Achieving business benefits from improving the quality and integrity of location-based data

A DNF White Paper

Responsibility for Content

The DNF Expert Group is responsible for the content of this document.

Change record of this document

Version	Date	Summary of change
V1.0a	February 2008	First draft
V1.0b	February 2008	Minor corrections
V1.0c	April 2008	Minor changes following Expert Group review
V1.0d	April 2008	Minor changes following Comms Group review

Content

This document consists of 13 pages

Location of this document

http://www.dnf.org/documentation/library.asp?ID=DNF0052

Approval for issue

This paper has been approved by the DNF Expert Group

This publication

1 Improving the quality and integrity of location-based data

This DNF White Paper is one of a series addressing leading issues relating to location-based or spatial data in digital form. The subject of this paper is spatial data quality and integrity and how it can be improved by adopting DNF principles. Inaccurate, inconsistent, incomplete or out-of-date data impacts on the services or processes that use the data and leads to increased costs, inefficiencies and user dissatisfaction. It can lead to a failure to deliver on projects and programmes.

The purpose of this paper is to:

- Explain what is meant by data quality and integrity in the context of location-based data;
- Highlight the key topics and challenges for users of location-based data;
- Say how DNF may be able to help users to improve the quality and integrity of their data..

Readers unfamiliar with the Digital National Framework and the terminology used are recommended to go to the DNF website at http://www.dnf.org

2 Data quality and integrity

Data quality has been defined in a number of ways. Most simply, it is defined as, "*fitness for purpose'*. In other words, it is a measure of the degree to which the data meets the needs of the particular application. A more precise way of putting this is, "*performance against specification*" i.e. how closely does the data fit to the specified requirements for the job. A more formal definition used in international standards is, "*the totality of characteristics of a product that bear on its ability to satisfy stated and implied needs*"[1]

The important thing to note about these definitions is that quality is a relative term not an absolute; it is the relationship between the properties of the data, the purpose for which it is being used and the degree to which the requirements (whether explicitly stated or implied) are being met. Statements such as , "this is quality data", or, "my data is 100%" are meaningless. Users need to define exactly what their quality criteria are and state how they are to be evaluated.

[1] International Organization for Standardization, ISO 8402: 1994 Quality management and quality assurance - Vocabulary

There will always be two views of data quality, the view of the data producer and that of the user. These are illustrated in Figure 1 which shows that the data producer and users' purposes and requirements will not be coincident (unless of course they are one and the same). Therefore it is going to be important for the user to understand the nature and quality of the data that they are proposing to use in their application. Understanding the context in which data is being used is basic to understanding quality.

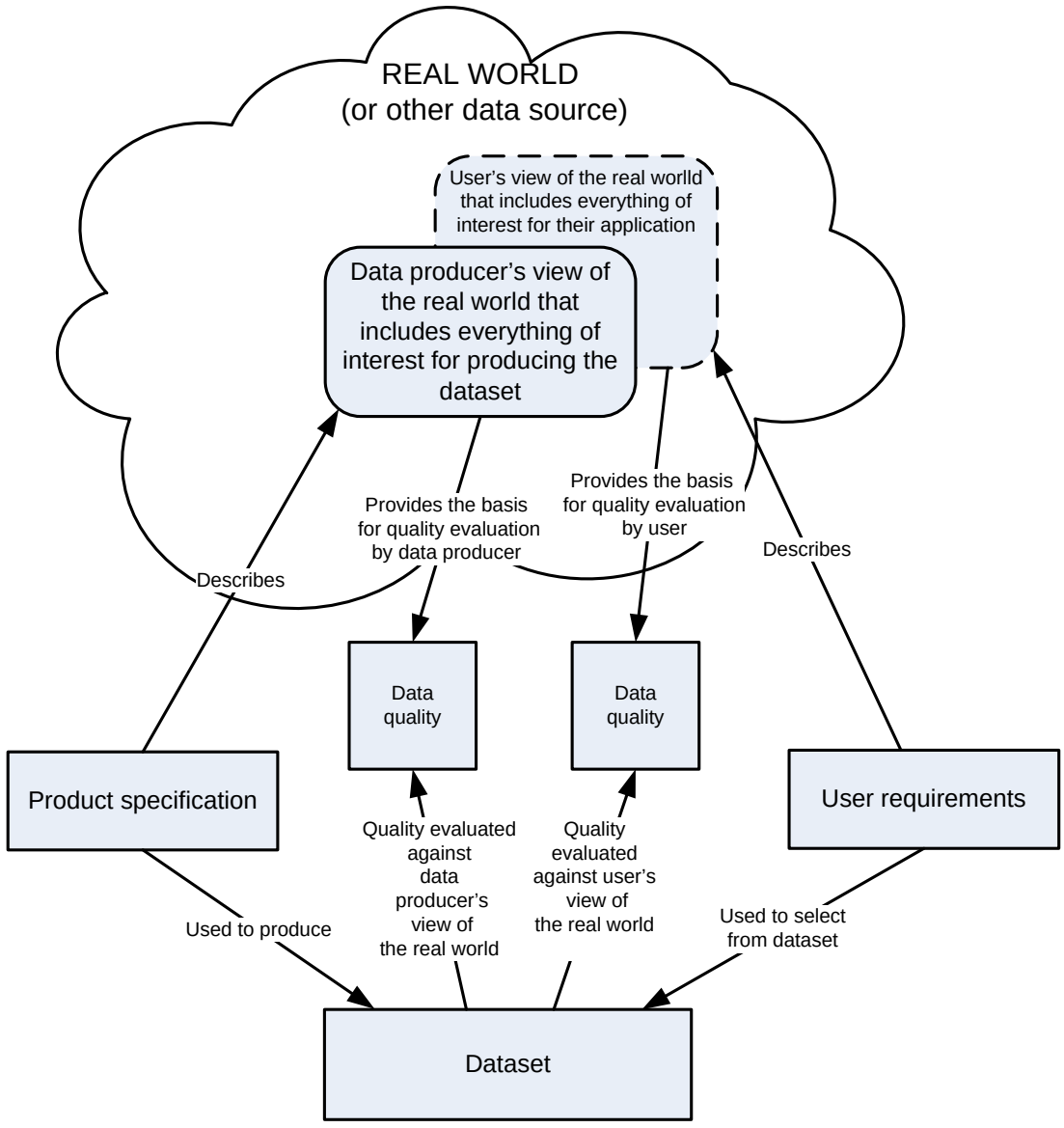

Figure 1: Data producers' and users' views of quality[2]

[2] Adapted in part from ISO 19113: 2002 Geographic information – Quality principles

For location-based data we can characterise data quality both descriptively and quantitatively. [3]

- **Descriptions** e.g. purpose, usage and lineage. These are non quantitative and tell potential users of the data why the data was captured, how it was created and subsequently modified or maintained and how it has been used They can give a useful indication of the suitability of a dataset for a particular purpose.

- **Data quality elements** and *sub-elements*. These are capable of measurement and can yield quantitative results:

  - Positional accuracy – this can be *absolute accuracy* - closeness of coordinate values to values accepted as being true or *relative accuracy* - closeness of the relative positions of features in a dataset to the relative positions accepted as being true;

  - Temporal accuracy - accuracy of time measurement. This can include *temporal consistency* - correctness of ordered events or sequences and *temporal validity* - the validity of the date assigned to a data item;

  - Thematic accuracy – accuracy of the attribution of the data. This can include *classification correctness* – comparison of the classes or attributes assigned to data items to the real world or other sources and *non-quantitative* and *quantitative attribute correctness*;

  - Completeness – this is either excess or missing data i.e. *commission* or *omission* when compared to the data source at the time of capture;

  - Logical consistency – this can include *conceptual consistency* - conformance to the data model or schema, *domain consistency* - adherence of values to the value domains, *format consistency* - degree to which data accords with the physical structure of the dataset and *topological consistency* – degree to which the geometry is correctly structured topologically.

Data integrity which is concerned with data accuracy, completeness and validity and its preservation during storage and transfer in an unaltered state overlaps in terms of descriptive usage with data quality.

---

[3] This broadly follows the approach in ISO 19113: 2002 Geographic information – Quality principles

Where an application draws on a number of datasets, whether from one organisation or from several, then the quality and integrity issues become more complex because of the interactions between the qualities of the two datasets. For example, their positional accuracy or completeness is likely to be different. Other factors also come into play such as the way that real world objects are abstracted and represented in the respective datasets e.g. streets may be abstracted as complete streets and represented by start and end coordinates in a gazetteer but as street segments and centre-lines in a network.

3 Leading challenges and issues

Common problems with the quality and integrity of location-based data include:

- **No quality statement** – the user has no information, whether in free text or metadata, about when or how the data producer created the data, why they created it or how accurate or complete it is.

- **Inconsistencies in reporting data** – quality statements do not conform to a particular standard such as that in ISO 19113 and ISO 19114[4].

- **Lack of quality evaluation and testing** – a quality statement should set out methods of quality evaluation and quality levels which are used to pass or fail data e.g. 95% of all addresses existing in the real world on a stated date are included in the data or 90% of real world change is included in data within 6 months.

- **Lack of data specification or feature catalogue** – the user does not know the specification used to capture and maintain the data – maybe the data producer has never produced a formal specification. Very commonly there is no catalogue with definitions of the types of feature or classes of geographic object contained within the data.

- **No statement of requirements** – the user has not detailed their requirements, or is unable to adequately define their requirements before acquiring the data so that even given a data specification and quality statement, they cannot really assess whether the data is fit for their purposes.

- **Lack of currency** – the data does not reflect the real world as it is today but as it was a number of years ago. This leads to errors of commission or omission e.g. buildings no longer present in the real world because they have been demolished or buildings now present because they were built after data capture. A dataset may not be consistent to one date in the past because it is subject to continuous revision albeit that only some classes of geographic object are regularly maintained.

[4] ISO 19114: 2005 Geographic information — Quality evaluation procedures

- **"Poor fit" across different datasets** – the data is captured to a different reference frame to other data or uses a more generalised model of the real world.

Inaccurate, inconsistent or incomplete location-based data results in:

- **Inaccurate, inconsistent, incomplete and misleading information** – the information can be no more reliable than the data from which it is derived meaning that analysis and decisions based on those data may be incorrect.

- **Lack of referential integrity in cross-referencing** of business and geographic objects in an application leading to inconsistent results e.g. the failure to maintain the cross-referencing of addresses to a street following a change of street name could result in a breakdown in business processes.

- **Problems with data sharing and interoperability** because of a lack of format consistency, errors of omission or positional inaccuracy are not apparent until data is combined, leading to problems only emerging at critical times such as during emergency response.

- **Inefficiencies in operations** because of missing, inaccurate or out-of-date data could mean extensive manual auditing and correction or having to perform tasks manually because they cannot be automated.

- **Costs resulting from invalid or incorrect results** which could be due to the extra effort required to perform a task or fines and litigation if the problems mean that organisations were not able to fulfil their obligations.

A further challenge will be the implementation of INSPIRE in the UK which will place an increasing demand on a number of organisations and agencies to improve the quality of the data that they produce[5]. (For a fuller discussion see the White Paper on INSPIRE[6].)

4 How can DNF help?

DNF can help overcome some of the challenges outlined above by providing a framework and approach for the linking of datasets to a common reference base thus avoiding some of the data quality issues described above. The benefits can be two-way: implementing DNF principles helps to improve data quality and improving data quality helps with the implementation of DNF.

[5] See http://www.ec-gis.org/inspire/

[6] Digital National Framework (2008) White Paper - Implications of the INSPIRE Directive,

DNF is developing a comprehensive set of technical guides and codes of good practice plus some supporting tools and services to enable users and user communities to share data and interoperate.

The documentation and services being developed by DNF are illustrated on the Roadmap shown in Figure 2. Existing documentation and services are available on the DNF website.

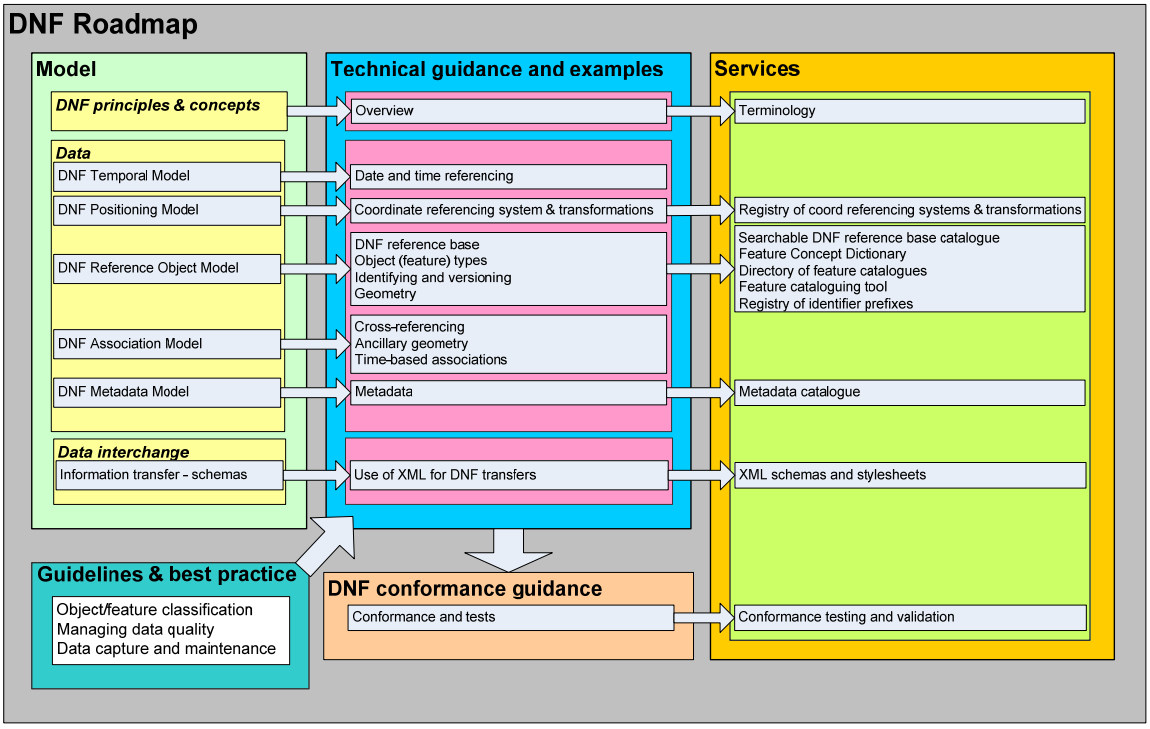

Figure 2: DNF Roadmap

For enterprise level spatial data quality management there is extensive expertise and a range of solutions available from organisations represented on the DNF Expert Group – see the DNF website.

The main benefits of adopting the DNF approach will come from:

- Linking different views of the world to a common reference base by cross-referencing using unique identifiers;
- Maintaining referential integrity between base reference objects and application reference objects. This enables cross referencing and also the identifying and fixing of geometric errors within a dataset;

- Sharing geometry between different application layers thus eliminating discrepancies, inconsistencies and inaccuracies;

- Using changes in one reference layer to provide change intelligence in another. This can be used to maintain the references and assess the impact of these changes.

The benefits can be two-way: implementing DNF helps to improve data quality and improving data quality helps with the implementation of DNF.

There are a number of DNF case studies that illustrate the points above, most notably:

- **Building a single-geometry land and property database at Dudley Metropolitan Borough Council using the Digital National Framework** (available at: http://www.dnf.org/Case%20Studies/Dudley%20DNF%20casestudy%20lores.pdf). The Dudley case study describes how Dudley Metropolitan Borough Council (DMBC) in partnership with Assist Applications Ltd was able to share and integrate land and property data held in disparate application databases. This required that (i) the information within each database was interoperable, (ii) there was a common standard for data accuracy and (iii) a common source could be used to facilitate cross-database comparisons, correlations and integration. Geometry in each application was associated to a single geometry source by converting existing polygon datasets to associate with OS MasterMap® Topography Layer. In the process, positional accuracy issues were eliminated. Further, because each of the applications carries cross-references to the OS MasterMap Topographic Identifiers (TOID®) and relies on OS MasterMap for its geometry, if the latter moves then, in effect, the boundaries of each of the geographic objects in the applications moves as well.

- **Achieving Interoperability at Staffordshire County Council using the Digital National Framework** (available at: http://www.dnf.org/Case%20Studies/Stafffordshire%20casestudy%20final.pdf) This case study describes how, using DNF principles, Staffordshire County Council worked with 1Spatial to prove and achieve an open interoperable solution to data management. By sharing regional mapping data across the organisation, a single repository was created for business and spatial data to be stored together, which met Staffordshire County Council's business requirement for improved spatial data access and sharing. The benefits of this have included a reference base using OS MasterMap linked to other data layers e.g. education, property, development via TOIDs thus providing reusable spatial data to a consistent quality standard.

- **Implementing the Countryside and Rights of Way Act 2000 using the Digital National Framework** (available at: http://www.dnf.org/Case%20Studies/DNF%20Countryside%20casestudy%20WEB.pdf) This case study describes how by re-using and cross-referencing the Topographic Layer of OS MasterMap it was possible for Black and Veatch (contracted to the Countryside Agency - now part of Natural England) to turn a project from one of data capture to one of information collection and attribution predominantly. The process of cross-referencing and land parcel aggregation ensured the information integrity at all times.

Below is a table summarising how DNF may help with particular quality elements.

Table 1: Table showing how DNF can help with specific aspects of data quality.

Quality elements ***Sub-elements***	***How DNF can help***
Positional accuracy	
Absolute accuracy	By defining a consistent reference base built on detailed national survey (whether over land or sea) which uses a recognised and fully documented coordinate reference system capable of transformation to other recognised reference systems By offering the possibility of using the geometry of the base reference layer thus removing duplication and inaccuracies
Relative accuracy	As above
Temporal accuracy	
Temporal consistency	By providing a framework within which datasets can be cross-referenced and hence allow time inconsistencies to be detected and rectified
Temporal validity	By providing a framework within which datasets can be cross-referenced and creating the potential for dates of capture or change to be cross-checked

Thematic accuracy	
Classification correctness	By providing a feature cataloguing tool (available on the DNF website) thus enabling definitions of geographic objects in scope to be improved and with it their correct classification
Non-quantitative attribute correctness	By providing for cross-referencing to other reference layers and the reference base and creating the possibility of corroborating attribute values
Quantitative attribute correctness	As above
Completeness	
Commission and omission	By providing for cross-referencing to other reference layers and the reference base, errors of omission or commission may be detected and corrected By providing intelligence of change through changes to the base reference layer
Logical consistency	
Conceptual consistency	By aiding the development of data specifications and providing the DNF feature cataloguing tool
Domain consistency	Through software validation – not specific to DNF
Format consistency	By publishing DNF schemas
Topological consistency	By using the topological integrity of the reference base where feasible and appropriate. Maintaining cross-references enables this consistency to be maintained.

5 Conclusions

A failure to deal with quality and integrity in location-based data can lead to increased costs, inefficiency and user dissatisfaction. Quality is not an absolute but many aspects of quality can be measured and tested. Data producers and data users have different views and often different requirements for data quality.

DNF is all about data sharing and interoperability. This can bring about quality improvement by highlighting inaccuracies and inconsistencies between different datasets. Documentation in the form of technical guides and codes of best practice and tools and services are being developed to assist users in adopting DNF principles and, in the process, take the opportunity to improve data quality.

6 References

Publications

1994	International Organization for Standardization,	ISO 8402: 1994 Quality management and quality assurance - Vocabulary
2002	International Organization for Standardization	ISO 19113: 2002 Geographic information – Quality principles
2005	International Organization for Standardization	ISO 19114: 2005 Geographic information — Quality evaluation procedures
2008	Digital National Framework	White Paper - Implications of the INSPIRE Directive
2005	Digital National Framework	Case Study - Building a single-geometry land and property database at Dudley Metropolitan Borough Council using the Digital National Framework
2007	Digital National Framework	Case Study - Achieving Interoperability at Staffordshire County Council using the Digital National Framework
2006	Digital National Framework	Case Study - Implementing the Countryside and Rights of Way Act 2000 using the Digital National Framework

Web References

Year,	Reference	Web Address
2008	Digital National	www.dnf.org

	Framework	
2008	Digital National Framework Registry	http://www.dnf.org/Pages/registry/
2008	DNF Demonstrator	http://www.mgeomatics.com/DNFDemo
2008	INSPIRE website	http://www.ec-gis.org/inspire/