

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Washington, D.C., United States, 28-30 November 2000)

Topic (i): Statistical metadata for dissemination

**DEVELOPING A CROSS-GOVERNMENT METADATA STANDARD
FOR THE UNITED KINGDOM**

Submitted by Office for National Statistics, United Kingdom ¹

Contributed paper

SUMMARY

The introduction to this paper describes the importance attached to metadata as an information discovery tool within the UK Government, and the development of a national standard for discovery metadata to underpin the UK Government's 'Modernising Government' and 'e-Government' initiatives. The paper goes on to describe the pioneering work done by the Office for National Statistics to disseminate statistical metadata. The paper also illustrates how the ONS's system underpins the other two main 'UK standards' for discovery metadata and compares all three with the main international standard known as the Dublin Core. The paper also looks at the 'entities' or 'objects' described by each system, and the metadata fields (and semantics) used to describe those entities or objects. The paper concludes by presenting an amalgam of all four of these separate standards as a first-cut proposal for a single UK government-wide mandatory standard for resource discovery metadata, and also suggests a few prerequisites for its successful implementation.

I. INTRODUCTION

1. One of the key aims of the UK Government's Modernising Government initiative is to make the UK a world leader in the Information Revolution. This overall strategy embraces several strands of work such as the complementary Information Age Government and e-Government initiatives, each of which has a number of key components. From a government statistician's point of view, the most significant features of these initiatives are the following major imperatives:

- ◆ The creation of a national communication infrastructure which will allow every citizen to interface with Government through the Internet;
- ◆ A top-down drive to make all government services available and deliverable over the Internet by 2005;
- ◆ A shift away from silo-based, producer-centric government services to more 'joined-up' and customer-centric government services tailored to meet the needs of citizens and businesses.

2. The provision of comprehensive metadata about all government services and, in particular, all government resources (i.e. information about government data) is seen as a key process within the e-Government initiative. Metadata provide a vital communication bridge between Government and the

¹ Prepared by James Denman.

citizen by fostering awareness and understanding, and by turning raw data into information. Metadata also fuel the Information Economy by opening up the vast treasure trove of information held by government departments for commercial exploitation by the private sector. This particularly true of government statistics.

3. All of these wider government initiatives have a particular relevance for UK government statisticians and have given rise to the following statistical initiatives:

- ◆ the development of internet-based data collection systems;
- ◆ a shift in emphasis away from static paper-based products (push) and towards more dynamic electronic dissemination media (pull);
- ◆ the generation of statistics geared more closely to the needs of the citizen e.g. statistics which are disaggregated to lower levels of geographic disaggregation;
- ◆ the development of more harmonised, cross-cutting statistical analyses;
- ◆ greater electronic interoperability between different statistical systems;
- ◆ more customised advisory and delivery services.

4. The drive to spread the metadata gospel throughout all Government Departments is being spearheaded by the Information Age Government (IAG) Metadata Working Group (MWG) chaired by the Cabinet Office which is drafting a Metadata Framework to accompany a Government-wide metadata standard. The work of this standards group is complemented by another body, the e-Government Interoperability Framework (e-GIF) working group, which is taking forward an international drive towards more technical metadata standards, and greater integration of national information resources. The e-GIF mandates all UK public sector organisations, including local authorities and the health service to use Internet standards and communication technologies and transfer protocols such as the eXtensible Mark-Up Language (XML), and XSL (eXtensible Stylesheet Language), and is also exploring complementary systems such as Z39.50, and information structures such as the Resource Description Framework (RDF). Another body is developing a common search vocabulary and Pan-Government Thesaurus (PGT).

5. The overarching metadata standard being drafted by these pan-government bodies is embodied in the Government Metadata Element Set (GMES) and the accompanying Metadata Element Content Rules (MECR). These map very closely to three existing metadata standards which are already well-established within the UK, two of which have been pioneered by Statisticians. The three systems are:-

- ◆ StatBase – StatSearch (administered by the Office for National Statistics);
- ◆ the National Geographic Data Framework (NGDF) (administered by the Ordnance Survey);
- ◆ the Information Asset Register (IAR) (administered by the Cabinet Office).

6. Furthermore, each of these three metadata systems maps very closely to the international standard known as the Dublin Core (<http://www.purl.org/dc/>) and their degree of complementarity is shown in Annex A attached.

I.1 StatBase (<http://www.statistics.gov.uk>)

7. The StatBase system which is sponsored by the thirty or more Government Departments that make up the Government Statistical Service (GSS), was launched in 1998 and is managed by the Office for National Statistics. StatBase is only concerned with the documentation of statistical resources and the dissemination of linked statistical data. The metadata component of StatBase is incorporated in the StatSearch system and consists of two types of metadata – Discovery metadata and Evaluation metadata. To date, no work has been done to extend the system to incorporate Micro-metadata. The metadata ‘elements’ or ‘fields’ which make up StatBase represent an amalgam of the metadata fields most

commonly recommended by the major international organisations including the ISO, UN, EU, IMF, as well as those most commonly used by a range of other Government Statistical Organisations.

8. Discovery Metadata - the discovery metadata elements within StatBase are spread across several templates:

- ◆ one metadata template of 47 metadata fields covers ‘Sources’ (i.e. any means of gathering data such as the quarterly Labour Force Survey (LFS) or the decennial Population Census);
- ◆ a second template of 32 metadata fields covers ‘Analyses’ (any value-added analysis of raw data gathered via any ‘Source’ e.g the International Labour Organisation definition of unemployment derived from the LFS, the monthly Retail Price Index, or the National Accounts system);
- ◆ a third template of 28 fields covers ‘Products’ (e.g any Database, Publication, CD-Rom, Enquiry Service, etc.);
- ◆ a fourth template of 11 metadata fields covers any associated ‘Contacts’- (i.e. the human beings responsible for each of the above).

9. Of all these StatBase metadata fields, a total 41 relate to the Dublin Core and these are shown in Annex A. The full StatBase metadata specification can be found in the ONS document entitled ‘StatBase – Metadata Assistant – A Guide to Populating Metadata Screens’ which is available from the ONS.

10. Evaluation Metadata - StatBase also includes a more extensive template incorporating a set of metadata fields covering any ‘Source’ defined as a sample survey. These extra fields cover such matters as ‘sample framework’, ‘sample size’ and ‘sampling methodology’ and are designed to provide users with sufficient information to allow them to judge whether a particular data source meets their requirements for coverage, disaggregation, precision, etc.

11. Thematic Interface - these templates currently identify over 1,000 ‘Sources’, a few hundred ‘Analyses’, and nearly 2,000 ‘Products’ all of which are linked to an hierarchical Search Directory. This Directory is based on thirteen high-level Themes, several hundred Subjects within Themes and a few thousand Topics within Subjects. Many of the templates are also linked to the relevant data. Planned developments include a geographic interface which will allow users to search for information and data relating to a particular geographic area.

12. Synergies - the ‘Discovery metadata’ contents of StatBase also serve to furnish the ONS’s contribution to two other major national metadata initiatives - the NGDF and IAR metadata gateways described below.

I.2 (UK) National Geographic Data Framework (NGDF) (<http://www://askgiraffe.org.uk>)

13. As its name implies, the NGDF system which is managed by the Government’s Ordnance Survey and sponsored, amongst others, by the UK’s Intra-Governmental Group on Geographic Information (IGGI) is concerned with cataloguing any resource which can be referenced to the earth’s surface – which, in practice, means almost everything. The NGDF have established an Internet-accessible Gateway which provides a central point of access to a wide variety of metadata covering commercial as well as governmental information resources. To facilitate linking the NGDF have produced a set of guidelines for documenting data resources. These guidelines are based on the draft ISO Metadata Standard 15046-15 and the intention is to make the guidelines a profile of the ISO standard in time. There are a total of 42 metadata fields in the NGDF metadata template and the 30 NGDF fields which relate to the Dublin Core are shown in Annex A. The full NGDF metadata specification can be found in the document entitled ‘NGDF - Discovery Metadata Guidelines – Version 1.1’ which is available via their website (<http://www.ngdf.org.uk/index.htm>). The NGDF repository includes a subset of the metadata contents of StatBase.

I.3 The Information Asset Register (IAR) (<http://www.inforoute.hmso.gov.uk/>)

14. The IAR system which is managed by Her Majesty's Stationery Office (HMSO) in the UK Government's Cabinet Office, was originally designed to catalogue all unpublished Government material. However, its coverage has since been extended to embrace published material as well. The IAR Metadata template has a total of 17 metadata fields, 15 of which relate to the Dublin Core, and these are shown in Annex A. The full IAR metadata specification can be found in the HMSO document entitled ' Guidelines for the Preparation of IAR Records' which is available on the inforoute website. In due course, the Inforoute gateway will provide a link to the contents of StatBase.

II. ENTITIES DESCRIBED

15. Each of the three complementary systems described above concentrates, in the main, on a particular subset of the full range of entities which come under the heading of 'Government resources' but there is a considerable degree of overlap between all three systems. The sort of entities covered by each of these three metadata systems are:

- ◆ Textual material, e.g. Books, Newspapers, Documents, Presentations, Press Releases, Files, Microfiche, Microfilm, Survey Questionnaires, etc.;
- ◆ Audiovisual material:- Audio Cassettes, CDs, Minidiscs, Records, Video Cassettes, CD-Roms, DVDs, Films, Photographs, Maps, etc.;
- ◆ Electronic products and services:- Databases, Websites, Spreadsheets, Powerpoint demonstrations, Floppy discs, other electronic documents, software modules, etc.;
- ◆ Systems:- Administrative Systems, Benefit Systems, etc.;
- ◆ Events:- Conferences, exhibitions, etc.;
- ◆ Services (Free or Commercial):- Enquiry points, Information Services, Analytical Services, Bookshops, etc.;
- ◆ Physical Resources:- Boreholes, Scientific Sites, museum collections, etc.;
- ◆ Personnel:- 'Contacts' such as Owners, Creators, Sponsors, Contributors, Publishers, Distributors, etc.

III. A COMPOSITE METADATA STANDARD

16. The government-wide metadata standard which is being developed by the IAGMWG represents, in effect, a composite of the three existing metadata standards described above. This standard is still in its drafting stage but is likely to embrace the following set of 'mandatory' minimum, or core metadata elements:

- ◆ As well as relevant 'Contact Details' and 'Organisation Details' (Address, Telephone/Fax number, e-mail address, etc.).

Proposed Metadata Fields	Specification
Title of Entity	The name by which a resource is formally known
Alternative Title/Acronym	
Identifier	A controlled and unambiguous reference to the resource within a given context e.g. the ISBN
Mandate	The legal or other basis which requires the resource to be created or provided
Language(s)	The language in which the content of the resource is expressed
Summary Description and History	An account of the resource's genesis, purpose, etc
Subject Coverage	The subject matter of the content within a resource. This can be a search vocabulary based on a standard keyword nomenclature
Geographic Coverage (several fields)	The geographical extent or scope of the content of the resource, based on for example: <ul style="list-style-type: none"> • administrative boundaries • co-ordinates
Type(s)	The nature or genre of the resource
Format(s)	The physical or digital manifestation of the resource
Reference period(s)	Any period of time associated with the content of the resource, e.g. <ul style="list-style-type: none"> • Reference period • Frequency of update
Date(s)	Any date associated with an event in the life cycle of a resource. Examples are: <ul style="list-style-type: none"> • Date first created • Date last created • Date issued
Links	Other complementary resources e.g. <ul style="list-style-type: none"> • A source resource • A related resource
Constraints	Limits on the availability of the resource
E-commerce details	Details of price, availability venues, etc
Owner/Creator/Originator*	The entity with primary responsibility for a resource
Sponsor/Contributor/Contractor *	Other stakeholder entities
Publisher/Supplier/Distributor *	The entity responsible for making the resource available

IV. IMPLEMENTATION

17. The formulation of a standard set of metadata fields which are both widely applicable and widely acceptable is an important first step in the process of establishing a vibrant and dynamic metadata repository. However, ONS's experience with setting up a metadatabase suggests that the creation of a mandatory standard is, perhaps, far less important than the creation of the right climate to allow that standard to take root and thrive. The two most important requirements are for:

- ◆ a Supply Route – it is important for managing organisations to provide potential contributors with an easy-to-use, and preferably on-line, means of supply – this is necessary if content contributors are to be persuaded to overcome their natural disinclination towards, or in some cases outright aversion to, the task of supplying metadata;
- ◆ Strategic Commitment – it is also important for senior managers in the contributing organisations to provide top-down support, encouragement, exhortation and direction – a necessary prerequisite if coalface contributors are to be persuaded to supply their metadata in the first place, and to keep their metadata continually refreshed.

18. Only when the metadata habit has caught on and the resultant metadata are comprehensive, reliable, coherent, consistent and timely can the metadatabase managers move on to the next important tasks in the metadata cycle:

- ◆ Extending the metadata element set to include micro-metadata or variable-level metadata;
- ◆ Devising satisfactory search paths to the metadata collection – whether thematic or geospatial;
- ◆ Providing fast links from the metadata to the associated data.

ANNEX A

Semantic Synergies – Dublin Core / NGDF / IAR / StatBase

DUBLIN CORE	NGDF	IAR	StatBase
1. Title	1. Title	2. Title	Name / Title
2. Creator	3. Originator 34. Contact Name or Title plus CONTACT DETAILS (see below)*	14. Creator	Owner plus CONTACT DETAILS and ORGANISATION DETAILS (see below) #
3. Subject	14. Keywords	6. Subject	(i) Subject Search Directory Links (ii) Demographic variables used (S) (A) (iii) Other specific variables used (S) (A)
4. Description	4. Abstract 31. Additional Information Source	5. Description	Summary Description
5. Publisher	33. Supplier plus CONTACT DETAILS (see below)*	15. Contact / Distributor	CONTACT DETAILS and ORGANISATION DETAILS (see below) #
6. Contributor	34. Contact Name or Title plus CONTACT DETAILS (see below)*		(i) Sponsor (ii) Contractor (S)(A)
7. Date	7. Start Date of Capture. 9. End Date of Capture	8. Date (created) 10. Date last updated	(i) Reference Period (S)(A) (ii) First Available Date (S)(A) (iii) Latest Available Date (S)(A) (iv) Most Recent Year (P)
8. Type	11. Presentation Type		(i) Method (S) (ii) Type (P)
9. Format	29. Supply Media 30. Data Format	12. Format	(i) Primary Medium (P) (ii) Secondary Medium (P)
10. Identifier		1.DTD 3. IARN (Unique Identifier) 4. Identifier/ Acronym	Reference Number (P)
11. Source		11. Source	(i) Sources from which data obtained (P) (ii) Linked Sources (S)
12. Language		13. Language	
13. Relation	32. Dataset Association		(i) Linked Sources (ii) Linked Analyses (iii) Linked Products (iv) Bibliographic material (v) Associated Publications (P)

14. Coverage	(Geographic Coverage) 17. System of Spatial Referencing by Coordinates 19. West Bounding Coordinate 20. East Bounding Coordinate 21. North Bounding Coordinate 22. South Bounding Coordinate 24. National Extent 25. Administrative Area Extent 26. PostCode Area Extent	7. Coverage	(General Coverage) Summary of data coverage (Geographic Coverage) (i) Summary of geographic coverage (ii) National/Sub-National Coverage (iii) Extent of National Coverage (iv) Disaggregation (v) Units of Data Collection (S) (vi) Smallest units for which data available (S)
15. Rights	12. Access Constraints 13. Use Constraint	16. Rights	Indicator – shows whether deposited with Data Archive
	* CONTACT DETAILS 35. Postal Address 36. Postcode 37. Tel. Number 38. Fax Number 39. Email address 40. Web address		# CONTACT DETAILS Tel. Number Fax Number Email Address # ORGANISATION DETAILS Organisation Name Group Name Division Name Branch Name Address Room Number

StatBase Notes:

S = 'Source' metadata field

A = 'Analysis' metadata field

P = 'Product' metadata field

- Up to 3 Contact Types allowed for each 'Source', 'Analysis', 'Product'