# The Equation between Semantics and Data Quality

**Stephen Wood**
Research and Development Department
Infoshare Ltd
Millennium House, 21 Eden Street
Kingston upon Thames, Surrey, KT1 1BL
United Kingdom
stephen.wood@infoshare-is.com
School of Crystallography
Birkbeck College, University of London
Malet Street
London WC1E 7HX
United Kingdom

**ABSTRACT**

I examine the equation between semantics and data quality and discuss suitable methods for data integration. As a rare example of a data quality based semantic technology, I take the ClearCore product from Infoshare Ltd.

I guide the reader through the process of using ClearCore and give examples from financial and retail services projects. The term ontology is found useful in describing the process of data integration, which comes to be seen as uncovering the ontology that underlies the data. The stages of validation and matching involve the specification of different kinds of business rules, constraining and generative respectively. The software helps to elicit these rules from human analysts and thus behaves as an expert system, accumulating a domain-specific knowledge base.

**Keywords**

Data integration – data quality – semantics – ontology – business rules – expert systems.

## INTRODUCTION

The equation between semantics and data quality is key to the veracity of business intelligence. Examples of data quality based semantics are rare. Infoshare's ClearCore product dates back to the company's start up in 1996, long before the current wave of interest in semantics. The core technology is found at more than 200 sites covering banking, telecomms, central and local government, and public safety. The technology's success demonstrates the relevance and market value of semantics and the importance of data quality. It also provides a vehicle to explore the semantics/data quality equation.

## THE MEANING OF DATA – SPURIOUS OR GENUINE?

According to an IDC report, 'An enterprise employing 1,000 knowledge workers wastes $48,000 per week, or nearly $2.5 million per year, due to an inability to locate and retrieve information' [1]. Semantic technologies offer a solution to this problem, by focusing attention on the meaning and context of the data, and links between different data items. In designing a new application, we would wish to put semantics in: to construct a system that makes the meaning of data items, business rules and relationships between different parts of the system explicit. If we are faced with making the best of existing systems, then we need to look at how to get the semantics out, at how to elicit meaning. Tony Picardi at IDC tells us that 15 years ago, firms spent 75% of their IT budget on new hardware and software and 25% on fixing the systems that they already had. Now 70-80% of IT spending goes on fixing things rather than buying new systems [9]. Therefore, the market for software to uncover the meaning of existing systems is likely to be much larger than the market for semantically aware replacements.

Dave McComb describes the process of uncovering meaning as part archaeology and part anthropology [15]. The archaeologist must look to users, requirements documents, existing systems, existing electronic data, existing paper data, industry literature and regulations. The existing data themselves are the most useful source of clues about meaning: 'The data in databases is a great clue to how people are really using the system … The existing systems often contain a wealth of data that the users are not aware of or forget to bring up' [14]. The anthropologist contributes insights in the basic types and distinctions involved, often highlighted by the consideration of boundary cases.

Many companies are looking at how to bring a new customer focus to their business. They have seen how getting a better hold of customer relationships enables them

to manage and retain those relationships in such a way as to maximize revenue. If a company is asking questions such as 'How many customers do we have?', 'What do our best customers buy?' or 'Why do customers leave?' then they are looking to uncover meaning that is currently locked away in the data held in existing systems. Getting to know your customers better is essentially a question of semantics.

In this context, the Organization for the Advancement of Structured Information Standards (OASIS) has formulated a set of standard ways of presenting customer information: names, addresses as well as miscellaneous information such as date of birth, telephone contact details and financial account details [18]. Customer names may be personal or organisational, and may involve complex relationships such as:

- Sakthisoft Pty. Ltd "TRADING AS" Mantra Corporation, or
- Mrs Mary Johnson and Mr Patrick Johnson "IN TRUST FOR" Mr Nick Johnson.

OASIS also identifies a key problem that besets any customer-focused initiative: There is no point in gathering data together on a customer from the different points of contact that he or she makes with the company, if those data are inaccurate: 'While data within [an] individual line of business systems may be fit for the purpose for which they were collected, combining data with different structures threatens the effectiveness of entire customer relationship initiatives. Adding even further to customer information quality degradation are errors that occur during data entry. As each error occurs, the ultimate effectiveness of customer data is reduced' [18, see also 2]. So we have to be aware that poor quality data will lead us to uncover spurious meanings and if we do we will hamper the customer relationships that we are trying to improve. Semantics is inseparable from data quality.

It is in this space that Infoshare positions itself. Any project that relies on data integration – single customer view, solvency risk evaluation, or business intelligence – will only prove cost-effective if the data quality problems are solved first. How can we uncover meaning, if we do not know which meanings are spurious and which are genuine? 'Infoshare focuses on measuring existing data quality in any IT and factually determining the gap between a company's current data accuracy and 100% accuracy. Only then can you determine how this "gap" translates to revenue loss, risk increase or degraded IT performance' [8].

## ONTOLOGIES AND SEMANTIC VERACITY

The concept of an ontology has emerged as very important one in the field of semantic technology. According to Tom Gruber, 'an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents' [3, see also 4]. The agents are said to 'commit' to the ontology. The ontology functions as a contract of shared meaning between participating agents, which allows them to communicate in a coherent and consistent manner. Agents written in different programming languages and operating on different platforms can all share the same ontology, which is IT-independent.

The World Wide Web Consortium has provided a number of use cases for web ontologies [7]. One such case is the construction of a multimedia catalogue. To allow easy searching of the catalogue, in ways that will make most sense to the user, items in the catalogue need to have rich semantic annotations. An ontology would then define a contract for these annotations to be agreed between the storage and search agents. Annotations could be content-specific, describing the subject, the setting or the participants, or media-specific, describing the media type or the length of the clip.

It would be possible to reuse a content-specific ontology to permit searches of all documents in the relevant domain. Thus if publicity announcements, reviews and academic criticisms subscribe to the same ontology as the annotation of the work itself, then all such sources could appear in the search results. So we can see how ontologies can be useful when putting the semantics in at the beginning, but how do they feature in getting the semantics out?

'Every (symbolic) information system (IS) has its own ontology, since it ascribes meaning to the symbols used according to a particular view of the world' [5]. From this perspective, the aim of uncovering meaning in legacy systems can be rephrased as an attempt to uncover the ontology shared by those systems.

The fact that OASIS has established standards for customer information shows that customer databases do tend to share a common ontology. (For the similarities and differences between an XML schema and an ontology, see [20].) The customer has a name, an address and one or more miscellaneous data items, such as date of birth, loan number, bank sort code and account number. Each personal name may specify a title, initials, forenames (first name, middle name), surname and suffix and have an identifiable gender. Organizational names may be compound, where one company "trades as" another, or acts as "administrators of" another. Addresses consist of the geographical components of building, street, locality and country, together with various annotations specific to the postal service, such as PO Box numbers and postcodes.

To properly recognize the different components of the customers' names, addresses and miscellaneous data, ClearCore consults various independent ontologies. In this way, the product assesses the semantic veracity of the data, the correspondence between the data and their real-world referents [10]. ClearCore cleanses address data by comparing them to entries in a gazetteer, which, for the UK, is generated from the Ordnance Survey AddressPoint file [17]. If a customer's address is given in one database as 110 Peter Hilton Court and in another database as 5 Agard Avenue, it is the gazetteer that can reveal that Peter Hilton Court is in fact a building located on street called

Agard Avenue. Cleansing also involves corroborative validity checks, revealing cases, for example, where the street is not associated with the postcode given in the customer address. ClearCore validates data items to controlled vocabularies, such as lists of valid ethnicity codes, male and female names, or genuine titles. Audit trails show the extent to which data items agree with the ontology, and the extent to which inconsistencies or omissions exist. Only accurate data take part in the integration process. An integration project that combines data of high precision, yet low veracity, is a recipe for garbage in-garbage out [11]. This is the reason that many integration projects fail [6].

In the context of the Semantic Web, ontologies put the semantics into the application from the start. However, in the context of data integration, an ontology, which embraces all distinguished instances of the company's customers, emerges as the *outcome* of the process.

## BUSINESS RULES WITHIN AN APPRECIATIVE SYSTEM

McComb describes two kinds of business rules: constraints and generative rules. 'Constraints are rules that prevent things from happening, such as updates to a database' [12, see also 16]. Commonly, a rule would constrain the number of occurrences, the type, the range or the logical connection of data items. ClearCore implements validation constraints out-of-the-box through its various configuration options. Constraints may be tailored to exact needs through scripting utilities.

In a retail services project, validation rules pick out illegal phone numbers and illegal dates of birth. Illegal entries are removed from the data to avoid spurious matches between customers. The first rule specifies that the age of the customer be in the range 18 to 100. There are two entries in the sample data highlighted below of customers aged 5 and 104, again both are probably data entry errors, especially as 1900 is the standard default year on many CRM systems. This may create issues with customers if their date of birth is requested for security reasons when calling the call center (see Figure 1).

The second rule detects any invalid telephone numbers such as all "1" or all "2" etc. An operator might enter such a number when the customer has left that field blank on their application form. This rule is implemented by comparing each number to a list of illegal combinations held in an external file.

Figure 2 gives a summary of the illegal values discovered by the two rules in the retail services project.

Now, let us turn to generative rules: 'The way applications, and specifically flexible applications built from business rules, create meaning is not through their ability to constrain updates; this merely ensures that whatever is presented is consistent. The way applications create meaning is through the judicious use of generative rules. The act of "creating meaning" consists of soliciting information from the environment (primarily users) or from inferences that allow the creation of additional information' [13].

The way ClearCore generates meaning is through matching rules. These describe the degree of similarity between two data items that an expert analyst would consider an acceptable guarantee of customer identity. For example, the analyst decides whether matching to an old address, or to a date of birth that is less than a week different, or to a telephone number that is one digit different identify customers within acceptable limits.

Figure 3 shows two matches from a financial services system. You will be able to see how the process has picked out matches between individual components of the address. The match between the addresses for Travelex is much better than the match for J A S Bowman, since many more components agree. Notice also how abbreviations in the second Travelex entry have reduced the match score on the name.

In deciding what is an acceptable match, an analyst makes a value judgment based on his appreciation of the quality, completeness and compatibility of the data sources. 'This 'appreciation' may be developed upon a wide range of activities such as past experience, personal preference, intuition, rules-of-thumb and what may be considered to be bias and even prejudice' [22]. The messy origins of the matching rules do not matter. The ClearCore GUI facilitates the entry of rules, their review and the visualization of their consequences. Not only are the software's judgments of fact (what has matched) transformed by the analyst's judgments of value (the matching rules), but the analyst may revise his judgments of the value of particular rules in the light of the software's judgments as to what has matched. This intertwining of judgments of fact and of value is characteristic of appreciative behaviour: 'The relation between judgments of fact and of value is close and mutual; for facts are relevant only in relation to some judgment of value and judgments of value are operative only in relation to some configuration of fact' [21].

The ClearCore software is not simply a tool for 'archaeology', for data mining alone. It also helps with 'anthropology', the elicitation of the business rules – for validation and matching – that underlie the system's effective use. As the software automates more and more of the analyst's rules, ClearCore comes to represent an expert system, in that it mimics the decisions made by human analysts to link data together from different sources. Daune West has shown how eliciting experts' knowledge of an application domain can best be understood as a process of appreciative inquiry [23]. This process can take account of factors of experience and intuition, which need not be reduced to pure logic.

| SURNAME | FORENAME | DOB | TITLE | GENDER | ADDRESS1 | ADDRESS2 | ADDRESS3 |
|---|---|---|---|---|---|---|---|
| SMITH | CAROL | 1900-02-25 | MISS | F | 299 COPPERFIELD | CHIGWELL | ESSEX |
| SMITH | KATE | 1999-12-28 | MISS | F | CHESTNUT COTTAGE | CAREW NEWTON | KILGETTY |

Figure 1. Invalid dates of birth from a retail services project.

| *Report* | *Sample 1* | *Sample 2* | *Sample 3* | *Sample 4* | *Sample 5* |
|---|---|---|---|---|---|
| Records processed | 40000 | 40000 | 40000 | 40000 | 34992 |
| Invalid age [less than 18] | 129 | 163 | 221 | 4509 | 7107 |
| Invalid age [greater than 100] | 1 | 1 | 2 | 1 | 5 |
| Invalid home phone number | 97 | 56 | 144 | 7703 | 5267 |
| Invalid mobile phone number | 47 | 27 | 73 | 6339 | 2185 |

Figure 2. Summary of illegal values in the retail services project.

| | |
|---|---|
| Name | TRAVELEX GLOBAL AND FINANCIAL SERVICES LIMITED |
| Address | PO BOX 36 WORLDWIDE HOUSE THORPE WOOD PETERBOROUGH CAMBS PE3 6SB |
| Name | TRAVELEX GLOBAL AND FIN SERVS LIMITED |
| Address | WORLDWIDE HOUSE PO BOX 36 THORPE WOOD PETERBOROUGH PE36SB |
| Analysis | Address match on PO Box, building name, postcode, street and locality |
| | Name match 68% |

Figure 3a. A name and address match from a financial services system.

| | |
|---|---|
| Name | JAS BOWMAN AND SONS LIMITED |
| Address | ARLESEY ROAD ICKLEFORD HERTFORDSHIRE SG5 3UN UNITED KINGDOM |
| Name | J A S BOWMAN AND SONS LIMITED |
| Address | ICKLEFORD MILL ICKLEFORD HITCHIN HERTS SG53UN |
| Analysis | Address match on postcode and locality |
| | Name match 100% |

Figure 3b. A second name and address match from the financial services system.

I elaborate on four ways in which ClearCore facilitates the process of appreciative inquiry:

- Learning from data which is processed
- Providing facilities for learning about new data
- Providing facilities for easily changing, adding and enhancing rules
- Identifying situations where new rules or information is required

**Learning from Processed Data**

The best matching and data integration is delivered not from a snapshot of the data at one time but also from a knowledge of the history of the data. ClearCore has the ability to save the history and record all alternatives of all data so when it matches data together, it takes into account past versions of the data and any other variations. Examples are person name changes, address changes, nicknames, colloquial names, known-as names, old phone numbers and old e-mail addresses.

### Learning about New Data

Users often come across new sources of data that they would like to integrate into the matching process. This may be new address locations or a list of locally used aliases; it may be a list of nicknames; and for phone numbers it may be a list of code changes. ClearCore provides facilities to easily incorporate this type of data to accommodate changes to the content or use of source data.

### New Rules or Data Required

ClearCore continuously generates statistics when processing and has the ability to monitor trends. It is possible [1]to configure ClearCore to take action if the results of processing are outside expectations. For example, if one finds that the percentage of valid email addresses falls unexpectedly, this could indicate a problem in the source data, such as missing data or a new data format. Changes in data over time can be identified and appropriate action taken – automatically, manually or interactively.

### Changing, Adding and Enhancing Rules

ClearCore processes data based on rules at every stage. Many of the rules have been developed over many years through Infoshare's experience in data quality. This experience has also led to the understanding that rules need to be easily changed and updated to reflect changes in the operating environment. One of the challenges is identifying those rules that need changing and then testing them. ClearCore provides many facilities for this process by readily identifying data which fall outside current rules or are on the borderline of rules. It is also possible to see the difference when rules are changed. In this way, when the need for a new or changed rule is identified, it can be implemented and tested using real data without affecting the data itself.

## CLOSING THOUGHT

Clay Shirky has pointed out that the Semantic Web demands considerable agreement between agents on what to communicate and how [19]. He then suggests that such a degree of agreement is unlikely except in narrow domains: 'like many visions that project future benefits but ignore present costs, it requires too much coordination and too much energy to effect in the real world.' If a prerequisite for the Semantic Web is completely clean, accurate data and universal consensus among all interested parties, it is unlikely that the vision will be realized. If we allow that we start from incomplete consensus and messy data and put in place a process where data quality can be improved and consensus reached, then we bring the vision of a Semantic Web closer to fulfillment.

## CONCLUSION

ClearCore from Infoshare is a tool for semantic elicitation, for uncovering meaning in existing information systems. Data quality is at the heart of this process. The tool may be used for any project that relies on data integration – single customer view, solvency risk evaluation, or business intelligence.

Data integration divides into two main stages:

1. Validation and cleaning of the source data against a range of input ontologies (for names, addresses and any miscellaneous data items).
2. Matching data items from different sources to create a single customer ontology, which embraces all distinguished instances of the company's customers.

ClearCore facilitates an iterative process, where analysts encode more and more of the rules they use to link data. This is a process of appreciative inquiry, in which the software's reality judgments interact with the analyst's value judgments in such a way that the software learns about the enterprise domain. Through its learning capability, ClearCore adapts to changes in the way data are used, to new data and to new business requirements.


## ACKNOWLEDGMENTS

This paper would not have been possible without input from Adrian McKeon, John Mole and Myles McKeown. I am very grateful to Daune West at Paisley University for sending me her papers on expert systems.


## REFERENCES


1. Feldman, Susan and Sherman, Chris. *The High Cost of Not Finding Information*. IDC, 2001. http://www.knowledge-wave.com/scripts-include/en-us/downloads/idcinfo2996.pdf
2. Friedman, Ted, Nelson, Scott D. and Radcliffe, John. *CRM Demands Data Cleansing*. Gartner, 2004.
3. Gruber, Tom R. A translation approach to portable ontologies *Knowledge Acquisition* 5(2): 199-220. 1993. Available at http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.
4. Gruber, Tom R. Toward principles for the design of ontologies used for knowledge sharing *International Journal of Human-Computer Studies* 43(5/6): 907-928. 1995.
5. Guarino, Nicola. Formal ontology in information systems. In N. Guarino (ed.) Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Italy, June 6-8, 1998. IOS Press, Amsterdam, 1998, pp. 3-15. Available at http://www.loa-cnr.it/Papers/FOIS98.pdf.
6. Hancock, Ellen: *A Recipe for Success*. Talk presented to CHAOS University, 1995. Standish Group, 1999.
7. Heflin, Jeff. *OWL Web Ontology Language Use Cases and Requirements*. Available at http://www.w3.org/TR/webont-req/.
8. Infoshare Ltd. *Data quality tools providing a route map of your data*. http://www.infoshare-is.com/company/index.html

9. Kluth, Andreas Make it simple *The Economist* 28th October, 2004. Available at http://www.economist.com/displaystory.cfm?story_id=3307363.
10. McComb, Dave. *Semantics in Business Systems*. Morgan Kauffman, 2003, p. 31.
11. McComb, Dave. *Semantics in Business Systems*. Morgan Kauffman, 2003, figure 3.7.
12. McComb, Dave. *Semantics in Business Systems*. Morgan Kauffman, p. 151.
13. McComb, Dave. *Semantics in Business Systems*. Morgan Kauffman, p. 152.
14. McComb, Dave. *Semantics in Business Systems*. Morgan Kauffman, 2003, pp. 159, 172.
15. McComb, Dave. *Semantics in Business Systems*. Morgan Kauffman, 2003, p. 167.
16. Nelson, Scott D. Rule *Engines Are the Next Battleground in CRM*. Gartner, 2004.
17. Ordnance Survey. *Address Point: Ordnance Survey's map dataset of all postal addresses in Great Britain.* http://www.ordnancesurvey.co.uk/oswebsite/products/addresspoint/.
18. Organization for the Advancement of Structured Information Standards (OASIS).*Xml Standards for "Global" Customer Information Management* http://www.oasis-open.org/committees/ciq/ciq.html.
19. Shirky, Clay. *The Semantic Web, Syllogism, and Worldview*. Available at http://www.shirky.com/writings/semantic_syllogism.html
20. Smith, Michael K, Welty, Chris and McGuinness, Deborah L. *OWL Web Ontology Language Guide*. Available at http://www.w3.org/TR/owl-guide/.
21. Vickers, Geoffrey. *The Art of Judgment*, Chapman and Hall, 1965, p. 40.
22. West, Daune. Knowledge elicitation as an inquiring system: towards a 'subjective' knowledge elicitation methodology. *Journal of Information Systems*, 2: 31-44. 1992.
23. West, Daune. The Appreciative Inquiry Method: a systemic approach to information systems requirements analysis. In Stowell, F. (ed.) *Information Systems Provision: The Contribution of Soft Systems Methodology*, McGraw Hill, London, 1995, pp. 140 – 158.