research and innovation, while also it provides a major boost for technology adoption and supporting actions for the successful implementation of the Big Data economy.	https://twitter.com/BIG_FP7
PUBLICAMUNDI - Scalable and Reusable Open Geospatial Data	Geospatial data account for an estimated 80% of public sector information and are the most significant category of open data due to their high production, procurement and update costs, as well as their relevance in multiple domains. PublicaMundi will deliver the required methodologies, technologies and software components to leverage geospatial data as first-class citizens in open data catalogues, and deliver reusable software components and tools enabling the development of scalable, responsive, and multimodal value added applications from open geospatial data.	http://publicamundi.imis.athena-innovation.gr/ https://twitter.com/PublicaMundi
GEO KNOW	Geoknow is adding spacial dimensions to linked data and making the web an exploratory space for geospatial knowledge	http://geoknow.eu/ https://twitter.com/geoknow
ALIADA - Automatic publication under LInked DAta Paradigm of library DAta	ALIADA will automatize the publication in the Linked Open Data cloud of open data hosted by different Library or Collection Management Software. ALIADA will be an open source plugin for the library or collection management software, initially for the ones developed by the SMEs in the consortium and already installed in the public bodies, opening the SMEs' market to new business opportunities. Usability in ALIADA solution will be a key aspect, as the final users will have little or no experience in	http://www.aliada-project.eu/ https://twitter.com/aliadaproject

	Linked Data technologies and processes. Librarians are used to manage the collections, they are responsible of, using Library or Collection Management Software that follows norms such as MARC21 and LIDO resulting in high quality annotations. ALIADA will make possible libraries and museums interoperability, so they can share their collections and offer them to the general public, by means of the linked open data cloud, allowing new interaction experiences for the general public.	
LEO - Linked Open Earth Observation Data for Precision Farming	A lot of remotely sensed data coming from satellites has become available at no charge in Europe and the US recently, and there is a strong push for more open Earth Observation data. Open Earth Observation data that are currently made available by space agencies (e.g., ESA and NASA) are not following the linked data paradigm. ICT STREP project TELEIOS recently introduced the linked data paradigm to the Earth Observation domain and developed prototype applications (wildfire monitoring and burnt scar mapping, semantic catalogues and rapid mapping) that are based on transforming Earth Observation products into RDF, and combining them with open, linked geospatial data. However, TELEIOS did not consider the whole life cycle of linked open Earth Observation data, and concentrated mainly on developing scalable storage and query processing techniques for such data. In LEO, the core academic partners of TELEIOS (UoA and CWI) join forces with 2 SMEs and one industrial partner with relevant experience (SA, VISTA and PCA) to develop software tools that support the whole life cycle of reuse of linked open EO data and related linked geospatial data. Finally, to demonstrate the benefits of linked open EO data and its combination with linked geospatial to the European economy, a	http://www.linkedeodata.eu/

	precision farming application is developed that is heavily based on such data.	
FALCON - Federated Active Linguistic data CuratiON	FALCON will integrate the resulting web of linked localisation and language data into localisation tool chains using existing data query and access control standards. a commercially sustainable localisation process that allows SME Language Service Providers (LSPs) to continuously reuse the output of web content translation projects. By integrating these outputs with language resources from public bodies, SME LSPs can easily and cheaply optimise and tailor open source machine translation and automated terminology extraction components to the needs of their customers. Today, the curation of parallel text involves the publication of the undifferentiated results of translation projects in differing formats or via centralised repositories. This restricts SMEs in leveraging these resources to train machine translation engines to match incoming translation projects.	http://falcon-project.eu/
HOMER - PSI in the Mediterranean space	unlock the full potential of the Public Sector Information in the Mediterranean space - facilitate the wider deployment of PSI in Spain, Italy, France, Malta, Greece, Slovenia, Cyprus and Montenegro. Starting from exposing 5 strategic sectors, characterising the MED political agenda of the next decades (Agriculture, Tourism, Environment, Energy and Culture), consistent with the NSRF strategies of the partners and linked to the commitment of their internal departments, a dedicated Task Force composed of IT and Open data experts. In this way, during its first phase, HOMER	http://homerproject.eu/

	will be able to open hundreds of public datasets, enhancing digital heritage transparency across the Mediterranean.	
SDI4Apps	SDI4Apps-Uptake of Open Geographic Information Through Innovative Services Based on Linked Data, The main target of SDI4Apps is to bridge the 1) top-down managed world of INSPIRE, Copernicus and GEOSS and 2) the bottom-up mobile world of voluntary initiatives and thousands of micro SMEs and individuals developing applications based on GI	
LOD2 project	The project aims to contribute high-quality interlinked versions of public Semantic Web data sets, promoting their use in new cross-domain applications by developers across the globe. The new technologies for enabling scalable management of Linked Data collections in the many billions of triples will raise the state of the art of Semantic Web data management.	http://lod2.eu/Welcome.html https://twitter.com/lod2project
RECODE	The Policy RECommendations for Open Access to Research Data in Europe (RECODE) project will leverage existing networks, communities and projects to address challenges within the open access and data dissemination and preservation sector and produce policy recommendations for open access to research data based on existing good practice. It will provide over-arching recommendations for a policy framework to support open access to European research data. The RECODE partners will identify relevant stakeholders, build upon and strengthen existing stakeholder engagement mechanisms. It will conduct studies of good practice and exchange good practice principles with relevant stakeholders and institutions during networking activities. The RECODE project will culminate in a series of policy recommendations for open access to	http://recodeproject.eu/events/upcoming-events/ https://twitter.com/RECODE_Project

	research data targeted at different stakeholders and policy-makers.	
PlanetDATA	Aims to establish a sustainable European community of researchers that supports organizations in exposing their data in new and useful ways. The ability to effectively and efficiently make sense out of the enormous amounts of data continuously published online, including data streams, (micro)blog posts, digital archives, eScience resources, public sector data sets, and the Linked Open Data Cloud.	http://www.planet-data.eu/ https://twitter.com/PlanetData_NoE
City Pulse Project	Real-Time IoT Stream Processing and Large-scale Data Analytics for Smart City Applications	http://www.ict-citypulse.eu/page/ https://twitter.com/ictcitypulse
PRELIDA project	Targets stakeholders from Digital Preservation and Linked Data communities	http://www.prelida.eu/ https://twitter.com/PRELIDA_project
LinkedUP Project	Pushing forward the use of linked & open web data for educational purposes. About to launch the Vidi competition.	http://linkedup-project.eu/ https://twitter.com/linkedupproject
EUCLID project	Facilitating professional training for data practitioners, who aim to use Linked Data in their daily work. EUCLID delivers a curriculum implemented as a combination of living learning materials and activities (eBook series, webinars, face□to□face training), validated by the user community through continuous feedback.	http://www.euclid-project.eu/ https://twitter.com/euclid_project
LATC	Supporting both institutions as well as individuals with tutorials and best practices	http://latc-project.eu/

	concerning Linked Data publication and consumption.	https://twitter.com/latcproject
Share-PSI 2.0	The network for innovation in European public sector information.	http://www.w3.org/2013/share-psi/partners
Open Data Support	Open Data Support is a 36 month project of DG CONNECT of the European Commission to improve the visibility and facilitate the access to datasets published on local and national open data portals in order to increase their re-use within and across borders. To achieve its objective, Open Data Support provides to (potential) publishers of open datasets, three types of services to local, regional and/or national public administrations publishing open data: ● Data and metadata preparation, transformation and publication services ● Training services in the area of (linked) open data ● IT advisory and consultancy services	https://joinup.ec.europa.eu/community/ods/description https://twitter.com/OpenDataSupport
ePSI platform	The ePSIplatform is a European Commission (DG CONNECT) initiative with the objective of promoting a dynamic Public Sector Information (PSI) and Open Data re-use market across the European Union. • news on European PSI and Open Data developments; • legal cases surrounding the re-use of PSI; • good practices and examples of new products and services created through Open Data re-use. • events, workshops and webinars around Europe.	http://www.epsiplatform.eu/ https://twitter.com/epsiplatform
Collaboration in Research and	An important objective of the CROS portal is to support collaboration within communities	http://www.cros-portal.eu/content/fp7-projects

Methodology for Official Statistics	related to research and methodology in Official Statistics.	

5.2 Important catalogues

Here is the list of most important national data catalogues and other interesting catalogues maintained in countries of COMSODE consortium members:

Link	Covered area	Maintainer
http://catalog.data.gov/dataset	USA	National catalogue
http://data.gov.uk/data/search	United Kingdom	National catalogue
http://data.gov.sk/	Slovakia	National catalogue
http://cz.ckan.net/	Czech Republic	National catalogue
http://datanest.fair-play.sk/datasets	Slovakia	Fair-play alliance - Public watchdog NGO
http://datacatalog.worldbank.org/	World	The World Bank

You can find overview of national catalogues of the EU member states in the work of Tomáš Šedivec.

5.3 Licenses

We can recommend these licenses, which are commonly used for Open Data publication:

- Creative Commons CCZero (CC0)
- Open Data Commons Public Domain Dedication and Licence (PDDL)
- Creative Commons Attribution 4.0 (CC-BY-4.0)
- Open Data Commons Attribution License (ODC-BY)
- Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA-4.0)
- Open Data Commons Open Database License (ODbL)

6 Executive summary

Deliverable 5.3 of COMSODE is a supplemental documentation for D5.2 giving publishers (mainly public bodies) broader context about international Open Data standards and best practices. The text is **addressed to several target groups** and by sharing the different perspectives, it should help them understand each other: Decision-makers (perspective of policy), IT professionals responsible for publishing data (perspective of technology) and Open Data users (perspective of day-to-day utility).

The **Decision-makers** will see Open Data in a broader context and learn about the benefits (to the government, private sector, as well as civil society, among others) and learn about the value of open data. Legal context is discussed and **key legislation is introduced**. These chapters aim to provide food for deeper thought.

The **IT Professionals** working with the Decision-makers will learn the basics about formats, security and licensing and **users / citizens** can read about how to get the data if it hasn't been published. The details are discussed in other COMSODE deliverables, so **pointers to external resources** are used (an overview of COMSODE deliverables is also included).

The **final chapters** list related Open Data activities, as well as resources and references that could be of interest. The activities of the Ministry of Interior (which has primary responsibility for D5.3) are discussed in more detail.