METADATA: AN OVERVIEW AND SOME ISSUES

Keith G Jeffery

Head Information Systems Engineering Division, CLRC-RAL, UK

kgj@rl.ac.uk

1. INTRODUCTION

1.1 Problem Statement

There are large problems for information systems today. There is a need to somehow manage / exploit the explosion of information appearing on multiple WWW sites with very variable standards of data quality and currency - and of course there is the need to know such data sources exist. This has the major aspects of:

(a) data quality

(b) query quality

(c) answer quality

(d) integration of heterogeneous sources

Let us consider each briefly in turn.

1.1.1 Data Quality

Data quality can be improved by better data collection facilities (including help and explanation with examples) and better validation controlled by constraints, with automated conversions of unit values if required or necessary.

1.1.2 Query Quality

Query quality can be improved not only by classical query optimisation using knowledge about the database size and structure, but also by assisting the user to formulate the query best to meet the requirement - by means of online help, explanation, examples.

1.1.3 Answer Quality

The answer to a query commonly includes values and structures that are unfamiliar to the user; explanations and help, hyperlinked descriptions of units, precision, calibration or of similar terms could help the user to understand better the results.

1.1.4 Integration of Heterogeneous Sources

First, there is the need to know that a source exists and to know something of its characteristics. Heterogeneous data sources commonly have disparate schemas and there is a need to

understand the differences, even when apparently reconciled by one of the integration techniques.

1.2 The solution - Metadata

For all of the above to be realised, there is one essential and common ingredient: metadata. Let us consider briefly how it may be used in each of the cases:

1.2.1 Metadata for Data Collection

Metadata is necessary for validation through schema and constraints, using value-sets and domain range limits and even more sophisticated logic tests [GoGlJe93]. It is necessary for online help / explanation, and - in the form of a multilingual thesaurus - for translation.

1.2.2 Metadata for Queries

Metadata is necessary for validation through schema and constraints, for online help / explanation, for translation, and for optimisation: both user assistance in proposing more appropriate terms (synonyms, super- / sub-terms) and performance optimisation [JeHuKaWiBeMa94] since the metadata stores the structural indexes into the databases, optimal access paths, optimal query segmentation and distribution for parallelism and minimal network transfers.

1.2.3 Metadata for Answers

Metadata from the schema and associated metadata as domain ontology information (in a KBS) is necessary for answer consolidation, for online help / explanation, and translation [MaBeWiJeKaHu95].

1.2.4 Metadata for Integration

Metadata can catalogue sources of information at a high level so that they become visible. The well-known web indexing systems such as [AltaVista] or [Excite] do this in a very general way. An example in the field of CRIS (Current Research Information Systems) is the Bergen system [BergenCRIS] which points to structured informaiton systems for CRIS. Metadata, when used with an inferencing mechanism, is the key resource to find matching data structure and content despite heterogeneous representations and languages so allowing integration across heterogeneous data sources [JeHuKaWiBeMa94], [KoJe95].

Similarly, metadata provides the information necessary for customisation of standard products allowing integration into the desktop / office environment.

2. METADATA

2.1 Introduction

At present most Information Systems make very limited use of metadata and - since it supports all the user-friendly easy-to-use features and extends the range of available information features outlined above - perhaps this explains why these Information Systems have been less successful the few information systems really using metadata. Having outlined above how useful

metadata could be for Information Systems, let us consider exactly what metadata is. The aim is to decide what kinds of metadata are useful for Information Systems and how best to generate, maintain and use metadata for the benefit of end-users of Information Systems.

Metadata is data about data. It is therefore of great utility:

(a) to any Information System which aspires to be more than a simple, inflexible unfriendly information source - use of metadata can allow dynamic optimisation and flexibility and allow integration over heterogeneous distributed information;

(b) to any end-user requiring help, explanation, data quality assurance, assistance in finding relevant information, assistance in integrating information from heterogeneous sources.

Distributed RDBMSs use metadata extensively [Je96]. WWW indexing systems (such as [AltaVista], [ExCite]….) are based on sophisticated metadata. Metadata is clearly of great importance. Perhaps the earliest use of metadata was in computerised library catalogue systems based on IR techniques where the catalogue card record is metadata describing the 'real data' in the book or other primary publication. Sadly this same field of endeavour is where metadata has hardly been developed further, and yet this is the very area of Information Systems technology where metadata could exert the greatest leverage.

2.2 Standards

There have been many attempts to standardise metadata structure and content for specific application areas. In the world of libraries the [MARC] standard for catalogue records allows some interworking. Unfortunately there are more than 50 major variants and so interworking is not as easy as one might expect. Similarly, in many scientific areas -e.g. space science, particle physics - there are metadata standards. In the world of commerce there is [EDI] / [EDIFACT]. There have been attempts to agree a standard European Patient Medical Record. Perhaps the most successful is in the field of Engineering: the EXPRESS language describing the STEP data exchange format with commercial support [STEPTOOLS].

The increasing requirement for interworking among systems handling grey electronic literature has caused the internet community to propose as a metadata standard the [Dublin Core] and, subsequently to provide convertors between the standards [UKOLN]. In the field of CRIS a common metadata form for exchange has been proposed [van Woensel] and is now used for metadata catalogs in the ERGO Project [Finch].

The great spread of WWW has increased dramatically the requirement for metadata standards to allow a global browsing and querying capability. The creation of [W3C] (World Wide Web Consortium) provided the forum for intense work on metadata [W3Cmetadata]. The main results have been PICS (Platform for Internet Content) which allows categorisation of WWW information in a way similar to film censorship, and following the Netscape MCF (meta content framework) and Microsoft XML-Data proposals, the W3C standard named RDF (Resource Description Framework - which is XML based) has gained widespread acceptance and subsumes PICS.

2.3 Kinds

Here we propose that there are three main kinds of metadata: schema, navigational and associative.

2.3.1 Schema

Schema metadata is an intensional description of extensional instances. Typically a schema consists of: database {name, size, security authorisations}, attributes {name, type, constraints}. Some of the constraints concern the attribute domain, some are inter-attribute and as such may express relationships.

The schema intension has a formal logic relationship to the data instances. This is important in ensuring data quality. It also provides a formal basis for systems.

2.3.2 Navigational

Navigational metadata provides information on how to get to an information resource. Mechanisms include: filename, DB name + navigational algorithm, DB name + predicate (query), URL (Uniform Resource Locator), URL + predicate (query) or various combinations of them. They may also be obtained via a web-indexing mechanism (such as [AltaVista], [ExCite]…) which themselves make extensive use of metadata. Navigational metadata has no formal logic relationship to the data instances.

2.3.3 Associative

Associative metadata provides additional information for application assistance. The assistance may improve performance, accuracy or precision of the system and / or provide assistance to the end-user through a domain aware supportive user interface. The main kinds of associative metadata are:

(a) descriptive: catalogue record (e.g. [Dublin Core])

(b) restrictive: content rating (e.g. PICS) or security, privacy (cryptography, digital signatures) [W3C]

(c) supportive: dictionaries, thesauri, hyperglossaries [VHG], domain ontologies e.g. [PROTÉGÉ]

Associative metadata usually does not have a formal logic relationship to data instances although there may be systematic association relationships.

3. METADATA AND DATAWEB

3.1 Dataweb Technology

In order to combine the benefits of universal access (WWW) with the benefits of data managed and with structure and quality in a database various teams have worked on linking WWW and Database systems. CLRC-RAL was early into this field and experimented with several techniques since 1993, currently basing the departmental web [DCIweb] on Microsoft ASP

technology. Now much of the information available over the web is held and managed within databases linked to the web through CGI (Common Gateway Interface) and scripts in a language such as Purl or Tcl.

3.2 The Problem

The data in these structured databases behind a web interface is essentially invisible to web indexing systems such as [AltaVista] or [Excite]. Since this is usually structured, managed, high quality data its use might be preferable to authored html pages. The problem is how to make it visible to web-indexing or information-cataloguing systems, in a way that is universally acceptable and utilised.

4. CONCLUSION

The key to the Future of Information Systems is Metadata. However, there are serious issues to be addressed:

(a) standard form for metadata: the W3C RDF is general and uses XML as the language - is this sufficient?

(b) sub-forms of metadata by application domain: will they all be based on the same basic data model and language to allow cross-domain interoperation?

(c) Progressively dataweb technology is being adopted; is there a standard mechanism for making such structured and hopefully quality information sources visible on the web through metadata?

References

[AltaVista] http://www.altavista.com/

[BergenCRIS] http://www.nsd.uib.no/english/research/eucris/

[DCIWeb] http://www.dci.clrc.ac.uk/

[Dublin Core] http://purl.oclc.org/metadata/dublin_core/

[EDI] http://www.dsii.com/edihome.html

[EDIFACT] http://www.r3.ch/sjwg/

[ExCite] http://www.excite.com/

[Finch] For details of this project please contact peter.finch@lux.dg13.cec.be

[GoGlJe93] Goble C A;Glowinski A;Jeffery K G: 'Semantic Constraints in a Medical Information System' Proceedings BNCOD-11 'Advances in Databases' July 1993 pp40-57 Edited by Worboys,M and Grundy,A F; Lecture Notes in Computer Science Series 696, Springer Verlag, 1993

[Java] http://www.javasoft.com/

[Je96] Jeffery,K.G: ‘Distributed Database: The Issues’ Invited Lecture DATASEM’96. Datasem’96 Proceedings pp 65-92, Ed Sbornik Prednasek, CS-Compex a.s., Olomoucka 84, 61800 Brno, Czech Republic. October 1996

[JeHuKaWiBeMa94] Jeffery,K G; Hutchinson,E K; Kalmus,J R; Wilson,M D; Behrendt, W; Macnee, C A: 'A Model for Heterogeneous Distributed Databases' Proceedings BNCOD12 July 1994; LNCS 826 pp 221-234 Springer-Verlag 1994

[KoJe95] Kohoutkova, J; Jeffery, K G : Hypermedata: Interoperability for Healthcare Systems ' Proceedings MBB’95 Slovakia, September 1995

[MaBeWiJeKaHu95] Macnee,C A; Behrendt, W; Wilson,M D; Jeffery, K G; Kalmus, J R; Hutchinson, E K: 'Presenting Dynamically Expandable Hypermedia' Information and Software Technology 37 (7) pp 339-350 1995

[MARC] http://minos.bl.uk/services/bsds/nbs/marc/commarcm.html

[PROTÉGÉ] http://smi-web.stanford.edu/projects/protege/Hpkb-web/MusenWestKickoff/tsld001.htm

[STEPTOOLS] http://www.steptools.com

[UKOLN] http://www.ukoln.ac.uk/metadata/interoperability/

[vanWoensel] van Woensel, l: ‘CERIF Manual’ October 1988 ; for more information contact lieve.vanwoensel@lux.dg13.cec.be

[VHG] http://www.venus.co.uk/vhg/

[W3C] http://www.w3.org/

[W3Cmetadata] http://www.w3.org/Metadata/