"In Cyberspace No One Can Hear You Scream …"

Dr. Christopher J. Pound

British Telecommunications plc.
Enterprise House, 84-85 Adam Street,
Cardiff CF24 2XF
United Kingdom
chris.pound@bt.com

Abstract

As the telecommunications industry endeavours to reinvent itself, the effective management and exploitation of *information*, data delivered in *context*, is now the key weapon in gaining and retaining customers. The data management challenges in an environment of massively growing data volumes and complexity introduced by distributed processing are outlined. A framework and methodology for the management of information is presented and the term *Context Data* is introduced.

1. Introduction

"I have a simple but strong belief. The most meaningful way to differentiate your company from the opposition…is to do an outstanding job with information. How you gather, manage, and use information will determine whether you win or lose…You need a fast flow of good information to streamline processes, raise quality and improve business execution." - Bill Gates[1]

As globalisation takes hold against a backdrop of accelerated convergence of the internet, communications and entertainment industries, an unprecedented wave of mergers and acquisitions amongst telecommunications providers (telcos) is occurring. These telcos are currently endeavouring to reinvent themselves – a future based solely on the transmission of bits (POTS – Plain Old Telephony Services) is no longer tenable. As network bandwidth becomes commoditised, telcos now regard networks as platforms for hosting value-added products and services offering higher margins than simple POTS.

Against this background, arguably the greatest assets that telcos possess are not the physical switches, access and transmission networks comprising copper twisted pairs and optic fibre, but the *information* that these networks and associated systems generate. It is this information that enables telcos to gain competitive advantage by understanding usage patterns and behaviours in order to refine and target product and service offerings to existing and potential customers.

2. 'The Challenge Facing Telcos'[A]

Telcos generate immense volumes of data e.g. British Telecommunications (BT) currently holds a staggering 60 Terabytes of data on disk (excluding PC's and LAN servers) and 1.9 Petabytes on tape. As complexity increases in the networked world and vast numbers of data-generating devices are introduced, e.g. through growth in mobile computing, data volumes and data distribution are set to explode. The situation is compounded by the fact that most traditional telcos are battling to break free from a legacy of product-centric platforms and systems, where data was viewed and managed as a by-product, in order to satisfy the rapid growth in demand for customer-centric communications solutions. In the spirit of the tagline from Ridley Scott's movie *Alien*: 'In cyberspace no one can hear you scream …'

Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, 1999.

[A] The views expressed do not necessarily reflect or imply BT policy or strategy.

Whilst the volumes of data persisted in the network are relatively low, the volumes of data stored long-term in operational data stores (ODS's) and data warehousing (DW) environments is large e.g. BT collects and stores in excess of one terabyte of raw call data records (CDR's) per month. Telcos have made substantial progress in consolidating data in ODS's to provide a customer-centric view. As e-commerce 'zero latency' channels open up, customers, suppliers and partners are increasingly being provided Internet access to telco ODS and DW data to place orders, view billing data etc. Transaction rates will be high. BT's Customer Service System, which covers PSTN service, experiences 80 million transactions a day, 8 million transactions per peak hour.

As data transmission volumes overtake voice, IP-based networks will increase substantially. The IP equivalent of CDR's will need to be captured (e.g. from WEB-server logs), stored, summarised and analysed in ODS and DW environments. These trends will require technologies that improve the synchronisation of network data (typically held in object databases) with ODS and DW data (typically held in relational databases). Longer-term, distributed DBMS capable of serving both transactional and decision support query access will be required in order to avoid unnecessary data duplication.

Telco networks will increasingly be viewed as an integral part of extended information supply chains. Existing approaches to data management will be replaced by concepts such as subscription to parameter-driven data services with associated data service tariffs. Data consumers will select the quality, accuracy, currency etc. level they are prepared to pay for. The data service will then select the appropriate protocol to deliver the data.

3. *'Data into Information'*

In recognition of the role of information as a key business asset, a framework and methodology has been developed within BT to support the translation of data into information. The requirement was for an environment that promoted the delivery of data accompanied by sufficient contextual detail to enable both business and technical data consumers to translate the data into information. Operational implementation of solutions developed in BT's Marketing Systems environment in 1996 has led to subsequent wider adoption across the business.

3.1 *'The Problem With Meta-data'*

From the outset it was recognised that 'traditional' meta-data, with its emphasis on technical systems definition and description, was too narrow in scope. Furthermore, meta-data management tools and repositories available in the marketplace at the time were firmly aimed at the information systems development professional. Indeed the very term 'meta-data' sounds overly technical and confusion surrounds its precise definition.[2] Furthermore, what is regarded as meta-data by one person may be regarded by another person as data. The word Metadata® has even been trademarked[B] – hence the usage of the hyphenated form 'meta-data' in this paper.

Devlin[2] defined *information* as "data in context". The following example serves to illustrate the range of contextual data that a market analyst may require access to in interpreting a report containing response data from a marketing campaign: descriptive 'look-up' values for 'Standard Industry Classification' code values[+]; definition of 'Region'*; data type and range details for 'Percentage Response'*; business rule used for calculating 'Percentage Response'; data source, accuracy and currency; data quality issues applying to the data analysed; Campaign Brief document; video clip of the TV commercial etc… Where: [+] = reference data; * = 'traditional' meta-data.

The technology exists today to access all of these items in a digital form from hyperlinks embedded 'in context' within a WEB-browser delivered report. It is proposed that the term *'Context Data'* is adopted for this class of data and is defined as 'anything that provides meaning to data'. The definition is deliberately open; the key point is that context data is broader in concept than, but can include, 'traditional' meta-data. As with meta-data, context data is perspective-based and multi-tiered: what is regarded as context data by one person may be regarded by another person as data.

3.2 A Framework for Delivering Data in Context

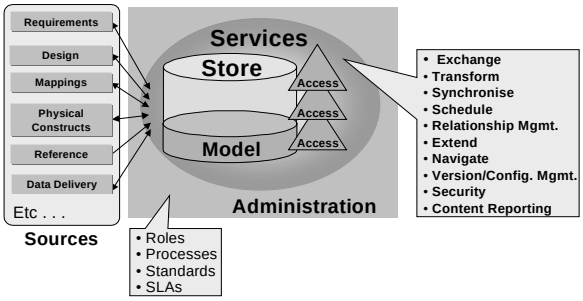

Fig. 1 – Context Data Management Framework

With reference to figure 1, the following principles were adopted during the design and development of the solution:

[B] The mark METADATA® was registered in 1986 in the United States of America Patent and Trademark Office as U.S. Trademark Registration No. 1,409,260 by Megadyne Information Systems; the Metadata Company is the current owner of the trademark. The trademark was granted "Incontestable" status in 1991.

Access via a WEB-browser was a given due to the pervasive use of a corporate intranet within BT. A single 'plain vanilla' user interface was considered inadequate; the user interface must reflect the role and objectives of the information consumer and dynamically adapt to reflect navigation and usage patterns over time.

Sources encompass any potentially relevant digitally available content available to the browser. A survey by the Meta Group in 1997 revealed that between 80% and 95% of large organisation's information was held in an unstructured format. As can be seen from the marketing campaign analysis example given above, access to unstructured context data sources is essential.

The *store*, which architecturally may be physically distributed, is analogous to a traditional meta-data repository and includes a meta-*model* schema / information medol. The *store* contains both context data *content* and *relations*, and has the following characteristics:

- The *content* is minimal, tending to be restricted to 'traditional' meta-data representing entity-relationship models and associated DBMS schemas. The reasons for keeping *store* content minimal are two-fold: (i) the *store* presents a performance bottleneck within the constraints of current technologies; (ii) organisational culture frequently militates against central 'control' of information.
- The *relations* are of two types: (i) *tightly-coupled*, facilitating navigation between instance context data items held *within* the *store* – typically to navigate between meta-data instances within entity-relationship and DBMS schema models; (ii) *loosely-coupled*, to facilitate the linkage of disparate context data sources e.g. a hypertext link from a data item reference embedded within unstructured text to associated data quality, currency and accuracy details applying to the data item but physically stored elsewhere.

A range of *services* are used in order to synchronise, exchange, transform etc. context data between point *sources* via the *store*. This service-based approach facilitates architectural separation between server-side functionality and storage from the browser-based user interface within a component assembly development environment.

The success of the project overall was largely not due, however, to the relatively simple technology deployed – the key proved to be in ensuring that the *administration* environment was in place i.e. roles, responsibilities, standards and Service Level Agreements were clearly defined and adhered to over the long-term. Particular emphasis was placed on ensuring the most appropriate person, irrespective of business or technical organisational alignment, was assigned responsibility for content creation and update. Root-cause analysis of meta-data repository implementation failures across industry revealed that inappropriate assignment of the content creation and update responsibility to the IS Department frequently mitigates against long-term success.

The design framework adopted can be distilled into three dimensions:

- *Functional* areas supported e.g. data definitions, data quality issues, operational metrics …
- *Content* held by system and/or business subject area
- *Role* and objectives of people accessing data in context – a spectrum ranging from knowledge workers 'surfing' context data in a discovery mode, through to data consumers who require the ability to drill-through from data in WEB-delivered reports to associated context data.

3.3 Methodology

The framework outlined above facilitated delivery of the technical solution and population of content to targeted roles via incrementally building outwards along the function / content / role dimensions. A methodological approach, adopting the central tenets of the Dynamic Systems Development Method[3], was evolved which has the following characteristics:

- continuous data consumer involvement during the design, development, testing and deployment phases
- iterative prototyping
- rigorous requirements prioritisation
- time-boxing

Delivery phases of three months duration with fixed end-dates were set. Rigorous prioritisation of requirements by the executive sponsors enabled requirements to be added / modified during a development phase; delivery end-dates were always met as lower priority requirements were rolled-over into subsequent delivery phases.

3.4 Context Data Futures

The following issues will need to be tackled if the context data framework is to be further developed:

Connecting Context Data With DSS / Data Warehouse Data: Although techniques and tools are now available for the sophisticated versioning of context data, in the DSS / data warehousing environment data and context data frequently become disconnected. A DSS table row will generally comprise fragments of data that reflect a

long lineage of data merged, spilt and transformed during its passage through multiple source systems. Looking down a DSS column, meta-data is not constant until you get to the level of row-sets where all the rows have a common temporal lineage. Data warehouse designers are not always successful in their attempts to 'normalise'/baseline data in columns to achieve a single column meta-data definition. Although a full history of meta-data versions over time may be available, source systems (typically OLTP) data will have long since been archived / deleted. In other cases, current tools simply do not enable meta-data to be recorded at the row-set level within columns.

Recovering Business Rules from Legacy Information Systems: When analysing the context and lineage of data, gaining an understanding of the business rules that have been applied to the data is crucial. In most cases these business rules are 'locked' in legacy system source code. Techniques and tools to recover these business rules are in their infancy[4].

Versioning and Access Authority Granularity: Techniques and tools have been developed that allow meta-object instance level versioning within current meta-data repositories. Consider the following scenario. In ensuring that the most appropriate people are authorised for content creation and amendment of context data (see above): for 'Product Group', a business person would maintain the meta-object property 'Description', a technical person the meta-object property 'Data Type', whilst different Product Managers may be responsible for 'Allowed Value Descriptions' … by range i.e. person one for code values 1 thru' 20, person two for code values 21 thru' 40, etc. Meta-object property (by range) instance versioning granularity is beyond the capabilities of currently available meta-data repositories, particularly those with relational DBMS engines where *n*-way joins typical of meta-data repository queries impose severe performance penalties.

XML-based Meta-data Standards: Significant progress has been achieved recently in converging meta-data standards. A joint press release issued by the Meta Data Coalition (MDC - which now also maintains the Microsoft produced Open Information Model) and the Object Management Group on 20th April 1999 announced plans for a "co-operative relationship to build consensus on metadata standards" – an initiative that spans all the key commercial vendors of meta-data solutions. It remains to be seen whether full meta-data interoperability will be achieved between the CORBA-based XML Metadata Interchange (XMI) and COM-based XML Interchange Format (XIF) standards. Finally, a gulf still exists between the MDC-OMG axis and the W3C XML Resource Description Framework (RDF) meta-data standards. Work to define RDF schemas is ongoing e.g. concerns surround the relatively simple constraints currently applied to the Dublin Core schema[5].

4. The Challenge Ahead

In a survey of major companies, McKean[6] established that of the major determinants of competency (people / process / organisation / culture / leadership / information / technology) in those who had successfully implemented customer information initiatives, technology only accounted for 10% of the competency determinant total; paradoxically, 82% of historical investment was in technology. The message is clear: IT designers and developers must re-focus their approach and provide solutions that more readily address the constantly changing people / process / organisation / culture / leadership environment that technologies will be embedded in. Of critical importance, investment in IT will be increasingly measured in terms of its impact on improving and exploiting the information assets of companies and organisations – speed to market and improved functionality are no longer sufficient to guarantee long-term survival.

Acknowledgements

I have been fortunate in working with a wide range of people who have contributed to the development of the ideas presented in this paper. Jeff Lee, Nigel Turner, Mike Kellett and Darryl Benjamin at BT and Subhash Chowdary (VIT Inc.), Prof. Keith Jeffery (CLRC Rutherford Appleton Labs.) and James Jonas (Oracle Corp.) have been particularly influential. Thanks are extended to Dr. Mike Revett and Dave Freestone of BT and Dr. Steve McKearney (Bournmouth University) for discussion of VLDB challenges facing the telecoms industry.

References

[1] Gates, W. H. & Hemingway, C. 1999. **Business @ The Speed of Thought – Using a Digital Nervous System**, Warner Books, p. 3.

[2] Devlin, B. 1997. **Data Warehouse: From Architecture to Implementation**. Addison Wesley Longman, Inc., 432 pp.

[3] DSDM Consortium. 1995. **Dynamic Systems Development Method – Version 2**. Tesseract Publishing, Farnham, UK., 240 pp.

[4] Shao, J. & Pound, C. J. (in press). Extracting business rules from information systems. *BT Technology Journal.*

[5] Jeffery, K. G. An Architecture for Grey Literature in a R&D Context, accepted for publication in **Proceedings GL'99 Conference**, October 1999.

[6] McKean, J. 1999. **Information Masters : Secrets of the Customer Race**. John Wiley & Sons Ltd., 287 pp.