METADATA:

The Future of Information Systems

Keith G Jeffery

Director, Information Technology, CLRC-Rutherford Appleton Laboratory

Chilton, Didcot, Oxfordshire, UK

Abstract

Information systems are increasing in complexity. There are greater volumes of data, users, processes and transactions. There are greater interdependencies between components. The range of available storage, user interface and computing devices is increasing so producing heterogeneity at the physical system level. The utilisation of multiple information sources to solve a problem (or create an opportunity) creates a need for homogeneous access over heterogeneous information sources. The optimal utilisation of multiple computing resources demands the creation of a uniform computing landscape. The key to homogeneous access to heterogeneous resources (not only information) lies with metadata. The future of advanced information systems depends on metadata. Metadata is the core of the emerging UK GRIDs project.

1 Introduction

The title makes an assertion that metadata is the future of information systems. The purpose of this paper is to support that assertion. The premise is that metadata is an essential, and the most important, component in advanced information systems engineering. The topic of Metadata has recently found the limelight, largely due to a sudden realisation of its necessity in making the WWW (World Wide Web) usable effectively. Metadata (data about data) is essential for WWW to scale, for finding information of relevance and for integrating together data and information from heterogeneous sources. Metadata is essential for refining queries so that they select that which the user intends. Metadata is essential for understanding the structure of information, its quality and its relevance. Metadata is essential in explaining answers from ever more complex information systems. Metadata assists in distilling knowledge from information and data. Metadata assists in multilinguality and in multimedia representations. The engineering of systems from components (data, processes, software, events, subsystems) is assisted by metadata descriptions of those components.

Metadata has been used in information systems engineering for many years – but usually in a specialist, one-off and uncoordinated way. Commonly the metadata has been human-readable but not specified sufficiently formally, nor accepted sufficiently widely, to be interpreted unambiguously by IT (Information Technology) systems. The ubiquity of WWW, the increasing need for access to

heterogeneous distributed information and the increased use of multilingual and multimedia sources all demand some common representation of, and understanding of, metadata.

Metadata is attached to data to aid in its interpretation. Metadata processing systems interpret the data using the attached metadata. In addition to information systems such as WWW (update, retrieval) and systems engineering as described above, metadata is essential for electronic business from advertising through catalogue information provision through initial enquiry to contract, purchase, delivery and subsequent guarantee or maintenance.

Metadata is like the Rosetta Stone – which provided the multiway translation key between Greek, Demotic and Hieroglyphics – or, with an associated processing system, like the Babel fish [1]. To quote from [2] "The Babel fish is small, yellow and leech-like, and probably the oddest thing in the universe. It feeds on brainwave energy received not from its own carrier but from those around it. It absorbs all unconscious mental frequencies from this brainwave energy to nourish itself with. It then excretes into the mind of its carrier a telepathic matrix formed by combining the conscious thought frequencies with nerve signals picked up from the speech centres of the brain which has supplied them. The practical upshot of all this is that if you stick a Babel fish in your ear you can instantly understand anything said to you in any form of language."

2 What is it?

2.1 Metadata

Metadata is data about data. Metadata can describe a data source, a particular collection of data (a file or a database or a table in a relational database or a class in an object-oriented database), an instance of data (tuple in a relational database table, object instance in a class within an object-oriented database) or data associated with the values of an attribute within a domain, or the particular value of an attribute in one instance. Metadata can describe data models.

Metadata can also be used to describe processes and software. It can describe an overall processing system environment, a processing system, a process, a component of a process. It can describe a suite of software, a program, a subroutine or program fragment, a specification. It can describe an event system, an individual event, a constraint system and an individual constraint. It can describe a process and /or event model.

Metadata can describe people and their roles in an IT system. It can describe an organisation, a department, individuals or individuals in a certain role.

The process of standardisation of metadata – models, semantics and syntax – is only just beginning, and then mainly in the data domain. Particular application domains have their own metadata standards to assist in data exchange e.g. engineering [3], healthcare [4], libraries [5]. An attempt at a more general exchange metadata for internet resources – the Dublin Core - has been proposed [6] but unfortunately it is insufficiently formal to be really useful [7]. A general metadata model, RDF (Resource Description Framework) has been proposed [8] with the implementation language XML (eXtended Markup Language) [9].

This paper concentrates on the traditional data / information / knowledge aspects of metadata; however, there are clear linkages to processing (including events) and people – especially from the object-oriented and logic-based viewpoints.

2.2 A Classification of Metadata

Metadata is used for several purposes;

(a) describing data for the purposes of data exchange;

(b) describing data for the purposes of global access from query (including update) to optimise recall and relevance;

(c) describing data for the purposes of query optimisation;

(d) describing data for the purposes of answer integration and explanation;

(e) describing data for the purposes of correct analytical processing or interpretation, representation or visualisation.

(f) describing the data to overcome multilinguality and multimedia heterogeneities

All of these purposes require that the data be described:

(1) such that the resource is constrained formally to ensure integrity;

(2) such that the resource is reachable by automated means;

(3) such that there is sufficient description for the purposes to utilise the resource.

This requirement leads to a classification orthogonal to purposes but serving all of them proposed in a tutorial in 1997 and published in [7] (Figure 1: Metadata Classification):
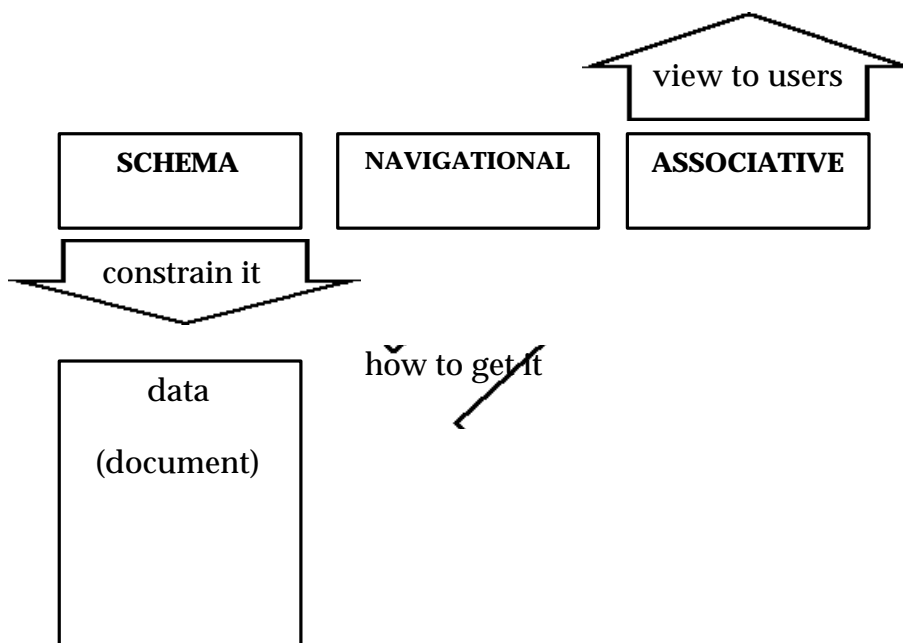
Figure 1: Metadata Classification

It should be noted that, whereas this classification is suitable for data it can also be used - and is sufficiently general for - metadata about processes, events, organisations or people.

1.1.1   Schema Metadata

Schema metadata constrains the associated data. It defines the intension whereas instances of data are the extension. From the intension a theoretical universal extension can be created, constrained only by the intension. Conversely, any observed instance should be a subset of the theoretical extension and should obey the constraints defined in the intension (schema). One problem with existing schema metadata (e.g. schemas for relational DBMS) is that they lack certain intensional information that is required [10]. Systems for information retrieval based on, e.g. the SGML (Standard Generalised Markup Language) DTD (Document Type Definition) experience similar problems.

It is noticeable that many ad hoc systems for data exchange between systems send with the data instances a schema that is richer than that in conventional DBMS – to assist the software (and people) handling the exchange to utilise the exchanged data to best advantage.

1.1.2   Navigational Metadata

Navigational metadata provides the pathway or routing to the data described by the schema metadata or associative metadata. In the RDF model it is a URL (universal resource locator), or more accurately, a URI (Universal Resource Identifier). With increasing use of databases to store resources, the most common navigational metadata now is a URL with associated query parameters embedded in the string to be used by CGI (Common Gateway Interface) software or proprietary software for a particular DBMS product or DBMS-Webserver software pairing.

The navigational metadata describes only the physical access path. Naturally, associated with a particular URI are other properties such as:

(a) security and privacy (e.g. a password required to access the target of the URI);

(b) access rights and charges (e.g. does one have to pay to access the resource at the URI target);

(c) constraints over traversing the hyperlink mapped by the URI (e.g. the target of the URI is only available if previously a field on a form has been input with a value between 10 and 20). Another example would be the hypermedia equivalent of referential integrity in a relational database;

(d) semantics describing the hyperlink such as 'the target resource describes the son of the person described in the origin resource'

However, these properties are best described by associative metadata which then allows more convenient co-processing in context of metadata describing both resources and hyperlinks between them and – if appropriate - events.

1.1.3   Associative Metadata

In the data and information domain associative metadata can describe:

(a) a set of data (e.g. a database, a relation (table) or a collection of documents or a retrieved subset). An example would be a description of a dataset collected as part of a scientific mission;

(b) an individual instance (record, tuple, document). An example would be a library catalogue record describing a book ;

(c) an attribute (column in a table, field in a set of records, named element in a set of documents). An example would be the accuracy / precision of instances of the attribute in a particular scientific experiment ;

(d) domain information (e.g. value range) of an attribute. An example would be the range of acceptable values in a numeric field such as the capacity of a car engine or the list of valid values in an enumerated list such as the list of names of car manufacturers;

(e) a record / field intersection unique value (i.e. value of one attribute in one instance) This would be used to explain an apparently anomalous value.

In the relationship domain, associative metadata can describe relationships between sets of data e.g. hyperlinks. Associative metadata can – with more flexibility and expressivity than available in e.g. relational database technology or hypermedia document system technology – describe the semantics of a relationship, the constraints, the roles of the entities (objects) involved and additional constraints.

In the process domain, associative metadata can describe (among other things) the functionality of the process, its external interface characteristics, restrictions on utilisation of the process and its performance requirements / characteristics.

In the event domain, associative metadata can describe the event, the temporal constraints associated with it, the other constraints associated with it and actions arising from the event occurring.

Associative metadata can also be personalised: given clear relationships between them that can be resolved automatically and unambiguously, different metadata describing the same base data may be used by different users.

Taking an orthogonal view over these different kinds of information system objects to be described, associative metadata may be classified as follows:

(a) descriptive: provides additional information about the object to assist in understanding and using it;

(b) restrictive: provides additional information about the object to restrict access to authorised users and is related to security, privacy, access rights, copyright and IPR (Intellectual Property Rights);

(c) supportive: a separate and general information resource that can be cross-linked to an individual object to provide additional information e.g. translation to a different language, super- or sub-terms to improve a query – the kind of support provided by a thesaurus or domain ontology;

Most examples of metadata in use today include some components of most of these kinds but neither structured formally nor specified formally so that the metadata tends to be of limited use for automated operations – particularly interoperation – thus requiring additional human interpretation.

3 Why is it Important?

3.1 Complexity

It is observed that the number of available information sources increases, the number of users increases and the number of information requests increase in both number and complexity. The complexity arises because of several factors:

(a) the heterogeneity of the information sources, including character set, language, media, content quality (accuracy, precision), structure and semantics;

(b) the increased required expressivity of queries including more complex syntax and semantics, the use of graphical interfaces, query improvement or refinement to improve relevance and recall;

(c) the increased complexity of the logic of processes acting over the information sources where the query (or update) may include inline functions (e.g. the concept of 'inexpensive' requires a function involving price of the required object and person salary to be inline in the query);

(d) the increased complexity of integrating information from multiple sources, resolving different values or sets of values for the same required object and explaining the choices made to provide the answer, and the values in the answer itself.

The increasing number of information sources and increased number of users is due to the reduced cost of a person joining the world information society and the increased commercial and non-commercial opportunities for marketing information either for itself or as a step towards purchasing or obtaining traditional goods and services. The increased expressivity and complexity of queries is caused by increasingly educated end-users demanding more of the information systems than previously especially in relevance and precision of answers, structuring of answers, associated explanation and multimedia representation.

3.2 Utility

Metadata increasingly becomes essential to be used in optimising queries, explaining answers, mediating between information sources and between those sources and the querying client and in handling access rights and possible associated payments. Metadata, with associated processes to use it, becomes the glue that holds together the rich diversity of information, suppliers and consumers on the internet.

3.3 Problem

Unfortunately, this sudden realisation generally of its overwhelming importance comes too late; already there are multiple sectoral standards for metadata and attempts to find a commonly agreed set of standards have so far failed to be accepted widely. Even the RDF [8] recommendation from W3C (The World Wide Web Consortium) [11], which is a basic model for describing 'things' and 'connections between them' without semantics, has itself failed to obtain universal acceptance. Various proprietary models – some loosely related to RDF and commonly using XML [9] as the implementation language – have appeared, such as XMI [12].

4   What Exists Now?

4.1 Introduction

There are many good and usable metadata systems in operation every day. Usually, they are specific to a particular organisation (internal data exchange standards, internal IT System documentation standards), a pair of organisations (agreed data exchange standards) or organisations in a particular business sector where a common standard for data exchange or accessing each others' systems is agreed – for commercial benefit. Some of these latter metadata systems have reached international standard status, notably EDI [13] and STEP/EXPRESS [3].

Most of these systems are successful because they are implemented in a narrow domain where the syntax and structure of exchanged datasets have been agreed and where the semantics are well understood in that circumscribed community.

4.2 Some Specific Initiatives

More recently, the explosive growth of WWW has caused several interesting initiatives concerning metadata:

(a) PICS [14]: a method of tagging pages on WWW with content classification information such that compatible processing elements can prevent the pages being displayed. This system is targeted at privacy and parental protection of minors from accessing unsuitable material. This is a kind of associative restrictive metadata;

(b) DC (Dublin Core) [6]: an initial attempt to provide a general associative descriptive metadata element set for the description of content in a WWW page. The original 13 element set was extended to 15 by the Warwick Framework and subsquently there has been much discussion between those who wish to keep the DC simple and human-readable and those who wish to make it more formal and computer-readable;

(c) RDF [8]: The Resource Description Framework General Model for metadata proposed by W3C [14]. This proposal is based on a simple binary relational model such that it can be used universally as a descriptor. The problem is the potential diversity of content, structures and semantics placed upon this basic model – and such diversity is appearing already, especially since the implementation language is XML [9] which is very flexible, providing a syntax but no semantics unless declared externally;

(d) XMI (XML Metadata Interchange) [12] is a standard accepted by OMG (Object Management Group) [15] and brings together XML [9], UML (Unified Modeling Language) [16] and MOF (Meta Object Facility) [17] to provide a metadata facility for information exchange between information systems;

(e) XIF (XML Interchange Format) [18] which may be seen as a competitor to XMI from Microsoft with a consortium of independent repository vendors;

(f) A host of application domain or business domain initiatives such as: numeric and statistical data [19], geospatial information [20], music [21], works of art [22] (and, because in this cultural heritage area there are several standards; a useful crosswalk is provided at [23]), scientific metadata [24] [25], biosciences [26], healthcare [27], education [28] and a host of others. Digital library metadata has already been mentioned [5], [6];

(g) A major use of metadata is in electronic business: the UN (United Nations) EDI standard [13] is widely adopted and the XML/EDI initiative [29] [30] is gaining popularity. ICE (Information and Content Exchange) [31] is being implemented and utilises various security features based on W3C initiatives such as P3P [32] which is an example of associative restrictive metadata.

It is unclear exactly how these initiatives will develop and inter-relate. Some are proprietary, and there are parties with commercial or other interests in the groups defining open standards. Many of these application initiatives concern data exchange, but increasingly there are groups working on the underlying associative supportive metadata in the form of terminological thesauri or domain ontologies. The latter developments are particularly significant because such resources provide maximum flexibility for systems built using cooperating intelligent agents e.g. [33] and also provide greater support in both query refinement and answer integration and explanation e.g. [10].

It should be noted that the Information Systems Engineering community has utilised metadata for many years in attempts to improve systems construction management and systems maintenance. The

major objective was to have well-understood communication between designers but also the metadata was used to drive tools which assisted in systems engineering. An early attempt was the extension of schema metadata with the IRDS (Information Resource Dictionary System) [34] followed by several attempts - such as Conceptbase [35] - to capture metadata for the purpose of describing systems.

4.3 Systems Using Metadata

Systems utilising metadata are similarly diverse. Basically, they may be classified into:

(a) systems with extensive human interaction to make choices based on metadata information (e.g. web browsing, use of web portals or query refinement systems accessing heterogeneous information sources [10])

(b) systems relying on profiles input by the client-user and the server(s) which then are used by mediating agents (e.g. electronic business systems utilising P3P for security [32], or CORBA-based systems accessing compliantly-wrapped information sources [33]);

(c) totally automated systems (e.g. automated sensor systems in scientific experiments or regular data exchange between earth observation devices).

5 Future

5.1 Metadata

Metadata has moved centre-stage as the most important component of the solution to the application requirements of the architecture and construction of modern information systems. Most modern systems are web-based, either within the organisation (Intranet) or public. In the latter case, especially, metadata is utilised to improve communication between heterogeneous information systems – for the purposes of obtaining and providing information, for communication between the user client workstation and the information servers and for electronic business between information systems.

The concept of separating the primary information resources from data and processes (metadata system) providing access to those resources is extremely important. This allows changes of access policy – such as changes in access restrictions for certain kinds of users in certain roles, changes in categorisation and classification and changes (additions) in descriptive metadata depending on viewpoints of different authorised users – without accessing the data resource itself.

5.2 Scale

The rapidly expanding internet community, and the ever increasing demand for services – largely WWW-based – demands that solutions must be scalable. Ever-increasing computer power, storage capacities and networking speeds only mitigate the problem – the expansion and consequential demand outstrip supply of the technological services. The technology, however, has predicted limitations varying from the need to develop a technology other than CMOS for processors through the need to develop faster and denser storage devices to the need for provision of inexpensive and faster communications technology than even that based on fibre. Thus the solution must lie with better systems engineering – the 'brute force' methods will not provide the whole solution.

A major component of that systems engineering solution has to be intelligent utilisation of resources. This implies better refined queries, better constructed databases, better utilisation of distribution and parallelism for algorithms acting on data resources, and better concurrency. For all of these aspects quality metadata – accessed and used by intelligent agent technology, is the basis for the solutions.

5.3 GRIDs

5.3.1 Background

There has emerged through 1998-1999 in North America the concept of a Computation Grid [37] closely followed by the same concept in Europe [38]. In UK the concept of GRIDs was first articulated completely in the summer of 1999 (but was dependent on much internal work before that finding its roots in the Distributed Computing Systems programme of the late 1970s) and captured succinctly in [39] which described the 3-layered Computation ∕ Data, Information and Knowledge grids architecture as proposed by the author. By September 1999 the North American community had also considered data access [40] and overall architecture [41] so moving from computation (linking supercomputers for compute-power) to the world of data. The 'Grid Bible' published in July 1998 gives some flavour of the challenges [37] although rooted in computation. However, the North American and European architectural view is less comprehensive than that in [39] which overviews underlying detailed considerations of access, security and rights strategies as well as uniform information access over heterogeneous sources and a uniform computation landscape.

The UK view of GRIDs has been driven by requirements in science, engineering and technology and is being promoted through the UK Government Office of Science and Technology with the label 'e-Science'. It is expected that this pull will lead to solutions later (but quickly) applicable to general commercial and business processes, especially e-Commerce. The author has coordinated a meeting of leading UK academics and industrial representatives who endorsed enthusiastically the architecture and who are now working with scientists in the application areas to refine specific requirements and implement component GRIDs systems. An early application will be the management of data streaming from the LHC (Large Hadron Collider) at CERN (Centre Europeen pour la Recherche Nucleaire) in Geneva where it is proposed that each member country will need to support a large data centre with data cascading to its scientists. In UK the GRIDs architecture will be used. Similarly, UK groups working on biosciences – and especially genomics, environmental systems, advanced materials science, engineering modelling and social science systems are active.

5.3.2 GRIDs and Metadata

The architecture envisaged by the UK community attempts to bring together the (upward) refinement of data to information and knowledge and the (downward) application of knowledge to information handling and data collection through feedback loop control (Figure2 : GRIDs Architecture). The computation ∕ data grid has supercomputers, large servers, massive data storage facilities and specialised devices and facilities (e.g. for VR (Virtual Reality)). The main functions include compute load sharing ∕ algorithm partitioning, resolution of data source addresses, security, replication and message rerouting. The information grid resolves homogeneous access to heterogeneous information sources. The knowledge grid utilises knowledge discovery in database technology to generate knowledge and also allows for representation of knowledge through scholarly works, peer-reviewed (publications) and grey literature, the latter especially hyperlinked to information and data to sustain the assertions in the knowledge [7].
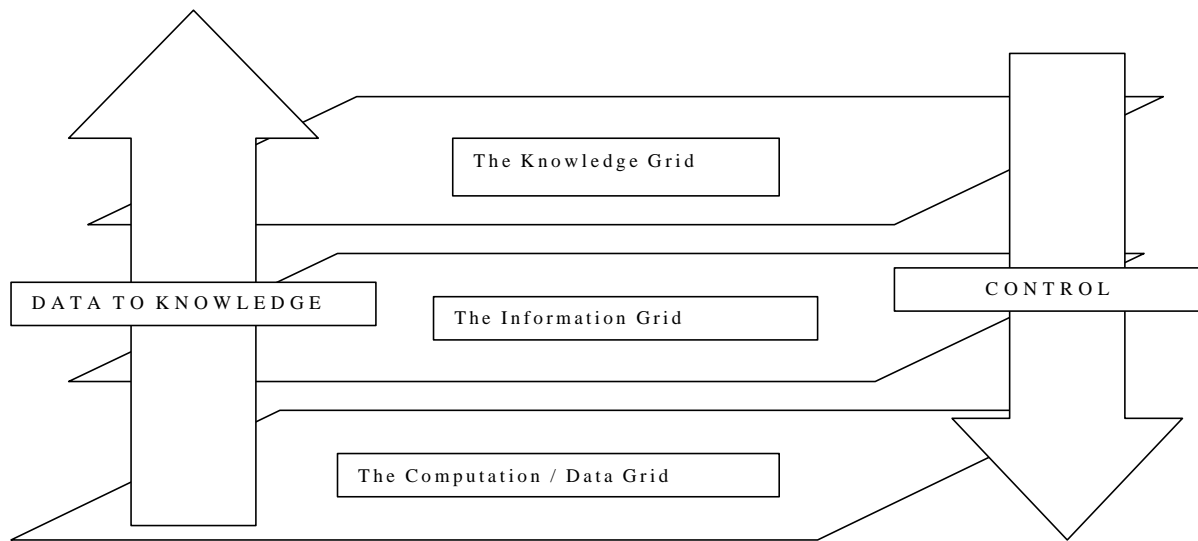
Figure 2: GRIDs Architecture

The concept is based on the idea of a uniform landscape within the GRIDs domain, and external appliances - ranging from supercomputers, storage access networks, data storage robots, specialised visualisation and VR systems, data sensors and detectors (e.g. on satellites) to user client devices such as workstations and WaP enabled Mobile phones. The connection between the external appliances and the GRIDs domain is through agents representing the appliance (and thus continuously available to the GRIDs systems). These representative agents handle credentials of the end-user in their current role, appliance characteristics and interaction preferences (for both user client appliances and service appliances), preference profiles and associated organisational information. These agents interact with other agents in the usual way to locate services and negotiate use. The key aspect is that all the agent interaction is based upon available metadata.

5.3.3 GRIDs , the CAiSE Conference Series and Janis Bubenko

Key functionalities in the GRIDs will be based on computer science results (some stretching back 30 years or more) in database, hypermedia, visualisation & VR, information retrieval, document systems, workflow-supported processes, knowledge management (including inferencing and dealing with uncertain and incomplete information), knowledge discovery in database (including data scrubbing, warehousing and mining) cooperative working and related topics. The building of the GRIDs will be a continuous, incremental systems development process based on results from the CAiSE conference series among others. The unique feature of the project is the size, complexity and open aspects of this systems development exercise, which has the problems also of facilitating interoperation of legacy systems. It is noteworthy that many of the technologies being used are based on the work of Janis Bubenko, especially in information systems modelling and systems development methods.

References

1. Pound C. Personal communication 1999

2. Adams D. The Hitch-Hikers Guide to the Galaxy. Pan Books, London,   1979

3. http://www.nist.gov/sc4/www/stepdocs.htm

4.   http://www.chime.ucl.ac.uk/HealthI/GEHR/Deliverables.htm

5. http:/lcweb.loc.gov/marc/marc.html

6. http://purl.oclc.org/dc/

7. Jeffery, K G: 'An Architecture for Grey Literature in a R&D Context' Proceedings GL'99 (Grey Literature) Conference Washington DC October 1999 http://www.konbib.nl/greynet/frame4.htm

8. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222

9. http://www.w3.org/TR/1998/REC-xml-19980210

10. 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

11. http://www.w3.org/

12. http://www.oasis-open.org/cover/xmi.html

13. http://www.harbinger.com/resource/edifact/

14. http://www.w3.org/PICS/

15. http://www.omg.org/

16. http://www.omg.org/uml/

17. http://www.omg.org/cgi-bin/doc?ad/99-06-05

18. http://msdn.microsoft.com/repository/technical/xif.asp

19. http://www.icpsr.umich.edu/DDI/

20. http://www.opengis.org/

21. http://www.oasis-open.org/cover/smdlover.html (music)

22. http://www.getty.edu/gri/standard/cdwa/HOMEPAGE.HTM

23. http://www.getty.edu/gri/standard/intrometadata/crosswalk.htm

24. http://www.getty.edu/gri/standard/intrometadata/crosswalk.htm

25. http://www2.echo.lu/oii/en/science.html#science

26. http://www.nbii.gov/tools/non-spatial.html

27. http://www.mcis.duke.edu/standards/HL7/sigs/sgml/index.html

28. http://www.manta.ieee.org/p1484/

29. http://www.geocities.com/WallStreet/Floor/5815/index.html

30. http://www.geocities.com/WallStreet/Floor/5815/guide.htm

31. http://www.w3.org/TR/NOTE-ice

32. http://www.w3.org/TR/P3P/

33. http://img.cs.man.ac.uk/tambis/index.html

34. http://burks.bton.ac.uk/burks/foldoc/72/56.htm

35. http://www-i5.informatik.rwth-aachen.de/CBdoc/index.html

36. http://www.privaseek.com/

37. Foster I and Kesselman C (Eds). The Grid: Blueprint for a New Computing Infrastructure. Morgan-Kauffman 1998

38. http://www.egrid.org/

39. http://www.itd.clrc.ac.uk/Activity/GRIDs

40. http://www.sdsc.edu/GridForum/RemoteData/Papers/architecture.html

41. http://www.gridforum.org/building_the_grid.htm