Initial Deployment and Methodology for Quality Assessment

MARCH 2015

DELIVERABLE

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

D5.1INITIAL DEPLOYMENT AND METHODOLOGY FOR QUALITY ASSESSMENT

Revision no. 05

Authors: Barbora Musilová (University of West Bohemia)
Otakar Čerba (University of West Bohemia)
Tomáš Mildorf (University of West Bohemia)
Václav Čada (University of West Bohemia)
Karel Charvat Junior (Czech Centre for Science and Society)
Karel Charvat (Czech Centre for Science and Society)

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	03/01/2015	Barbora Musilová	UWB	Initial draft - data quality
02	03/03/2015	Tomáš Mildorf	UWB	Recommendations on data quality
03	12/03/2015	Tomáš Mildorf	UWB	SDI4Apps methodology for quality assessment
04	17/03/2015	Tomáš Mildorf	UWB	Reviewers comments on data quality incorporated
05	18/03/2015	Tomáš Mildorf	UWB	Minor changes
06	25/03/2015	Otakar Cerba, Karel Charvat Junior, Karel Charvat	UWB, HSRS, CCSS	Added data deployment, Final Version

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

DCAT	W3C Data Catalogue Vocabulary
DIS	draft international standard
GEOSS	Global Earth Observation System of Systems
GI	geographic information
ICT	information and communication technology
ISO	International Organization for Standardization
INSPIRE	Infrastructure for Spatial Information in the European Community
OGC	Open Geospatial Consortium
ORNL	Oak Ridge National Laboratory
OSM	OpenStreetMap
OSMCP	OpenStreetMap Collaborative Project
SDI	spatial data infrastructure
SME	small and medium enterprise
VGI	volunteered geographic information

EXECUTIVE SUMMARY

The aim of this report is to document open and where possible INSPIRE compliant data with relevant metadata which are integrated/harmonised in the SDI4Apps platform. In addition to that, the deliverable is addressing used data models and metadata profiles and describing mechanisms for initial publication and harmonisation of open data and metadata including the quality assessment of voluntary data which will be collected during the course of the project. This is not limited only to voluntary data. All other data sources are taken into account.

For initial deployment was selected two initial data sets OpenLandUse data sets for Europe, based on INSPIRE LandUse data models and Point of Interest, which will be base for SmartTouristData

The report describes the complexity of data quality with special focus on voluntary data. Different aspects of data quality are introduced and structured according to the ISO standard ISO/DIS 19157 Geographic information – Data quality. These aspects include lineage, completeness, logical consistency and positional, thematic and temporal accuracy. Based on these aspects, key basic principles were drawn. The data quality control is concluded by an example of processes and quality control mechanisms on an example of a transport network.

1 INTRODUCTION

1.1 About the Project

The potential of geospatial information (GI) collected by various actors ranging from public administrations to voluntary initiatives of citizens is not fully exploited. The advancements of ICT technologies and shift towards Linked Open Data give an excellent foundation for innovation based on reuse of GI. The establishment of SDI has largely been driven by the "traditional" GI community and the national and European policies governing this sector. However now 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 use and 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[1], Plan4business[2] and EnviroGrids[3], SmartOpenData[4] (hereinafter referred to as related projects), 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. SDI4Apps aims to ensure that users profit from INSPIRE, and that INSPIRE profits from different voluntary initiatives. SDI4Apps will build a "WIN-WIN" strategy for building a successful business for hundreds of SMEs on the basis of European SDIs.

1.2 Structure of the Report

The report is divided in two parts:

- Chapter 2 – Initial data publication.
- Chapter 3 – Data quality assessment – this chapter includes different aspects of data quality with focus on volunteered geographic information (VGI), basic principles for quality control in SDI4Apps and an example of process and quality control on a transport network data which will be collected and maintained by SDI4Apps.

[1] www.inspiredhabitats.eu
[2] www.plan4business.eu
[3] www.envirogrids.net
[4] http://www.smartopendata.eu/

2 INITIAL DATA PUBLICATION

For initial data deployment we decide to deploy large scale Open Data sets covering all Europe, eventually all World. The firs data sets are initial data for OpenLandUse pilots. There were selected more sources of data with different spatial and temporal extend and all this data were transformed into data models designed on the base of INSPIRE Land Use model. Second data sets are Point of Interest (PoI) with focus on tourism. For this purpose was defined data models based on RDF scheme, all data was transformed into this model and published in Viruoso. Initial data sets are based on regionals partners data set and PoI from OpenStreetMap. This data set is base for SmartTouristData pilot.

2.1 OpenLandUse deployment

2.1.1 Data model

The Open Land Use (OLU) data model joins two basic data model of the INSPIRE Land Use specification – existing land use and planned land use.
The main difference among INSPIRE data models and OLU model has been caused by the fact that OLU data model connects planning and existing land use data. In the OLU the different attributes are used for both types of land use data. The OLU model also follows INSPIRE land use specification (uses same data attributes; the set of used attributes is larger than in the case of Land Use Database Schema), but it works with more simple view on data. Both models are transformable to each other and it is also possible to migrate data from these model to or from other data sets that are in harmony with INSPIRE specification. The main reason for above-mentioned differences is determine by different usage of data and data models. OLU will be used for any land use (and land cover) data, Land Use Database Schema serves just to spatial planning data as a special part of land use data.
The OLU data model contains following components:

Model (LU Object)	Type	Multiplicity	Description
inspireId	Identifier	1	Identifier
geometry	GM_MultiSurface	1	Geometry
hilucsLandUse	HILUCSValue	**1**	HILUCS value (Land use classification)
regulationNature	*RegulationNatureValue*	***0..1***	*Nature of regulations*
<<lifeCycleInfo, voidable>>			
beginLifespanVersion	DataTime	1	Date related to beginning of existence of data
endLifespanVersion	DataTime	0..1	Date related to ending of existence of data
<<voidable>>			
hilucsPresence	HILUCSPresence	0..*	Representation (in percentage) of the HILUCS value in the area. Because majority of land use and land cover data sets do not deals with this construct, the hilucsPresence attribute will not be used in majority of cases.
specificLandUse	LandUseClassificationValue	**1..***	Value of original land use classification

specificPresence	SpecificPresence	0	Similarly to hilucsPresence attribute
observationDate	Date	1	Date of finding of situation represented by data (for example date of aerial photography).
processStepGeneral	*ProcessStepGeneralValue*	*1*	*Step of spatial planning, which caused creation of the object.*
backgroundMap	*BackgroundMapValue (combination of data of origin, link to the map as a string, optional URI)*	*1*	*Link to the map, that contains planned land use objects.*
dimensioningIndication	*DimensioningIndicationValue (interger, real or string)*	***0..1***	*Specifications about the dimensioning that are added to the dimensioning of the zoning elements that overlap the geometry of the supplementary regulation.*
validFrom	Date	0..1	Date related to ending of existence of the object represented by data
validTo	Date	0..1	Date related to ending of existence of the object represented by data
note	**Citation**	**0..1**	For example link to original data, link to a relevant spatial plan etc.
landCoverValue	**LandCoverClassificationValue**	**0..***	Value of original land cover classification

Table 1 OpenLandUse data model

Notes:

- Elements written by italic (for example *regulationNature*) are related only to planning land use areas.
- Bold values show changes in comparison to INSPIRE specification.
- Attributes with specific data types contains values that are defined in INSPIRE specification.
- There is also RDF version of OLU data model.

- There have been created following mappings between HILUCS classifications and nomenclatures used in other land use and land cover data sets: CORINE Land Cover, Urban Atlas, GlobCover, GeoBase Land Cover (Canada), Land register (Czech Republic), data from Prague City (Czech Republic).

```
geo:hasGeometry         rdfs:Literal
lu:hilucsLandUse        → URI (INSPIRE registry)
lu:regulationNature     <<ennumeration>> (bindingForDevelopers | bindingOnlyForAuthorities | generallyBinding | nonBinding | definedInLegislation)

<<lifeCycleInfo, voidable>>
gcm:beginLifespanVersion    xsd:date
gcm:endLifespanVersion      xsd:date

<<voidable>>
lu:specificLandUse      rdfs:Literal (in the future there could be URI to original Land use classification)
lu:observationDate      xsd:date
lu:processStepGeneral       <<ennumeration>> (adoption | elaboration | legalForce | obsolete)
lu:backgroundMap        → lu:BackgroundMapValue
lu:dimensioningIndication   rdfs:Literal
validFrom           xsd:date
validTo             xsd:date
olu:note            rdfs:comment



lu:BackgroundMapValue
---------------------
lu:backgroundMapDate        xsd:date (or xsd:dateTime, for example in case of satellite images)
lu:backgroundMapReference   rdfs:Literal

<<voidable>>
backgroudMapURI         →URI



Namespaces
----------
gcm:    Generic Conceptual Model    http://inspire.jrc.ec.europa.eu/schemas/gcm/3.0/
geo:    GeoSPARQL           http://www.opengis.net/ont/geosparql#
lu: Land Use            http://inspire.jrc.ec.europa.eu/schemas/lu/3.0/
olu:    Open Land Use
xsd:    XML Schema          http://www.w3.org/2001/XMLSchema#
```

Figure 1 RDF model for OpenLandUse

2.1.2 Initial OpenLandUse deployment

From the above mentioned data sets we are currently working with CORINE Land Cover (CLC), Urban Atlas, Land register of Czech Republic and data from Prague City. These data sets are not focuses exclusively on land use. Each of them is some kind of mix between land use and land cover data sets.

The covered area and temporal coverage varies between these data sets.

- Urban Atlas covers cities and large urban zones (over 100 000 inhabitants) in EU Currently there are over 300 cities or zone in the data sets.
- CLC covers whole EU.
- Czech Land register covers the whole Czech Republic but not all area are available in digital form at this moment. Currently approximately 85 % of cadastral areas are available.

The source data of various sets are available in various formats. For purposes of other work and analysis of the data we are importing all data sets into PostgreSQL database with PostGIS extension. The structure of all tables for the data sets is based on described data model, however in some cases we add several table columns to improve efficiency of various queries.

The numbers of rows in Open Land Use tables based on different data sets are following.

Data Set	Number of rows
Urban Atlas	6033225
CORINE Land Cover	2213233
Land register of Czech Republic	15176609
Prague land use	13128

Table 2 Initial OpenLandUse data sets

The level of detail of land use classification, which can be extracted from the data sets also varies between individual sets. HILUCS classification consists from three level. Level 1 is very general.

- Primary production
- Secondary production
- Tertiary production
- Transport networks, logistics and utilities
- Residential use
- Other Uses

Each of the next two levels is more detailed than previous level. Various categories in classifications used in original data sets are linked to various levels of HILUCS classification. For example Urban Atlas involves 22 Categories. 7 of them is linked to first level of HILUCS classification, 4 of them to second level and 5 to third level. 6 Urban Atlas categories can't be clearly identified in HILUCS classification, so land use of this area is described as Not Known according to the HILUCS.

Urban Atlas	HILUCS
1.1.1 Continuous urban fabric (S.L. > 80%)	5_ResidentialUse
1.1.2 Discontinuous urban fabric (S.L. 10%-80%)	5_ResidentialUse
1.1.2.1 Discontinuous Dense Urban Fabric (S. L. 50%-80%)	5_ResidentialUse
1.1.2.2 Discontinuous Medium Density Urban Fabric (S. L. 30%-50%)	5_ResidentialUse
1.1.2.3 Discontinuous Low Density Urban Fabric (S. L. 10%-30%)	5_ResidentialUse
1.1.2.4 Discontinuous Very Low Density Urban Fabric (S. L. <10%)	5_ResidentialUse
1.1.3 Isolated structures	5_ResidentialUse
1.2.1 Industrial, commercial, public, military and private units	6_6_NotKnownUse
1.2.2 Road and rail network and associated land	4_1_TransportNetworks
1.2.2.1 Fast transit roads and associated land	4_1_1_RoadTransport
1.2.2.2 Other roads and associated land	4_1_1_RoadTransport
1.2.2.3 Railways and associated land	4_1_2_RailwayTransport
1.2.3 Port areas	4_1_4_WaterTransport
1.2.4 Airports	4_1_3_AirTransport
1.3.1 Mineral extraction and dump sites	6_6_NotKnownUse
1.3.3 Construction sites	6_6_NotKnownUse
1.3.4 Land without current use	6_6_NotKnownUse
1.4.1 Green urban areas	3_4_CulturalEntertainmentAndRecreationalServices
1.4.2 Sport and leisure facilities	3_4_CulturalEntertainmentAndRecreationalServices
2 Agricultural areas, semi-natural areas and wetlands	6_6_NotKnownUse
3 Forests	1_2_Forestry
5 Water	6_6_NotKnownUse

Table 3 Comparision of Urban Atlas and HLUCS classification

Following table shows numbers of rows according to the HILUCS classification in data set based on Urban Atlas

HILUCS	Rows
ResidentialUse	3299552
NotKnownUse	1891527
Forestry	502717

CulturalEntertainmentAndRecreationalServices	267820
RoadTransport	43523
RailwayTransport	21528
WaterTransport	5314
AirTransport	1241

Table 4 HILUCS classification in data set based on Urban Atlas

In following table, there is the same information for data set based on CORINE Land Cover

HILUCS	Rows
Forestry	554228
LandAreasNotInOtherEconomicUse	538621
ResidentialUse	142409
NotKnownUse	81707
CulturalEntertainmentAndRecreationalServices	17813
MiningAndQuarrying	9811
WaterTransport	3408
AirTransport	1630
WasteTreatment	1427
AbandonedAreas	686

Table 5 HILUCS classification in data set based on Corine Land Cover

These two examples shows, that some of the Categories, which appears in sets based on Urban Atlas don't appear in sets based on CORINE Land Cover and vice versa. It has two main reasons.

- Urban Atlas covers less area than CORINE Land Cover, but it contains much more objects in covered areas.
- Some HILUCS categories, which can be recognized from Urban Atlas Classification can't be recognized from CORINE Land Cover classification and vice versa.

In other data sets there are also some objects with not clear equivalent in HILUCS classification.
Currently there is no data set covering whole Europe, which would enable to clearly recognize land use of every place. To be able examine land use of every place, it is necessary to combine data from various sources. Even having data from different sources don't guarantee the possibility to clearly identify land use of every point. For many places following problems occur.

- There is no land use information for the place in any of the available data sets.
- The land use information for the place is different according to different data sets. In this situation, there is no clear rule which data set should be preferred.

Let's suppose there are two data set's covering some place, where you are trying to recognize the land use. First data set is more detailed, but older and the other one is less detailed, but newer. In situation like this the choice of the most suitable data set would probably depend on the purpose of examination of the place.
For this reason further Open Land Use work will focus on possibility to compare land use information from various data sets for each place.

2.2 Tourist Point of Interests

2.2.1 Data model

POI base is a part of Open Smart Tourist Data pilot activity. The final version of the POI base will contain POI (Point of Interest) over the Europe. The data are collected from various heterogeneous resources. The current

version contains selected data from OpenStreetMap, data provided by non-profit organization Úhlava (south-west Bohemia), data from experimental ontologies (ski resorts in Europe and sights in Rome) developed in the University of West Bohemia. The data are kept as RDF triples in Virtuoso tool. The current version contains more than 1,2 millions of triples.
The data model of POI base includes following components:

- Identifier of the object
- Labels (in various languages) of obejcts – rdfs:label
- Geometry – latitude and longitude published according to http://www.w3.org/2003/01/geo/wgs84_pos# – geo:lat, geo:long
- Category of POI (adopted from 1st level of OpenStreetMap classification) – poi:category
- Complete OpenStreetMap classification – poi:category_osm
- Link to DBpedia and GeoNames.org representations of relevant country – geo:sf-within (adopted from standard topological relations)

This basic structure (which is same for all data) was extended by several attributes which are contained in several input data sets:

- poi:address
- poi:email
- poi:fax
- poi:phone
- poi:www
- poi:opening_hours
- poi:access
- Links to Wikipedia, Wolfram Alpha or piste maps – redfs:seeAlso
- Links to relevant objects in other knowledge bases – skos:exactMatch

2.2.2 Deployment

Transformation and mapping of classifications were realized by XSLT templates and XSLT processor. As the output the RDF file (554,7MB) is provided which is uploaded to the Virtuoso tools. Currently there is a SPARQL endpoint (http://ha.isaf2014.info:8890/sparql) and simple visualization (http://ha.isaf2014.info/wwwlibs/hslayers-ng/examples/vectorSparql/?hs_x=1482015.3749640775&hs_y=6342528.247547781&hs_z=14&hs_panel=&visible_layers=Base%20layer%3BPoints%20of%20interest).

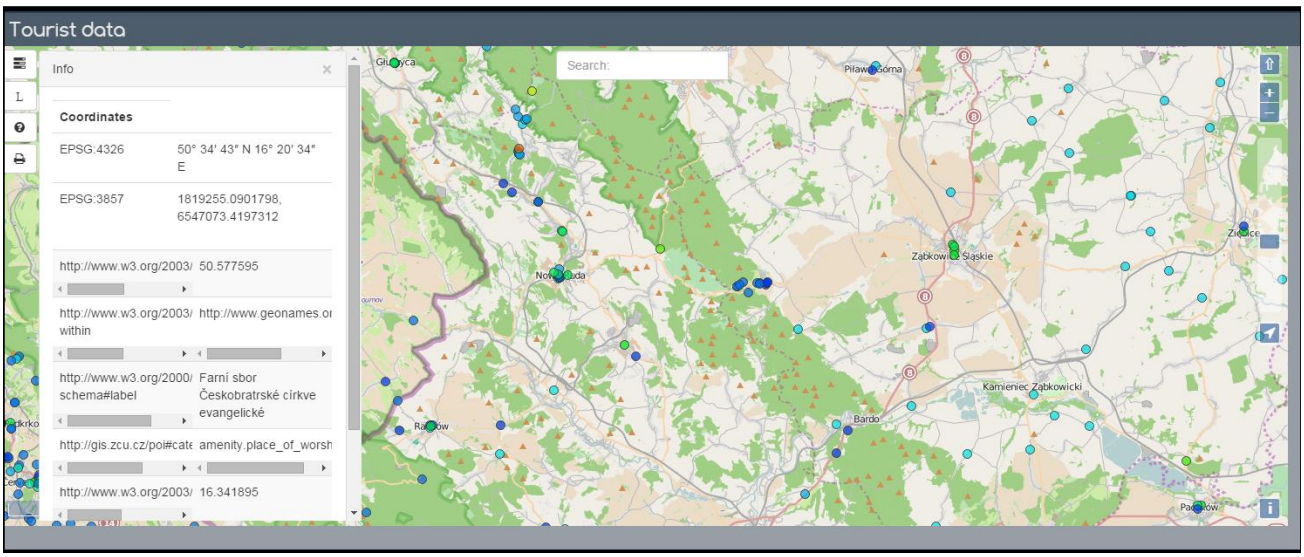

Figure 2 Visualisation of PoI triples using HSlayers NG

3 DATA QUALITY ASSESSMENT

3.1 Different Aspects of Data Quality

Outcomes of data quality assessment provides an important information about the possible use of data for particular needs. Evaluation of data quality is a subject facing various methodological approaches and ambitions, taking into consideration maturation of the GI domain as well as other domains addressed by the scope of the SDI4apps project, particularly Semantic Web technologies.

In order to ensure comparability of the data quality aspects between VGI and other GI resources (e.g. INSPIRE (Tóth, et. al. 2013)[5], GEOSS[6]), relevant standardisation activities have to be taken account (ISO, OGC). From the ISO/OGC perspective the structure of the main classification of particular aspects of data quality is based on the ISO standard ISO/DIS 19157 Geographic information – Data quality. The data quality aspects in this report are structured according to the components of data quality in this ISO standard.

Similar activities are taking place from web of data direction driven by the W3C efforts (Data on the Web Best Practices[7], DQ Notes[8]) as well as via some projects supporting publication of linked open data (e.g. C OMSODE[9]).

There is no quality aspect to be omitted, as well as it is not possible to consider any of them to be more important than any other. Data should supply all quality aspects as much as possible, weakness in any of them can lead to biased representation of the real situation.

In general, there are two main categories of evaluation of a dataset quality (according to Cho, 2013).

- **Internal evaluation** – the quality aspect of data is investigated in the dataset.
- **External evaluation** – the quality aspect of data is assessed from comparison with "true situation", usually represented by a dataset with higher quality level (e.g. national or military surveys).

For the purpose of external evaluation, a comparison of investigated dataset with "real world situation" would lead to the most appropriate quality determining. However, the "real situation" is not always easily accessible and therefore comparison with a dataset with higher quality should be used instead. For that purpose, official national surveys or data from military services (or any other surveys fulfilling data quality requirements) fit appropriately.

Some methods of external evaluation require assessing features from investigated and reference dataset. The matching should be always done on a principle: a feature with lower quality matched to the feature with higher quality.

The easiest way to assign the features is according to name or another key attribute – however, in volunteered datasets, these attributes are often omitted and therefore different way to do the matching has to be applied.

There exist several matching algorithms, suitable differently according to the feature type. For meaningful results of all of them, sufficient positional accuracy is required (small differences or errors can be easily compensated by using buffer):

- One of the most common is based on "buffer zone", when a buffer with specified width is made around an investigated feature, and the closest feature from reference dataset (ideally with the same name or ID, or another key attribute) is assigned. This method suits well for point features, modified version could be used also for linear features.

[5] http://inspire.ec.europa.eu/documents/INSPIRE_/JRC83209_Online_Data_quality_in_INSPIRE.pdf
[6] http://wiki.ieee-earth.org/Documents/GEOSS_Tutorials/GEOSS_Provider_Tutorials/Data_Quality_Tutorial_for_GEOSS_Providers/Section_0%3a_Table_of_Contents
[7] http://www.w3.org/TR/dwbp/
[8] https://www.w3.org/2013/dwbp/wiki/Data_quality_notes
[9] http://www.comsode.eu/wp-content/uploads/D3.2_-_Requirements_summary1.pdf

- For areal (polygon) features, "centroid method" suits better. It is detected, whether the geometrical centre of the reference feature lies within the area of the investigated areal features – if so, these two features are matched. (used by Hecht, 2013)
- Also, the "overlap method" could be used for areal features, investigating overlap area of each two features. When the overlap is at least 50% of the area of reference feature (i.e. no other investigated feature would have bigger overlap), the features are matched. (used by Hecht, 2013)

For all quality aspects, the methodology of measuring the quality level can be based on:

a) the entire database (e.g. comparing all features to any other dataset),

b) samples (stratified selection, ideally).

Using samples makes sense for those areas and data, when features are located with logical manners (i.e. not chaotically), and when the dataset is sufficiently complete.

In some of the further described methods, a dividing the area of interest into subareas is used (Figure 3). These subareas can be defined by several keys (according to Hecht, 2013):

- geometrical raster (e.g. square grid, hexagonal grid),
- administrative units,
- centric circles (e.g. around an important point/place – capital, harbour),
- centric shape reflecting shape of the area of interest.

Before selecting a way to define subareas, it is always important to consider characteristics of features and the investigated area (i.e. heterogeneity/homogeneity of features, density of features, shape of the area, locations of big cities).

(a) Geometrical raster (square, hexagonal); (b) Administrative units (municipalities); (c) Concentric circles; (d) Buffering of municipal borders.

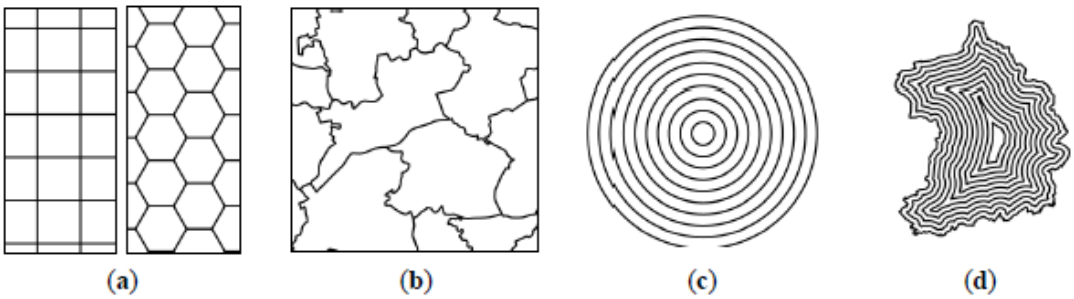

Figure 3 Division of the area of interest (Hecht, 2013)

3.1.1 Lineage

Lineage is about history of a dataset – how it was collected, including the information about the author and the date, the source of the data or the method used to obtain the data.

Even that lineage is not a standard quality aspect (according to the ISO standards), and the level of quality is hardly measurable, it can influence the data quality in all conditions, and therefore the attention should be paid to lineage as well as to other aspects. Data can be split into two obvious groups – those with lineage, and those without. In all cases of use, data with lineage are preferred and in most cases the known history should be mandatory for further processing.

The existence of lineage, and the level of information about it, is itself an indicator of data quality. Based on some lineage information, levels of other quality aspects can be obtained (e.g. positional precision, temporal quality). Also, when any error in a dataset is found, any systematic aspects (e.g. weather condition, "a problem user") can be disproved or confirmed according to the lineage.

Possible lineage attributes can include:

- information about the author (e.g. identification, being citizen in the place, age, occupation),

- source of data (e.g. own survey, a paper map, other dataset),
- method of collecting (e.g. digitisation of paper maps, GPS tracking, “a la vue”, methods/software used to input in the dataset),
- date and date connected conditions (e.g. date of survey, date of input, weather).

Recommendation: All datasets should be accompanied by metadata of good quality. Where possible also for every feature.

Note: Geospatial data should follow the INSPIRE specifications for metadata, non-spatial data the W3C Data Catalogue Vocabulary (DCAT) standard.

3.1.2 Completeness

Completeness is a condition where presence or absence of features and their attributes and relationships is investigated (in comparison to the real world situation).

The procedure includes several aspects:

- **Completeness of the whole dataset**, i.e. what layers are in the database included and what features they contain. It should be inspected, if all types of features are included, or whether the omitting of any type is meaningful (e.g. no occurrence of the tree species in the area of interest at all, or systematic selecting on higher classification level).
- **Completeness of features**, i.e. if the database of features of some type is complete. Omissions can be caused by systematic selection (e.g. based on size or further categorisation), or it can be a result of non-systematic evidence (e.g. omitting in areas with lower population density, or an incorrect classification of the feature)
- **Completeness of attributes**, i.e. whether mandatory attributes of features are filled (e.g. author, date, type of road), and level of completeness of optional attributes.

Unrecorded systematic absence of any features or key attributes can lead to biased results based on the data, e.g. incorrect correlations. Working with dataset with several random omissions can cause misleading results of analyses based on spatial relations.

Recommendation: The completeness of VGI data in terms of the above mentioned aspects should be as much as possible ensured automatically by the SDI4Apps platform.

Note: For example, the user providing data input will be requested to fill in mandatory attributes.

When evaluating level of completeness, there can be used two main categories of methods

- Internal evaluation – the complements of a dataset are investigated particularly.
- External evaluation – completeness of data is assessed from comparison with “true situation”, usually represented by dataset with higher quality level (e.g. national surveys).

Internal evaluation

In case of volunteered information, completeness cannot be easily measured as ratio of completed objects to all objects planned to be surveyed, since the process of building the database is unclosed (and also, it is not usually specified, which all objects should be involved in the survey). However, there exists another internal approximation of completeness:

- After dividing the area of interest into several comparable subareas, a level of completeness of particular subarea can be calculated as: Number of objects in the subarea/maximum number of objects of any region (Arago, 2011). This comparison would lead to meaningful results only when the subareas would have similar characteristics (especially size and aspects affecting density of investigated features).

Hence, using the dataset itself to obtain the level of completeness is recommended only for areas with uniform feature distribution or/and when no data with significantly better quality are accessible.

Evaluation based on external dataset

For obtaining the level of completeness, a comparison of investigated dataset with the "real world situation" would lead to the most appropriate quality determining and should be made before further data processing.

In comparison, three main categories of methods can be applied. Choosing the right way/method should be conditioned by tracked features characteristics (e.g. point/area/line feature, dislocation in the area, number of objects overall).

- **Comparison of individual features** (one-to-one) is based on matching features from the investigated dataset to features from the reference dataset. This methods can be used either for assessing completeness of total dataset (by comparison of number of matched features) or, after matching, to obtain level of attribute completeness. An example from this kind of matching is in Table 6.

Match Method	ORNL-OSMCP		ORNL-OSM	
	Record Count	Percent	Record Count	Percent
1-4, 6-9 Automated Matches	329	82%	225	56%
5, 10 Manual Matches	28	7%	62	16%
11 No Match	44	11%	114	28%
Total	401	100%	401	100%

Table 6 Matching results for ORNL and OSMCP Data of public and private schools. ORNL - Oak Ridge National Laboratory, OSMCP - OpenStreetMap Collaborative Project, OSM - OpenStreetMap (Jackson, 2013)

- **Comparison of features as a complex** is based on comparison of numeric summaries of features in both datasets. However, it is necessary to remember, that in both datasets can be omitted different features (e.g. systematic exclusion of path-walks in official dataset of streets, vs. random omissions of all types of roads in VGI dataset), which can lead to similar quantities of features, but do not mean completeness of the investigated one.
    - For all types of features, the easiest way is to compare numbers of features in the area of interest.
    - For linear features, the "length" comparison could be used, when summary of total lengths of roads/rivers/… in investigated dataset to summary in reference dataset is compared. (used by Goodchild, 1997 or Kounadi, 2009)
    - For area features, this can be modified as summary of total areas of the features.
    - For all types of features, attribute completeness can be easily determined by comparing number of attributes in both datasets (e.g. name completeness)
- **Dividing the area of interest into grids** and comparing results in them (grid-to-grid) is modification of methods described above (used by Kounadi, 2009). The advantage is, that the level of completeness is determined for each grid separately, so positional heterogeneity in quality is more possible to reveal (e.g. differences urban vs. suburban areas).

Cells	Area (km^2)
Empty cells	17 632 (14.3%)
Meridian 2 more detailed than OSM	80 041 (64.7%)
OSM more detailed than Meridian 2	26 041 (21.0%)

Total	123 714 (100%)

Table 7Street data - lengths comparison with attributes between OpenStreetMap (OSM) and Ordnance Survey Meridian 2 (Haklay, 2010)

3.1.3 Logical Consistency

Logical consistency is a quality element focused on internal consistency of dataset, detecting failures and defects in logical rules (including topological correctness and rules resulting from feature characteristics). In this aspect could be also involved differences in attribute labelling (e.g. low/upper case, names with or without space, more spaces in attribute) and positions (e.g. building as point/polygon, placing building in middle of parcel or in the street adjacent to the building) caused by not standardised and unified collection process.

Errors in logical consistency can be represented by several aspects:

- topological and geometrical validity (e.g. correctness of topological relations, such as continuity of linear objects, not parcel overlapping, closed polygons),
- positional validity (e.g. overlapping ways, buildings in water areas),
- attribute logical consistency (e.g. dead ended one-ways),
- checking for level of inconsistencies arising from lack of standards (e.g. building as point/polygon, variations of local names).

Some of events, investigated for logical consistency, overlap with other quality aspects – some problems with logical rules can reflect defects in attribute accuracy, some with positional accuracy. Hence, checking for logical errors is important step that should precede accuracy evaluation.

For the purpose of indicating errors in logical consistency, internal evaluation is sufficient (i.e. investigating the inner relations and values in dataset).

In some cases, error detecting mechanics are integrated in the VGI tool (e.g. checking attribute values, e.g. the attribute “date” cannot be a text, “number of floors” has to be positive integer) or some another error detecting tool is accessible for contributors (e.g. for OSM exist several tools, checking for topological consistency, null attributes, inconsistency in attributes).

However, it is not a rule, so before further work with VGI datasets, they should be screened, ideally by computerised checking tool(s) and detected failures should be corrected.

The queries mechanics depend on the database system of the dataset and feature types and characteristics. There will be different queries for features only with positional and attribute components, and features with positional, attribute and relative components (relative component defines relation with other features in touch). For a dataset with wide range of features and attributes, the queries will be more complicated than for simpler datasets. The investigation should be done over the whole dataset, since the logical errors have to be detected and corrected before further work with the dataset, and sampling would not provide sufficient results to do so.

Logical inconsistency in attributes relation (e.g. discontinuity of roads or rivers, gaps/overlaps = more than one parcel line, crossing roads without bridge/crossroad/tunnel) would lead to biased exploration above the dataset (e.g. incorrect shortest way, multivalent parcel owner).

Recommendation: Discrepancies in logical consistency can be minimised using standardised data specifications and automated error detecting mechanisms integrated in the SDI4Apps platform.

3.1.4 Positional Accuracy

Positional accuracy is one of the most obvious quality aspects. It reflects closeness of localisation of the feature to the real situation (real position on the ground).

One of the aspects, closely connected with level of positional accuracy, is precision of measuring (and also placing) locations of the features. This depends on the method of investigation:

- using an equipment (e.g. GPS tracker): precision determined by the device precision,
- redrawing from photography or another map: precision derived from the map resolution,
- "a la vue", i.e. estimation based on what contributor see/know: precision indeterminable, dependent on the contributors carefulness, knowledge.

Due to lack of standards for VGI, and therefore multivariate of placing of some features, some other failings in positional accuracy can occur (e.g. placing building as point in middle of the building outline vs. in the street adjacent to the building), which should be considered while evaluating.

Datasets	Distance (m)		
	< 30	30 - 150	< 150
ORNL-OSMCP	164 (45.8%)	178 (49.7%)	15 (4.2%)
ORNL-OSM	90 (31.4%)	186 (64.8%)	11 (3.8%)

Table 8Percent of schools with spatial error in each threshold - 30m = average size of one half of school building side, 50m = side of school area (Jackson, 2013)

In order to investigate the level of positional accuracy, a comparison to a dataset with higher quality (as the closest representation of reality one can obtain) should be done. Nevertheless, there exist some minor methods of internal evaluation, for those cases, when datasets with higher quality are not available, or for general overview and detecting obvious positional errors in a pre-evaluation.

Before choosing the right method, characteristics of data should be considered to determine desired accuracy, as for some datasets the position is required to be more accurate than for others (e.g. streets in cities should be located more accurately than field trails).

Internal evaluation

A comparison of the investigated dataset with the reality (represented by better quality dataset) is the best way to determine the positional accuracy. When the reference dataset is not possible or not necessary to use, a few internal evaluation methods exist:

- **Evaluation based on logical rules** can be used for primary evaluation of positional accuracy and for detecting gross positional errors. This procedure should be already included in methods of logical consistency evaluation, i.e. automated checking of database on topological errors.

  Specially, for positional accuracy purposes

  - checking for logical coordinate values (depending on coordinate system, e.g. all features from dataset of land use in Poland should be placed on northern hemisphere),
  - checking for logical relations (e.g. a building should have an access road, in lake can be placed only some types of features, …),
  - dualities (multiple positions of one feature point out that probably at least one of those positions is incorrect),
  - or geometry validity (e.g. closed polygons).

  This method detects potential errors that should be checked and corrected individually (e.g. restaurant in lake can be badly located, but also it could be really placed on lake, in houseboat). Larger number of errors can detect more complex problem in dataset or in dataset implementation (e.g. inconsistency in positional data conversion).

- **When history of features editing is available**, edits in coordinate system can be tracked. For this purpose, objects should be decomposed to points (when the dataset structure enables it), and positional changes of these points are tracked. Usually, not all edits of the point have to be monitored – important is the current localization that can be compared either with the last previous localization, or with the position at first placing of the feature. Also, in some cases, comparison to some middle-term situation should be done (e.g. situation after collective changes in

dataset). Big differences between compared values could reflect some positional inconsistencies in the dataset.

All in all, on the base of results of the internal data evaluation, potentially problem places should be checked (compared to real situation) and eventually corrected.

Evaluation based on reference (external) dataset

For obtaining the quality of positional accuracy, a comparison of the investigated dataset with the "real world situation" would lead to the most appropriate quality determining. Real situation should be represented by reference dataset with enough high quality.

For comparison, several methods can be used:

- **Buffer zone** method is based on a proportion of the tested features that lie in buffer zone around objects from reference dataset. For linear features, proportion of length within buffer can be calculated, for point features, number of points. Also, a buffer zone around the tested features can be created (tighter than around the reference features), that reflects precision of input and real width of the object. When implementing this method, different widths of buffer zone should be used to see the positional accuracy. A distance, within e.g. 95% of tested features lies, can be then identified easily, as well as only proportion of features lying within a distance can be obtained to further comparison. (used e.g. by Goodchild, 1997, Hunter, 1999, Coleman, 2010, Haklay, 2010)

  Buffer zone method was designed for linear features accuracy evaluation, and hence for point and area features is not so well convenient.

Buffer width [m]	Length of coastline within the buffer [km]	% of coastline within the buffer
20	17.2	9,6
40	32	17,9
100	75.7	42,3
150	105.7	59,0
200	126.7	70,8
300	156.8	87,6
450	172.2	96,2
1000	175.6	98,1

Table 9 Case Study results showing the length and percentage of coastline inside the buffer in terms of its width (Goodchild, 1997)

- **Points positions comparison** can be used either for point, linear or areal features. Before computing, points from investigated and reference dataset have to be matched. For point features, simple difference in positions is the indicator of quality. For linear and areal objects have to be firstly chosen points to be compared – it can be edges or nodal points, or points sampled with a constant interval (e.g. one point each 50m) (used by Coleman, 2010). Than the matching points are compared and for the feature the average point difference can be calculated.

Spatial Error (m)						
Dataset	**Count**	**Minimum**	**Maximum**	**Mean**	**St. Deviation**	**Media**
ORNL-OSMCP	357	2	487	47	50	33

ORNL-OSM	287	2	1848	190	314	43

Table 10 Spatial Error for matched schools (Jackson, 2013)

- For area objects, the **overlap method** can be used as well, investigating overlap area of each two features. Proportion of overlapping area (should be bigger than 50% to exclude matching more features to one feature) reflects positional accuracy and precision.
- Comparisons can be made for whole dataset, or the investigated area can be divided into grids and comparison can be made on the grid level. The advantage of this dividing is ability of detecting heterogeneities in positional accuracy in the tested dataset.

Area	Average (m)
Barnet	6.77
Highgate	8.33
New Cross	6.04
South Norwood	3.17
Sutton	4.83
Total	5.83

Table 11Street comparison (Ordnance Survey vs. OpenStreetMap). Average difference of 100 samples in each area (Haklay, 2010)

3.1.5 Thematic Accuracy

Thematic accuracy is correctness of attributes, including proper spelling and classification. The quality results from knowledge and carefulness of contributor, and also from existence/lack of controlling mechanisms (e.g. auto check while inserting the feature). The level of quality can be investigated both by internal and external methods.

Also thematic accuracy in terms of usage and purpose of data should be checked before further work with the dataset. This should be evaluated individually, according the best knowledge and conscience rather than by any universal mechanic.

Thematic accuracy is important. Thematically unsuitable dataset disables meaningful investigation, and improper classification or wrong attributes lead to incompleteness of dataset, and to potential errors in work over the dataset.

Internal evaluation

Because of the evaluation of thematic accuracy is mostly composed of checking fulfilment of logical, typographical, and other rules, main evaluation should be based on internal methods. As the indicator of level of accuracy can be used a proportion of number of error features to exposed features (i.e. number of features with the attribute filled). The indicators should be computed for each attribute or kind of error separately:

- Checking for typographical errors – e.g. lower/upper case, misspelling;
- Checking for logical inconsistencies in attributes, including wrong attributes - e.g. one-way dead-end road, incorrect direction in round-about, wrong data type, data out of logical range (e.g. dates, some quantities), checking for existence in reference database (e.g. tree species), comparing with theoretical distribution (e.g. tree species, architectural style), comparing overlapped areas (e.g. residential land use in industrial area);
- Checking for unfilled mandatory attributes

When history of features editing is available, edits in attributes can be tracked.

- Comparing attribute versions - Not all of them have to be monitored usually, comparing last two versions (the current and the last previous) should be sufficient enough to detect potential errors.
- Number of edits of an attribute can also show problem features. An editing of attribute indicates apparently necessity to correct or refine the attribute. In some cases (i.e. for some attributes, which are not supposed to be updated), more than few editing can reflects potential problem, specially, when the edits were made by more people.

Version	Number of Objects	Total %	Cumulative %
1	2 246 369	59.211	59.211%
2	780 320	20.568	79.779%
3	329 342	8.681	88.460%
4	169 831	4.477	92.937%
5	91 488	2.412	95.349%

Table 12Version number distribution for all objects in the OSM dataset for the UK and Ireland. (Mooney, 2012)

Evaluation based on external data source

For the purpose of external evaluation, a reference dataset with relevant attributes has to be obtained and assessing of investigated and reference database has to be done firstly. When the identification attribute (i.e. ID, name) is investigated, the assignment should be based on positions of features.

The thematic accuracy is evaluated by comparing relevant attributes in each matched pair. Then, the proportion of correct/incorrect attributes in dataset is computed. As the denominator, total number of features with the attribute (i.e. not null) should be used rather than only total number of features.

3.1.6 Temporal Quality

Temporal quality is a parameter of currency of features in the dataset and how much are the data up-to-date. It reflects, how quickly and accurately were changes made in the dataset, when something changed in the "real world". The positional component of features usually does not change, so the changes are mostly related to attribute component. Also, it should be considered, that no dataset is possible to be absolutely current, as the changes can be proceeding in every moment. A method used to evaluate the temporal accuracy should reflect these two circumstances.

Internal investigation of the dataset is for obtaining the level of temporal quality sufficient enough. Before evaluating, it is important to ensure, that all features have a temporal date/attribute. If there is not any information about the date when the data were collected (i.e. to which date are current), or at least the publication date, no evaluation method would make sense and a dataset like this should be excluded from further investigation.

For evaluation of temporal quality, several indicators can be used.

- Currency of the whole dataset/layer in general is based on statistical indicators of temporal attributes (both cases – dates of publication or date of collection) – e.g. from which date are the oldest features, and when were made the last edits, distribution of adding/editing in time.
- Currency of features shows temporal accuracy of individual features in the dataset. The most important, and sufficient for quick overview, is the date of the last edit of the feature. More complex investigation should be based on tracking changes in the dataset (e.g. how big the changes were, how often, the number of changes) together with information about date of the change (e.g. when was the previous or last edit) and information about a contributor (who edited the feature – e.g. was he/she the same person as the last editor?)

- In both cases, the same result values of temporal quality evaluation can be for some datasets sufficient, whether for some they can be lacking. This results from the fact, that some features (and some attributes) should be updated more often than some others, according to expectations and frequency of the changes (e.g. number of pupils in a school would change at least periodically every year, in contrast to tree species, where no change is expected).

For further investigation of a dataset, it is always important to know the date, to which is the information current. Then, considering purpose of the investigation, the fulfilment of temporal quality requirements should be checked.

Recommendation: For every feature and dataset, temporal aspects including the date of creation and/or edit and who did the change should be registered in the SDI4Apps platform.

3.2 SDI4Apps Quality Control

3.2.1 Basic Principles

The SDI4Apps methodology for quality control of volunteered geographic information (VGI) and other information will slightly differ between different pilot applications, depending on the data collected and their intended use. Each pilot, which is considering collecting data from various sources including VGI, will define mechanisms for quality control. The proposed methodology should ensure that quality of datasets is well documented and monitored.

Figure 4 shows general data flows and key processes while integrating VGI and other open data into a common repository.

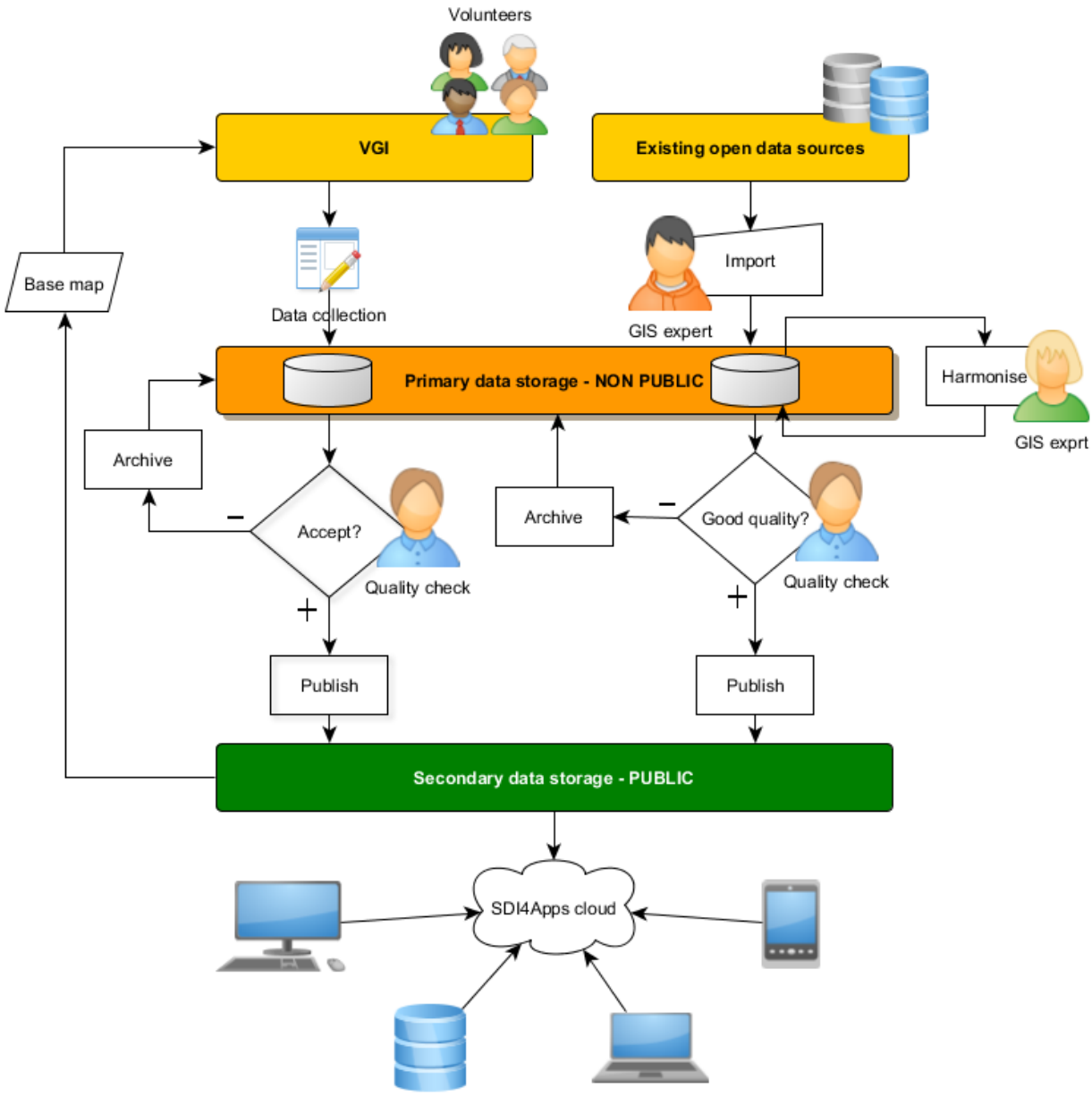

Figure 4Data flows and key processes

The SDI4Apps platform will serve to anyone who seeks access to data – open and mainly geographical data. The intended use and the respective quality requirements are not known. Therefore, the SDI4Apps quality control will focus mainly on documenting the quality of existing and newly collected data. The quality of data should be clear for the user of the SDI4Apps platform.

General recommendations on data quality mechanisms are included in Section 3.1 . These are:

- All datasets should be accompanied by metadata of good quality. Where possible also for every feature.
- Discrepancies in logical consistency can be minimised using standardised data specifications and automated error detecting mechanisms integrated in the SDI4Apps platform.
- The completeness of VGI data in terms of the above mentioned aspects should be as much as possible ensured automatically by the SDI4Apps platform.

- For every feature and dataset, temporal aspects including the date of creation and/or edit and who did the change should be registered in the SDI4Apps platform.

It is vital that any data collected voluntarily or imported from other open data sources are not published through the SDI4Apps platform before a manual quality check is done by local operators. All collected data will be firstly stored in a primary database, which is non-public. After approval, data will be published through the secondary data storage which will be publicly accessible. In the case that data are rejected by the local operators due to non-sufficient quality, data will be archived in the primary data storage for statistical purposes and future improvements in the quality check mechanisms.

Another important feature is automation of certain types of quality check. In the case of VGI, certain quality check should be ensured by the application for data collection. For example by setting certain data types for user's input and mandatory elements.

These are general principles that should be followed when designing the quality control mechanisms for a particular pilot application. The following section shows an example of a concrete plan for quality control that will be applied for road network. This dataset will be used in several pilot applications.

3.2.2 Transport Network as an Example

The aim is to create a database of transport network. There will be three layers of the transport network:

1. N**on-guaranteed voluntary INSPIRE OSM** map which will be created by simple transformation of OSM data into the data model based on the INSPIRE specifications.
2. **Guaranteed INSPIRE OSM map** which will have its own local administrators that will be responsible for keeping the map in their corresponding regions up-to-date and who will do the quality check. There are several features that need to be accomplished in order to provide this service:
    a. The administrator will transfer and edit data from other (other than OSM) sources to be fully (topologically and by attributes) merged into the existing database. This method is semi-automated. The starting point is the non-guaranteed INSPIRE OSM map.
    b. The administrator will select and then transfer selected changes from the non-guaranteed INSPIRE OSM database to this guaranteed layer. In the process of selection the admin needs to make sure, that the data he/she wants to transfer will not overwrite the data that were already updated.
    c. An editor for VGI data collection will be developed. The user will be able to make edits to the network. The changes made however again need to be checked and approved by the local administrator.
    d. A tool, which will provide capability to extend/edit the network by data collected from GNSS devices (e.g. in cars) will be developed.
    e. It is necessary to elaborate the way the data can be transferred from our map to OSM database to be useful for OSM users too.
3. **Other regional maps** that are taken from different commercial and public sources (guaranteed maps) and that will be converted into the data model based on the INSPIRE data specifications. The decision whether to use these maps in a pilot or other applications will be based on licensing and data quality that fits the intended use.

By taking these steps (depicted in Figure 5) we want to develop and maintain maps (transport network) for each region in Europe that will be in the same data model. For different applications thus it will be possible to have maps that are using the same data model. These maps will be of different quality depending on the types of the map available for a certain region.

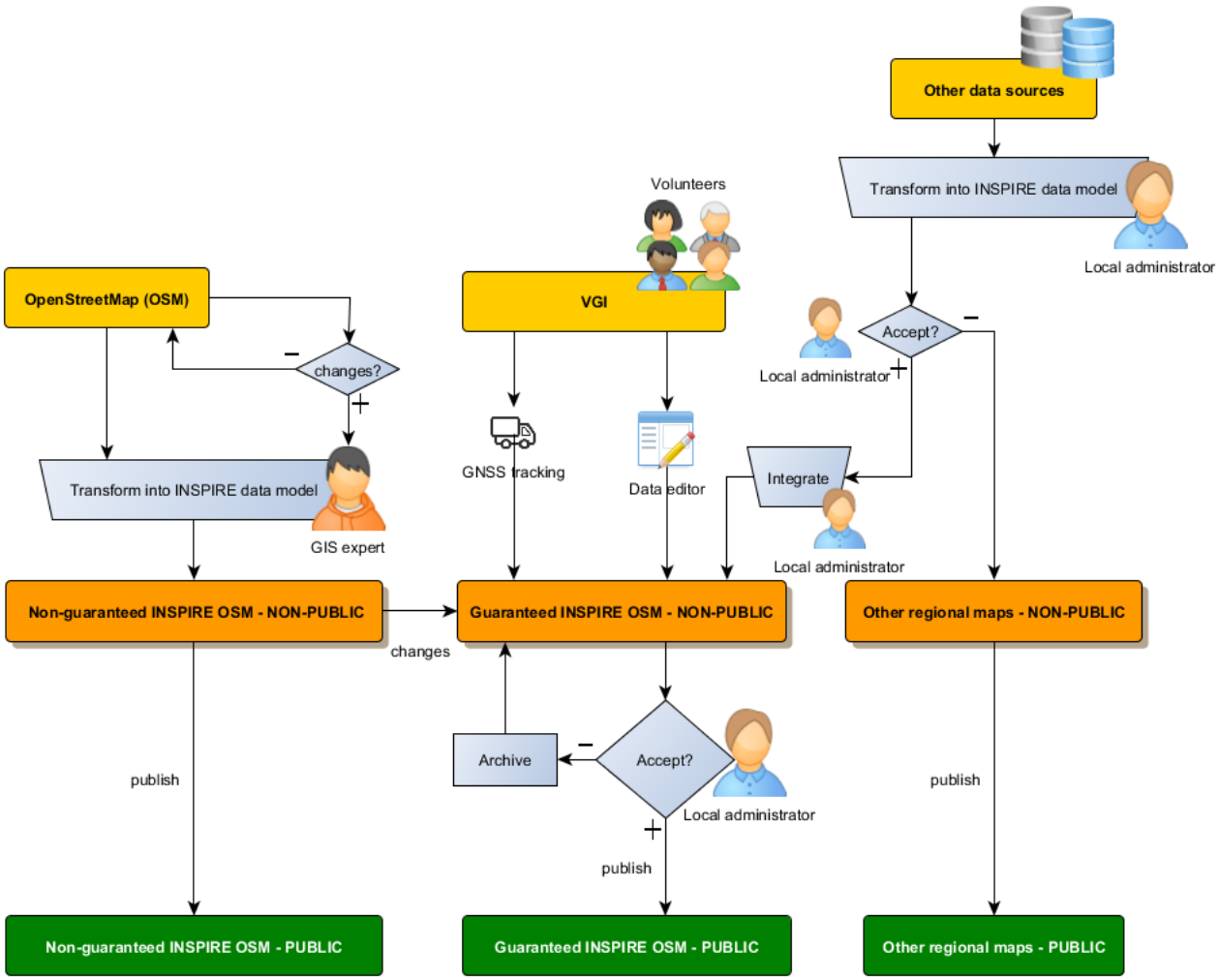

Figure 5 The process of creating transport network

4 CONCLUSION

Data quality is a complex issue. The report shows how complex it can get when taking into account different aspects of data quality. The SDI4Apps approach is pragmatic and rather than improving quality of existing datasets, it focuses on proper documentation of the quality through metadata and standardisation of data models and other aspects of data interoperability. Many issues regarding for example data completeness can be monitored automatically through enforcing user to fill in certain information in a certain format.

A key role play local administrators who understand local conditions, legislation and language. Local administrators will decide whether a data update coming from VGI or other sources will be accepted or not taking into account the needs of the pilot sites and the aim to improve the data quality in terms of the different aspects presented in Section 3.1 . The SDI4Apps platform will offer users the same type of data in different quality and coverage, as shown on the example of the transport network (Section 3.2.2) when three layers will be maintained.

REFERENCES

ARAGÓ GALINDO, P., DÍAZ, L., & HUERTA, J. (2011). A Quality approach to Volunteer Geographic Information (pp. 109–114). Presented at the 7th International Symposium on Spatial Data Quality (ISSDQ 2011)

COLEMAN, D. J. (2010). Volunteered Geographic Information in Spatial Data Infrastructure: An early look at opportunities and constraints

GOODCHILD, Michael F. and HUNTER, Gary J. "A Simple Positional Accuracy Measure for Linear Features." International Journal of Geographical Information Science 11, no. 3 (1997)

HAKLAY M, 2010, "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets" Environment and Planning B: Planning and Design 37(4) 682 – 703

HECHT, R., KUNZE, C. & HAHMANN, S. (2013): Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. In: ISPRS International Journal of Geo-Information, 2, 4, 1066-1091.

HUNTER GJ (1999) New tools for handling spatial data quality: moving from academic concepts to practical reality.

CHO Woosung, LEE Ki-Joune, KIM Jun-Sung. Data Quality Control: Crowd-Sourcing Geospatial Information. HANGZHOU FORUM ON UNITED NATIONS GLOBAL GEOSPATIAL INFORMATION MANAGEMENT. 2013

JACKSON SP, et al. Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information.ISPRS International Journal of Geo-Information. 2013; 2(2):507-530.

KOUNADI Ourania, “Assessing the quality of openstreetmap data,” Msc geographical information science, Uni-versity College of London Department of Civil, Environmental And Geomatic Engineering, 2009

MOONEY P, CORCORAN P. Characteristics of Heavily Edited Objects in OpenStreetMap. Future Internet. 2012; 4(1):285-305.