

The Importance of Data Quality Control

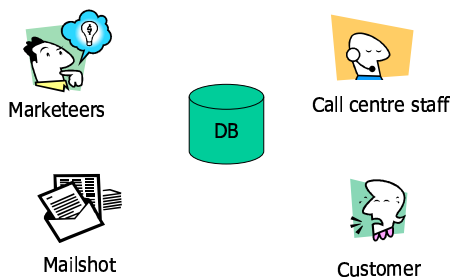
Suzanne M. Embury
Department of Computer Science
University of Manchester

Data: Asset or Liability?

- Data is an asset
 - Very long-lived
 - Can be used in many unforeseen ways
 - Data mining, mass customisation, optimisation
 - But much data is of poor quality
 - Costly to diagnose and assess
 - Costly to repair
 - Barrier to new uses of data
- Cost in lost business through customer dissatisfaction?*

NERC Microarray Data Quality Control and Analysis Workshop – March 2004

A Cautionary Tale



NERC Microarray Data Quality Control and Analysis Workshop – March 2004

Cost of Poor Data Quality

- Most costs are hidden and hard to quantify
 - Increased costs through wasted resources (e.g. under-billing)
 - Increased costs through need to correct and deal with reported errors (rework)
 - Increased costs through inability to optimise business processes
 - Lost revenue through customer dissatisfaction
 - Lost revenue through lowered employee morale
 - Lost revenue through poorer decision making

NERC Microarray Data Quality Control and Analysis Workshop – March 2004



DQ in E-Science

- Early indications
 - Same kinds of data quality problems found in scientific data
 - Just as difficult to quantify costs
 - Just as difficult to detect and diagnose
 - Just as expensive to correct
- The first problem is ...

NERC Microarray Data Quality Control and Analysis Workshop – March 2004



Beyond Accuracy...

- Data quality is about much more than just accuracy
- By analogy with quality in general, we say:
 - Data is of high quality if it is *fit for purpose*
- What does this mean?
 - No global definition of high data quality
 - Domain dependent
 - Application dependent

NERC Microarray Data Quality Control and Analysis Workshop – March 2004



What is Good Quality Data?

- Data that is an *accurate* representation of the part of the "real world" that it models
- What about the following examples:
 - Annotation for a microarray on manufacturer's website is out of date
 - SWISS-PROT entry without either an RP line or a DE line
 - NASA Mars Climate Orbiter error (imperial units treated as metric)

NERC Microarray Data Quality Control and Analysis Workshop – March 2004



Data Quality Dimensions


- Accuracy
- Precision
- Completeness
- Currency
- Non-redundancy
- Portability
- Credibility

>500 spellings of the Nat West Bank

Multiple spellings of Escherichia coli (RDS)

Chromosome "22" in mice... (CH)

NERC Microarray Data Quality Control and Analysis Workshop – March 2004

 **Data Quality Dimensions**


- Accuracy
- Precision
- Completeness
- Currency
- Non-duplication
- Portability
- Credibility

Rounding errors and vanishing data

Experimental error

Other forms: e.g. GO terms

NERC Microarray Data Quality Control and Analysis Workshop – March 2004

 **Data Quality Dimensions**


- Accuracy
- Precision
- Completeness
- Currency
- Non-duplication
- Portability
- Credibility

Address matching is a sizeable industry in itself

Functional annotation that is not updated

Gene names can change or obtain synonyms, without this being reflected in the data

NERC Microarray Data Quality Control and Analysis Workshop – March 2004

 **Data Quality Dimensions**


- Accuracy
- Precision
- Completeness
- Currency
- Non-duplication
- Portability
- Credibility

Within the recorded data set

Beyond the recorded data set

Often measurable against a second source
e.g. SWISS-PROT against GenBank

NERC Microarray Data Quality Control and Analysis Workshop – March 2004

 **Data Quality Dimensions**

- Accuracy
- Precision
- Completeness
- Currency
- Non-duplication
- Portability
- Credibility

Duplication has benefits and downsides

Possibility of inconsistency

Possibility of non-unique keys or identifiers

SWISS-PROT/TrEMBL overlap

NERC Microarray Data Quality Control and Analysis Workshop – March 2004

Data Quality Dimensions

- Accuracy
- Precision
- Completeness
- Currency
- Non-duplication
- Portability
- Credibility

Can the data be used outside the context of its creation

Dates: 03/04/03

Archaeological data

Lack of provenance, incomplete metadata

NERC Microarray Data Quality Control and Analysis Workshop – March 2004

Implications

- Data stored by computer is just as likely to be wrong as data stored by any other means
 - Perhaps even more likely because of scale and degree of automation

NERC Microarray Data Quality Control and Analysis Workshop – March 2004

Data Quality Dimensions

- Accuracy
- Precision
- Completeness
- Currency
- Non-duplication
- Portability
- Credibility

V. important but v. hard to quantify

Some labs/scientists/equipment are believed to produce better data

In science, can also use "weight of evidence"

NERC Microarray Data Quality Control and Analysis Workshop – March 2004

Implications

NERC Microarray Data Quality Control and Analysis Workshop – March 2004



Positives

- Lots of lessons to be learnt from industry
- Chance to catch the problem early
- Tools and techniques are beginning to emerge
- Community support for data quality improvement (standards, data centres, data quality policies)
- This workshop!

NERC Microarray Data Quality Control and Analysis Workshop – March 2004



Acknowledgements

- Thanks to
 - Nigel Turner, BT Exact
 - Dawn Field, CEH Oxford
 - Joe Wood, CEH Oxford
 - Robert Stevens, University of Manchester
 - Mike Cornell, University of Manchester
 - Cornelia Hedeler, University of Manchester

NERC Microarray Data Quality Control and Analysis Workshop – March 2004