

DELIVERABLE

Project Acronym: E.L.F.

Grant Agreement number: CIP 325140

Project Title: The European Location Framework

D 2.3.6: Data Quality

Revision: final

Authors:

Sonja Werhahn (BKG)

Antti Jakobsson (Maamittauslaitos)

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 and Statement of Originality

Revision	Date	Author/Editor	Organisation	Description
0.3	23/10/2014	Sonja Werhahn	BKG	draft
0.4	30/10/2014	Antti Jakobsson	Maanmittauslaitos	comments and QM part
0.5	3/11/2014	Sonja Werhahn	BKG	After comments by NDE and AHO
0.6	13/11/2014	Sonja Werhahn	BKG	After revision by WP2 and WP4
1.0a	12/01/2015	Sonja Werhahn	BKG	Resolved comments by PMG
1.0	30/01/2015	Anja Hopfstock	BKG	Final editorial changes

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This deliverable contains original unpublished work except where clearly indicated otherwise.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.

Data Quality

References

Ref	Title/Version/Publication Date/Author
[1]	D2.3 Data Maintenance and Processing specification http://www.elfproject.eu/documentation/specification/data-maintenance-and-processing
[2]	D2.9 Data Specification Regional/Global
[3]	D2.10 Data Specification Master
[4]	ISO quality standard: ISO 19157 Geographic information - Data Quality
[5]	Rule Speak http://www.rulespeak.com/en/
[6]	ERM D51_DataValidationSpecifications_V47_20130508.doc (ERM internal document)
[7]	ESDIN_ExM_DataSpecification_MediumSmallScale_v2.0 http://www.esdin.eu/sites/esdin.eu/files/D6.1a%20ExM%20data%20specification%20small%20and%20medium.pdf

Figures

Figure 1 ELF data quality approach.....	10
Figure 2 Data quality reporting process in ELF	11
Figure 3 Example of the ELF Quality Label with full ISO 19158 quality assurance	16

Table of content

1	Introduction.....	6
1.1	Scope	6
1.2	Terminology.....	6
2	Data Quality in the ELF framework.....	6
2.1	ELF organisational framework.....	6
2.2	Background.....	6
2.3	Data Quality at different Levels of Detail	7
2.3.1	Data quality requirements for Master LoD0, Master LoD1 and Master LoD2	7
2.3.2	Data quality requirements for Regional and Global LoD.....	8
2.4	Data Quality Use Cases.....	8
2.4.1	Use Case National Data Producers	8
2.4.2	Use Case Generic ELF.....	8
2.4.3	Use Case Generalisation	8
2.4.4	Use case EuroRegionalMap	9
2.4.5	Use case Insurance Risk Analysis.....	9
2.5	Metadata	9
3	Data Quality evaluation process.....	9
3.1	Reports by other geotools.....	10
3.1.1	Transformation:.....	10
3.1.2	Validation.....	11
3.1.3	Change detection.....	11
3.1.4	Edge matching	12
3.1.5	Generalisation	12
3.2	National Data Providers.....	12
3.3	Coordination body.....	12
3.3.1	Final acceptance	12
3.3.2	Expert reviews	12
4	Data quality tools.....	13
4.1	Input for tools	13
4.1.1	National Producers (NMCAs).....	13
4.1.2	Coordination body	13

4.2	Output of tools	13
4.3	Data Provider Profile	14
5	Data Quality Management	14
5.1	Organisation and Responsibilities	14
5.2	National Quality Organisation	14
5.3	Expert Reviews	15
5.4	Data Provider Quality Assurance	15
5.5	ELF Quality Label in the GPF	16
Annex A: Data Quality rules.....		17
Annex B: Discovery Metadata		17
Annex C: Metadata template for reporting national specifics.....		17

I Introduction

I.1 Scope

The primary purpose of this document is to specify the details and requirements of the data quality evaluation process and quality management within the context of the ELF project in the operational stage. It also specifies when data quality evaluation processes should be performed within the overall ELF workflow.

This document will evolve as the tools are developed.

I.2 Terminology

The ELF Glossary **Feil! Fant ikke referansekinden.** defines basic terms relevant for the ELF project. Beside this generic glossary, the following terms are used in this document:

Acronym	Term	Definition
ATS	Abstract Test Suite	
DQ	data quality	
LoD	Level of Detail	
ERM	EuroRegionalMap	EuroGeographics topographic data set at 1:250.000 scale
EGM	EuroGlobalMap	EuroGeographics topographic data set at 1:1.000.000 scale

2 Data Quality in the ELF framework

2.1 ELF organisational framework

Data quality (DQ) is one of the components of the ELF organisational framework. The general data supply chain, from national data to the ELF platform, is described in the deliverable D2.3 Data Maintenance and Processing specification **Feil! Fant ikke referansekinden..**

The basic intention is to apply the DQ evaluation concept to ELF data; that is, national data transformed to the ELF data specifications, see D2.9 Data Specification Regional/Global **Feil! Fant ikke referansekinden.** and D2.10 Data Specification Master **Feil! Fant ikke referansekinden..**

2.2 Background

The ELF data quality rules specify quality requirements based on identified user requirements, some INSPIRE consistency rules and considering what kind of data can be provided by the NMCAs. Quality requirements are described for several use cases.

The ELF quality rules concentrate on the requirements needed to achieve better harmonisation of the ELF data and identify errors in the data's topological structure as well as feature and property compliance with the ELF data specification.

The ELF data quality rules contain only automatically testable quality criteria. Data quality requirements that cannot be tested automatically will be stated in the ELF data specifications, e.g. concerning information included in the metadata files.

Therefore, quality criteria that can only be validated against reality or other reference data are considered in the responsibility of the NMCAs and out of scope for ELF.

Rules testing the data against the application schema were not considered part of the quality evaluation. The compliance to the ELF application schema is considered to be assured through the ELF transformation and is checked at the end of the ELF transformation process.

In practice, the quality rules to be checked by the quality validation tool are mainly about:

- Consistency within a theme
- Cross-theme consistency
- Respect of selection criteria
- Minimum content

Where possible the rules were assigned to the appropriate data quality element of the ISO quality standard [4].

The data quality rules were written in RuleSpeak [5], a set of practical guidelines and sentence forms for expressing business rules in clear, unambiguous and well structured English. This approach provides an implementation independent description of the data quality rules that everybody (users, producers and tool programmers) can understand and thus supports a common understanding of the rules.

2.3 Data Quality at different Levels of Detail

2.3.1 Data quality requirements for Master LoD0, Master LoD1 and Master LoD2

For Master LoD there are no ready defined data quality rules available from ESDIN or any other EuroGeographics project. In ESDIN data quality measures for large scale were stated only for Cadastral Parcels. These were integrated into the ELF data quality table.

For all other themes new data quality rules had to be created. The frame to do this was given by the following general requirements:

- For Master LoD0 and 1 there should be as little data quality rules as necessary to insure that as much data as possible can be included in the ELF platform. For Master LoD0 and 1 the defined data quality rules should be seen as guidelines for the data producer. Consistency rules should be respected, but rules about minimum content should be seen as reference to the expected data content in ELF. For these LoD the national data delivered to ELF will be heterogeneous and only very simple rules can be applied. It will have to be discussed which of the rules should be made mandatory for Master LoD0 and 1.
- Master LoD2 is the result of the generalisation process from Master LoD1. Master LoD2 is generalised from Master LoD1 and is more homogenous and harmonised. For now only the simple requirements regarding the generalisation process to Regional LoD have been considered. Generalisation parameters and data quality requirements must use the same values.

- In future (after the ELF project) it is envisaged to generalise Regional LoD from Master LoD2 (or Master LoD1). All necessary information needed for this generalisation process should be available in Master LoD2.
- Master LoD0 has only themes Buildings, Addresses and Cadastral Parcels.

2.3.2 Data quality requirements for Regional and Global LoD

Data quality rules for Regional and Global LoD are based on the EuroRegionalMap (ERM) Validation Specifications [6] and the experience gained in the ERM validation process.

In a first step the quality rules from the ERM validation specification were translated into the ELF data model. Most rules could be transformed, some had to be changed slightly due to differences in the data model and some were dropped because they are dealt with in other geoprocessing tools such as the transformation tool or are not relevant in the ELF context.

In the next step data quality requirements stated in ESDIN [7] and INSPIRE were considered and added if necessary.

2.4 Data Quality Use Cases

Data quality requirements depend on the intended use of the data. Within the ELF project the use cases National Data Producers, Generic ELF, Generalisation, EuroRegionalMap, and Insurance Risk Analysis will be considered by the data quality tool. There is an option to develop further data quality use cases in the future. For each use case different data quality requirements are necessary in order to check the usability of the data for a specific purpose.

2.4.1 Use Case National Data Producers

All rules apply to the data delivered to the ELF platform by the national data producers. Some rules are only for information but most rules return errors that should be corrected. The use case Data Producers summarises all other use cases investigated in this document. The data delivered by the data producers should be suitable for all the described use cases and should comply with all stated requirements.

Rules that are for information only apply to data at Master LoD0 and 1. These rules deliver results that can give an overview over the situation in the data set (how many features in a feature type? How many empty attributes?...). Some rules describe requirements that would be nice to have but are not essential or may not be achievable at the moment but important in the future. These rules are mostly related to minimum content, selection criteria and geometric consistency.

2.4.2 Use Case Generic ELF

The set of rules that define minimum requirements that have to be met in order to be accepted as ELF data. This means full compliance with the ELF data specification. Besides the automatically testable requirements some additional requirements have to be met concerning metadata content.

2.4.3 Use Case Generalisation

For generalisation the geometric and topological correctness of the data is important. Additionally the properties used in the generalisation process should have valid values and not be empty.

The use case Generalisation should apply only on the data that may be used to derive a generalised data set, i.e. on Master LoD0 and 1 (that may be used to get Master LoD2 by generalisation) and on

ELF Regional (that will be used to get ELF Global by generalisation).

2.4.4 Use case EuroRegionalMap

Although EuroRegionalMap will be derived from ELF Regional in future the data specifications of the two datasets are not identical. EuroRegionalMap requires some special quality rules that go beyond what is needed for ELF Regional. Requirements that go beyond the ELF specifications are due to special customer needs, technical constraints and the production process that result from the long term production of EuroRegionalMap. On the other hand ELF Regional includes more themes, feature types and properties than those included in EuroRegionalMap.

2.4.5 Use case Insurance Risk Analysis

tbd

2.5 Metadata

The content of the metadata files provided to the ELF platform are defined in the ELF Data specifications.

For the data quality evaluation process especially the information about the currency or last update, the positional accuracy and the scale of the national data used as source for the ELF data is important. This information will be shown on the GeoProductFinder to give the user an easy overview for each data set. For more details see Annex B.

National specifics that are not conform to the ELF data specifications have to be detailed in the metadata. This applies especially to missing data or properties, problems with matching national classifications to European classifications or specifics originating from the national source data. The metadata template to report these national specifics is attached as Annex C.

3 Data Quality evaluation process

The data quality evaluation process in ELF consists of several steps. In the first step the national data providers, the NMCAs, are responsible for validating their national ELF data. In the second step the Regional Coordination of ELF checks the ELF compliance of data presented at the ELF platform and provides expert reviews to the user. This process has to be rerun when new or updated data is provided to the ELF platform. As part of the ELF Quality Management expert reviews will be performed regularly to see whether the data provider meets the requirements.

ELF data quality approach

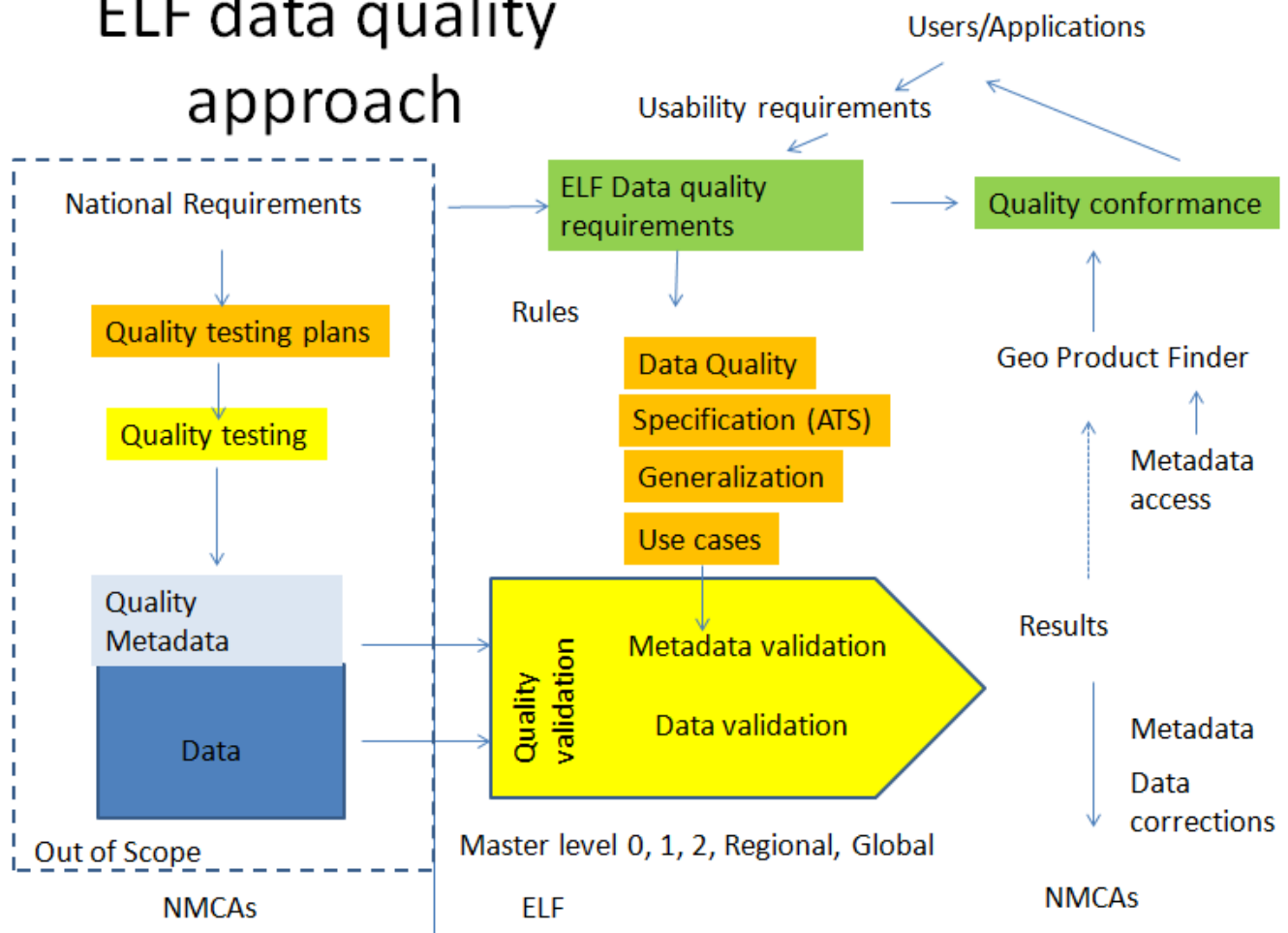


Figure 1 ELF data quality approach

3.1 Reports by other geotools

All ELF geotools create reports when data producers utilise them. These reports are used in the ELF DQ evaluation process to verify which geoprocessing steps have been applied successfully. Additionally the data producers have to state any national specifics for geoprocessing step in the metadata.

3.1.1 Transformation:

In order to be able to transform the national data to ELF data the data producers have to set up matching tables. In this step national specifics like missing information, mismatch in classifications or other known deviations from the ELF data specifications have to be stated in the metadata.

After the transformation the tools should report if the output schema is valid or if any problems occurred during the transformation.

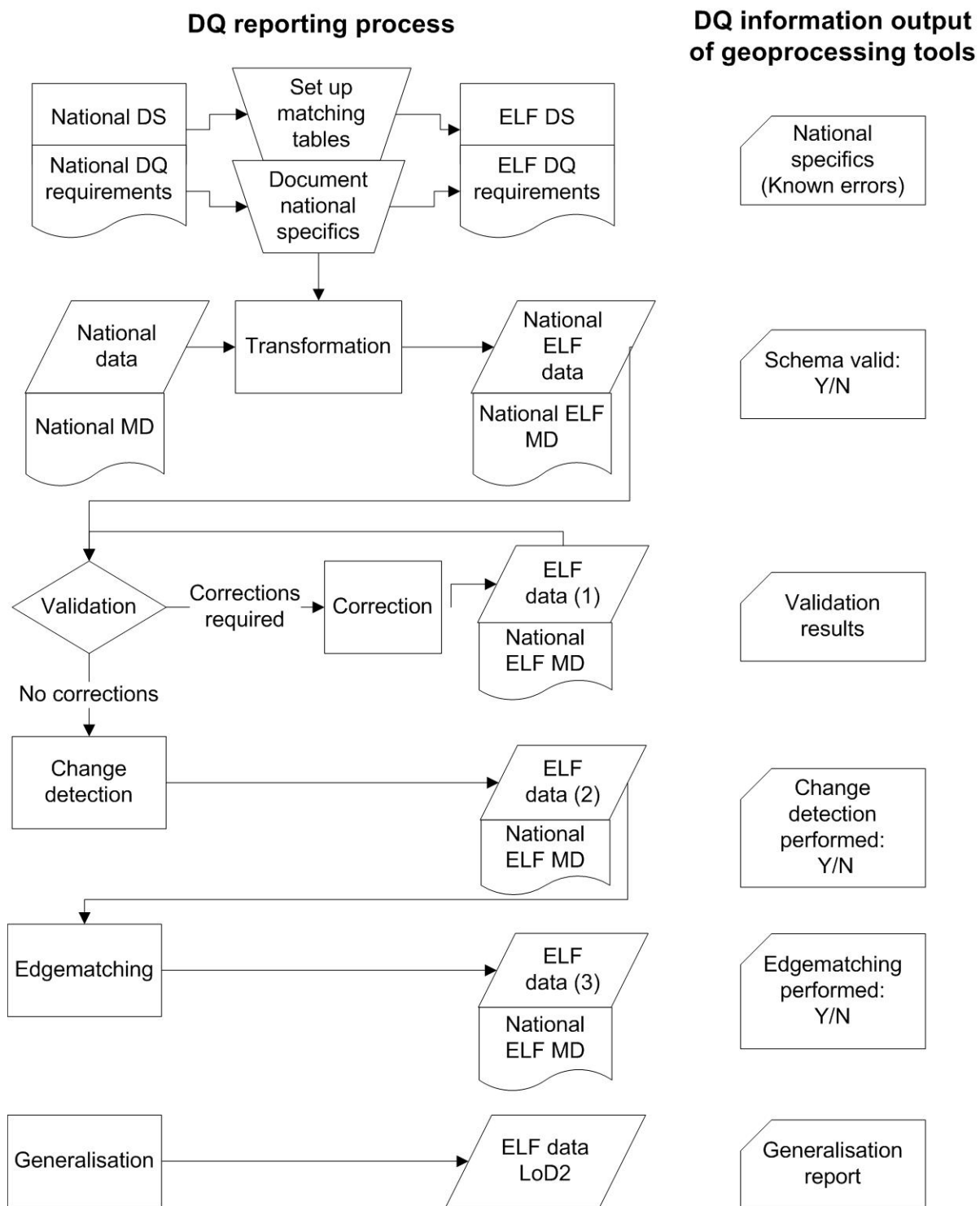


Figure 2 Data quality reporting process in ELF

3.1.2 Validation

The data validation process delivers several reports and error tables described in chapter 4.2 Output of Tools

3.1.3 Change detection

The change detection tool reports if it has been run successfully or if there were any problems. The change detection tools has to be run in case persistent identifiers and life-cycle information is not available.

3.1.4 Edge matching

The edge matching tool reports if it has been run successfully or if there were any problems. It also reports errors that could not be solved by the tool. The edge matching tools has to be run if there are neighboring countries (not necessary for islands).

3.1.5 Generalisation

The generalisation tools report if they have been run successfully or if there were any problems. Additionally the tools produce a report stating which feature types and how many features were processed. National specifics during the generalisation process (e.g. parameters changed for national specifics) have to be reported in the metadata.

3.2 National Data Providers

The NMCAs are responsible for checking their national data after the transformation to the ELF specification. Completeness and positional accuracy should have been check for the national source data; this is out of scope for ELF. Errors found by the ELF data quality evaluation tools should be corrected or identified as national specifics in the metadata. After the correction the evaluation should be rerun to verify the corrections. The error statistics tables and feature statistics tables should be sent to the Coordination Body. They will be used as source for creating data quality reports and expert reviews.

NMCAs should also state in the metadata if the change detection tool and the edge matching tool were used and send the reported statistics to the Coordination Body.

3.3 Coordination body

3.3.1 Final acceptance

In order to confirm the final acceptance of the data to the ELF platform the Coordination Body decides the methodology needed to accept NMCA data based on metadata, performed checks on test areas and visual checks. If testing is used number of test areas may vary depending on the size and topography of the countries. Systematical errors can be detected with this approach.

The structure of the coordination body is proposed in Chapter 5 consisting of several regional coordinators and the ELF Quality Manager.

Additionally the error statistics, feature statistics and metadata from the national producers are reviewed and accepted or rejected.

The Coordination Body decides if the data is accepted for the ELF platform or rejected. The Coordination Body can ask NMCAs to improve the quality of their ELF data if necessary. The acceptance of a data set can be combined with the demand to improve certain quality elements for the next update.

3.3.2 Expert reviews

Expert reviews will be written by ELF data experts nominated by the Regional Coordination. An Expert review concentrates on one special use case. It gives an overview over the relevant data quality requirements considered important for the specific use case. An Expert review evaluates the data quality from the perspective of the use case and may give guidance for the usage of the data.

Expert reviews are compiled using the result of the evaluation of the test areas, the features and error statistics tables sent by the NMCAs and the provided metadata. The expert reviews will be published in the GeoProductFinder on the ELF platform.

The detailed process is described in Chapter 5.

4 Data quality tools

4.1 Input for tools

4.1.1 National Producers (NMCAs)

The data quality tools will use the national ELF data at all LoD as input to the data quality evaluation process.

The tools will work on individual countries.

4.1.2 Coordination body

The data quality tools will use the ELF data provided by the national ELF Web Feature Services at all LoD as input to the data quality evaluation process.

The tools will work on test areas or individual countries. Additionally an aggregated data set for the whole of Europe can be used as input.

4.2 Output of tools

The structure of the output of the data quality tools are the same for NMCAs and for the Regional Coordination.

The data quality tools will create an error dataset including

- appropriate error geometries that make it easy for the data producers to find the error
- Id of the data quality rule
- Short description of the data quality rule
- Other features involved in the same error if necessary

Additionally the tools will output two statistic tables. One table will refer to the number of features and attributes in the validated data set and the other will refer to the number of errors found.

The features table should state:

- Processed country (two-character country code according to ISO3166-2char)
- Processing date
- Processing time
- Number of input features per theme, feature type and property/attribute
- Number of empty features per theme, feature type and property/attribute (empty features are features that have no value, Null value or are void for a certain property/attribute)

The errors table should state:

- Processed country (two-character country code according to ISO3166-2char)
- Processing date

- Processing time
- Number of errors per data quality rule per theme, feature type and property
- percentage of errors per data quality rule per theme, feature type and property

4.3 Data Provider Profile

The tools provide the possibility for the data provider to modify parameter values. The tools are delivered with default parameter values that can be adapted easily by the data provider to their national specificities. The acceptable parameter value range is stated in the data specifications. National specifics have to be stated in the metadata supplied to the ELF platform.

The default values present an average value for Europe and were defined based on a survey among NMCAs. These average values may not be appropriate for all themes or features in all NMCAs. In order to allow for national specificities the default parameter values can be modified inside a given value range. The default parameter values should be modified as little as possible in order to maintain the homogeneity of the output data.

5 Data Quality Management

National data processes and quality testing are not considered in the ELF. ELF Data Quality Management considers quality validation processes for ELF data at national and European level and the organisation of the Quality Management (QM).

5.1 Organisation and Responsibilities

Following roles and responsibilities are identified:

- ELF data providers are responsible for validating their ELF data and setting the ELF services. In most cases this is the NMCA but there might be cases where there are multiple data providers. In case there are multiple data providers, it should be agreed how ELF data is validated nationally.
- ELF Regional co-ordinators will be managing the expert reviews for their regional data provider's data using the ELF data quality validation tool(s).
- ELF Quality Manager will be responsible for overall quality management within ELF.
- ELF Quality Manager will be assuring the ELF regional co-ordinators using the proposed approach below.
- ELF Quality Manager will together with ELF Regional co-ordinators manage the quality assurance of data based on Data Producer Agreement.

5.2 National Quality Organisation

Each Data provider should nominate a responsible person(s) for ELF QM and data quality. These person(s) have following responsibilities

- To perform INSPIRE ATS as specified by INSPIRE

- To ensure the ELF quality evaluation will be run for the national data. As this is an automated test it should be run for each update or on regular basis.
- To report national quality evaluation results within ELF metadata
- To ensure that the organisation meets ELF ISO 19158 quality assurance as specified by the ELF Quality Manager
- To be the contact person for possible quality reclamations and give answers to the quality feedback (from the GPF)

5.3 Expert Reviews

Expert Reviews shall be performed to determine overall quality of data provided through ELF. These reviews may also be used to check how well certain user requirements have been achieved. Expert reviews are done at the national, regional and European level.

As a result of the expert review data can be classified as follow:

Rejected: Data does not meet the INSPIRE ATS and cannot be published in ELF. This test should be done by the National Data Provider and expert review is confirming this.

INSPIRE schema: Data meets INSPIRE requirements and can be published in ELF. Only schema level quality check has been performed. These checks are not provided by ELF, they have to be performed with other tools and reported in the metadata.

Generic ELF: Data meets ELF minimum requirements, ELF national validation was performed, ELF metadata is available

Generalisation: Data meet generalisation requirements

Edge Matching: edge matching tool has been run or edge matching done nationally; Data can be used for cross-border use cases

In addition there might be other ELF quality results for a specific use case.

Results will be presented using ELF Quality Label in the GPF.

5.4 Data Provider Quality Assurance

As the European quality validation will be done when national data is already ready, it is too late to correct errors at this stage. Therefore it is utmost essential that national data provider's production processes for ELF are assured by utilization of ISO 19158. Assurance is needed at the national level but also at the European level.

This assurance at the national level will consist of three steps:

- 1) Basic Level Assurance: Assuring that a process appears to be capable of creating or maintaining ELF products at the required quality. This includes identification of production processes and utilization of ELF validation or some other tools.
- 2) Operational Level Assurance: This will be achieved working together with ELF regional coordinators, checking that validation results are what is expected, training programmes have been implemented.
- 3) The Final Level Assurance: This is achieved when the data provider for ELF is capable of maintaining the quality achieved at the operational level over a period of time.

The ELF Quality Manager is responsible to see that ELF data providers have achieved the operational level of assurance and through experts reviews maintain their assurance level. If the ELF data provider has not reached the operational level or loses their assurance level even after corrective actions have been taken, appropriate management actions should be taken.

It is suggested that in ELF Data Provider Agreement there is a requirement for reaching the final level assurance. Also financial consequences not meeting the requirements should be considered.

5.5 ELF Quality Label in the GPF

The GeoProductFinder will be the place to store information on the ELF quality results. All information on the quality will be stored within ELF metadata. In addition a ELF Quality Label will show a summary of the results utilizing stars. There will be also indication whether the organisation has reached basic/operational or full ISO 19158 quality assurance.



Figure 3 Example of the ELF Quality Label with full ISO 19158 quality assurance

Annex A: Data Quality rules

Table containing data quality rules: D2.3.6_Quality_AnnexA.xls

Annex B: Discovery Metadata

Instructions for providing discovery metadata for the ELF.doc (by Kai Koistinen, NLF FI):
D2.3.6_Quality_AnnexB.xls

Annex C: Metadata template for reporting national specifics

tbd