DATA QUALITY MANAGEMENT OF REFERENCE DATASETS – PRESENT PRACTICE IN EUROPEAN NATIONAL MAPPING AGENCIES AND A PROPOSAL FOR A NEW APPROACH

Jakobsson, A.[1] and Marttinen, J.[2]

[1]National Land Survey of Finland, PO Box 84, FIN-00521 Helsinki, Finland. Tel: +358 20541 5177. Fax: +358 20541 5070. E-mail: antti.jakobsson@maanmittauslaitos.fi
[2]Geoaudit Ltd, Askaistenpolku 2B 20, FIN-00300 Helsinki Finland. Tel: +358 9 43693030, Fax: +358 9 43693031. E-mail: jorma.marttinen@geoaudit.fi

ABSTRACT

This paper introduces a *model for data quality management* that could be utilized for reference datasets. We use topographic datasets as examples, and the same principles can be applied to other reference data as well. The model is based on practical experiences developed among national mapping agencies (NMAs) in Europe, as well as standards on quality management (ISO 9000 series), geographic information (ISO 19100) and other international standards.

The model had to be cost effective, be based on international standards, provide representative and reliable quality results, have definable confidence levels and utilize differing test methods for different quality elements.

The data quality management model includes:
- Identifying the user requirements
- Developing the specifications, setting quality requirements – e.g. conformance levels
- Controlling quality during data production
- Quality inspection by the producer or the user
- Reporting quality results in metadata
- Improving the model

The paper presents a number of common practices among the NMAs. The National Land Survey of Finland (NLS) experience in data quality evaluation is explained and some improvements to the standards are discussed.

1. INTRODUCTION

The geographic information (GI) community is accustomed to using available datasets. The importance of metadata has gradually been recognized. Data quality is an important aspect of metadata (data about data). Several initiatives including Global Spatial Infrastructure [1], European efforts such as INSPIRE [2] and the National Spatial Data Infrastructures (Australia, USA) are emerging in order to increase the use of geographic information. The Open GIS Consortium (OGC) [3] has developed methods to change data between applications (e.g. web mapping initiative). The term 'interoperability' has been used to describe the process of using the same data across different applications and/or the same application using data from different sources.

Both national and local-level agreements on common spatial data content have been prepared. The content is known as 'framework' data in the United States and 'fundamental' data within the Australian Spatial Data Infrastructure (ASDI). We use the term 'reference data'. The concept is a collaborative effort to create a common source of basic geographic data, providing the most common data themes that geographic data users need, as well as an environment to support the development and use of these data.

In the 1990s, the standardization of GI began in Europe. The results were published as pre-standards in 1996. At the same time, the International Organization for Standardization (ISO) began its operations. The ISO 19100 series now has over 30 standards for GI, and the number is increasing.

1.1 National Mapping Agencies produce reference data

Data producers are currently investigating the next steps in data management. In Europe, national mapping agencies (NMAs) are the key producers of reference data within the European Spatial Data Infrastructure. For this purpose the NMAs have created EuroGeographics [4], whose mission is to represent Europe's NMAs and to co-ordinate their efforts in developing the European Geographic Information Infrastructure. At present, EuroGeographics is looking into

the possibilities of producing the reference datasets for Europe. This requires common procedures for process management, quality evaluation and data/service management. These common procedures should be based on best practices among the NMAs as well as standards and models such as ISO 19100, the OGC specifications, the EFQM and ISO 9000. These procedures would ultimately become part of the European Spatial Data Infrastructure.

Datasets that are included in reference data vary. In Europe, the EteMII project [5] has described reference data as follows:

- It is a series of datasets that everyone involved in geographic information uses to reference his/her own data as part of their work.
- It provides a common link between applications and thereby provides a mechanism for the sharing of knowledge and information.

The INSPIRE (Infrastructure for Spatial Information) project aims to make relevant, harmonized and quality geographic information available for the purpose of formulating, implementing, monitoring and evaluating EU policy-making. One of the major tasks is to determine the relevant data needed.

The working group on reference data and metadata [6] has described the main functional requirements that geographic reference data must fulfil:

- Provide an unambiguous location for user information
- Enable the merging of data from various sources
- Provide a context to allow others to better understand the information that is being presented.

The listed reference data components include: geodetic reference system, units of administration, units of property rights (parcels, buildings), addresses, selected topographic themes (hydrography, transport, height), orthoimagery and geographic names [6].

Eurogeographics organized a workshop in March 2003, in which the draft results of a survey conducted among the NMAs were presented. The results demonstrated that reference data themes are available in Europe at reasonably accurate resolution (scale of 1:10 000) and that the positional accuracy will be improved over the next 10-year period. The survey covered 191 reference datasets produced in 68 organizations in 26 European countries. A topographic database was available in 19 countries and all had topographic data elements. The majority of reference data themes were produced by a NMA [7].

2. QUALITY OF REFERENCE DATASETS

Quality is described in the ISO 9000:2000 standard as the "degree to which a set of inherent characteristics fulfils requirements". Requirement is described as a "need or expectation that is stated, generally implied or obligatory". Quality characteristic means an inherent characteristic of a product, process or system related to a requirement. Inherent means "existing in something, especially as a permanent characteristic".

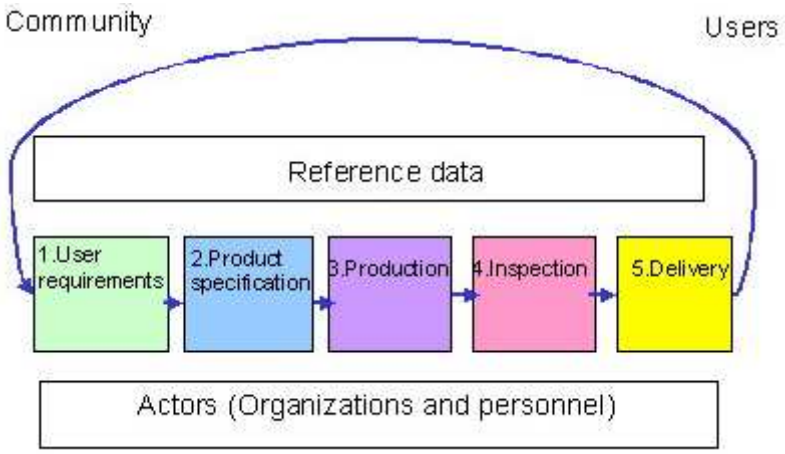

Figure 1. Process approach to quality (modified from [8])

A common process for producing geographic datasets is illustrated in Figure 1. Here we divide geographic datasets into reference datasets and datasets based on user requirements. The latter can either be based on reference datasets or they

can georeference them. The quality requirements for reference datasets should therefore be more strict. Quality management procedures are similar in both dataset types.

Process management is especially important for reference datasets because production usually requires the use of many resources. Personnel training and know-how of are vital. Producing geographic data requires human analysis and interpretation and therefore has unique characteristics compared with other types of production. The conception process is quite complex, and automation has not so far been very successful if we consider vector-based datasets. Ensuring the capability of human resources is therefore essential.

Figure 2 illustrates how different actors influence the quality of reference datasets. Two actors are identified on the *producers side*. Reference datasets are produced by subcontractors in many European countries, and the NMA then assumes the role of subscriber. The subcontractor is then the actual data producer, who tries to fulfil the agreement between the two parties. Several actors can be identified when we consider the *user side*. There might be a value-added producer, who uses a reference dataset to compile a product for the end-user (the end-user then has the role of customer). There might also be users, who don't actually pay or are otherwise obliged to use a reference dataset (end-user in the role of private citizen). We can conclude that describing the quality is dependent on the viewpoint selected (e.g. producer or user), and therefore the quality parameters that the producer uses do not necessarily fit the end-user.

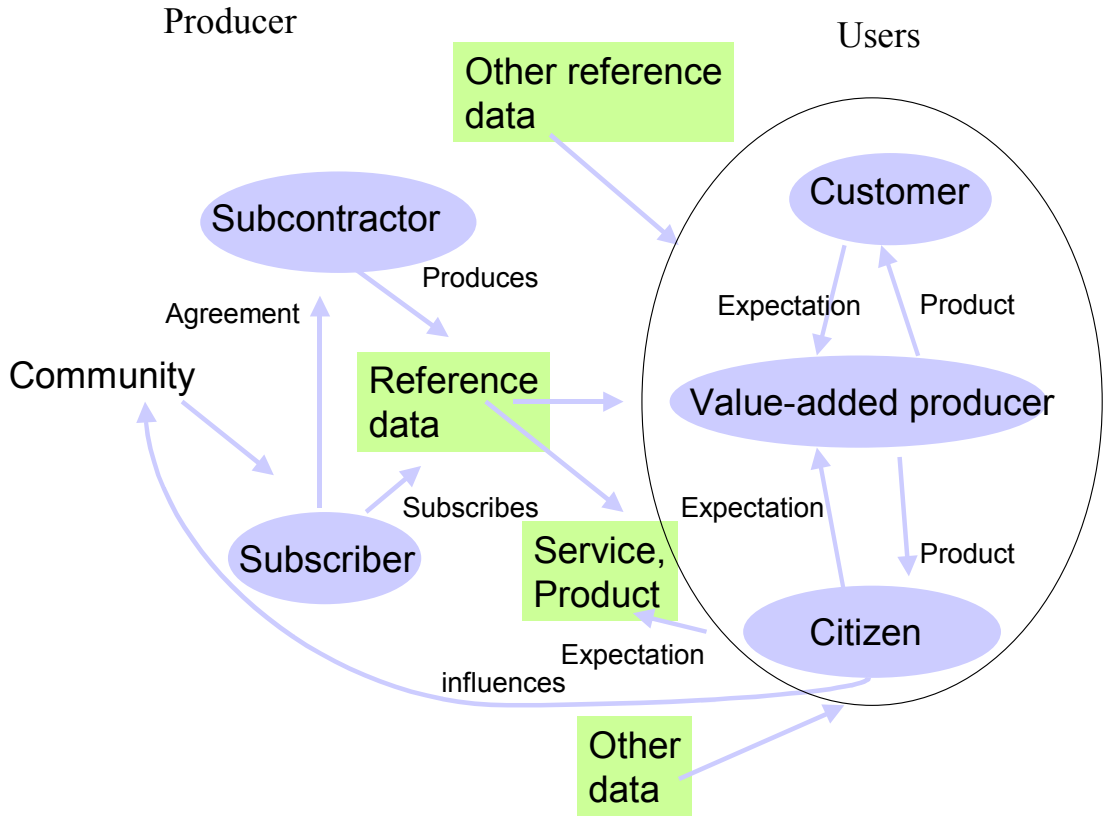

Figure 2. Actors and roles in the value-chain

2.1 Quality results

A study carried out in Finland [9] on the use and importance of geographic datasets showed that the Topographic Database (TDB) was the most important map dataset in Finland, and that users utilized it in conjunction with other datasets (thematic, orthoimagery). Though the study concentrated on map datasets and the number of respondents was quite small (n=27), it clearly illustrates the role of the TDB as a reference dataset. The reference themes were (in order of importance): hydrography, transportation, metadata, buildings, height, geographic names, cadastral information, geodetic reference, addresses, orthoimagery, depth (of waters), navigation, satellite imagery and others. The study also identified the relevance and urgency of some of the development projects. Of eight development projects listed, the top three were as follows: describing the data quality requirements, describing the content and harmonization of reference datasets, and organizing metadata services. Figure 3 illustrates the quality problems identified. A list of 16 possible quality problems was given, together with a set of reference datasets in Finland. From the list, the respondents identified problems in each of these datasets. Positional accuracy, price and currency made up in the top three.

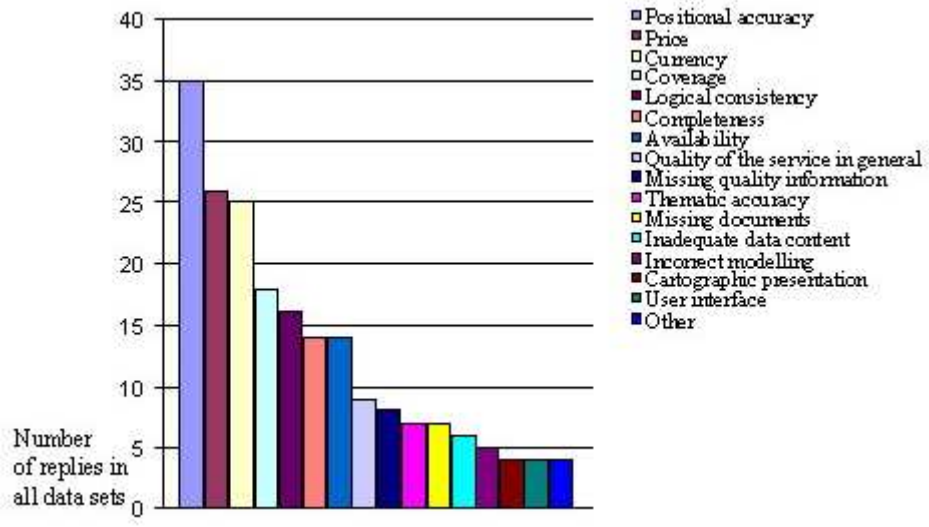

Figure 3. Quality problems in reference datasets

Eurogeographics' Expert Group on Quality has carried out several studies related to quality among the NMAs. A recent survey [10] of NMAs revealed the following needs:

- improvement in knowledge of quality evaluation standards and procedures
- better process management, especially in quality evaluation processes
- data management improvement, especially in data specification
- better metadata, e.g. quality results.

To have accurate and meaningful metadata and data quality results, the producers must implement procedures for data quality evaluation.

3. MODEL FOR DATA QUALITY MANAGEMENT

Our paper introduces a *model for data quality management* that could be utilized for reference datasets. We use topographic datasets as examples, and the principles can be applied to other reference data as well. The model is based on practical experiences developed among NMAs in Europe, as well as standards on quality management (ISO 9000), geographic information (ISO 19100), and other international standards.

The model had to be cost effective, be based on international standards, provide representative and reliable results of quality, have definable confidence levels and utilize differing test methods for different quality elements. Data quality evaluation is often neglected because it is very expensive; in fact, it can cost more than the actual production. Data.quality evaluation must therefore be cost effective. Standards must be used to ensure reliable results and to assure the users that the producer has met the requirements. If standard methods are applied, the producer can more easily have representative results and define confidence levels. Different test methods must be applied for different quality elements. For example, the test method for logical consistency can be a direct evaluation method using full inspection and, for completeness, semi-random sampling.

The data quality management model includes:

- Identifying the user requirements
- Developing the specifications, setting quality requirements – e.g. conformance levels
- Controlling quality during data production
- Quality inspection by the producer or the user
- Reporting quality results in metadata
- Improving the model

3.1 Identifying the user requirements and developing the specifications

Different users require different datasets. In the case of For reference datasets, we must take into account the fact that these should meet many user requirements. The ISO 19100 series does not cover user requirement studies, but we can employ *a human-centred design approach*. The approach is used in the ISO 13407 standard – Human-centred Design Processes for Interactive Systems – and in the ISO TR 18529 technical report.

According to this report, a human-centred design consist of five types of activity:
- planning the human-centred process
- specifying the context of use
- specifying user and organizational requirements
- producing design solutions
- evaluating designs against user requirements.

This approach was developed by the INUSE (Information Engineering Usability Support Centres) project, as part of the Telematics Applications Programme of the European Union [11], and an improved version is now published as ISO TR 18529.

ISO 19131 provides guidance on how to create data product specifications. It identifies what a product specification should contain:
- Overview
- Scope of each specification
- Identification
- Data content and structure
- Reference system information
- Data quality information
- Data capture information
- Maintenance information
- Portrayal information
- Data delivery information
- Additional information
- Metadata to be included as part of the product.

Here we concentrate on data quality information, for which ISO 19113 and ISO 19114 should be used to describe applicable data quality elements, sub-elements and quality measures. This should also include acceptable conformance quality levels.

When establishing the conformance quality levels in a data product specification, the following factor should be taken into account:
- different quality evaluation methods may be applied to different parts of the dataset (different data quality scopes)1
- for the same data quality element, different results with different confidence intervals can be achieved with different quality evaluation measures
- conformance quality levels can differ for different features in the dataset.

3.2 Controlling quality during data production

Quality control is part of an organization's quality management process, and is the most important part of the quality management of reference datasets. Commonly several tools, can be utilized during data production based on the principles of statistical process control (SPC). These include process descriptions, control charts, etc.

Quality control includes the testing of personnel before and during production. For example, a data interpretation result might be checked by an other experienced employee or some random samples might be taken. ISO 19114 and ISO 2859 might also be applicable between process steps. Process lead-time may often be several years so it is important to measure process performance during production. We will discuss this further in the following chapter.

3.3 Inspection of quality by the producer or the user

Further on, production quality must be evaluated. ISO 19114, examined in an overview by Joos [12], provides principles how quality evaluation should be organized. Quality evaluation methods can be classified into direct and indirect evaluation methods, and direct evaluation can be divided into full inspection and sampling. Logical consistency should be evaluated using full inspection, while sampling is selected for other data quality elements, based on economic reasons. This phase provides the quality results, and producers can confirm that conformance quality levels are met.

ISO 2859 (Sampling Procedures for Inspection by Attributes), which contains four parts, might be utilized. ISO 2859-1 is applicable when a continuing series of lots is submitted and two parties have agreed on a limiting quality level (AQL), which is the worst tolerable process average. Part 4 (Procedures for Assessments of Stated Quality Levels) is suitable for formal, systematic inspections such as reviews or audits. The procedures have been devised in such a way that there is only a small, limited risk of contradicting the declared quality level, when, in fact, the actual level conforms to the declared level. However, this standard should not be applied to verify a quality that has been declared for some entity, because there is a somewhat higher risk of accepting a nonconforming sample. The standard uses the term "Declared Quality Level" (DQL) and the values correspond to the AQLs in part 1. In the audit process, three different

Limiting Quality Ratio (LQR) levels can be applied depending on the desired sample size. Normally Level II should be applied. For example, if we want test completeness and we have set the DQL to 1%, then at LQR Level II the sample size should be 80 and only two nonconforming items are allowed. In this plan there is a 10% risk of failing to contradict this DQL when the actual quality level is 6.52%. If the actual quality level is 1%, then there is a 4.7% risk of falsely contradicting this correct DQL.

Selecting samples is quite problematic for geographic datasets. Geographic datasets typically cover quite large areas and if the producer wants to ensure the representativeness of the sample, then a careful planning of the sampling procedure is required. Standards normally require a random sample, which in many cases is not practical for geographic datasets. ISO 19114 offers guidance on design of a sampling procedure. However, no standard procedures exists today. The geographic database comprises several feature types and if the producer wishes to test all the attributes then the sampling procedure will be quite complex. We will discuss the experience of the National Land Survey of Finland later.

3.4 Quality results in metadata

Quality results should be reported as part of the metadata. ISO 19115 offers guidance on how to report quality results.

4. DATA QUALITY EVALUATION EXPERIENCE IN FINLAND

Topographic database production in Finland is decentralized in 13 District Survey Offices. During a decade of external quality control of decentralized production, the participants have learned many things. These lessons are to be utilized in developing the quality control process. The data quality evaluation process is described by Jakobsson [13] as well as in the Topographic Data Quality Model (QM) [14].

To achieve data quality results, sufficient items in the population should be included into the sample. Defining sampling units for geographic data is quite complex. Until now, we have used 1 km squares as sampling units. The advice of the QM is to take enough sampling units in order to get sufficient number of items for the defined sampling ratio. During the 2001 control mission, the extremes were a square with 32 nonconformities in densely populated area and a square with 4 nonconformities in forest.

While the densely populated square had 562 items to inspect the forest square had only 10 items with an AQL to control. The 4 commissions and 21 omissions of the first square gave 4.4% nonconformity in completeness and the 7 misclassified items led to 1.2% thematic accuracy. Unfortunately, the forest square did not contain the path that split the misclassified track into two separate items. The two commissions would give 20% nonconformity in completeness and the two track parts 20% nonconformity in thematic accuracy if only this square would have been used to determine the quality. The QM requires enough squares to be taken to have a sufficient number of items. This is not very economical, so to improve the sampling process, we suggest that instead of squares a constant amount of items should be used in future. One of the reasons for this is that the coverage of the TDB comprises 91% of the country, and consequently production has moved to the updating phase. This will mean that the changes in the database will not be concentrated in certain production areas but be scattered across the database instead.

The second lesson from the densely populated square deals with the process lead-time. The aerial photograph of the controlled survey area was taken in 1999. The survey was made in 2000 and the control done in autumn 2001. The majority of the 21 omissions were buildings in their finalizing phase. None of them were seen on the photographs and they probably did not exist when the field checking took place. Figure 5 illustrates a typical situation, in which a building process takes maybe several years. We have since found out that it is impossible to separate errors according to production date. It can be assumed that the customer is not interested when the data was collected. He or she wants to use it as a model of the present – not the past. In the future, as the database items receive individual dates, the data quality element *completeness* turns into a question of sub-element *temporal validity*. The redundancy between ISO 19113 data quality elements, such as completeness and temporal accuracy, is not yet clear. How to manage with them, is something that still needs more experience.

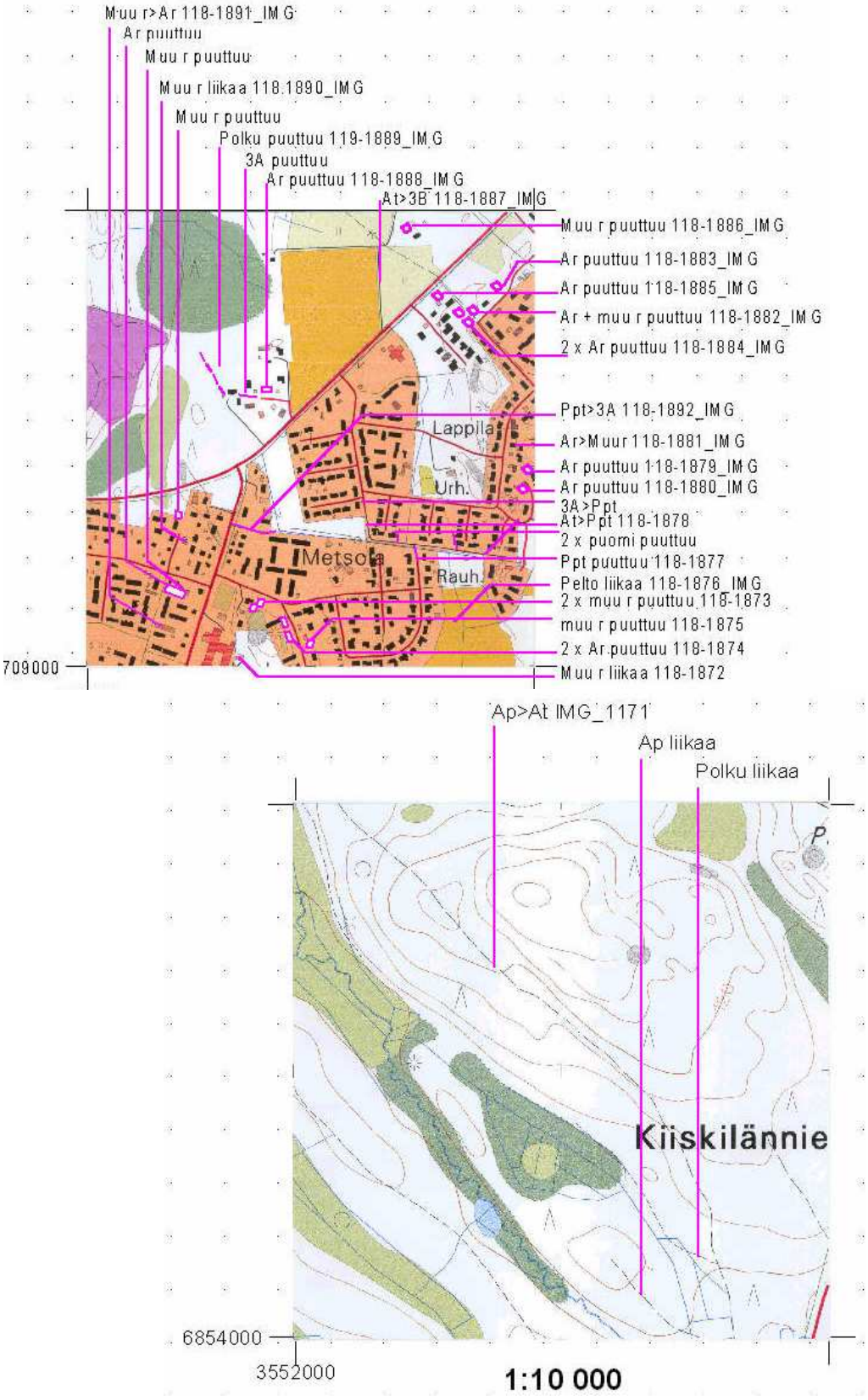

Figure 4. The two extreme cases of control squares in 2001 Quality Control Mission

Figure 5. A building with completed exterior, and insulation waiting to be installed – example of continuous change causing nonconformity to completeness or temporal validity?

5. DISCUSSION

The data quality management model we presented sets a framework. Experience from the NMAs in Europe demonstrates that we can apply standards to enable quality results for geographic datasets.

However, much more experience and research are required in the following areas:

- Common quality evaluation procedures for reference datasets
- Common data quality measures
- Harmonized data quality management procedures
- Quality reporting to users (metadata)

The main user interest is to ensure that the product fits the purpose. The question is whether or not the present approach will meet the requirements. In the ISO 19100 series, the basic assumption is that quality is declared using quality measures. However this might not be the user requirement. We don't suggest that quality measures are not required, but we see these as more in the nature of background information from the user's viewpoint. Quality measures are needed mostly when something goes wrong. For example, the producer wants to protect himself from user complaints. This brings us to conclude that producers want the product to conform to the specification. In ISO 9000, an audit means a systematic, independent and documented process for obtaining audit evidence and evaluating it objectively to determine the extent to which audit criteria are fulfilled. If producers want to claim conformity then an independent audit might be reasonable. The term "geoaudit" might be used when evaluating geographic data management processes, but still common criteria needs to be developed.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge the experts of Eurogeographics' Expert Group on Quality and the National Land Survey of Finland for financial support.

7. REFERENCES

[1] http://www.gsdi.org
[2] http://www.ec-gis.org/inspire/
[3] http://www.opengis.org
[4] http://www.eurogeographics.org
[5] T. Hancock, Reference Data White Paper (v.1.1a), ETeMII project consortium, 48 p. (2001)
[6] D. Rase, A. Björnsson M. Probert, M-F. Haupt (eds), Reference Data and Metadata Position Paper, Eurostat (2002), at http://inspire.jrc.it/reports/position_papers/inspire_rdm_pp_v4_3_en.pdf (accessed February 10th, 2003)
[7] http://www.eurogeographics.org/Projects/EuroSpec/EurSpec%20WS%201/WS%201_presentations/session%203/Reference_datasets.ppt (2003) (accessed April 23th,2003)
[8] L. Dassonville, F. Vauglin, A. Jakobsson, C. Luzet, Quality Management, Data Quality and Users, Metadata for Geographical Information, In Spatial Data Quality, (edited by Wenzhong Shi et. al.), Taylor & Francis 313 p., 202-215 (2002)

[9] Jakobsson, P. Takala (eds.), Use and importance of geographic data sets in Finland, Report of national council on geographic information (2003), in Finnish
[10] Jakobsson, F. Vauglin, Status of Data Quality in European National Mapping Agencies, Proceeding of the 20th International Cartographic Conference, Volume 4, 2875-2883, (2001)
[11] O. Daly-Jones, C. Thomas and N. Bevan, Handbook of user centred design. EC Telematics Applications Programme, Project IE 2016 INUSE, NPL Usability Services, National Physical Laboratory, Queens Road, Teddington, Middlesex, TW11 0LW, (1997) at http://www.ejeisa.com/nectar/inuse/6.2/3-3.htm, (accessed April 23rd, 2003).
[12] G.Joos, Standardization of Data Quality Measures, In Proceedings of the 2nd International Symposium on Spatial Data Quality, 205-209 (2003)
[13] National Land Survey of Finland, Topographic Data Quality Model,1995
[14] Jakobsson, Data Quality and Quality Management – Examples of Quality Evaluation Procedures and Quality Management in European National Mapping Agencies, In Spatial Data Quality, edited by Wenzhong Shi et. al., Taylor & Francis 313 p., 216-229, (2002)

DATA QUALITY MANAGEMENT OF REFERENCE DATASETS – PRESENT PRACTICE IN EUROPEAN NATIONAL MAPPING AGENCIES AND A PROPOSAL FOR A NEW APPROACH

Jakobsson, A.[1] and Marttinen, J.[2]

[1]National Land Survey of Finland, PO Box 84, FIN-00521 Helsinki, Finland. Tel: +358 20541 5177.
Fax: +358 20541 5070. E-mail: antti.jakobsson@maanmittauslaitos.fi
[2]Geoaudit Ltd, Askaistenpolku 2B 20, FIN-00300 Helsinki Finland. Tel: +358 9 43693030,
Fax: +358 9 43693031. E-mail: jorma.marttinen@geoaudit.fi

Biography

The main author was Mr Antti Jakobsson, who is employed by the National Land Survey of Finland (NLS). He is working on the development of national mapping, data quality and quality management. He is also the chairman of the EuroGeographics Expert Group on Quality. Mr Jakobsson has participated in the work groups of ISO TC 211 developing the ISO 19113 and 19114 standards. He is a postgraduate student at the Helsinki University of Technology, and is currently preparing a doctoral thesis on data quality and organizational and quality management issues regarding topographic datasets.

Mr Jorma Marttinen is the Managing Director of Geoaudit Ltd, whose main business area is independent quality evaluation for governmental agencies and municipalities among other customers. He possesses decades in the supervision of surveying and mapping projects both at the National Land Survey of Finland and abroad. He wrote Chapter 4.