Information and Data Quality in Spreadsheets

Patrick O'Beirne
Systems Modelling Ltd
Tara Hill, Gorey, Co. Wexford, Ireland
Tel +353-53-942-2294 Email pob@sysmod.com
www.sysmod.com

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

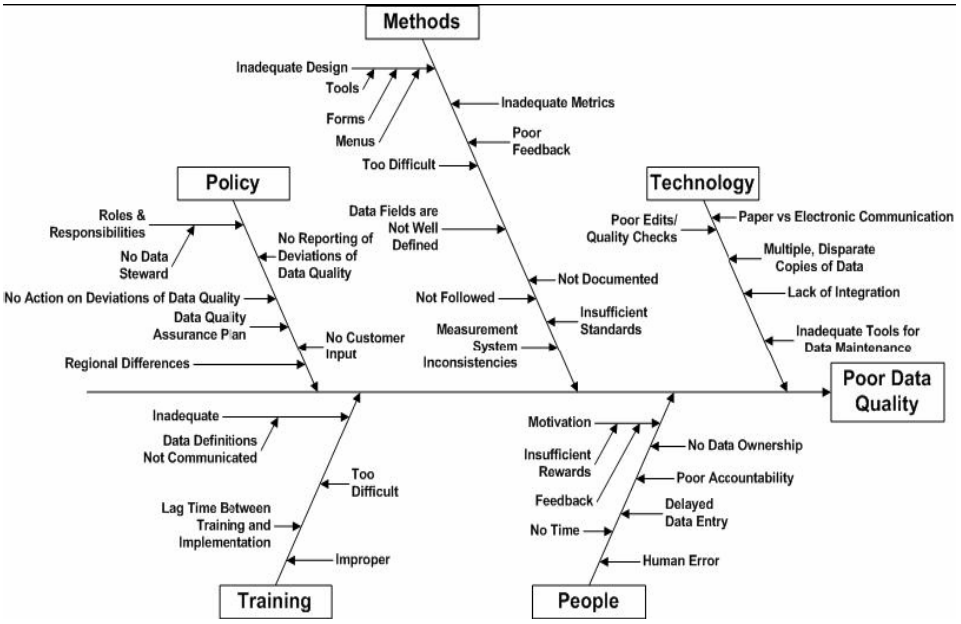
Information Quality Attributes

- 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

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

Process Checks, Tests (1)

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

Process Checks, Tests (2)

Source	Where did the data come from?
Use	Where is the data used?
Linked data	Control automatic links among spreadsheets and data sources for completeness, accuracy and appropriateness of data transfer.
Continuous	Are there gaps in some field sequences?
Duplicates	Are records duplicated?
Statistics	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.
Ambiguity	Is there more than one field of the same name with different meanings?
Error statistics	Statistics on the expected occurrence of random errors; for example transcription errors.
Calibration	Validating a measuring instrument against a standard
Sampling	Where there is more data than can be checked, a sample must be taken following standard statistical sampling techniques.

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

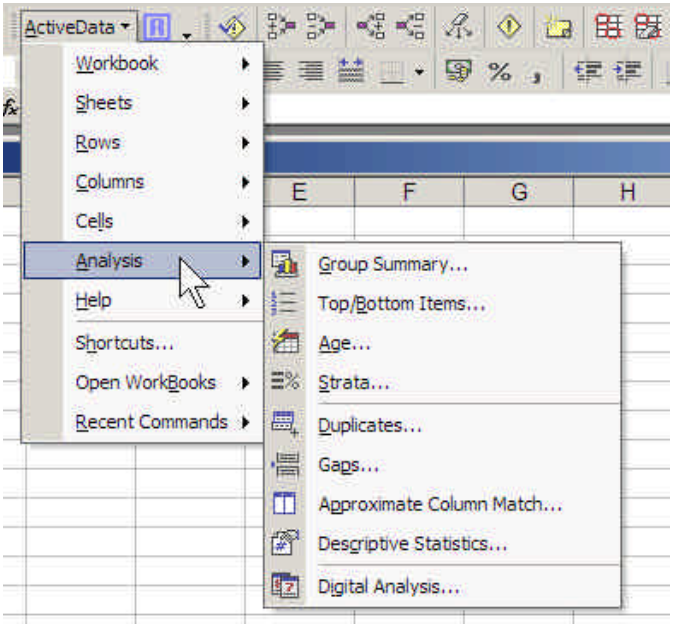
Software tools

- EXChecker (Compassoft)
- Prodiance Spreadsheet IQ
- ClusterSeven (change log+)
- Lyquidity (change log+)
- ExSafe (security service)
- ScanXLS (Inventory, Links)
- XLTest 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 CAATTS

- 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?"

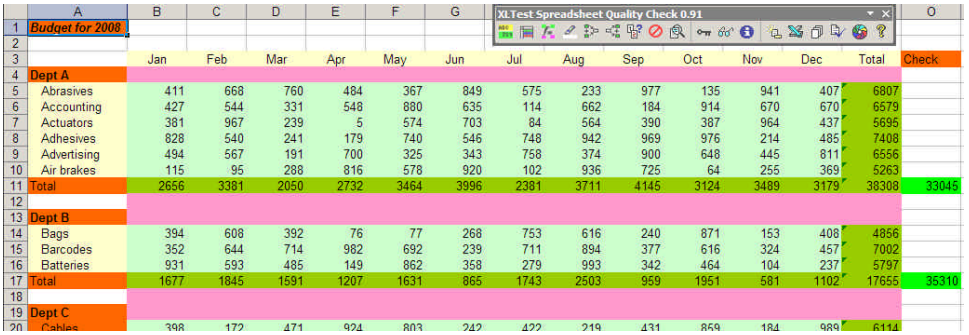
Recommendations for users

- Spreadsheet Check and Control book
- Spreadsheet Safe certification
  - Set out conventions used
  - Isolate constants
  - Validate Imported data (eg 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

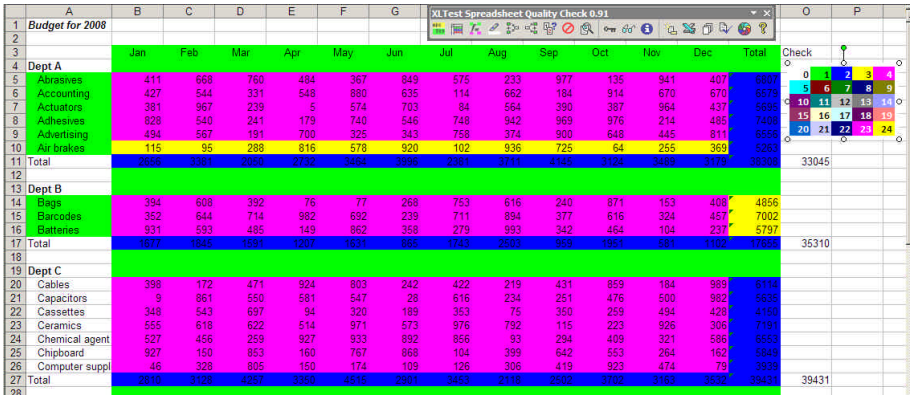
Colouring Data Type and usage



Budget for 2008	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total	Check
Dept A														
Abrasives	411	668	760	484	367	849	575	233	977	135	941	407	6807	
Accounting	427	544	331	548	880	635	114	662	184	914	670	670	6579	
Actuators	381	967	239	5	574	703	84	564	390	387	964	437	5695	
Adhesives	828	540	241	179	740	546	748	942	969	976	214	485	7408	
Advertising	494	567	191	700	325	343	758	374	900	648	445	811	6556	
Air brakes	115	95	288	816	578	920	102	936	725	64	255	369	5263	
Total	2656	3381	2050	2732	3464	3996	2381	3711	4145	3124	3489	3179	38308	33045
Dept B														
Bags	394	608	392	76	77	268	753	616	240	871	153	408	4856	
Barcodes	352	644	714	982	692	239	711	894	377	616	324	457	7002	
Batteries	931	593	485	149	862	358	279	993	342	464	104	237	5797	
Total	1677	1845	1591	1207	1631	865	1743	2503	959	1951	581	1102	17655	35310
Dept C														
Cables	398	172	471	924	803	242	422	219	431	859	184	989	6114	

Data Type:	Number	Date	Text	Logical
Unused Constant				
Input Constant				
Input Formula				
Intermediate Formula				
Output Formula				
Empty input cell		Error		

Colouring number of dependents



Budget for 2008	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total	Check
Dept A														
Abrasives	411	668	760	484	367	849	575	233	977	135	941	407	6807	
Accounting	427	544	331	548	880	635	114	662	184	914	670	670	6579	
Actuators	381	967	239	5	574	703	84	564	390	387	964	437	5695	
Adhesives	828	540	241	179	740	546	748	942	969	976	214	485	7408	
Advertising	494	567	191	700	325	343	758	374	900	648	445	811	6556	
Air brakes	115	95	288	816	578	920	102	936	725	64	255	369	5263	
Total	2656	3381	2050	2732	3464	3996	2381	3711	4145	3124	3489	3179	38308	33045
Dept B														
Bags	394	608	392	76	77	268	753	616	240	871	153	408	4856	
Barcodes	352	644	714	982	692	239	711	894	377	616	324	457	7002	
Batteries	931	593	485	149	862	358	279	993	342	464	104	237	5797	
Total	1677	1845	1591	1207	1631	865	1743	2503	959	1951	581	1102	17655	35310
Dept C														
Cables	398	172	471	924	803	242	422	219	431	859	184	989	6114	
Capacitors	9	861	550	581	547	28	616	234	251	476	500	982	5635	
Cassettes	348	543	697	94	320	189	353	75	350	259	494	428	4150	
Ceramics	555	618	622	514	971	573	976	792	115	223	926	306	7191	
Chemical agent	527	456	259	927	933	892	856	93	294	409	321	586	6553	
Chipboard	927	150	853	160	767	868	104	399	642	553	264	162	5849	
Computer suppl	46	328	805	150	174	109	126	306	419	923	474	79	3939	
Total	2810	3128	4257	3350	4515	2901	3453	2118	2502	3702	3163	3532	39431	

Conditional Format formula consistency

A1 fx Client

	A	B	C	D	E	F	G	H	I	J
1	Client	Date	PaymentTyp	BenefitAmou	Vendor	City	Coun	Poverty	DOB	Se
2	A006342	01/02/1999 00:00	Benefit	285	50081	GARY	2	0.215688318	15/01/1965 00:00	F
3	A006427	01/02/1999 00:00	Benefit	240	50081	GARY	2	0	06/02/1973 00:00	F
4	A006328	01/02/1999 00:00	Benefit	240	50081	GARY	2	0.59633702	20/04/1967 00:00	F
5	A003919	15/03/1999 00:00	Benefit	210	50081	LAKE STATION	2	1.120073318	14/12/1958 00:00	F
6	A006465	24/03/1999 00:00	Benefit	240	50081	GARY	2	0.54635942	24/09/1962 00:00	M
7	A006323	24/03/1999 00:00	Benefit	270	50081	GARY	2	0.649216592	11/04/1954 00:00	F
8	A001221	24/03/1999 00:00	Benefit	225	50081	GARY	2	0.873540401	08/06/2018 00:00	F
9	A006543	24/03/1999 00:00	Benefit	255	50081	GARY	2	0.670783401	03/01/1956 00:00	F
10	A006308	25/03/1999 00:00	Benefit	270	50081	GARY	2	0.557811558	13/08/1960 00:00	F
11	A001201	26/03/1999 00:00	Benefit	255	50081	GARY	2	0.918260872	30/11/1918 00:00	F
12	A006432	26/03/1999 00:00	Benefit	270	50081	GARY	2	0.722981393	30/09/1931 00:00	F
13	A006334	27/03/1999 00:00	Benefit	270	50081	GARY	2	0.737888217	11/02/1939 00:00	M
14	A017058	27/03/1999 00:00	Benefit	225	50081	GARY	2	0.787204981	03/02/1966 00:00	M
15	A006458	27/03/1999 00:00	Benefit	255	50081	GARY	2	0.761739135	16/01/1927 00:00	F
16	A017023	27/03/1999 00:00	Crisis	200	50081	GARY	2	0.517142832	04/12/1970 00:00	F
17	A006431	27/03/1999 00:00	Crisis	100	50081	GARY	2	0.615896106	17/12/1958 00:00	F
18	A006574	28/03/1999 00:00	Crisis	200	50081	GARY	2	0.673347712	04/06/1955 00:00	F
19	A006346	28/03/1999 00:00	Benefit	240	50081	GARY	2	1.208198786	23/10/1923 00:00	F
20	A017075	28/03/1999 00:00	Benefit	255	50081	GARY	2	0.989515543	02/07/1925 00:00	F

	A	B	C
1	$CF	First use	Conditional Format Formulas
2	1	D2	1) > 300 Interior.ColorIndex=46(Orange)
3	2	I2	1) < 14611 Interior.ColorIndex=45(Light Orange) 2) > 36526 Interior.ColorIndex=4(Bright Green)
4	3	I8	1) < 14611 Interior.ColorIndex=45(Light Orange) 2) > 43831 Interior.ColorIndex=4(Bright Green)
5	4	I11	1) < 3654 Interior.ColorIndex=45(Light Orange) 2) > 36526 Interior.ColorIndex=4(Bright Green)
6			

Data Validation formula consistency

	A	B	C	D	E	F	N
1	Budget for 2008						
2							
3		Jan	Feb	Mar	Apr	May	
4	Dept A						
5	Abrasives	411	668	760	484	367	6807
6	Accounting	427	544	331	548	880	6579
7	Actuators	381	967	239	5	574	5695
8	Adhesives	828	540	241	179	740	7408
9	Advertising	494	567	191	700	325	6556
10	Air brakes	115	95	288	816	578	5263
11	Total	2656	3381	2050	2732	3464	38308
12							
13	Dept B						
14	Bags	594	608	392	76	77	4856
15	Barcodes	358	644	714	902	692	7002
16	Batteries	725	593	485	149	862	5797
17	Total	1677	1845	1591	1207	1631	17655
18							
19	Dept C						
20	Cables	398	172	471	924	803	6114
21	Capacitors	9	861	550	581	547	5635
22	Cassettes	348	543	697	94	320	4150
23	Ceramics	555	618	622	514	971	7191
24	Chemical agent	527	456	259	927	933	6553
25	Chipboard	927	150	853	160	767	5849
26	Computer suppl	46	328	805	150	174	3939
27	Total	2810	3128	4257	3350	4515	39431

A1 fx $DV_Data Validation in 'Budget08 (2)'

	A	B	C	D
1	$DV	First use	Data Validation	
2	1	B5	Stop Decimal > 0	
3	2	B14	Stop Decimal 0 .. 1000	
4	3	B20	Stop Decimal 0 .. 980	
5	4	M20	Stop Decimal 0 .. 980 Invalid!	

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 … ?