UK Surveillance: Adding Value to Data by Conforming Domains and Deriving Additional Attributes

Smith, L. H., Paiba, G.A., Holdship, S., Lysons, R., Lawton, S., Hicks, J. and Roberts, S.
Department for Environment, Food and Rural Affairs (DEFRA), 1A Page Street, London, SW1P 4PQ, England

Abstract

The UK Government launched it's strategy for veterinary surveillance (VS) in October 2003. A key component is the development of a new, surveillance information management system, 'RADAR' (Rapid Analysis and Detection of Animal-related Risks). RADAR has developed a common data architecture and mechanism to support the Extraction, Transformation and Loading (ETL) of surveillance data from multiple source systems. This process adds value to animal health data by combining data from different sources, extrapolating additional information and making data readily comparable.

This paper describes the RADAR system and approach to handling data, and discusses some of the innovative solutions and challenges remaining, such as the development of algorithms and specification of common domains in a manner that is likely to accommodate data from future sources as well as those presently known.

Introduction

The fundamental objective of RADAR is to increase the value of animal health data by combining data from different sources and making such data readily available. In the UK, information of relevance to veterinary surveillance is often collected in different computer systems to differing nomenclatures, collection standards and coding systems. RADAR takes data from these existing systems, transforms it into a common coding system, sometimes referred to as a 'conformed dimension' and quality assures it so that it can be used in ways that have not been possible before. It also derives additional information from the raw source data automatically using a variety of complex calculations and algorithms. For example, monthly cattle population statistics are automatically calculated from movement reports captured by the Great Britain (GB) Cattle Tracing System.

RADAR is being developed and released in phases between now and 2013. The first release of data happened in 2005 and information now available includes the GB cattle and poultry populations, laboratory-reported incidents of Salmonella and information relating to exotic disease control in birds. Phase 2 will run throughout 2006 and 2007 and will provide population data about other livestock (sheep and pigs), livestock movement information (cattle, sheep, pigs and goats), information from statutory surveillance programmes (Brucellosis and EBL) and information about the occurrence of bovine TB. High level analysis of the phase 2 data sources and their different coding systems has already begun.

Phase 3 onwards, will continue to expand the surveillance data available in RADAR. However, some of the veterinary conditions of relevance to the UK may yet be unknown. Consequently, the challenge is to specify common domains in RADAR in a manner which will accommodate data from future sources as well as those presently known.

Materials and Methods

Data incorporated into RADAR

Phase 1 and preparatory work for phase 2 of the RADAR development, has led to the analysis of an extensive amount of surveillance data from various source systems, including, data on the GB cattle population from the Cattle Tracing System (CTS), cases of Salmonella from the Veterinary Laboratories Agency, GB poultry population from the Poultry Register, movement data on sheep and pigs from the Animal Movements Licensing System (AMLS), livestock population data from the Agricultural Census, bovine brucellosis and TB testing data from the Vetnet System used by the State Veterinary Service.

RADAR data loading and processing

The RADAR system is based on well-established software technologies including the Oracle database management system (Oracle Corporation), DataStage (IBM, formerly Ascential Software Corporation) for the extraction, transformation and loading of the data into the data warehouse and QuickAddress (QAS Ltd.) for address cleansing. User access to the RADAR warehouse is provided by a Business Objects (Business Objects Ltd) query tool and ArcGIS geographical information system (ESRI Ltd).

First Stage of Data Processing

Raw data is extracted from source systems by DataStage into the Reception area of the RADAR system. This holds a verbatim copy of all incoming data and is retained as part of the audit trail of data.

Second Stage of Data Processing

Data in the Reception then undergoes a second stage of processing to load the RADAR data warehouse. This second stage of processing has three distinct elements:

a) individual ‘on’ and ‘off’ cattle movement records from CTS are matched into pairs to create a movement history for each animal. Each record of an animal’s life history represents a ‘stay’ at a particular location between two specified dates. Where doubt exists about the animal’s whereabouts, for example if it was in transit overnight or because of a missing record, then a ‘stay’ at an ‘unknown’ location is recorded.

b) raw address records from all source systems are passed though QuickAddress and matched to records in the Post Office Address File (PAF). Where an address match is found, QuickAddress uses the Ordnance Survey AddressPoint dataset to assign a map reference to the matched location.

c) the various source system codes are mapped to a common set of RADAR codes

Third Stage of Data Processing

The warehouse data is transformed into a structure that is more suitable for presentation to the users via Business Objects or GIS. The output of this processing step is a series of data ‘marts’ where the data of interest (e.g. number of animals) can be analysed by a number of domains (i.e. dimensions e.g. breed, age group, location). This third stage of data processing carries out a number of typical tasks:

a) converts a coded value in the warehouse to a more meaningful value for presentation to the users e.g. the cattle breed code ‘HFX’ would be converted to ‘Hereford Cross’.

b) other additional attributes are derived or inferred from coded values in the warehouse e.g. a cattle breed code is converted into a code indicating the normal purpose for which that breed is farmed – beef, dairy or dual purpose.

c) the life history data on individual cattle is processed to generate animal population counts, including counts of animals present at a location on the first day of the month, count of the number of animal days spent at a location during the month, count of number of births at a location during a month etc.

Incremental Load Data Processing

Data in RADAR is refreshed once a month by the RADAR incremental load process in order to keep it up to date. This process is designed to deal with both new records from source systems and amendments to existing records.

Design and Development of the Cattle Algorithm

Approximately 20 million cattle movements are reported each year to the British Cattle Movement Service (BCMS) and recorded in the CTS database. Each movement report records either: a birth, a death, a movement off a location or a movement on to a location. The 'on' and 'off' components of a movement between locations are independently reported and captured (i.e. the 'off' movement is reported by the seller and the 'on' movement by the purchaser). There therefore exists the possibility that the two reports which make up the end-to-end movement can be missing or misreported

At a summary level, the RADAR rules to deal with these inconsistencies are as follows:

- Duplicate movements are discarded
- Movements earlier than the date of birth or later than the date of death are discarded
- Where an 'on' movement has no corresponding 'off' movement for the previous location, or the 'off' movement is reported to be at a later date than the 'on' movement, then the date of the 'on' movement is taken as the date that the animal left its previous location.
- Where an animal moves off its current location but has no subsequent 'on' movement (or the next 'on' movement is at a later date) then the animals is recorded as being in an 'unknown' location for the period between the 'off' and the next 'on'.

Where an animal moves on and off a location in the same day (for example at a market) this event is recorded as a 'transient stay'. No information about the time of these movements is provided so it is not possible to time sequence these transient stays where there are multiple transient events on the same day.

Results

Preliminary analysis of data

Analysis of different source systems has resulted in the division of surveillance data into four classes which represent the generic concepts from which every query on the RADAR system is constructed. These four classes are defined in table 1.

Table 1 Four classes of surveillance data

Class	Definition
Surveillance Activity	A process by which the presence or absence of a disease is determined. Each disease has its own type of surveillance activities which collect similar information on the number and type of animals, the location of the animals and the date of the activity etc. e.g. periodic brucellosis blood test of a beef herd bovine TB herd test BSE test on fallen stock examination of samples for Salmonella serotyping

Surveillance Outcome	A result of performing a Surveillance Activity expressed in terms of the presence or absence of a disease, the category of disease found and the number of instances of the outcome. A surveillance activity may result in more than one outcome. e.g. a bovine TB herd test may result in outcomes expressed as number of reactors, number of inconclusive reactors and number of confirmed cases. Alternatively, a Salmonella isolation could result in outcomes characterised by the *subgenus*, *serotype* and *phagetype*.
"Risk Factor"	Any factor that could be viewed as affecting an animal's risk of contracting a disease or the risk of a disease being spread. e.g. age, species, location, housing, weather, soil type, season, number of animal movements etc. Presenting signs are also included in this category, i.e. the signs observed in a clinical case. This includes all explanatory and confounding variables.
Population at Risk	A measure of the total number of animals at risk of contracting a disease. The way this is measured differs according to the intended purpose of the query. e.g. average number of animals during a period, number of unique animals during a period, number of locations or number of animal days etc.

Within the RADAR analysis environment *Risk Factors* are normally used to constrain the population measures. So a typical query on the RADAR system may ask - What is the occurrence of **Factor X** relative to the occurrence of **Factor Y** categorised by **Factor Z**?

e.g. What is the incidence of Salmonella in cattle relative to the size of the cattle population categorised by cattle production type? In this example, Factor X is a *Surveillance Outcome*, Factor Y is the *Population at Risk* and Factor Z is a *Risk Factor*.

Obviously other combinations of factors are possible, and the list of potential *Risk Factors* itself is long and varied, but for use within RADAR it can be conveniently sub-divided into further classifications:

- Aspects of the animal – species, breed, age, sex, production purpose, presenting sign
- Aspects of the location – type of premises, disease history, species kept, geography, geology, economic activities
- Animal movements – number of movements, size of group, number of previous movements, duration of movement etc.
- Surveillance method – the reason for surveillance or laboratory used etc.

Conforming Domains

In particular, from the analysis of data in numerous source systems, RADAR has identified the need to conform the domains associated with the animal-related *Risk Factors*, including

- Species
- Breed
- Age group
- Purpose
- Sex
- Presenting Sign

The age group domain presents particular problems because of the diversity found in the construction of age group bands in different source systems. The issue is illustrated in table 2 which shows age bands for cattle from 3 different source systems. The table also shows a 'conformed' set of age bands that are applicable across all three data sources.

The conformed age bands provide only the basic ability to distinguish between calves and adults. This may be adequate for some applications but other uses will require more granular analysis, so in addition to providing the conformed Age Group dimension, RADAR also retains the original categorisations against the relevant source system data.

Table 2 Conforming cattle ages from 3 different source systems

Cattle Age Groups			
Conformed Domain	**Census Data**	**VLA Data**	**RADAR - CTS Data**
Calf	< 1 year	Neonatal	0-6 days
		Immature	1-4 weeks
			1-2 months
			3-6 months
			7-12 months
	1 – 2 years		13-24 months
Adult	> 2 years	Adult	25-30 months
			>30 months
Unknown		Unknown	Unknown
Mixed		Mixed	
Not Given		Not Given	

The Cattle Algorithm

The cattle algorithm processes over 1 million movement records each month, updating the life histories of over 28 million animals, and resolving a number of inconsistencies in the data including duplicate reports and non-matching movements, so that the entire cattle population can be analysed. During data investigations in 2004, the animal life history data was analysed to generate a series of life history patterns. Some 28 million life histories were examined and approximately 210,000 unique patterns were identified. The ten most frequently occurring patterns, shown in table 3, accounted for around 66% of the life histories analysed. The top 100 patterns accounted for 90% of the total.

Table 3 Top ten most frequently occurring cattle movement patterns

Rank	Date 1	Date 2	Date 3	No of animals
1	Birth A			5,307,371
2	Birth A	Death B		5,228,726
3	Birth A	Off A – On B – Death B		1,920,314
4	Birth A	Death A		1,570,941
5	Birth A	Off A – On B		1,099,137
6	Birth A	Off A – On B	Off B – On C – Death C	1,093,514
7	On A			784,509
8	Birth A	Off A – On B – Off B – On C	Off C – On D – Death D	591,892
9	Birth A	Off A – On B	Death B	524,999
10	On A	Death A		497,712

In the table, life history events are grouped by the date of occurrence. Taking the 3rd most frequent pattern as an example, the animal is born at Location A and then sometime later moves from Location A to Location B where it dies on the same day. It can be seen that the majority of the top ten patterns represent the expected life history of cattle. Pattern 3 is consistent with animals reared on the farm of birth and then being sent directly to slaughter. Pattern 8 is consistent with animals being sent to market and then being reared at a second location before they are dispatched for slaughter.

The shaded patterns represent anomalous life histories (2, 7, 10) which have missing events. Numbers 7 and 10 would be consistent with animals that were already alive when CTS began collecting data as these animals were 'placed' at their locations with a special 'on' movement. Pattern 2 would be consistent with the expected history for animals that were born after the introduction of cattle passports but before movement reporting was introduced. This conclusion is confirmed by further analysis of the Pattern 2 animal life histories, which shows that the vast majority of them occurred for animals born between July 1996 and September 1998.

Efforts have been made by BCMS to improve the accuracy of CTS data and to cleanse erroneous cattle movement data. These have improved the quality of the data in RADAR, which can be seen in a year on year reduction in animals residing in unknown locations (i.e. those that do not have full PAF details available) on the 1st March each year (Figure 1).

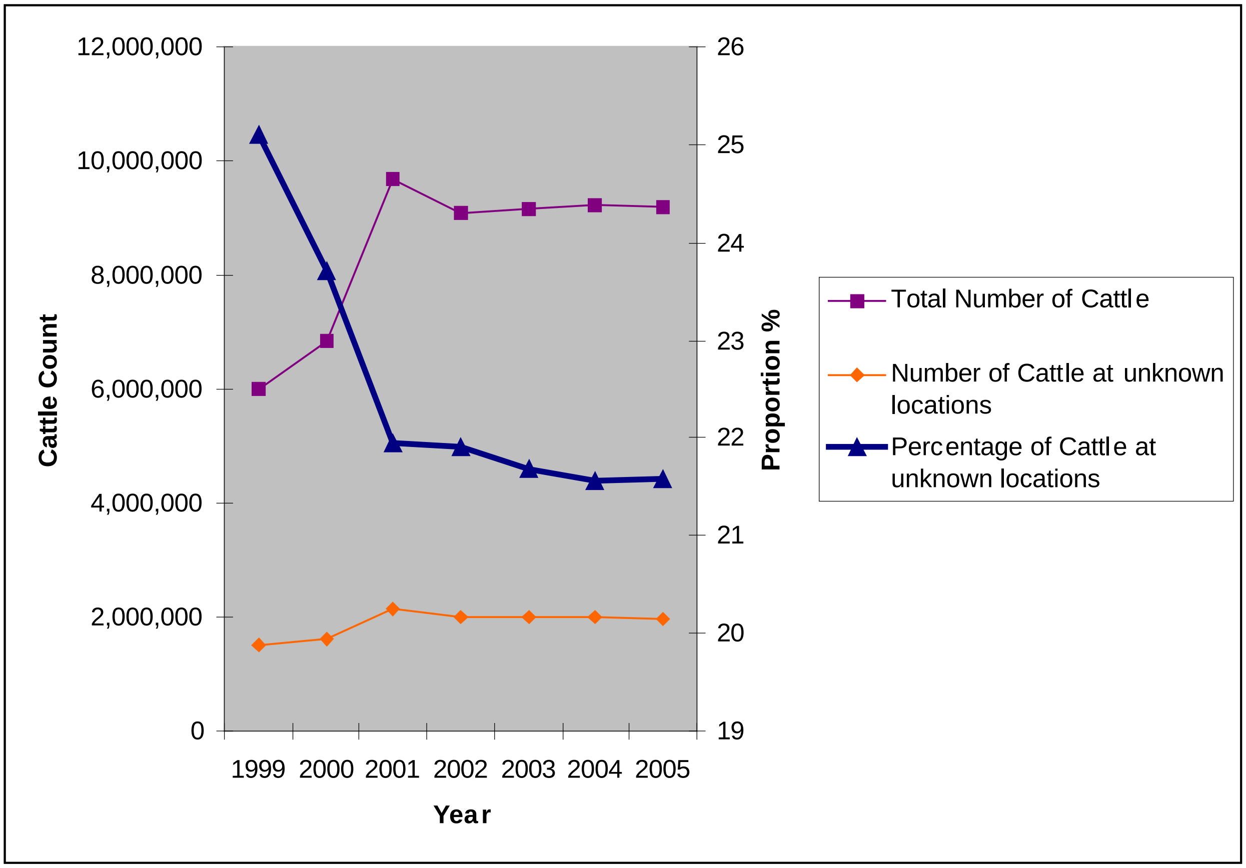

Figure 1 Proportion of cattle in an 'unknown' location on the 1st March each year

Discussion

The extensive scope of the RADAR system means it will take a number of years to bring all of the key source systems online; RADAR is currently in year 3 of a 10 year development plan. Consequently, it is not possible to predict every structure of data that RADAR will ever be required to hold, some source systems may not even yet exist and some diseases which may be important to GB in the future, may not yet be known to science.

However, RADAR has managed to identify some common concepts and inter-relationships, which can be applied to a wide range of surveillance data. This has lead to the categorisation of data into 4 generic classes, which provide the founding principles on which the RADAR system is built and mean that the RADAR warehouse is designed in such a way that data from future source systems can be accommodated without the need for structural change. Data from the 4 classes are also conformed into a common coding system wherever possible to allow the cross comparison of data from different sources.

As the scope of RADAR data increases it also becomes increasingly important that users have an understanding of the quality of data in the system and can be assured that the data is suitable for their intended purpose. RADAR performs a number of validity checks during its multistage data processing to try and identify erroneous data. In common with much source data, the processing of CTS data has shown that the consistency of the dataset is subject to human error and is ultimately dependant on correct information being recorded by or captured from animal keepers. In recent years CTS has made provision for this potential source of error by automatically detecting certain types of anomalies and by the insertion of both manually and automatically inferred movement reports. A consequent year on year improvement of data quality has now been seen. To communicate this, and other information about the quality of data in RADAR, a data quality framework has been developed which reports on the quality of the data from all sources to all users of the RADAR system (Paiba, G. A *et al*, 2006).

In the immediate future, the priorities for phase 2 of the RADAR development will be to incorporate further surveillance data into the common architecture and to use the results and experience gained from phase 1 of the development to make improvements to the internal processing of data e.g. to enhance the cattle algorithm so that allowance is made for the type of holding (e.g. farm, slaughterhouse or market) involved in the movement of cattle, so that the sequencing of same day movements can be improved.

References

Paiba G.A. *et al* (2006) UK Surveillance: combining datasets to provide quality assured surveillance measures. Proceedings of ISVEE XI, Paper 751.
IBM Corporation, 1133 Westchester Avenue, White Plains, New York 10604, United States
Oracle Corporation, Corporate Headquarters, 500 Oracle Parkway, Redwood Shores, CA 94065, USA. Tel: 650-506-7000. www.oracle.com
QAS Ltd, George West House, 2-3 Clapham Common North Side, London, SW4 0QL Tel: +44 (0)20 7498 7777, Fax: +44 (0)20 7498 0303
Business Objects SA, 157-159, rue Anatole France, 92309 Levallois-Perret France. www.businessojects.com
ESRI, 380 New York Street, Redlands, CA 92373-8100, USA. Tel: 909-793-2853, Fax: 909-793-5953. www.esri.com.

Acknowledgements

The authors are grateful for the IT consultancy services supplied by Medley Business Solutions Limited, AtosOrigin and IBM. They would also like to thank Andrew Mitchell at the VLA, Rowland Kao at Oxford University and the members of the RADAR Animal Population Working Group for the review comments they provided on the design of the cattle algorithm.