W4.2

What stands in the way of e-Government and e-Commerce? Bad data quality!

Adrian McKeon, Managing Director, Infoshare Ltd

Abstract

The world is becoming evermore data-centric. Previously standalone databases are now being linked together in greater numbers as the public sector embraces e-government, putting the customer at the centre. The same thing is happening as private sector partnerships exploit e-commerce.

Merging variable-quality data sets from multiple sources usually increases dirty data issues. Using an 80% accurate address index to validate an 80% accurate client list generates only 64% reliable intelligence i.e. 80% of 80%. With each data source added, inaccuracy further increases.

Huge amounts of public and private sector data relates to people and places i.e. who people are, where they live, and what services are delivered to them. Unfortunately for organisations wanting to target their services, customers' details change frequently – on average 1,000, 000 people, 100,000 postcodes and 60,000 organisations move each month.

21st century applications require precise information. In the 1990's, data about people and places accurate to postcode level sufficed. Today, "one to one" applications like eCRM, data warehousing and e commerce, require "one to one" data accuracy to sub building level. This demand highlights the importance of BS7666 (locations) and BS8766 (names) which underpin data accuracy at sub building level. These standards are local government biased yet Central Government and the private sector will be the biggest users.

Total solutions to the 21st century data quality problem must validate the currency of information in addition to defining locations correctly, and must support ongoing revision and maintenance of data. Case studies show how technology can be used to secure typical accuracy gains of 25-40% in matching public sector data sets - local authorities and emergency services - and private sector data sets - telcos and utilities. The paper ends with a look to the future.

How dirty data derails ecommerce/e government

The performance of even the most elegant IT infrastructure is directly dependent upon the quality of data held by the system. Whether the IT is a data warehouse, an e-CRM, or the latest thing in data mining and gis, the result is the same. Junk data in equals junk intelligence out. Indeed, according to CIO Magazine 50 – 70% of all CRM projects fail and 92% of data warehouses under deliver. Invariably end users/budget holders blame the IT when usually it is a data quality fault.

The trouble is that most 21st century solutions depend on information extracted from various 20th century data-clumsy legacy systems. If such legacy data is to be relied on it needs to be cleaned, corrected and validated. Otherwise, whilst high quality data in a new system doesn't guarantee success, poor quality data will go a long way to ensuring failure.

As local authorities around the world harness the power of the web to deliver the same 24/7 availability and convenience, fast delivery, customer focus and personalisation provided by the best online businesses, they are coming up against the same data quality barrier.

A good definition of data quality is “fitness for purpose”. For example, the London tube map serves travellers perfectly well with station names and tube routes but it would not help to buy or sell a house. This legal transaction requires precise location data. Data quality relates to how well data held assists in achieving 95% - 100% of objectives. Let us examine how “fitness for purpose” relates to data quality.

Fitness for Purpose

As data is moved from one application to another it is almost invariably found to be dirty, inconsistent and incomplete. What was “fit for purpose” in one application often needs a significant amount of “scrap and rework” for use elsewhere. This is the IT norm and most analysts have experience of it.

In the 1990’s, technology restrictions meant most end users only required precision to postcode level and above, and the accuracy achieved by validating internal data against common reference data sets such as PAF was more than “fit for purpose”.

(see Table 1).

Table 1

Example of reference data against which internal data is validated	Existing data quality level	Data quality level needed	Accuracy and precision needed to:	Type of software solutions available	Typical End users
	%	%			
ADDRESS-POINT Post Office Address File NLPG Electoral Roll Dun & Bradstreet business data Experian Consumer data	60 – 85	75 - 85	Postcode, EDU, Ward, Regional, National	Name and address merge/purge, data entry, geocoding routines i.e. QAS, Capstan, Hopewiser	Direct mailers, mailing list and other data providers, in house data entry/market analysis users
		85 - 99	Building	Solutions using off the shelf industrial strength components, consultancy expertise and requiring IT infrastructure changes i.e. Trillium, Vality, Innovative	In house teams and systems integrators (big 5, EDS, ICL, Cap Gemini) building data warehouse, crm, e commerce solutions
		85 - 99	Sub building	Industrial strength self learning, intelligent, automated data cleansing, validation and matching needing no consultancy or IT infrastructure change. i.e. Infoshare	In house teams and systems integrators Non experts with little technical expertise needing an affordable, effective off the shelf permanent solution

By 2001, significant advances in technology and massive reductions in cost had changed end user definitions of “fitness for purpose” of data. The affordability for all of sophisticated, computing power has seen an increase in demand for data which is “fit for purpose” and can support end user work at building and increasingly, sub building level. (see table 1) Certainly, if e government and e commerce is to work, data must be accurate, revised and maintained to sub building level. If popular analysis tools such as GIS are to deliver evidence based decision support, they’ll probably have to work off sub building level data, or data aggregated from that level – so it is possible to drill back down, when evidence is checked.

This shift in focus has served to highlight the technical complexity of what hitherto has been seen as a simple process – maintaining the currency of property level information. The media is full of failed IT projects where people “had a go” because it seemed easy. It has also served to highlight the importance of data standards, which underlie the supply of people/location data accurate to sub building level. BS7666 – the UK standard for defining the location of property and places – and BS8766 – the UK standard for defining the names of people – are both controlled by IDeA. IDeA focuses exclusively on local government. Yet these standards need to be developed to serve Central Government and the private sector, both of which exploit people and places data, and both of which are heavily involved in eCRM, data warehousing, advanced analysis and various web based activities requiring accurate intelligence.

Let us examine the barriers to data accuracy and currency.

Data quality barriers to “fitness for purpose”

As organisations can no longer rely on the cost of IT as a barrier to competition, they must increasingly rely for competitiveness, on the quality of intelligence they hold. Essentially, this means ensuring that data are accurate, precise, up to date and well maintained. As most data is about people and locations, this means getting names and addresses right at the name and address level of precision and maintaining the currency of this information i.e. who people are, where they live, and what services are delivered to them. Unfortunately customers’ details can change frequently. For example:

- Post Office Address File: 100,000 postcode changes each month
- AddressPoint: 50,000 addresses positioned every month
- Voters Roll: 1,000,000 people move each month
- Business Rates: 60,000 businesses change each month

Even if something as basic as getting a persons name and address right goes wrong, any initiative involving multi partnership/department co-operation is likely to fail, and any analysis using the data will deliver junk. Accepting without question the accuracy of data regarding people, and that taking action based on analysis of such data, is an unacceptable risk. All e-government/e commerce will do is speed up the transmission of incorrect information to a wider audience.

Three data sets currently exist in the UK against which addresses are cleansed and validated:

- A post office address file of all postal delivery points in the UK (Post Office)
- A geocoded address file of all addresses in the UK (Ordnance Survey)
- A national land and property gazetteer of all property locations (IDeA)

Increasingly rapid changes in our movements mean that, on average, none of the above exceeds 80 – 85% accuracy at any time. Their decline in “fitness for purpose” is proportional to increasing demands for intelligence that enables “one on one” targeting of people and places. Indeed, using an 80% accurate address index to validate an 80% accurate client list generates only 64% reliable intelligence i.e. 80% of 80%. With each data source added, inaccuracy further increases.

Typically, the more information a gazetteer contains, the greater the need to ensure data accuracy. Of equal importance is *how* a revision is made. To be concise, revisions must be in a common format or adhere to a common standard. For example, changes to "Acacia Avenue" input as "Acacia Ave." or “St Albans” which is often confused with “Albans St” may create whole new entries instead of revising old ones. Typical data problems include:

Data Issue	For example
abbreviations	*Male, female / m, f, / 1,2 / b, g / a, b, c* and so on
data entry accidents	Data in the wrong field or incorrectly spelt data
data hiding in data	Hidden characters embedded in a data record which, when output, automatically commands the operating system to do something e.g. send form to printer.
different phrases	ASAP; doing business as; round the corner from; c/o
duplicate records	The same data appears twice, or did the incident really occur twice?
incomplete records	Missing data whether deliberate or in error
irrelevant data	Data that adds no value to the intended process
localisation differences	Different departments use different location indicators i.e. postcodes, grid references, addresses, bus stops wards, edu's and so on
name conventions	Robert Smith Ltd, R Smith Ltd, RSL and so on, may lead to multiple database records for a single person
no data keys	No visible key to relate one record to anything else
non standard representations	There are in excess of 150 ways of representing the time and date, all are used regularly.
spelling variations	UK v US English terms
timing differences	Time dependent data may give two completely different data sets from the same database if downloads occur at different times
unique reference number systems	Different unique reference systems on different databases for the same record
unit differences	Different departments may describe labour in terms of man hours, weeks, months, full time equivalents and so on

Turbo charging applications by fixing the dirty data problem

First of all, you need to decide whether the intelligence you want to exploit is "fit for purpose". If it is then you are fine. If it is not, you need to define the problem. This could be a simple "lack of postcodes or geocodes". It could also be "I've data from 15 legacy systems to integrate, all the data formats are different and many records are full of errors and omissions". Then choose the tools you need to fix the problem. The market for data quality IT is split into 5 segments:

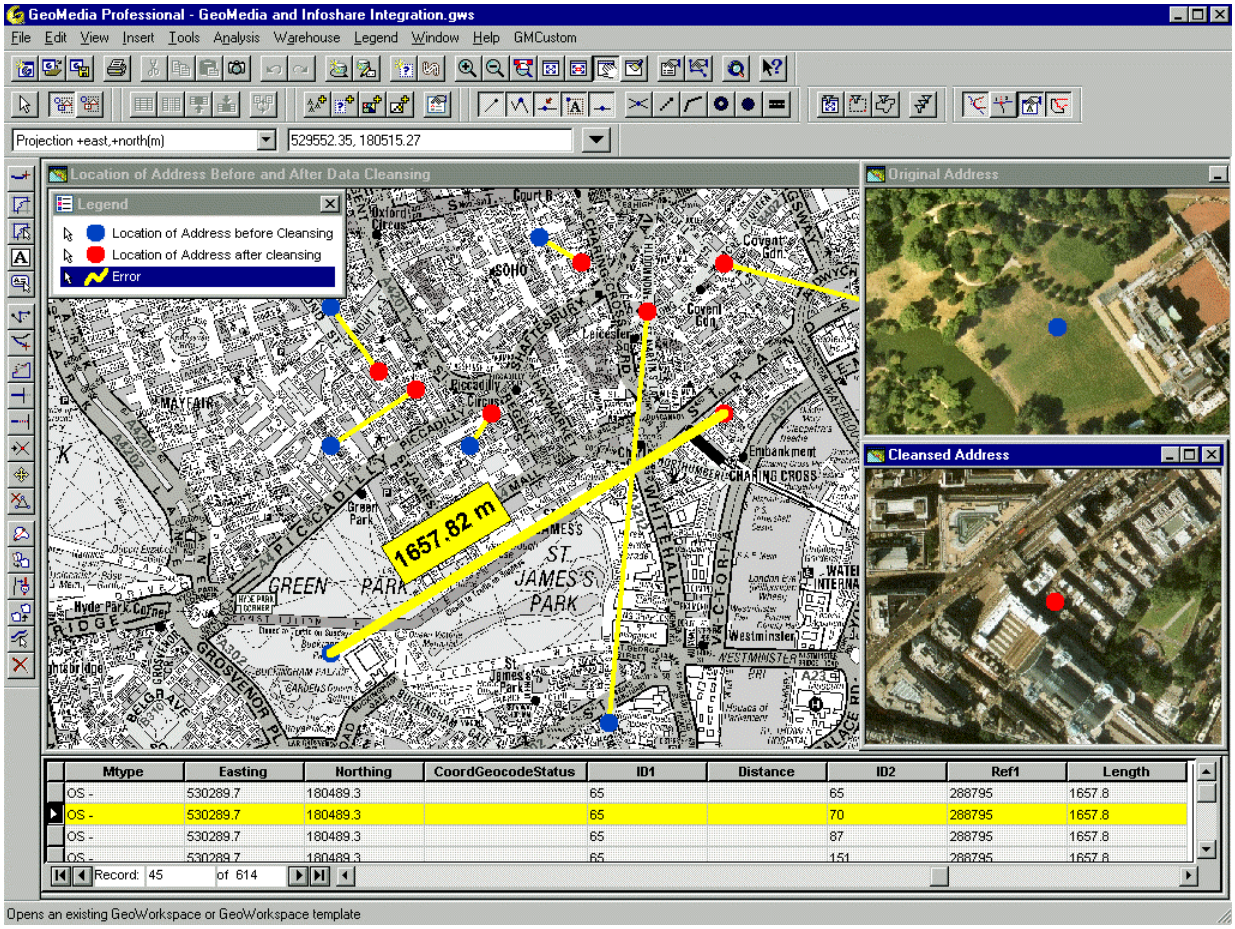
Classification		The sort of data quality problems each type of product can tackle	Typical suppliers
1	Information quality analysis products	Extract data, measure qualities such as validity or conformance to business rules, report analysis	Mobius Pine Cone Systems Rockwell Data Automation
2	Business rule discovery products	Analyse data to discover patterns and relationships, which define business rules as actually practiced.	Information Discovery Re-Genisys WizSoft
3	Data re engineering, cleansing and transformation products	Extract, standardise, correct, transform, enhance in preparation for data integration/migration to crm, warehouse etc	Infoshare Innovative Systems Trillium, Vality
4	Information quality defect prevention products	Merge/purge/de-duplication. Preventing data error at point of entry by applying business rules and quality tests.	AFD, Capstan, Hopewiser, QAS Systems
5	Metadata management and quality products	Managing the quality of data about data	Compedia Intellidex Kismet Analytic Corp

Some companies i.e. Innovative Systems operate in more than one segment so ensure the software you use is “fit for purpose”. Do you need a £250,000 data re-engineering tool to add postcodes or a £250 post coder? Ideally it should integrate seamlessly with your application thereby letting you focus on analysis instead of data processing.

Then choose the reference data against which you’ll validate your information i.e. PAF, AddressPoint. Bear in mind their currency and fitness for purpose. Your information quality advisor should be able to assist with this.

Finally, do not get too hung up on IT hardware/software, or about existing or planned new systems. The aim is to focus on data not on IT. Too often the purpose of a project to introduce evidence led decision-making is morphed by IT specialists into a project for *“linking disparate legacy systems via a network soup of technology and interfaces”*. Data quality is ignored; the result is “inaccurate data in: inaccurate intelligence out”, and a project fails or under delivers.

Resolve the data quality problem and then feed the intelligence into your application. Like the example below (with kind permission of Intergraph UK Limited) shows, it will turbo charge GIS analysis performance. This screen shot shows a map of Westminster, London, with the original and cleansed addresses and the distances between them. The highlighted record shows a discrepancy of 1657.82m. Without data cleansing, resources would have been targeted in the wrong area on the basis of this GIS analysis. Looking at the aerial photography we can see that the original address is in the gardens of Buckingham Palace!! The correct address, which has been validated, is actually just off the Strand.



The following public and private sector case studies from Infoshare illustrate the importance of resolving data quality issues.

Crime and Disorder: The Jupiter Project

Let us look at an application that relies on accurate address data. The Government Offices of the East Midlands and of the South West are piloting a crime and disorder project known as Jupiter. Jupiter regularly collects data from multiple agencies at local level (police, councils, health, ambulance, fire, probation etc). Using the unique, push button technology from Infoshare Ltd., the data is continuously cleansed, corrected, validated, revised and maintained to the highest level of quality (BS7666/BS8766 – national standards for locations and names) then integrated for analysis. Resources are then able to targeted based on hard evidence.

Data collected is accurate to household level and aggregated as necessary for data protection and data sharing purposes. Each record carries an audit trail and unique identifier recording its full history and origin, enabling aggregation for reporting purposes from local to regional to national level to be kept simple. Likewise, it is easy to follow the trail down from national aggregated statistics to individual local records.

The implication is that Jupiter offers a local test bed for developing and testing national policy using hard evidence. Feedback to inform the policymaking process can be daily, weekly, monthly or even hourly. The historical problem of data incoherence due to a dynamically changing local population change is therefore, no longer the difficult issue it once was.

That Jupiter currently focuses on crime and disorder related data is irrelevant; it clearly demonstrates the possibilities for processing any data related to citizens or service delivery. Jupiter is the only large scale regional project of its type in the UK whose evidence based targeting of resources is based on sub building data (where data protection does not apply) and postcode data aggregated from sub building data (where data protection does apply).

National Land and Property Gazetteer: Torbay Council

The drive toward local e government is encouraging local authorities to create a local land and property gazetteer. One function of an LLPG is to gain a single view of clients and services instead of accessing information from a range of, often conflicting, sources. The need to integrate variable quality data from multiple sources such as electoral register, council tax, planning, non-domestic rates, AddressPoint, National Street Gazetteer etc means this task is highly complex. Get it wrong and as well as the costs being huge, the e government agenda cannot be delivered. Cleansing and validating the data to sub building level (via Infoshare) has enabled Torbay Council to create a cross-reference index, which links all of its data together. Torbay was able to isolate good matches from questionable ones and focus on the later. It took 4 days to resolve data queries – instead of the 64 man-days, which had been allocated – and the LLPG was completed 2 months ahead of schedule.

National e government requires that the National Land and Property Gazetteer hub should match the version it has created of Torbay's gazetteer with Torbay's actual 95%+ accurate one. The hub has introduced around 30% more mismatches – probably due to duplicates and other inconsistencies. 50% of these NLPG errors can be handled by Torbay using automation, but unless the hub provides Torbay with resources to remove the other inconsistencies - which were, after all, created by the hub and not by Torbay - the NLPG will not be able to receive revision and maintenance data from Torbay, and will not be fit for purpose. The NLPG must be reprocessed to sub building level to remove such national:local discrepancies and avoid the extra costs involved.

Location Based Services

A major telecoms company planning to deliver location based services has discovered that data provided by content brokers is on average only 60-65% accurate and brokers are unable to tell which records are good, which are bad and what the data quality issues are. This directly risks the huge investment made in building the infrastructure to deliver the service – people will not pay for junk intelligence. Cleansing and validating the data to sub building level has enabled the company to pinpoint those records which are 100% accurate,

to build its service using those only, to return erroneous data to content providers for correction, to create a cross reference index linking data from multiple providers to a single feed, and to reduce the cost of content.

Motorsport mergers

Fierce competition for the leisure £ in the UK in recent years has required companies involved in motorsport to merge and gain economies of scale. One example resulted in 6 huge separate customer databases all providing the same marketing function. Managers did not know if private clients were replicated across the databases, which holding companies owned which subsidiaries in the corporate hospitality segments, which records were duplicates, or how much money was being wasted by replicating similar activities. Cleansing and validating the data to sub building level meant marketing activities could be based on 100% accurate data; records identified as queries could be corrected and added later, a cross reference index linking data from all sources could be built to feed the marketing effort and cut costs, and staff could be safely reallocated to different functions required as a result of the merger.

A look into the future

1 Revision and Maintenance of national data sets against which data can be validated

In the long term the ultimate reference database of people and places against which information can be validated will be a mixture of Ordnance Survey's digital national framework (including AddressPoint), PAF, the NLPG and the rolling electoral register.

In the short term, PAF and AddressPoint will remain the reference databases of choice for address locations, especially with Royal Mail's ongoing development of PAF and OS's move to quarterly updates for AddressPoint, and introduction of the DNF. IDeA's NLPG suffers from inaccuracy, revision and maintenance problems. If the same problems befall IDeA's rolling electoral register - which it aims to link to the NLPG - their "fitness for purpose" will be significantly reduced for the next few years. For example, ONS's Neighbourhood Renewal project will fail or significantly underperform if it relies on the NLPG alone.

2 Development of BS7666 and BS8766

The move in end user demand from Postcode level to building level to sub building level accurate data, driven by widely available sophisticated technology and moves toward e government and e commerce, raises questions about the direction of BS7666 and BS8766.

IDeA has developed these standards purely for local government, but as they become the basis for delivering accurate data to technology applications, the needs of Central Government and the private sector (each of which hold huge amounts of location/people data) need to be considered. IDeA decided not to include organisation names in the NLPG – thereby reducing it's usefulness to IDeA's own ER project, to a huge segment of UK plc and to a raft of other joined up government projects. Local Government beaurocrats will not be allowed to make similar decisions, which serve their own interests but which stifle private sector and central government use of these standards.

3 Data-led IT

The days of expensive big bang IT projects leading the exploitation of data are numbered. In future, data quality issues will be resolved first to create a database of validated intelligence.

Such intelligence will store original raw data, the cleansed and corrected version of the raw data, unique references to link data on disparate systems, audit trails recording changes made, business rules reflecting data owner policies, and data quality reports. This information will then be used by end users to produce extremely high quality evidence based analysis, and by IT experts to design and test IT infrastructures, inform business process decisions, and link legacy systems to the latest IT, thereby prolonging legacy lifespans.

4 Development of the legal framework

Data Protection, Freedom of Information and Human Rights will all have an increasing effect on what can and cannot be done with data about people and places. Neither government nor the private sector will be able to ride roughshod over the rights of private individuals. People will vigorously enforce their rights

using “no win: no fee” legal services either individually or via class actions. Organisations not tackling the data quality issue will increasingly face large litigation bills.

Although there will always be tension between the need for information and the need to respect personal rights, Infoshare’s experience is that legislation like Data Protection complements any data exploitation in the public interest. People are pragmatic. They will support action, which they believe to be in the public interest. Conversely, people will remove permission to use data about them if it is abused – already, people can opt out of any commercial use of the electoral roll.

5 Increasing levels of co-operation between public and private sector

White-collar crime is a huge problem. It is estimated to cost UK plc billions each year. The social costs in unemployment, lost opportunity, business failure, social deprivation and the higher cost of living are immeasurable.

Increasingly, data resources from the public and private sector will be combined and analysed to fight this problem. Such action will require increased co-operation and trust between the public and private sector and the continued introductio of supporting legislation such as RIPA to enable data sharing. This will be driven by public opinion, public interest and the need of highly influential national projects like ONS Neighbourhood Renewal to deliver. Bureaucratic inertia and bedded in vested interests against such co-operation will be swept aside. The drive to co-operate in tackling white-collar crime will initiate co-operation in other areas. The opportunities will be huge. They’ll all involve high quality data and high quality analysis including GIS and they’ll all be delivered over the web.

Adrian McKeon is Managing Director of Infoshare Limited, a data cleansing and matching solutions company. For further information please contact:

T: 020 8541 0111 F: 020 8541 4010

E: *amckeon@infoshare-is.com* W: *www.infoshare-is.com*