Continuous validation and feedback of data quality from a LIMS system used for surveillance

Introduction

Performing data validation on a regular basis on data entered and stored in a Laboratory Information Management System (LIMS) is difficult to do in a generic way. This poster presents a workflow to test and generate feedback on the data quality.

Methods

Unit testing in computer programming is a common and well established technique to check that functions behave and performs as expected. We have used the R package **testthat** as a base for our data validation. The **testthat** package has a test structure consisting of expectations, tests and contexts. We have developed an R package for data quality control using the structure and adding an expectation for data quality. We have also included a structure for adding meta information from the test function and use that for layout when displaying the results of the data validation.

To validate data quality, each expectation has a condition to test, and subsequently check that it passes the quality control. If it fails, the identified errors are displayed in the generated html status report.

All user tests are written in an R script file and can easily be run every time there is a new or updated data set to generate a status report.

Example: The code below shows a test on a dataset of foxes with expected geo-referenced data. The generated html status report is shown to the right. The title and headings are extracted from the test code.

```
qc_context('title{Quality control of data} author{qcdata}')

test_that("{section{Coordinates} subsection{Missing coordinates}", {
    foxes <- read.csv('foxes.csv')

    foxes <- subset(foxes,
                    is.na(Lat) & is.na(Lon),
                    select=c('SampleId',
                             'Region',
                             'Description',))

    expect_that(foxes,  qc_pass())
})
```

Layout

Condition

Quality control

Results

The data quality package has made it possible to code expectations in a generic and structured fashion. It has also made it possible to follow data and expectations in a LIMS system with daily updates and feedback of the data quality.

We believe that the presented workflow to perform quality tests on data is useful and can be further developed.

References

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Leisch, F. (2002a). Sweave: Dynamic generation of statistical reports using literate data analysis. In W. Härdle and B. Rönz (Eds.), Compstat 2002 - Proceedings in Computational Statistics, pp. 575{580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9.

Wickham, H. testthat: Get Started with Testing The R Journal, 2011, 3, 5-10

Quality control of data

qcdata

2012-03-12

1. Coordinates
   - 1.1. Missing coordinates
   - 1.2. Incorrect coordinates
   - 1.3. Missing region

1. Coordinates

1.1. Missing coordinates

SampleId	Region	Description
312	M	
511	E	
348	M	Fox

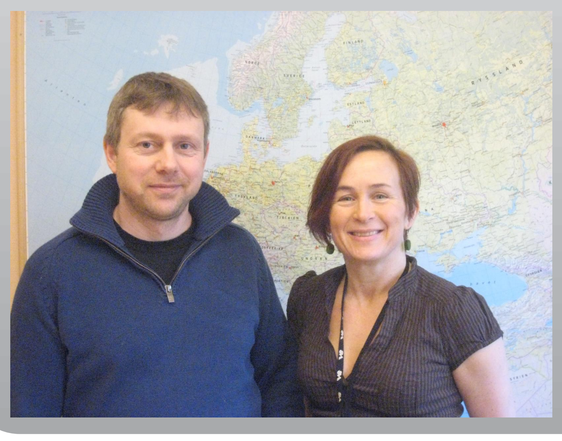
1.2. Incorrect coordinates

SampleId	Region	Lat	Lon
345	M	1385800	6244800
364	O	6496088	136452
516	E	650610	147745
480	E	6490689	
31	O	586254	116966

1.3 Missing region

OK

Generated 2012-03-12 10:00:40 by qcdata version 0.2



Stefan Widgren, MSc, DVM

Ann Lindberg, Assoc. prof, DVM, PhD, Dipl ECVPH

Department of Disease Control and Epidemiology
NATIONAL VETERINARY INSTITUTE
post. SE-751 89 Uppsala, Sweden
phone. +46 18 67 40 00 **fax.** +46 18 30 91 62
e-mail. sva@sva.se **web.** www.sva.se

SVEPM 2012, 28th - 30th March - Glasgow, UK