Data Integration 1

MARCH 2016

DELIVERABLE

Project Acronym: **SDI4Apps**
Grant Agreement number: **621129**
Project Full Title: **Uptake of Open Geographic Information Through Innovative Services Based on Linked Data**

D5.3.1 DATA INTEGRATION 1

Revision no. 03

Authors: **Otakar Čerba** (University of West Bohemia)
Dmitrii Kožuch (Help Service Remote Sensing)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

REVISION HISTORY

Revision	Date	Author	Organisation	Description
01	08/03/2016	Otaka Čerba	UWB	Initial draft
02	29/03/2016	Martin Tuchyna, Karel Charvat	SAZP/CCSS	Internal review
03	30/03/2016	Otaka Čerba	UWB	Final version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Disclaimer:

Views expressed in this document are those of the individuals, partners or the consortium and do not represent the opinion of the Community.

TABLE OF CONTENTS

LIST OF FIGURES

EXECUTIVE SUMMARY

This report describes particular data integration issues in the SDI4Apps project. The particular activities of data integration and harmonization are related with a requirement of combining data from heterogeneous resources, re-using of existing data or publishing data as open data or Linked data.

Due the fact, that particular data integration activities vary in particular pilots, the sections of this document are connected to SDI4Apps pilots – Easy data access, Open Smart Tourist Data, Open Sensors Network, Open Land Use Map Through VGI, Open INSPIRE4Youth and Ecosystem Services Evaluation.

1 INTRODUCTION

The potential of geographic information (GI) collected by various actors ranging from public administrations to voluntary initiatives of citizens is not fully exploited. The advancements of information and communication technologies and shift towards Linked Open Data (LOD) give an excellent foundation for innovation based on reuse of GI. The establishment of spatial data infrastructures has largely been driven by the "traditional" GI community and the national and European policies governing this sector. However, GI is no longer a separate information space but finds itself part of a larger European information space where the ultimate objective is the creation of value-added services based on reuse of public sector information as defined by the PSI and INSPIRE directives rather than exchange of "layers" between different GI software.

Establishing an infrastructure to meet this new and wider objective puts greater strain on local authorities and institutions that traditionally were users of GI but now find themselves in an environment where they are also expected to be data and service providers, a role that is far more demanding in terms of technical knowledge and resources.

The main target of SDI4Apps is to build a cloud based framework that will bridge the gap between:

1. the top-down managed world of INSPIRE, Copernicus and GEOSS, built by SDI experts, and
2. the bottom-up mobile world of voluntary initiatives and thousands of micro SMEs and individuals developing applications (apps) based on GI.

SDI4Apps will adapt and integrate experience from previous projects and initiatives such as HABITATS, Plan4business and EnviroGrids, to build its cloud based platform with an open API for data integration, easy access and provision for further reuse. The solution will be validated through six pilot applications focused on easy access to data, tourism, sensor networks, land use mapping, education and ecosystem services evaluation.

The aim of this deliverable is to report on ongoing external validation of the SDI4Apps solutions and pilot applications. This is conducted in cooperation with dissemination activities, organised hackathons and stakeholder management group.

2 DESCRIPTION OF DATA INTEGRATION ISSUES OF EACH PILOT

This section describes the current situation related to data integration in particular SDI4Apps pilots. The situation in different pilots is very miscellaneous. It illustrates various level of pilot development as well as distinct ways of data processing (including data integration) in pilots.

The final version of this report (Data integration 2) should have the same structure of each pilot description composed of following components: Requirements for data integration; Offer of reusable data which could be integrated for other partners; Examples of existing solution (if they are available).

2.1 Easy Data Access

The main innovative aspect of the pilot is to advance tourism for conservation as a European model of value to local communities. This aims to be a strong demonstration issue with pilot actions being stimulated to test the use of tourism for conservation in the Burren Geopark, Ireland.

The pilot deals with two main data sets, which could be integrated with other data related to tourism (see the 2.2. Open Smart Tourist Data section):

1. ETIS Dataset – the European Tourism Indicator System (Figure 1) for the Sustainable Management of Destinations (ETIS) to monitor and measure performance of destinations in Europe. This data will be transformed from Excel/CSV to LOD RDF format by MAC using the Enablers of the SDI4Apps Platform.

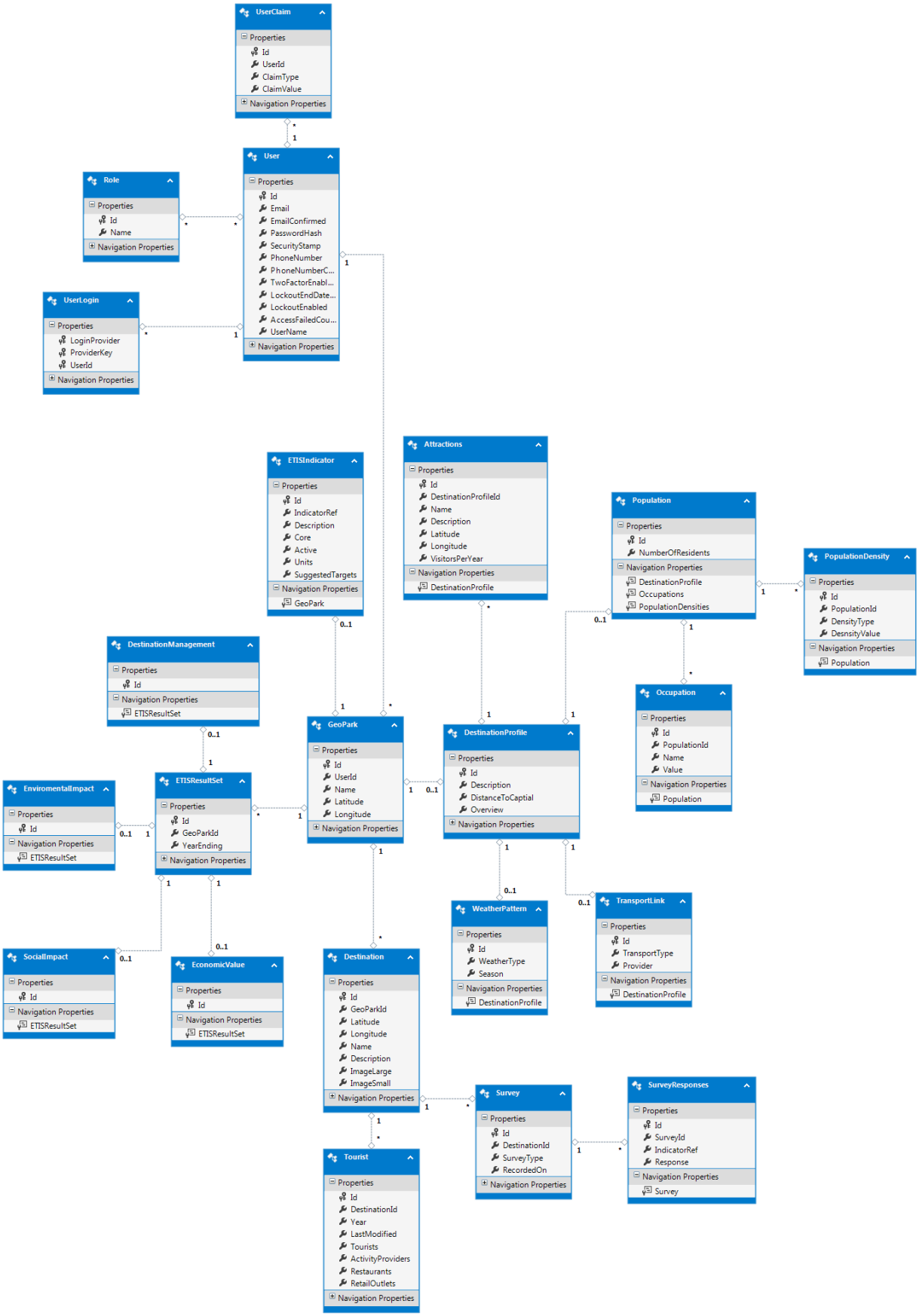

Figure 1 Tourism for Conservation Data Model

Potential Monuments Voluntary Geographic Information Dataset (Figure 2) – dataset to record Voluntary Geographic Information (VGI) reports from professionals, visitors and people interested in their local heritage, to seek out and ground truth potential Monument sites in the Burren and beyond.

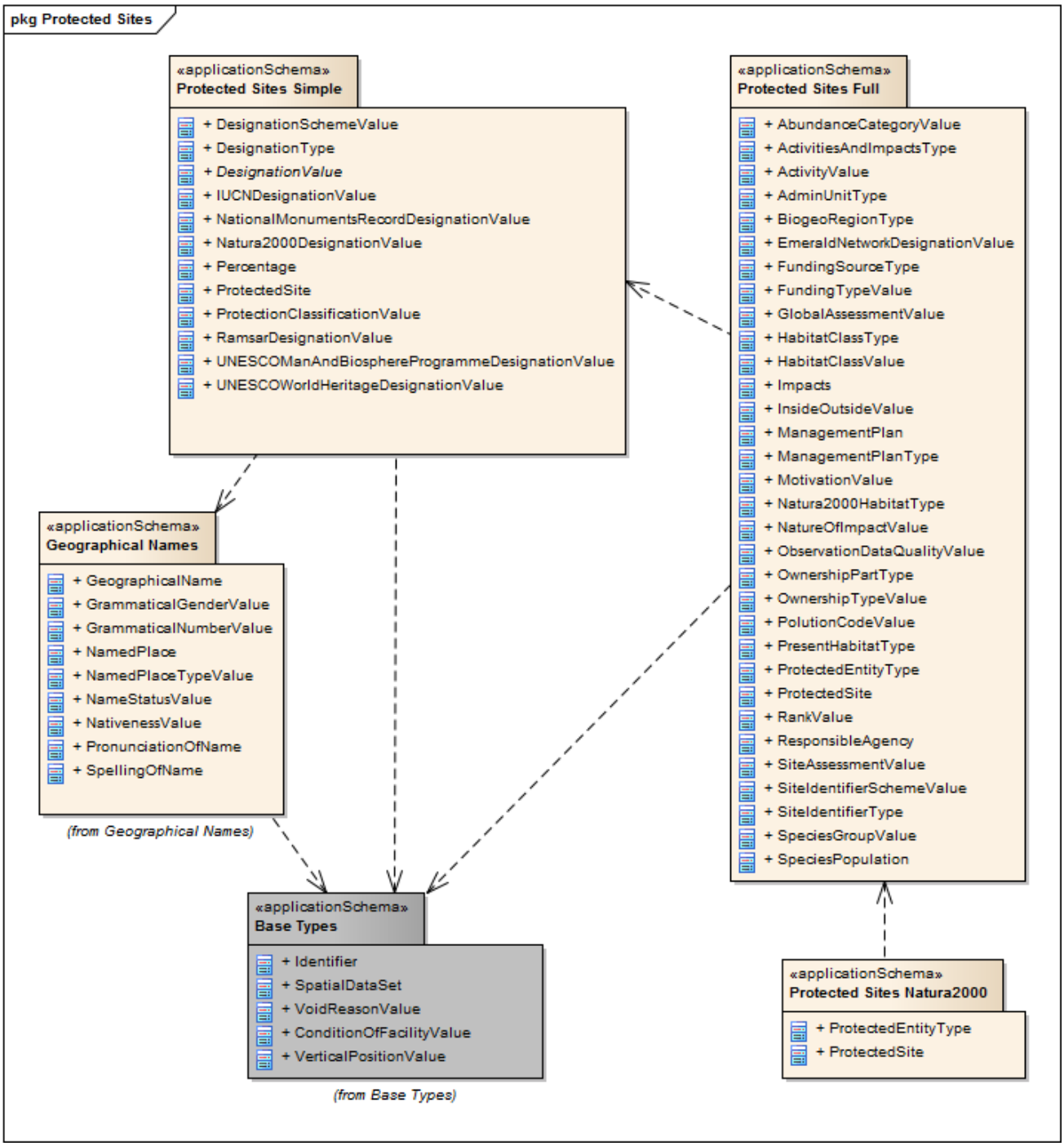

Figure 2 Protected Heritage Sites - Data Model for National Monuments

2.2 Open Smart Tourist Data

The data integration issues in this pilot are focused on Smart Point of Interest (SPOI) data development. This data set contains Points of Interest (POI) related to tourism from different data resources such as OpenStreetMap, GeoNames.org, Natural Earth, Citadel on the Move or several local data resources (Sicily, Pošumaví etc.). The data integration is composed of a few steps (Transcription to structured data,

Preparation of common vocabularies and mappings, Filtering data, Adding new information, Transformation to the common data model and Export to a common data format). It is necessary to mention that previous steps differ for each input data set. For example in case of OpenStreetMap was filtering very important process because of big size of data, but data sets from several local resources had not been filtered, because they were prepared directly for SPOI.

The following scheme (Figure 3) presents the particular harmonization issues realized for data resources.

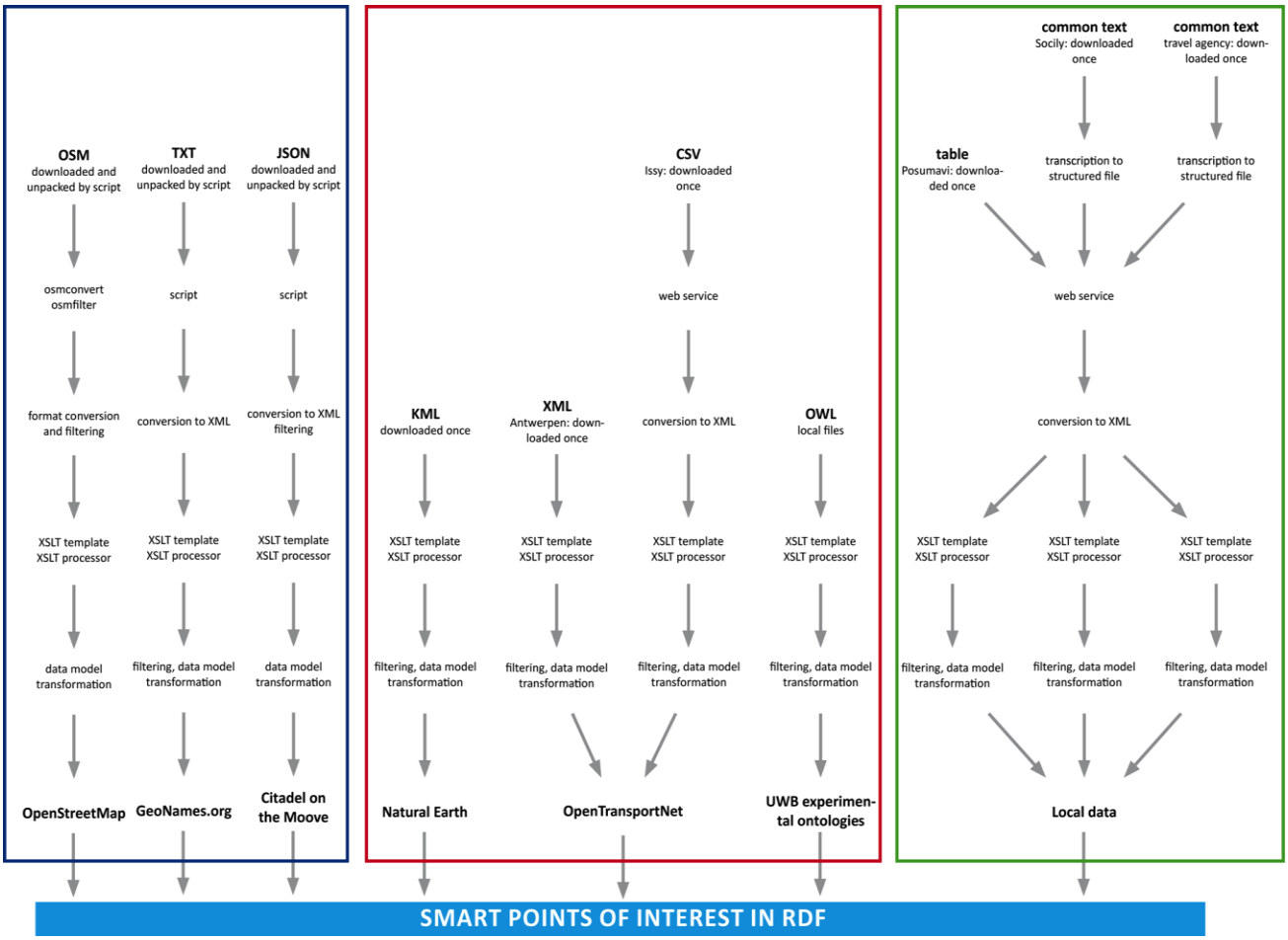

Figure 3 SPOI Data Integration

Figure 3 depicts several examples of different harmonization approaches depending on the structure, accessibility and level of harmonization of the source datasets. The particular colour frames mean different ways of data collection, processing and harmonization, depending on the source data format (e.g. TXT, JSON, KML, XML, CSV) and structure:

- Blue - global data downloaded by script (BASH – Bourne again shell) with necessity of pre-processing (filtering) data before transformation. Automation of the processes is used as much as possible.
    - OpenStreetMap (the main data source, about 86% data) – data are downloaded for particular countries with use wget software in tar.bz2 package. Therefore it is necessary to unpack (by tar and bunzip2 software) data in osm format (specific XML-based format). Because the several files are very large to process them by XSLT templates and java-based processor, the filtering (selection of fitting nodes and atrributes) is realized by osmconvert and osmfilter software components. The XSLT template is applied to transform data to the final RDF format. This seemingly complicated approach (combination of the script and XSLT transformation) was chosen, because there is used one “universal” XSLT template, which is slightly modified for particular datasets (this information concerns other input datasets, too).
    - GeoNames.org – similarly to previous point there are applied wget and unzip software to download and unzip original data. Than the symbol & is replaced by character entity and text

data transformed to temporary XML file, which is processed by XSLT template and XSLT processor (Saxon) as in previous case.

- Citadel on the Move – data (more than 30 suitable data sets) are downloaded with use wget software in JSON format. The pre-processing consists in replacing of inconvenient symbol (&) by character entity (&). Than the temporary XML file is created, which is transformed by XSLT template and processor to RDF format.

- Red - structured data downloaded and transformed just at once, or on demand. The processing of data belonging to this group is very similar to the "blue" group. The main difference consists in an absence of automated downloading. There are two reasons – (1) data are in structured format, but the update frequency is very low or it seems that a development of dataset is closed (e.g. experimental ontologies from UWB or Natural Earth); (2) Data is stored in local storage. The transformation process includes only above-mentioned implementation of XSLT templates (with exception of Issy data that has to be converted to a XML-based format at first).
- Green - non-structured data (texts or simple tables) is transcribed (manually converted to a structured data, usually csv tables), downloaded, pre-processed (several necessary attributes are added such as missing IDs), stored as temporary XML file (by a web service such as Zamzar converter) and transformed by XSLT template (similarly to previous cases) just at once, or on demand. These examples require manual work and is time consuming, and data update might be a costly issue. This approach was chosen, because the data belonging to the "green" group are usually not published online and also any frequent update is not expected.

2.3 Open Sensors Network

This pilot is focused on data collecting via sensors primarily. Therefore any data integration issues have not been realized. In the future there could be prepared data integrations related to interconnection of sensor data with some reference data or background maps (such as OpenStreetMap or environmental data). Also there are supposed data integration activities for a purpose of open and linked data publication (for example interlinking with existing vocabularies or topological interconnection to regions).

2.4 Open Land Use Map through VGI

The objective of the Open Land Use (OLU) Map is to create a land use map with as much detail as possible, and covering the whole EU territory. The primary conditions to establish this are to have:

- Spatial detail, ideally to the smallest distinguishable land use unit, such as land parcels;
- Hierarchical, layered and consistent categorization of land usage (HILUCS categorization) integrated or linked;
- Uniform data quality (ideally at the level of available data sources with highest precision);
- Up-to-date information;
- Use of unique IDs - making links to other relevant data sources possible, such as transport related information.

To create the OLU dataset, it is necessary to perform a spatial integration of the data coming from several sources, and secondly, to harmonize the land use attributes.

2.4.1 Step 1: Understanding the theory of spatial data harmonization

The spatial integration of datasets is performed in the order of their level of detail. The dataset with the most recent and most detailed land use information will replace the information of the less detailed dataset. This exercise needs to be done stepwise, as shown in Figure 4 for the Czech Republic.

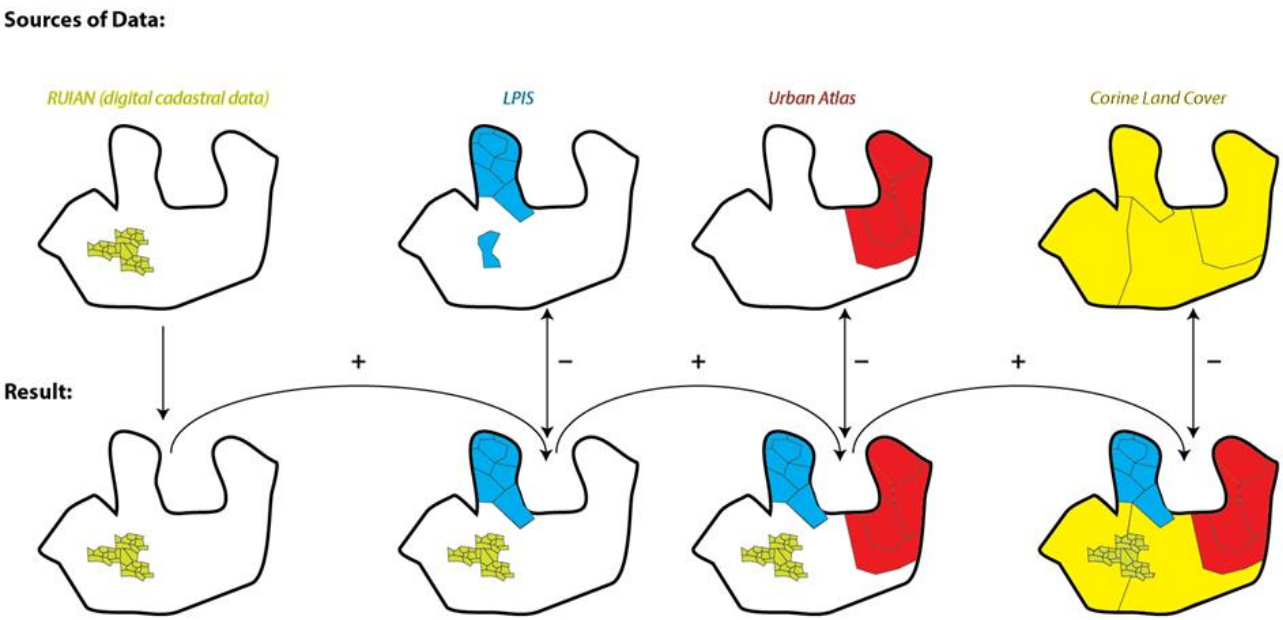

Figure 4 Creation of Open Land Use Map in the Czech Republic

When the spatial integration of datasets is finished, the next process is to harmonize the land use attributes.

To undertake this process the following are required:

- Reference layer with a geometry that is logically partitioned into specific units (for instance divided into parcels, houses, fields for agriculture, etc.). The use of artificial units like hexagons is theoretically possible, but not recommended.
- Land use information sources that can be linked to specific units like parcels, houses, etc. From a land use perspective it is ideal that the land use is already part of the reference layer, like it is the case in the Czech Republic cadastre. The situation in Flanders is more complex because the land use information is not part of the LRD parcels or buildings layer. The land use information needs to be derived from other LRD based layers like the company activity layer shown in Figure 5. From a user perspective, in Flanders you need to understand the LRD structure and the structure of the other linked data sources, which is more complex.

2.4.2 Step 2: Source Data Understanding - Understanding of Source Data Scheme up to the Level of Attributes

After the spatial integration, the individual polygonal features must have a link (ID and name of the original dataset) to the source, and also if applicable, a land use attribute or another attribute from which land use information can be derived (e.g. land cover attribute).

2.4.3 Step 3: Target Data Understanding

In this step, a proper data model is created for the open land use dataset. The data model used for land use classification is based on the INSPIRE HILUCS classification1. The HILUCS classification is part of INSPIRE and uses a three-level hierarchy that is suitable for most cases. If more detail is needed, extra levels can be added to the hierarchical structure.

The scheme below shows the hierarchy for Transport Networks Logistics And Utilities category. Besides this, the HILUCS scheme has five other top-level categories: Primary Production, Secondary Production, Tertiary Production, Residential Use and Other Uses.

The INSPIRE land use data model allows adding more than one land use category to a specific object, for example the land-use classification of a shopping centre area will match several land-use categories. In our case, only one land use category will be used for an object for the sake of simplicity and unambiguity. An example of a data model with attributes is shown in Figure 5:

Open Land Use Map - Vlaanderen Feature	
id	168765
hilucs_land_use	414
geometry_source	{"table": "gbg", "gid":"193420"}
attribute_source	{"table": "urban_atlas", "gid":"62705"}
municipal_code	BE231005
original_land_use	12300

Figure 5 Open Land Use Map – Data Model with Attributes

2.4.4 Step 4: Definition of Harmonization Steps

The harmonization process is nothing more than a translation and correspondence of the land use category from the original table to the land use class using HILUCS.

An example of such translation (or mapping) for land use categories in Czech Cadastre and HILUCS is given in the Figure 6:

1	skleník, pařeniště	112	FarmingInfrastructure
2	školka	332	EducationalServices
3	plantáž dřevin	12	Forestry
4	les jiný než hospodářský	12	Forestry
5	lesní pozemek, na kterém je budova	12	Forestry
6	rybník	632	WaterAreasNotInOtherEconomicUse
7	koryto vodního toku přirozené nebo upravené	632	WaterAreasNotInOtherEconomicUse
8	koryto vodního toku umělé	632	WaterAreasNotInOtherEconomicUse
9	vodní nádrž přírodní	632	WaterAreasNotInOtherEconomicUse
10	vodní nádrž umělá	432	WaterAndSewageInfrastructure
11	zamokřená plocha	632	WaterAreasNotInOtherEconomicUse
12	společný dvůr	53	OtherResidentialUse
13	zbořeniště	53	OtherResidentialUse
14	dráha	411	RoadTransport
15	dálnice	411	RoadTransport
16	silnice	411	RoadTransport
17	ostatní komunikace	411	RoadTransport
18	ostatní dopravní plocha	415	OtherTransportNetwork
19	zeleň	344	OpenAirRecreationalAreas
20	sportoviště a rekreační plocha	343	SportsInfrastructure
21	hřbitov, urnový háj	335	OtherCommunityServices
22	kulturní a osvětová plocha	341	CulturalServices
23	manipulační plocha	411	RoadTransport
24	dobývací prostor	133	OtherMiningAndQuarrying

Figure 6 Translation Table between the Czech Cadastre Classification

2.4.5 Step 5: Practical Realization

To do the spatial integration of data sources (which can be quite large, with millions of features), and to cover the whole area of interest, it is advised to do a partition of the data by single communes within the EU. The list of communes, together with their generalized geometries and IDs can be found at the pages of Eurostat.

For data storage the PostgreSQL RDBMS is used. Practical implementation of the spatial data integration was done in .sql script using spatial functions from the PostGIS extension of PostgreSQL. After spatial integration of the data, the translation tables were used for deriving land use HILUCS class of features.

Last, if data didn't have an attribute from which the land use could be derived, then the spatial relationship with the data sources that has this attribute was examined. For instance in Flanders, parcels and buildings don't have a direct attribute from which land use could be derived - so the spatial relationship between these objects and features of the Urban Atlas were examined. And if the object was covered by an Urban Atlas feature, the object was getting the same land use attribute as the Urban Atlas feature, and then translation to HILUCS took place.

2.5 Open INSPIRE4Youth

The pilot is focused on building of Environmental and Geographical Web based atlas based on utilization of Geospatial data, Linked Open data and other environmental data (maps) for educational and gaming purposes. The main components of the environment will be introduced - water, air, soil, forests, nature protection, climate information, landscape, waste management, forest management etc.

The pilot applications (map composer, thematic map viewer and semantic explorer) will integrate many free and open data sets covering above mentioned topics (for example protected sites) as well as general geographical data (such as borders) enabling better localization of thematic data. Data integration will be realized similarly to majority of other pilots on the basis of 5-steps harmonization framework. It is composed of Understanding the theory of spatial data harmonization; Source data understanding; Target data understanding; Definition of harmonization steps; Practical realization of data harmonization (based on interconnected particular harmonisation steps).

2.6 Ecosystem Services Evaluation

EcoSystem Services (ESS) Evaluation pilot combines datasets related to environmental protection. The used data include:

- CORINE land cover.
- Protected sites – dataset containing protected sites of Slovakia formatting according INSPIRE rules and requirements.
- Urban Atlas – pan-European comparable land use and land cover data for Large Urban Zones with more than 100.000 inhabitants as defined by the Urban Audit.
- Open Street Map – a map of the world, created by voluntiers and free to use under an open license.
- Relevant data from Danube reference data and service infrastructure (DRDSI) catalogue.
- Land Use and Coverage Area frame Survey (LUCAS) – EUROSTAT's survey to identify changes in land use and cover in the European Union.
- SK Open Land Use – Open Land Use Map is a composite map that is intended to create detailed land-use maps of various regions based on certain pan-Europen datasets such as CORINE Landcover, UrbanAtlas enriched by available regional data. The transformation issues of this dataset are similar to the section 2.4 Open Land Use Map Through VGI.
- Statistical data published as GeoTIFF – ESS Wood (Paper pulp) production, ESS Number of livestock per hectare of pasture, ESS Carbon sequestration, ESS Landscape quality from tourism perspective, ESS Biodiversity, ESS Overall assessment

The above mentioned data will be integrated into Ecosystem services portal. It is simple and unified responsive web application providing the possibilities to communicate, collect, search, display and access the ESS related data and information (documents, projects). In addition portal will allow users to create the zonal statistics from the outcomes of the ecosystem services evaluation.

A transformation of particular data to RDF (to be a part of Linked data structure) mainly the Smart Open Data INSPIRE Vocabularies (https://www.w3.org/2015/03/inspire) will be implemented. Data will be transformed via OpenDataNode (a platform for Open Data publication developed in COMSODE project).

3 CONCLUSIONS

This deliverable is the initial document focused on data integration activities in the SDI4Apps project. The particular sections representing pilot activities of the project seems to be a bit unbalanced, but it is necessary to mention that this fact reflects a complexity of the data integration issues on the one hand and on the second hand the pilot are very different from the view of progress as well as dealing with the data and need of integration.

During the next year of the project there will more intensive exchange of knowledge and experience. This fact as well as better sharing of the data among pilots will lead more uniform and comparable deliverable Data integration 2.