Lineage - what should be covered in the lineage statement?

This document provides further guidelines relating to resource Lineage based on a review of INSPIRE, ISO19115 and various other sources. The guidelines are followed by two real-world examples for spatial data resources provided by Terence O’ Rourke and Jochen Roller.

Data lineage is a cornerstone of metadata management. In the Lineage field, the metadata author should supply their best knowledge of who, why, when and how the data resource came into existence. Include statements on the following: source material, processes used to create the data, including resolution of measurement, the method of updating and any quality control processes applied. Knowing the original source and understanding what exactly happens to the data as it flows to another product such as a report helps to boost confidence in the results. Data transformation, migration, and replication imply dependency among different manifestations of the same original data. Knowledge of the lineage of a resource helps to clarify the implications of changes to datasets over time.

INSPIRE considers lineage to be part of resource data quality. The INSPIRE Metadata Implementing Rules v1.2 (16/06/2010) states: “Apart from describing the process history, the overall quality of the dataset (series) should be included in the *Lineage*metadata element. This statement should contain any quality information required for interoperability and/or valuable for use and evaluation of the data set (series)”. Please consider the following headings when providing details of the lineage of a resource.

1. The Source of the data

Describe where the data originated and what methods were used to capture the data.

1. The Process History

Describe the life cycle, from collection and acquisition through compilation and derivation to its current form. Information about any events or transformations in the life of a dataset should also be supplied, where appropriate.

1. The Overall Quality of the dataset (series)

Where appropriate, lineage may include a statement of whether the data set has been validated or quality assured, whether it is the official version (if multiple versions exist), and whether it has legal validity. Provide information on topological gaps, overlaps, edge-matching and positional accuracy for spatial data sets.

1. Provide a summary of the data structure of the resource

Identify key details of the format of the resource. Where applicable, indicate the data attributes, business rules (logic) and any source to target mappings.

1. Technical conditions for delivery

Supply the name and version details of the primary software required to access the resource content. Indicate any special requirements for setup and use.

## **Sample Shape File Resource Lineage statements**

Example 1: Spatial data related to the Inventory of Irish Coastal Lagoons (Terence O’Rourke)

The National Lagoon Dataset 2010 was based on the lagoons identified in the Inventory of Coastal Lagoons 2007. Included in the inventory were sketch maps for each lagoon. However, due the scale and lack of background mapping these sketch maps could only be used as a reference tool in the creation of this spatial dataset.

The OSi 1:5000 vector dataset was used as the basis for the creation of the National Lagoon Dataset 2010. The polylines from the OSi 1:5000 dataset, which corresponded to the extent of the lagoons were copied into the new lagoon dataset. In cases where the OSi 1:5000 vector dataset did not contain any spatial data for lagoons identified in the inventory, the OSi 2005 orthophotographs (1:40000) were used as the base layer to manually digitise the boundary of these lagoons. It should be noted that expert knowledge was used to correct instances where what was indicated on the OSi 1:5000 dataset did not match what was apparent on the ground from the orthophotographs.

The source datum is the Irish National Grid.

Example 2: Polygon shapefile resource related to the Coastal Monitoring Project 2004-2006 (Jochen Roller)

Habitat types are assigned to surveyed areas, using both EU Habitats Directive Annex 1 codes, as well as habitat codes established for the purpose of the project. Information on site names, areas and other (see fields listed below) is given.

ESRI ArcView 3.2  was used in the creation of the original file.  Aerial photographs of the sites from the most recent complete, or near complete, set (the 2000 Ortho-photo set) were printed and used in the field as an aid to mapping. A GeoExplorer handheld GPS unit (Trimble GeoXT) was used for recording the locations of boundary lines and areas.  Additional observations recorded on the printed aerial photographs and in target notes were used in the creation of this file. Field notes relating to various aspects of, among others, topography and vegetation composition were written at all sites.

The file was subsequently checked for topology and consistency within NPWS. During this process a number of amendments,  including the addition of site codes and removing overlaps, gaps, slivers and multipart polygons, as well as running the repair geometry tool using ArcGIS9.3, were implemented.  For this purpose, the paper copies of site maps submitted as part of the project were used for clarification in certain areas.  Although the present file is cleaned of topology errors, it must be noted that due to preceding processing steps (separate incongruent files have apparently been merged / clipped in the creation of the file submitted to NPWS) this file is not free from errors, especially when examined on a large scale.

The file's attribute table contains the following fields:

SITE\_CODE:

Type: Text

Length: 12

Description:  Coastal Monitoring Project site codes

CMP\_SITE:

Type: Text

Length: 31

Description:  Coastal Monitoring Project site name

COUNTY:

Type: Text

Length: 9

Description: Name of coastal county in which site occurs

NPWS\_ID:

Type: Text

Length: 12

Description: Site code of SAC in which site occurs

NPWS\_SITE:

Type: Text

Length: 44

Description: Site name of NHA or SAC in which site occurs

CMP\_HABIT:

Type: Double

Length:

Description: Coastal Monitoring Project habitat code

NPWS\_DESIG:

Type: Text

Length: 14

Description: NPWS Designation type (NHA or SAC where applicable) of area in which site occurs

SURVEYNAME:

Type: Text

Length: 23

Description: 'Coastal Monitoring proj' entered for all polygons

SURVEYDATE:

Type: Text

Length: 33

Description: Date on which field survey was carried out

SURVEY\_MET:

Type: Text

Length: 13

Description: Survey method applied (ground or desk)

CMP\_NAME:

Type: Text

Length: 30

Description: name of habitat type according to Coastal Monitoring Project habitat coding (in field CMP\_HABIT)

ANNEX\_I:

Type: Text

Length: 22

Description: marks areas containing EU Habitats Directive Annex 1habitats

AI\_SALTMAR:

Type: Text

Length: 12

Description: marks areas containing EU Habitats Directive Annex 1saltmarsh habitats

EU\_CODE:

Type: Text

Length: 10

Description: EU Habitats Directive Annex 1code

SHAPE\_LENG:

Type: Double

Length:

Description: Length of polygon boundary in metres

SHAPE\_AREA:

Type: Double

Length:

Description: Area of polygon in square metres