Introduction

- Patrick O’Beirne BSc MA FICS
- Systems Modelling Ltd. Ireland (sysmod.com)
- Current focus: spreadsheet quality and auditing.
  - ‘Spreadsheet Check and Control’ book
  - Software for assessing s/s (ScanXLS, XLTEST)
  - Other IT books and articles.
- Presentations to ICS, ISACA, EuSpRIG, the Excel User Conference, and other interest groups.
- Professional affiliations:
  - Irish Computer Society
  - European Spreadsheet Risk Interest Group (EuSpRIG)
  - Software Testing Interest Group in Ireland (SoftTest)
Outline

- A brief outline of interest in Data and Information Quality
- A review of the data attributes commonly described in the literature on data quality
- A review of papers and software tools
- Considerations specifically to do with spreadsheet data control

IQ Trainwrecks

- http://www.iaidq.org International Association for Information and Data Quality
- http://www.iqtrainwrecks.com/
  - An IQ Trainwreck is a problem that affects real people in the real world that has, at its heart, poor quality information or a failure to manage the quality of information.
- $125M: Mars Climate Orbiter lost in space in September 1999. One team used English units, the other used metric for a key spacecraft operation
## Information Quality Attributes

- Accessible
- Accuracy
- Appropriate Amount
- Atomic
- Credible
- Complete
- Concise
- Coverage
- Conformity
- Consistent
- Coherence
- Interpretable
- Meaning
- Objective
- Redundancy
- Relevant
- Reputable
- Secure
- Timely
- Understandable
- Usability
- Value
- Validity

## Information Quality Actions

- Access
- Measure accurately
- Satisfy
- Normalise
- Credit
- Complete
- Compact
- Collect
- Standardise
- Reconcile
- Calibrate
- Clarify
- Mean
- State fairly
- Economise
- Relate
- Verify
- Secure
- Update
- Communicate
- Facilitate
- Serve
- Validate
Root Causes of Poor Data Quality

Data item attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data type</td>
<td>Storage type used for the data element</td>
</tr>
<tr>
<td>Default value</td>
<td>If the user makes no entry this value is assumed</td>
</tr>
<tr>
<td>Error</td>
<td>keying mistakes include transposition of digits</td>
</tr>
<tr>
<td>Format</td>
<td>Presentation in a form to aid comprehension</td>
</tr>
<tr>
<td>Missing value</td>
<td>What the system does with empty values</td>
</tr>
<tr>
<td>Null</td>
<td>Whether Null is allowed</td>
</tr>
<tr>
<td>Precision</td>
<td>Measure of detail in which the quantity is expressed</td>
</tr>
<tr>
<td>Primary key</td>
<td>Unique record identifier</td>
</tr>
<tr>
<td>Range of values</td>
<td>Minimum to maximum valid values</td>
</tr>
<tr>
<td>Referential integrity</td>
<td>Primary and Foreign Keys must exist</td>
</tr>
<tr>
<td>Restricted value list</td>
<td>Discrete list of valid values</td>
</tr>
<tr>
<td>Size</td>
<td>Length in characters or scale</td>
</tr>
<tr>
<td>Unit of measure</td>
<td>For quantities</td>
</tr>
</tbody>
</table>
### Process Checks, Tests (1)

<table>
<thead>
<tr>
<th><strong>Procedures</strong></th>
<th>Are there defined procedures for processing the data and are they followed?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Controls</strong></td>
<td>Authorisation and separation of duties</td>
</tr>
<tr>
<td><strong>Audit trail</strong></td>
<td>Auditing answers the question what was changed or viewed, by what user and on what date and time.</td>
</tr>
<tr>
<td><strong>Audit checks</strong></td>
<td>A re-check against business rules for example to reconcile two accounts; to detect whether there are many payments just below a n authorisation threshold.</td>
</tr>
<tr>
<td><strong>Archive</strong></td>
<td>Secure access to backups and consistent versions</td>
</tr>
<tr>
<td><strong>Tags</strong></td>
<td>Data can be flagged with metadata to indicate conformity</td>
</tr>
<tr>
<td><strong>Match &amp; Merge</strong></td>
<td>Can sets of records be merged without contamination; or unmerged? Can different time series, scales, or types be integrated?</td>
</tr>
<tr>
<td><strong>Missing records</strong></td>
<td>How can you know when records are absent and what is done?</td>
</tr>
<tr>
<td><strong>Process</strong></td>
<td>Frequency and time characteristics</td>
</tr>
<tr>
<td><strong>Sharing</strong></td>
<td>What other systems have access, and to what level?</td>
</tr>
</tbody>
</table>

### Process Checks, Tests (2)

<table>
<thead>
<tr>
<th><strong>Source</strong></th>
<th>Where did the data come from?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Use</strong></td>
<td>Where is the data used?</td>
</tr>
<tr>
<td><strong>Linked data</strong></td>
<td>Control automatic links among spreadsheets and data sources for completeness, accuracy and appropriateness of data transfer.</td>
</tr>
<tr>
<td><strong>Continuous</strong></td>
<td>Are there gaps in some field sequences?</td>
</tr>
<tr>
<td><strong>Duplicates</strong></td>
<td>Are records duplicated?</td>
</tr>
<tr>
<td><strong>Statistics</strong></td>
<td>Measures such as Min/Bottom 10, Max/Top 10, Average, Frequency distribution; can indicate expected values and help in identifying outliers and unexpected or unlikely values.</td>
</tr>
<tr>
<td><strong>Ambiguity</strong></td>
<td>Is there more than one field of the same name with different meanings?</td>
</tr>
<tr>
<td><strong>Error statistics</strong></td>
<td>Statistics on the expected occurrence of random errors; for example transcription errors.</td>
</tr>
<tr>
<td><strong>Calibration</strong></td>
<td>Validating a measuring instrument against a standard</td>
</tr>
<tr>
<td><strong>Sampling</strong></td>
<td>Where there is more data than can be checked, a sample must be taken following standard statistical sampling techniques.</td>
</tr>
</tbody>
</table>
DQ in Spreadsheets

- As applied to modelling
- Forecasting, projection
- Data input list specifies for each input item:
  - Format,
  - Units,
  - Frequency of update,
  - Status of its authority,
  - Validation rule,
  - Source, and
  - other notes.

Data Errors in forecasting

- Multiplier effects when incorrect data undergo many and sequential manipulations
- Aggregation significantly dampens input error
- The expected value of a ratio is not equal to the ratio of expected values
- Hence, Monte Carlo simulation
- Flaw of Averages: Plans based on average conditions are wrong on average
**Data Manipulation**

- Supplement non-agile IT systems
- Transformation, presentation
- Stovepipe systems
- Dumb solutions
- Spreadmarts
- Multiple versions of ‘the truth’

**Computer Aided Audit Tools & Techniques**

- ACL (Audit Command Language)
- IDEA (Interactive Data Extraction and Analysis)
- ActiveData for Excel
- Data Mining (incl OSS)
- Many spreadsheet auditing tools
Software tools

- EXChecker (Compassoft)
- Prodiance Spreadsheet IQ
- ClusterSeven (change log+)
- Lyquidity (change log+)
- ExSafe (security service)
- ScanXLS (Inventory, Links)
- XLTTest add-in
- RedRover error-finding audit
- SpACE 3 (pending)

- Spreadsheet Detective
- Spreadsheet Professional
- Operis Analysis Kit
- Rainbow Analyst
- XLAnalyst
- XLSior test runner
- Code Tracer
- XLSpell style checker
- Navigator Utilities
- ActiveData data analysis

This list maintained at: http://www.sysmod.com/sslinks.htm
Very many other useful add-ins in the marketplace

Typical functions of CAATTTS

- Match and Merge: Combines columns where rows are matched
- Compare: compares two sheets.
- Extract (Demerge) & Sampling (Random, stratified)
- Generate: Fill cells with random, fixed or incremental values
- Convert: transform or reformat data formats or data types.
- Group: Subtotals, Top/Bottom Items, Date Aging
- Stratification by bands, Cross-tabulation
- Statistics: Descriptive Statistics, Summary
- Duplicates: duplicated rows (are primary keys still unique?)
- Gaps: missing rows, data items missing (empty cells), or invalid
- Find: suspicious data (all the 9s, 01/01/01, and similar)
- Spell-check: are there any spelling mistakes?
- Benford's analysis: used to detect fraud from the pattern of digits where amounts have been invented.
Avoiding GIGO

- All relevant data are input,
- No irrelevant or inappropriate data are input,
- Data are input accurately,
- Data are input for process at the correct time,
- Controls on completeness and accuracy of the transfer of data among sheets & files.
- Butler, 2000, "Is This Spreadsheet a Tax Evader?"
Recommendations for users

- Spreadsheet Check and Control book
- Spreadsheet Safe certification
  - Set out conventions used
  - Isolate constants
  - Validate imported data (e.g., CSV)
  - Validate links
  - Check for missing input values
  - Use IF() to test values expected
  - Review for data type mis-entry
  - Apply conditional formatting to highlight errors,
  - Apply validation criteria.
- These introduce another layer of double-check

Visualisation

- Charting
- Detecting outliers
- Colour differences between records
- Labelling & colouring of input cells
- Self-checking formulas for output cells
- Make reference structure visible
Colouring Data Type and usage

Colouring number of dependents
Conditional Format formula consistency

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Date</td>
<td>PaymentType</td>
<td>Beneficiary</td>
<td>Vendor</td>
<td>City</td>
<td>Value</td>
<td>Property</td>
<td>OCBI</td>
<td>Inv</td>
</tr>
<tr>
<td>2</td>
<td>01/02/1993</td>
<td>BonFire</td>
<td>200</td>
<td>5001</td>
<td>GARY</td>
<td>0.216103</td>
<td>123.45</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>01/02/1993</td>
<td>BonFire</td>
<td>200</td>
<td>5001</td>
<td>GARY</td>
<td>0.985237</td>
<td>123.45</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>01/02/1993</td>
<td>BonFire</td>
<td>200</td>
<td>5001</td>
<td>GARY</td>
<td>0.546789</td>
<td>123.45</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>01/02/1993</td>
<td>BonFire</td>
<td>200</td>
<td>5001</td>
<td>GARY</td>
<td>0.123456</td>
<td>123.45</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Data Validation formula consistency

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>First use Conditional Format Formulas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>=100 Interior ColorIndex=45 ligh Orange</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1461 Interior ColorIndex=44 Bright Green</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1462 Interior ColorIndex=44 Light Orange</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1463 Interior ColorIndex=44 Light Green</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>First use Data Validation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1014 Stop Decimal 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1020 Stop Decimal 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1025 Stop Decimal 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary

- Information quality serves decisions
- Not just Technology but also Policy, Training, Methods, and People
- Responsibility and ownership of data quality
- Data quality supports Information Quality
- Spreadsheets are fragile data structures
- Many checks possible, many available in software tools

Thank you!

- Any questions … ?