BCS
THE BRITISH COMPUTER SOCIETY

KIDMM: knowledge, information, data & metadata management
www.epsg.org.uk/KIDMM

Account of the KIDMM discussion meeting, 6th March 2006

On 6th March 2006, the BCS Electronic Publishing Specialist Group hosted a discussion meeting at the BCS meeting rooms in Southampton Street off The Strand in London. The topic was the management of knowledge, information, data and metadata – a topic that is a shared interest between a number of groupings within the BCS. The meeting was chaired by Conrad Taylor, Secretary of EPSG, who recorded the discussion and prepared this account from the recordings. This document and other supporting resources are available from the URL above.

Introduction to the day — how did we get here?

Conrad Taylor started the day with some words about how the meeting had come about, and where we might head afterwards. He hoped this would help steer the meeting, considering that we had no agenda. He also noted that although EPSG was running this event, it had an unusual status: an 'inter-SG-plus' event, between the Specialist Groups of the BCS – plus some other organisations, individuals and experts.

How we got here

The most recent lap of the journey, related from EPSG's point of view,[1] began at the second Specialist Groups Assembly, at Bletchley Park, at which David Penfold had given a talk on metadata. Conrad asked David to remind us about that occasion.

David observed that the content of his talk had been converted into an article and can be read on the BCS Web site.[2] He had described the problems that face people trying to find useful information on the Internet, and the role that XML could play in adding metadata to these resources. After that talk, people came up to him and said that it would be good to explore this further (a) because metadata is an area of great interest and (b) because other people have other interpretations of what the word means.

Afterwards, when David gave a talk on metadata to the Configuration Management Group, the same issue came up: CMG members thought about metadata in a different way from how David thought of it. He saw that it seemed to make sense to broaden that discussion, to try to understand what people mean when they talk about metadata, and to take it further to do something useful – either between the Specialist Groups, or more broadly within the BCS.

Conrad commented he was sure that metadata – and our different understandings of it – would be part of our focus for the day; but that there are other methods of information management with which we might equally be concerned. Information can also be managed by holding it in fielded containers, as it is in a standard database; it can be structured using tag vocabularies, as in structured SGML and XML documents; it can have metadata attached to it; and it can be indexed. Some of these approaches overlap, of course.

Conrad mentioned two workshops organised by the Developing Countries Specialist Group and chaired by John Lindsay. The first of these (21 January 2003)[3] was called to discuss the World Summit for the Information Society, and focused on the issue of information literacy; a later DCSG meeting (12 May 2004)[4] was about metadata and its relationship to issues of world development.

These meetings had been conducted in a style which Conrad dubbed 'the Lindsay method' – a series of fairly informal 'rounds' to which each person around the table had contributed. This process Conrad proposed to use as a model for this day's discussion.

At one of the Specialist Groups Assemblies (Oxford, 14 April 2004), Ian Herbert gave a short talk about Health Informatics,[5] in which it was evident that information management and metadata were becoming very important

[1] Other participants have had other journeys. John Lindsay refers to a meeting that took place during the 'Making Sparks' festival run at Imperial College, 7–8 September 2000, by the British Association for the Advancement of Science. That in turn led to the seminars run by the BCS Developing Countries Specialist Group on information literacy, metadata and world development.

[2] http://archive.bcs.org/BCS/review04/articles/multimediaandelectronicpublishing/dissectinginfosociety.htm

[3] See archived material including report at www.epsg.org.uk/dcsg/wsis-focus/

[4] www.epsg.org.uk/dcsg/mwd/

in the health services, with issues like the adoption of controlled vocabularies, and the efforts being made to ensure electronic patient records would be interoperable across the health services. Similar issues are arising, as we know, in matters of government.

The Health Informatics Committee later transformed into the Health Informatics Forum, reflecting an awareness within the BCS that there are topics that several Specialist Groups might have in common, and could usefully talk to each other about. In April 2005 at another Specialist Groups Assembly, there was a discussion about what other kinds of forum-like groupings there might be (another term bandied about was 'supergroups'). That was where the idea came up that several SGs have shared and overlapping interests in knowledge and information and data – and the management, organisation, retrieval and publishing thereof.

Out of that discussion came the formation of a 'KIMtec' email discussion list, which saw some sporadic discussion; but it seemed obvious that a face-to-face meeting would be more productive. So, during summer 2005, the EPSG Committee made a commitment to host and facilitate such a discussion, and so here we were today.

Where we might go from here?

Conrad suggested that out of today's discussions might come, for example:

- further rounds of discussions, leading to further shared understanding;
- reporting back to our colleagues in the various groups we belong to;
- reporting back to the Specialist Groups Assembly, which David had been asked to do;
- building bridges with XML UK, between whom and the BCS there seems to have been some kind of friction or misunderstanding;
- perhaps finding some common BCS position on what is important in the management of knowledge and information management;
- identifying other organisations with whom collaboration could be productive, such as CILIP or KIMNET;[6]
- an examination of BCS's commitment internally to managing its own knowledge and information resources, and what has happened to the Knowledge Services Board's efforts to establish a taxonomy for BCS subject classification; and
- the impact of this on the BCS Web site, and on BCS publishing in general.

Around the room – introductions, identification of concerns

Conrad asked those present to introduce themselves, their affiliations, and the interests they have in knowledge, information and data and metadata management; and also their hopes, or ambitions, about what might come from today's discussions.

Keith Gordon: Keith is Secretary of the Data Management Specialist Group. His working life was spent in the Army; he was a professional soldier for 39 years. He ended up being responsible for strategy and policy for data management for the Army. When he was declared 'too old for use to anyone' – as he put it – he set up his own company: training, but also consulting in data management. He works with the Open University as a tutor on courses about data and databases, and has an interest in business analysis and requirements engineering.

Keith is on IST/40, the British Standards Institution committee that represents the UK within the ISO on issues of data management and interchange.

Keith described a problem he faced when asked by the BCS to write a book about the principles of data management. He thought, 'Shouldn't the word *information* be in the title somewhere?' This got solved by adding to the book's title the strapline 'Facilitating information sharing.'

There is an issue about what is information and what is data, and what is the difference between the two, and Keith said that he finds it difficult to explain: it appears to be a grey area. When does information become data, when does data become information? There are some definitions around, for example in ISO statements, but it still seems like a grey area nonetheless.

Carl Harris: Carl is the BCS Web site manager, more commonly known as the 'Webmaster' at HQ. He looks after the overall development and day-to-day running of the BCS's main Web site.

[5] MP3 of Ian Herbert's talk available from www.epsg.org.uk/sga/

[6] CILIP is the result of the merger between the Institute of Information Scientists and the Library Association; KIMNET is a knowledge and information management specialist group of Aslib.

Recently a new content management system was implemented as the 'back end' of the Web site. The way in which content is being added to this system builds on the work which Judi Vernau undertook for the BCS Knowledge Services Board. This has led to a new taxonomy and metadata scheme being plugged into the content management system.

Carl explained that this means that the BCS is now in the position of having an underlying system in which all the content and data can be stored, and a scheme which can apply taxonomy and metadata to that content. The challenge now is to gather together all of the content. At the moment, the only content on the site is that held centrally by the Society. BCS Branches and Specialist Groups also have lots of content 'out there' which the BCS would like to capture and catalogue and record in the same way.

The other challenge which the BCS is having internally, having built up some experience of using the new system for a little while, is how to apply metadata and taxonomy consistently. In experiments so far, it has been found that staff in HQ find it very difficult to grasp the concept of what metadata is, how important it is, and how to choose their terms consistently. There is a very large collection of classifications, but they are finding that people either don't know enough about their subject, or they know only the specifics; and so the first attempts to use the CMS to generate metadata resulted in too many inconsistencies.

Therefore, one of the things that Carl hoped we might touch on today was how to overcome those challenges, so that the BCS can not only have a scheme that will work for us, but also one that is consistently applied by the many people who will eventually use the system.

John Alexander: John works in corporate communications and information design at Donovan Data Systems, and is on the committee of BCS-EPSG. He considered his position to be similar to Carl's: he has responsibility for the corporate Web site, and they have a lot of problems with implementing site search, and problems of vocabulary. A lot of John's day-to-day work is editorial in nature, and in this he tries to enforce consistency in how terms are used. This is reasonably possible for the UK side of things; but worldwide, Donovan Data Systems has no-one in that role.

Genevieve Hibbs: Genevieve described herself as 'very multi-disciplinary', and interested in cybernetics and information-handling concepts which have arisen in practical situations. At the moment she is involved in the development of a lexicon for a version of 'EasyEnglish', with a core vocabulary of somewhere around 3,500 words. She has also specialised in the health care of people at work, and in the health informatics issues associated with that. She is also a local elected councillor, which gives her quite another slant on a number of these issues.

Berin Gowan: Berin started out as a systems designer; then he had realised the importance of information design, or 'information architecture' as the current term has it. He has largely been applying this in the publishing arena, with large publishers in the UK.

What Berin is struggling with is that he doesn't feel he has ever really got to grips with information… It is as if he is trying to label something that he doesn't really understand. Part of the problems of definitions are: what are the properties, characteristics and behaviour of information? He thinks we will wrestle with all these other problems until we have made some progress; or maybe we will make progress with the labelling, if we begin to understand information a little bit better. Then, we might decide how we can manage it. Berin wonders if the 'information scientist' community might be able to help.

Berin said he had served his term on the Knowledge Services Board for three years, but is no longer involved. However, he was involved with KSB when the taxonomy project was being developed.

Adrian Walmsley: Adrian is on the Committee of the Oxfordshire Branch of the BCS, and is on the Knowledge Services Board. He manages the Engineering and Technology Forum of the BCS on a part-time basis. One of his current objectives is to explore how Forums can work more effectively and constructively with the Specialist Groups.

Adrian was with IBM for 36 years. Reading Conrad's pre-meeting paper[7] reminded him that he had been involved with one of the early IBM tools, initially called Script, which worked with GML, one of the forebears of SGML.

He remembers the days when relational databases were new, and the turf wars that had broken out between the old-style people who used to believe that maintaining data occurred naturally in hierarchies, as against 'relational' people who believed that it occurred naturally in tables. The truth, of course, is somewhere in between.

Trying to find things in the BCS Web site at the moment illustrates the difficulty – it is not obvious for example that 'Thought Leadership' comes under the Academic part of the site, rather than the Events part. You tend to be forced along the way the navigation works, to navigate it hierarchically. Really, you need both ways of finding stuff.

[7] *Knowledge, information and data management – and technology*. First drafted after the Spring 2005 SG Assembly, and revised and expanded ahead of the 6 March KIDMM meeting. Available from the KIDMM documents and resources page at www.epsg.org.uk/KIDMM/docs/

Picking up on the point that Keith had made about 'information' and 'data', Adrian said that while at IBM he was involved in IT architecture: he sat on the board that certified new architects. He also remembered discussions he had taken part in, when he had worked with ISEB a couple of years ago to create the BCS certificate in IT Architecture.

One of the things that an IT architect was supposed to produce, the first draft declared, was a 'data architecture'. That wording got changed in the final version to 'information architecture', on the grounds that Information Architecture was a superset of Data Architecture.

Like Genevieve, Adrian is a locally elected councillor. He set up the Web site for his local parish council in 1995.

Ken Moore: Ken is the Publishing Systems Manager for the academic division of Oxford University Press, which means that he manages the Content Management System used to compile the Oxford English Dictionary, the Oxford Dictionary of National Biography, and a broad range of trade dictionaries. He has worked in this area of computing and publishing for a long time.

Ken expressed his main ambition for the day as being to make sure that OUP stays up to date, or at least aware of the mechanisms by which they can make their content accessible in the most effective manner.

Zinat Bennett: Zinat is on the Committee of EPSG, and is responsible for IT systems at the library of Aston University. Her current challenge at work is in establishing an institutional repository. As an example of what that challenge involves, Zinat showed a printed and bound MPhil thesis that had not been handed in electronically. She is looking at how to scan it all in, including the colour graphics, and add tags, which will almost certainly on the basis of extended Dublin Core.

The workflow for this, how to generate the metadata appropriately, and then how to place the results in a database or some repository from which the thesis can be retrieved meaningfully, is where the challenge lies.

They already have problems managing the electronic journals in the library; it gets progressively worse with this sort of unstructured resource. So what Zinat said she really wants is an 'automatic metadata generato'r ; and something to retrieve the citations too, since she wants her students to be able to click on a citation reference and move directly to the relevant article. At the moment, this is at the pilot project phase.

A number of different approaches to this are being tried in different institutions, but it is difficult to get hold of the details because of intellectual property issues. People think this is like gold dust – and they don't want to give it away.

Martin Bryan: Martin is a member of the XML UK group. He started out as a printer and went into publishing, and from there into computing and metadata and ontologies. He was involved in the development of ASPIC (Authors' Symbolic Pre-press Interface Codes), a typesetting language developed by the British Printing Industries Federation (BPIF), in the days before we had structured mark-up. He was then asked in 1985 to review an ISO standard, ISO 8879 – in other words, SGML. Since then, he has been working with the British Standards Institution (BSI) and ISO, monitoring the work on Document Description and Processing Languages committees which have developed SGML, DSSSL, Topic Maps… all the languages we know which have formed the basis of much that is going on in the XML world these days.

Martin is convenor of the project on Document Schema Description Language (DSDL), which is looking at the next generation of schema languages. They are dealing with the XML schema language RELAX NG,[8] the XML schema validation language Schematron, a new datatype library language that Jeni Tennison is putting together called Datatype Language Library (DTLL), and the Document Schema Renaming Language.

For his daytime job, Martin works for CSW Informatics, which specialises in 'knowledge engineering'. They make knowledge management tools, mainly for medical and pharmaceutical purposes; the parent company CSW's main field is healthcare and informatics, and is involved in putting in a lot of the front-end software for many of the NHS systems that are currently being deployed. At present, therefore, Martin is working on the use of rules and ontologies within NHS systems.

John Lindsay: John prefaced his introduction by saying that are at least two more organisations we should add to our 'organisation soup' – one is the International Society for Knowledge Organisation (ISKO), which is one of the UNESCO consultative bodies. (BCS also is linked into this consultative relationship role with UNESCO through IFIP, the International Federation for Information Processing.) The other is the UK Academy of Information Systems, UKAIS[9].

About 40 years ago, John started calling himself an Information Systems Designer. Like Keith, he started off

[8] RELAX NG (REgular LAnguage for XML Next Generation) is a schema language for XML – in other words, a way of describing the pattern for the structure and the content for an XML document. Such schemas play a role that is broadly similar to that of a DTD in traditional SGML publishing applications.

[9] UKAIS is a charity that came out of a meeting in 1995 of UK academics in information science, and is concerned to promote recognition of IS as a discipline and to discuss issues in IS teaching and research. (www.ukais.org)

in the Army: he thinks that quite a lot of the issues we are working through now with information started out in the military. It would be very useful to have a proper history – or the best that we can do – about where some of our concepts have come from.

John said that what fascinates him is why people find concepts difficult to grasp. Recently he has been working with people who call themselves IT practitioners, museum curators, archivists, librarians, academics and scholars. What is missing is any sense at all of 'common sense' – by which he meant a slightly uncommon meaning: there is *no sense in common* between these communities. They have their own community languages and community discourses, but don't have any commonality.

What technology is forcing us to do is to pull things into connections and containers which are different from those they had previously. Either we will have to invent new words – in which case we will have to explain what those new words are – or we re-invent old words, which will cause absolute confusion.

That is what Boole did with the word AND… and if the words AND and OR don't mean anything to people, and people don't recognise that a hyphen is an ASCII character, and so is a space, then John thinks our starting point is some quite primitive issues that people don't really grasp the significance of properly.

What John hoped we might get out of the day would be a blog on which we could put up some papers, a space where we could pull together what we consider our concepts to be. If we have got the energy, we could then try and build a little toy system to show the interoperability of the concepts – and with a bit of luck we might end up where Aristotle was about 3,000 years ago!

Conrad commented that we do have amongst us some skills in wikis. Adrian runs a wiki for the Oxfordshire Branch. There is also an interesting content management system used by the Open Source Specialist Group; it was a pity that their website manager Paul Adams couldn't join us for the day. Maybe within our community we do have the ability to build such a thing.

Robin Kyd: Robin works for the Open University. His job title there is Electronic Publishing Adviser, and he is a member of the EPSG Committee. At the Open University they have a number of interests in this area. Just at the moment they are looking at introducing a content management system. Another buzzword is 'VLE' – meaning Virtual Learning Environment.

The Open University with its history and its sense that it is a leader in distance learning is beginning to realise that in the modern world you have to fight your corner – there are all sorts of people who now claim to be world leaders in distance learning. The OU has to move with the times and keep up with everybody else (or stay in front, if that is where they are).

Miltos Petrides: Miltos works for the University of Greenwich. He came to this meeting as a member of the BCS Specialist Group for Artificial Intelligence. Miltos' area of interest in research is case-based reasoning and machine learning, and he is very much interested in information that is encoded in structures such as images and temporal relationships, as well as more standard kinds of data.

In a lot of this AI research work, XML is becoming very important. In order to assist communication between intelligent agents, people are trying to work out a way through different sources of information, each of which has its own semantics. This basically brings us to the issue of the Semantic Web. In addition, Miltos is interested in the whole area of intelligent middleware, actively trying to reconcile the different meanings and semantics between different systems.

Andy MacFarlane: Andy is from City University in London, and he is secretary of the Information Retrieval Specialist Group of the BCS. His own personal interests in teaching and research are largely in information retrieval, a subject which he teaches to librarians and information scientists, information systems students – and also recently to computing students as well. As a group that focuses on information retrieval, the IRSG is obviously heavily interested in search, and the use of metadata to support search. That would also mean helping the likes of John Alexander and Zinat Bennett with their problems.

Andy's one ambition for the day was to get some idea of how information retrieval tools can help the community as a whole.

Andrew Tuson: Also from City University, where he is Director of Student Recruitment, Andrew is a member of the committee of the Specialist Group on Artificial Intelligence committee. In his research, he is interested in genetic and local search optimisations algorithms, and the knowledge-based design thereof. Andrew said that his main purpose for attending today was to see if there were any areas with which the SGAI might usefully engage.

Conrad admitted that he had already been halfway through the process of organising this meeting before he realised how little he know about the artificial intelligence community, and how it was researching into the role of AI in constructing metadata sets or searching more effectively.

Andrew said that he was also quite interested in seeing what would come out of the day about educational needs related to this area. There may be a need to change structures, and how people do things, and educationalists should be concerned about what the demands of industry were, not just to focus on supply of what education thought people *should* be interested in.

Terry Freedman: Terry is involved in many Specialist Groups. He is on Council, SG Executive, and is treasurer of the Data Management SG and of the Business Information Systems SG. He also plans the programme for the Project Management SG (PROMS-SG), and is involved with IRMA, FINSG, Information Retrieval, Sociotechnical, sometimes with the AI group, and the Law SG.

Terry has many interests in this area of metadata, because of his work for the National Archives, gathering electronic data from various government departments and exposing it to the public through a web site. The Web site is XML-based; they have finding tools, and metadata, and catalogues, and taxonomies – everything.

Terry sees no real difference between data and information – except for one step, which is that metadata, by describing data, makes it meaningful and therefore makes it information. He would welcome the advice of others about this.

He said there is an issue about who *owns* the data. That's because it is the owner of that data who defines what that data means. Once a piece of data 'escapes', it is no longer information, and that is when trouble starts because it can be interpreted in a different way. The person who owned that data attached certain values to it; now it is being used by somebody else, does that person understand the original intent of the acquisition of that information, or data? Terry was sure we would have much to discuss here.

Conrad commented that different people might have valid opinions about different meanings of the data, and that some of that might be quite political. Terry agreed, and said that was what he was implying in the concept of ownership, which is why it was so important an issue.

John wondered if we could begin to make charts to capture important points as they arose in discussion. For example, he thought that Terry's comment was the first time today that an absolutely fundamental point had arisen that required some moderation.

A document, as it exists in an archive, historically had an owner, had an author, and had an authority; and in some sense or other the attributes of that authority are part of a process. But the *reader* of the document is a different agent entirely; and the *information* is the association of those two components – it's not the property of any one of them. Terry replied that he was saying that the *owner* of the data has the authority to say what it means. John replied that he thought that was a point of disputation: that anybody can say what the data means. 'The question is, who's got a gun; and the gun decides what it is going to mean,' he said.

Nic Holt: Nic is a Systems Architect for Fujitsu, and is a member of the BCS Knowledge Services Board and the Engineering and Technology Strategic Panel. His main areas of interest are around knowledge management – search, categorisation, information discovery – finding and recognising items of information, entities, relationships within documents and so on.

Nic remarked that in the paper Conrad had circulated before the meeting, he had quoted a definition of information as being 'stuff which has meaning'. Nic would say that in the context of knowledge management, the significance of an item of information or an item of data depends on the context in which it is read, and in which it is used. Some of that context may be metadata – having that could provide a very valuable context; other parts of the context would be the context of creation, and equally the context of use.

Nic's experience in this field leads him to believe that we need to be careful about classification – it is potentially very dangerous. It tend to give a particular paradigm for looking at things which excludes all others, and therefore we may miss things which are really very important. That too is related to context.

The whole issue of classification and categorisation and search is really hard. Artificial Intelligence may be a key to this, but Nic had noticed that in the article by Nigel Shadbolt article which Conrad had pointed out,[10] Nigel says – and this could be contentious – 'AI is but embedded adapted smart software that can lift the burden of tedious, difficult and often repetitive jobs.' (Actually, Nic said, he thinks that AI has always aspired to something far higher than that.)

He would equally argue that it is correspondingly very hard, and that until we do crack the problem of really good automatic categorisation and automatic metadata, assignment, tagging or whatever – software will at best do an imperfect job of that. A lot of this is about making the best use of humans and machines together to achieve what we can. But we must not over-rate what can be achieved.

Nic concluded with a point about the BCS Web site. He said that we need to distinguish between classification schemes on the one hand, and navigation structures on the other hand. They may be closely related, but they are not the same thing – a point that Adrian and others had made. We are experimenting in developing this Web site, and it will be a little time before we know enough about the factors involved to get it how we would like it to be.

[10] *Web Intelligence* by Nigel Shadbolt, fir published in *IT Now* in May 2005 and available on the BCS Web site here: http://www.bcs.org/server.php?show=ConWebDoc.3043

Judi Vernau: Judi works for a consulting company called Metataxis, which specialises in information management and information architecture. She works rather more on the information architecture side. Judi comes from a publishing background; that is where she first encountered these ideas. She has also worked on metadata standards and taxonomy development in central government and local government.

Judi echoed what Berin had said earlier: working in information architecture is a constant voyage of discovery. She had recently been training within a large corporation; she started by explaining information architecture and then said, 'What you will find, as I do, is that when you are approaching this subject, you think you have got it, and you think, *This is so cool!* – and then the next minute it's… *Oh no, what is it? I've lost it again!*' Judi said that there is a constant process of discovery about what information is, and what you need to do to manage it. A common view of what these concepts and terms are, and what they mean, is something she would want to explore.

Judi said, 'The expression I hate the most is 'Content Management System. It seems that people can make it mean whatever they want it to mean!' However, she is very interested in where Information meets Technology, and identified it as an area that she needs to know more about. She reiterated Nic's point that not missing stuff when you are searching is very hard, and the classification-versus-navigation angle does tend to get confused.

Judi did some content analysis work for the BCS within the last year, on the basis of which they had put together a metadata scheme and a taxonomy. She was curious to know how it was getting on, but said she was sympathetic to the problems of implementation: it doesn't matter how beautiful your metadata scheme is – if it is not applied properly, it doesn't do you much good.

Francis Cave: Francis said he was present in two capacities. For one thing, he is Chairman of the XML UK. He thought he should first address a confusion that seems to exist. XML UK has had a long-standing relationship, through David Penfold, with the Electronic Publishing Specialist Group, who have had an obvious and logical interest in XML from an early stage; and the two groups have held a number of joint events. It had become obvious to Francis, and probably had also to David, that the common interests of the two groups might spread more widely to other groups within the BCS. Therefore Francis had approached the BCS with a view to finding out what kind of relationship would be appropriate.

XML UK already had a separate existence, having been formed originally as an UK activity subgroup within the international SGML Users' Group, which became the international XML Users' Group. XML UK has had a separate legally independent existence from the international group for only about a year and a half. For this reason, XML UK was not able to become a Specialist Group of the BCS.

The BCS therefore offered XML UK the status of an 'affiliate group', and that is what XML UK's status is today with respect to the BCS. However, there are people, both in XML UK and in the BCS, who sometimes forget the distinction, and that, if anything, has been the cause of some 'fractiousness'.

Francis is also the Chairman of the BSI Technical Committee IST/41, which is concerned with the standards that Martin Bryan had already described, and about which he would say more during his presentation.

In addition, he is an XML consultant; that is his main income-earning activity. One of his customers is Book Industry Communication, the communications standards organisation for the UK book and serials industry – books and journals publishers, and other players in the industry such as booksellers, libraries and so on. Working with other experts, they have developed the ONIX series of product information standards: this started off being about book product information and book metadata, and has now moved into the serials area, with an interesting series of standards for the exchange of metadata between suppliers of serial publications, and users of serial publications – principally libraries. These activities are international in scope. They work very closely with NISO in the USA, and with a number of other national groups.

ONIX has now extended into the area of license expression, which has brought them into contact with ontology development, and how you organise information as a publisher would wish: licenses to use materials on-line, and how that could then be interpreted in a standard machine-readable form by libraries and by users: how one may effectively communicate what are the terms within the rights to use information products.

David Penfold: David is the Chair of the BCS Electronic Publishing Specialist Group. He has been involved in the of SGML and XML in publishing, and in journal publishing. He also recently started teaching part-time at the London College of Communication, formerly the London College of Printing, where he teaches MA students in publishing about the ideas behind XML, and why they are important.

LCC also has a knowledge transfer partnership with a publishing company, which involves trying to organise the data for them. This means that there is an educational aspect within that company's work, so that they can learn to produce their information in the right format. David has a student working for him on that project who is also doing an MPhil around topics of information management and knowledge management.

David said he is really interested in the areas that are specific to publishing, but essentially are about how we can interpret metadata or information.

David had also chaired the BCS working group for the project which Judi and Carl had referred to, which had looked into how to manage the content of the BCS.

Conrad Taylor: Conrad's involvement with the BCS has been primarily through the EPSG, though in recent years he has branched out into some other BCS projects. His engagement with computers has followed the same kind of path as Martin's or Judi's – through publishing, and as a creator of information products. As well as being a practitioner, Conrad has taught publishing skills and knowledge of publishing technologies, which means he has been involved in processes of looking for information and transforming it in ways that help to communicate it to other people.

Conrad referred to the discussion document which he had made available as a prelude to this meeting. He said he could now see that it does have certain deficiencies – for example when he had written that paper, he had missed the important point that the significance of information is socially constructed.

Conrad said he couldn't claim to do much that was terribly original: his job is more to rearrange and transform stuff to help people to understand it. At one point in his life he did come close to having an original idea, after been exposed to SGML at a conference in Amsterdam of the Graphic Communications Association. Of 600 delegates there seemed to be only two graphic designers, himself and Sharon Adler, and he took part in a workshop on DSSSL with Sharon Adler and Anders Berglund. At that time he had been getting frustrated with an online email, bulletin board and database service provider called *The Manchester Host*, part of the GeoNet global system. It had a plain old ASCII VT-100 command-line interface.

He returned from the SGML conference all fired up, preaching to the Manchester Host user community, saying 'If we had client software for our desktop machines that could understand and interpret a markup language based on SGML, we could use a larger character set – distinguish between levels of headings, lists, various kinds of emphasis – maybe even have mark-up that would get client software to generate on-screen forms that you could interact with using a mouse, so we could choose between predefined categories on pop-up lists, and search for information better.' (A shame, said Conrad, he didn't also think of hypertext, or have the programming abilities of Tim Berners-Lee…)

For a long time he has been hooked on the idea that you can structure information through mark-up. He has always been a firm advocate of the use of stylesheets in document design, for practical reasons, and he still dreads receiving those Word documents where every paragraph tagged as 'Normal', making it difficult to get any kind of handle on the information content.

Conrad said that much of the explanatory media that he creates is in the form of illustrations, diagrams, sequences of images, animations, video and audio recordings. When we talk about knowledge and information management, we often keep things simple for ourselves by talking about text, and we have gained a certain amount of experience in how to handle that. But the meaning that one has in, say, a video programme, or a sequence of diagrams with accompanying narration, seems harder to handle. You can attach metadata to it, but how one can actually bring structure to such information products or get machines to understand them seems phenomenally difficult.

Summing up the introduction round

Summing up, Conrad said that this first round had not only allowed us to get to know each other better, but had allowed various issues to arise. He said that David Penfold had agreed to function as a kind of rapporteur or summariser, taking note of points to which we might return in later discussion. The discussion between Terry and John highlighted one of those issues. He now asked, had David picked up on any others?

David said he thought one was that information needs some kind of structure to be useful, and that some sort of metadata was necessary too. There was also the point that Zinat had made, about how one might be able to add structure in an automated way.

Another point was that a producer of information may attach metadata to it that reflect's the producer's view of what it is about, but the user may wish to have access to *different* metadata attached to the information, and would interpret it in a different way.

Conrad interjected that one thing we might take on board is some of the thinking that has been done around Topic Maps, of which Martin Bryan had made him aware. One might not only have an argument about whose classification scheme is 'right' – you might allow multiple classification schemes.

Nic pointed out that there was the issue that cropped up in Ian Horrocks' lecture in December 2005, about ways of establishing interoperability between Topic Maps, ontologies or what have you – and this was going to be a real challenge.

David reported that sometimes when he tries to explain to his students the principles behind XML, they say 'But what has this to do with publishing?' As for the interests of publishing on the one hand, and information retrieval on the other, it seems like an issue of 'push' and 'pull' but surely they were essentially related.

Conrad identified the 'classification versus navigation' issue as being another one to return to.

Andy MacFarlane said that one issue that really jumped out at him when the discussion took place between Terry and John, and Nic came in with that idea of context of use. How central that idea of 'context of use' is in information retrieval, and this is something we should discuss in detail. Martin added that another key concept would be that of relationships – they, he said, are what make knowledge.

Conrad remarked how noteworthy it is that whenever people come together to devise a metadata scheme, or an SGML Document Type Definition, there was much discussion trying to define what was important, and also trying to anticipate the ways in which different people might want to use the information in the future. This comes across in many of the case studies described by Liora Alschuler in her book *ABCD…SGML*, which Conrad recommended.

Nic questioned David's assertion that we would all agree that some metadata was useful. Organisations often find it difficult to justify the effort required and the difficulties that would be met in applying useful metadata. And if it isn't useful, what is the point anyway? Sometimes it is just a lot easier to throw the problem in the direction of Google than to work out which terms you have to put into an information resource to make something more structured.

Nic insisted that he not saying that metadata is not useful, but he says we shouldn't take it for granted that it is always the right way to approach an information management problem.

David said yes, it would be interesting to explore that, both in the context of the idea of context and in relation to the idea of relationship. But can one actually deal with context and relationship without using metadata?

Difficult words, useless words

John Lindsay said that we should also start to track words which are causing us trouble. The word 'relation' has a meaning in all sorts of different communities. Quite often we use the word loosely, and just to mean associations. We can *associate* things in any way that we liked. But when we define a *relation*, that has greater specificity. So, a parent and a child have a relation; it might be a biological relation, it might be a natural relation, it might be a legal relation. In comparison, around this table, all we have is *associations*. John would like us to separate those two concepts.

John also said that he had real difficulty with the way that David and Conrad were using the word 'information' – so much so, he said, that he doesn't know what it means when they use it. He wondered if we might even try having a moratorium – a list of words which we would not be permitted to use for the rest of the day – so that by being forced to use other words, we would begin to clarify for ourselves what it is we are avoiding. Out of which, we might get an idea of what it was we were really trying to talk about. He'd nominate 'information' for this curfew.

The metadata issue seems to be a profound one, because you have to make design decisions about whether metadata is defined as attributes of words, sentences, characters, spaces, images, sounds, paragraphs, chapters, sections, documents, volumes, libraries, collections, the universe, the galaxy… Unless you have got the architectural issues sorted out, it is impossible to make design decisions.

John also said that he had been hearing a silence about music – which seems to be quite an important industry, quite an important player in the game. What might the role of music be in the industries (not to mention the games) we were playing?

Finally, said John, he wasn't hearing it for the 'victims'. By and large, we are professionals whose job it is to mess up other people's lives. Only when we – as citizens, and as bounders of our public and private spaces – try to make our own worlds work do we realise how much damage we are causing. It is only at the 'victim support' level that everything comes together inside our own heads. And there are a lot of people out there who don't have the toys to play with that *we* have.

Martin Bryan picked up on two key words which he said had been used without real meaning. These were 'word' and 'punctuation'. We should not forget that what we are dealing with, in text, is a set of characters to which we apply meaning. He used the phrase 'set of characters' deliberately, because words on their own don't have meaning. It is the context in which words are used that gives meaning – that and the way in which words associate together. Unless you understand that context, and the impact that punctuation has on that context, you will be in trouble.

For example, you start to use a search engine to look for a term – you say 'look for term x' – and it finds the term x after the words '*and not*'. Now, because of the context in which the term is placed, it negates it. Information is the context within which words are used.

Conrad suggested we had now come to a point where we might usefully take a break, following which he would invite Judi and Francis to make the presentations which they had beeen invited to prepare.

He also read a quote from J E Gordon's book *The New Science of Strong Materials*. In the section in which he describes the development of composite materials, Gordon writes 'Our modern understanding of materials has been brought about by getting engineers and physicists and chemists to talk to each other, which they were rather reluctant to do.'

On that note, the group took a short break for refreshment – and to talk to each other.

Presentation by Judi Vernau on metadata and publishing

Judi had been asked to talk about metadata in publishing, or metadata *and* publishing. Not knowing what to expect, she decided to start with the very basics.

The first contentious question is, what is metadata? For the purposes of this meeting, she offered three definitions:

- Structured data about data
- Properties which describe an information object
- Attributes describing units of information

She said that she hates the definition of metadata as simply 'data about data' – it sounds so meaningless, hence her addition of the word 'structured'.

Metadata describes information objects, as examples of which she showed images of stone tablets, folios, books – or magazines like *IT Now*, the BCS journal, or CD-ROMs. These are published units of information. As soon as what we think or know comes out of our heads and is captured, then information is broadcast or published in some way; and if that is true, then we can add metadata to it.

John Lindsay said at this point that she might as well take the word 'information' out – it seemed completely redundant. Those items Judi had listed are *objects*, or *units*. He suggested that 'information' is an 'unword' invented to confuse people (and to make sure that consultants have jobs to do?)

A brief history of metadata

Judi displayed 'a brief and somewhat patchy history of metadata' that started with the ancient Chinese, grouping their texts by subject, and Callimachus in around 245 BC creating a subject catalogue of 250,000 scrolls for the great library at Alexandria. In the fourteenth century came the idea of adding information about where books were shelved, and in the sixteenth century came the first catalogue to be printed – that of Leiden University Library.

Present-day cataloguing could be said to have started in the middle of the nineteenth century, with Anthony Panizzi who as Chief Librarian of the British Library (1856–67) applied his Ninety-One Cataloguing Rules to the creation of a new catalogue for the Library. The Anglo-American Cataloguing Rules came in 1908.

Electronic automation arrived in 1966 with MARC records (MAchine Readable Cataloguing – not itself a cataloguing standard, but a framework for how to replicate electronically the cards which until that point had been the main store of catalogue content. The British Library again led the way with MARC, and sent the centrally- generated electronic records out to the other libraries – which of course very promptly printed them all out onto catalogue cards. These cataloguing efforts obviously focused on printed objects, but expanded to include other kinds of object such as microfilm, tapes, videos and disks.

Metadata considered from this perspective has been compiled by 'guardians of knowledge', rather than from the perspective of the publishers who physically produce the information objects, whether they be publishers by trade, or 'accidental' publishers such as government departments, or organisations building their own websites.

The picture gets more complicated when publishers start making money out of describing what *other* people publish. Some are even publishers of metadata and nothing else. In 1872 Richard Bowker started to produce *Publishers' Weekly* in New York, while in Britain Joseph Whitaker published the *Reference Catalogue of Current Literature*, which subsequently became *British Books in Print*. These people were creating metadata to help other publishers get their products known, but also of course to make money for themselves.

Additional services can be added onto the production of this kind of metadata, like abstracting and indexing services. Then from 1987 you find Amazon.com adding further layers, including book reviews.

Three functions for metadata

The function of metadata when used for cataloguing or bibliography is pretty easy to understand. It's about the properties that help people to know about the existence of information, and to be able to find it, which as Conrad had mentioned earlier is what we sometimes call 'discovery metadata'. Typically, discovery metadata is such things as: Creator, Title, Subject and Publication date. Other categories may be useful in particular contexts, such as Type, Language, Geographic coverage, Audience and Publisher.

Then there is metadata that can help publishers manage their content. When was the content created (as opposed to when was it published)? When was it last updated? Where are we storing it? Does it have an ID code by which we can identify and manage it? And what is the status of this information object – is it in draft form, is it published?

In addition there is metadata about utilisation, i.e. metadata from a user perspective. What do I need to know about this information object in order to be able to use it? What format is it in? If it is CD, I am going to need a CD drive for it. What size is it – if it is a book, is it large or small, and how many pages has it got? Does it occupy 77 Mbytes of storage capacity? If it is a video, what is the playing time? Then there are other aspects beyond those that have an impact of physical utilisation, such as permissions: are we actually allowed to use this?

Chunks and clumps

Up to this point, Judi said she was still talking about that object to which metadata would be added as a single unit: a published book, a CD, a video. The next stage in thinking about information comes when we delve deeper, and start to think about the smaller objects that make up the overall object.

It has taken most publishers a long time to get their heads around this. Even some of the reference publishers who started making electronic products available in the 1990s – such as Bowker, Reed – were still wrestling with understanding the concept of dividing information into manageable chunks. Some were still producing their content as one big slab of typesetting, which they then more or less just threw at a CD.

However, luckily (especially for Judi, because it has helped her make her living!) other publishers could see the value in analysing the content, breaking it down into more usable and more useful chunks, so that parts could be published separately or recombined with others.

Journals publishing is an obvious example of that: the individual articles that make up a journal are units that can be sold separately, and it makes sense to catalogue them separately. Services like Crossref interlink articles from across the different publishers so they can be searched by the end users. The goal now is to look at ways of standardising metadata, to allow interoperability, and to make shared availability more workable.

Just as journals can be broken down into articles and reviews, online publications can be broken down into articles, or even facts. Images from encyclopedias or newspapers may also be information units.

As reference publishers tried to get a competitive edge, and as users became more familiar with the technology and started to increase their expectations, we had the advent of this idea of a 'chunk' or information unit. Judi said she was quite fond of the word 'chunk' at the moment, and so are the people she is working with; it feels like something that they can comprehend.

Judi showed us this definition for a chunk: 'The smallest unit of content that is used independently and needs to be indexed individually'. That definition was offered by Jon Jermey and Glenda Browne. For 'indexed' you can read 'have metadata applied'. Indexing is not quite the same thing as metadata – or is it? That might be something we could argue about later.

The 'chunk' may be an article, a review, a news story, a poem or something smaller. Recently she gave someone the example of a crossword. That counts as a chunk, because the Across and Down clues and the grid all need each other to make sense and to be usable for its purpose.

Sometimes it is less obvious what is a chunk. Consider manuals: sometimes it is relatively easy to divide content into chunks of information – if you retrieve one of these chunks, it makes sense all by itself. But at the moment, she is trying to 'chunk' another set of manuals which have been written in such a way that it's hard to find a single piece of content that works at all on its own – each part needs everything else around it, needs its context to make it understandable; and it seems that it doesn't matter how much metadata you throw at this problem, it is not really going to work. In fact, said Judi, it looks as if they are going to have to rewrite the whole manual to achieve what they want.

Conrad remarked that this chunking potentially quite contentious, because if someone has written something for a particular method of exposition, then they may object to having it chopped up just because it makes it easier to add metadata. Yes, agreed Nic – particularly if the text is a narrative exposition, written so that it flows. True, said Judi, and in the case she was thinking of, the texts are part narrative and part 'inspirational' – the manual is supposed to inspire the company's staff with the joy of understanding company culture and know-how. Throughout the manual there are mottos that are intended to inspire them in their daily work. (*Cynical laughter.*)

John thought that the concept of a 'chunk' needs its antidote or antithesis – namely, the concept of a 'clump'. So you 'chunk' and 'clump' alternately – in a way it is not very different from cooking. First you go through a process in which you break things down, but then you assemble them again. (There seems to be a tendency in English culture to do one half and to forget about the other.)

Judi next showed a slide that shows kinds of metadata that might apply to chunks, to suit different purposes. Three of these purposes she had already talked about:

- Discovery metadata e.g. author, title, subject, publication date, type;
- Management metadata e.g. date created, date last updated, location, ID and status;
- Utilisation metadata e.g. format, size, playing time and permissions

…but to this she now added a fourth:

- Aggregation metadata e.g. parent publication, subject, type and what she labelled as 'glue'.

Aggregation and structure

Aggregation, she said, is indeed a process of clumping stuff back together again. John said that all four of these categories would apply as meaningfully to 'clumps' as to 'chunks'. Judi agreed. Also, we shouldn't assume that because she had given 'Subject' as an example of potential discovery metadata, that Subject wouldn't apply just as much for purposes of management or aggregation too.

Nic suggested that another word which might usefully be added would be 'structure', of which aggregation is a particular case. Then John suggested that structure is the synthesis, arising from chunking as the thesis, and clumping as the antithesis (and apologised for taking us too far into Hegel before lunch!); while Nic added that structure might be implicit metadata, in a way.

As an example of a 'chunk', Judi offered an information unit from some manuals she had been working on for an NHS project:

> Lighting in the operating theatre
>
> Advice on the provision of general and table lighting in the operating theatre is given in the CIBSE Lighting Guide. Semi-recessed or recessed luminaries to IP54 should be provided in the operating theatre.

This chunk was festooned with twelve categories of metadata: a title, an author, a unique numerical identifier for the chunk, a tag for which 'GCL subject' categories it belonged to (health; building and construction), creation and update dates, the intended audience, a status (guidance), a topic (lighting), and tags for 'space use' (operating theatre) and 'clinical discipline' (general). These last three metadata categories were supported in particular by controlled vocabularies.

Identifying chunks of information in this way is part of a process of making know-how much more flexible. These chunks of information serve as building blocks which can be aggregated using different criteria, and which are searchable using these criteria too.

The advantages of 'chunking plus metadata' are that you can create new publications using extracts from one publication, or indeed many publications; that you can create user-focused paths through the content – assuming that you have done your homework and know what kinds of path the user might want to take; and that you can work in a 'create once, use many times' fashion, re-using content in different publications, and on different kinds of media.

To finish, Judi presented us with something she had discovered in Wikipedia: that the term 'metadata' was coined in 1969 by Jack E Myers, and was even trademarked in 1986 – so maybe we should use it at our peril! There was some discussion about whether we should hyphenate it for safety's sake, but most felt that Myers would have a hard job stopping us from using it.

Discussing Judi's presentation

Conrad asked for responses to Judi's presentation.

John Lindsay said that we should acknowledge that not everyone in the universe is equally honest and truthful. Metadata is sometimes intended to hide and disguise. Catalogues in the eighteenth century and prior to that were quite often about private languages. That might be because Catholics were organising themselves to avoid being burned at the stake, while in other countries, Protestants were hiding themselves to avoid being burned at the stake. Quite a lot of cataloguing is in order to communicate to some people, without the people you don't want to communicate with knowing what was going on. It can be quite a task to work out what the stories actually are, because metadata is often there to tell a different story.

Adrian remarked that the ubiquitous example of people trying to tell lies with metadata is the common misuse of the `<meta keywords=" ">` tag in HTML pages, to try to get one's pages higher up the search engine pages.

John Alexander thought there is a grey area between navigation and classification, and Judi agreed. David said that in a sense, however, navigation and classification displayed the same kind of dichotomy as indexing and metadata. Indexing is intended as a navigation tool – it's the reason we do it – and navigation and indexing are subsets of metadata and classification.

John Lindsay thought that 'push' and 'pull' are more general terms: someone is trying to push you to see the world in *their* particular way, and you are trying to pull things round to the way *you* want to see the world, and so navigation and classification are 'push' and 'pull'. David said that he thought navigation was about both pushing and pulling, and John agreed.

Conrad suggested that we might develop these ideas later in the day by imagining a case study, related to the kind of information that we in the specialist groups have got – the kind of stuff that we produce on our Web sites, or in our own publications. We could think about how we would go about applying metadata structures to that, along the lines that the Knowledge Services Board were intending. That would be worth unpacking, he said, because he thought we didn't all know that story. He suggested that we move on to Francis' presentation.

Francis Cave's presentation on standards relevant to KIDMM

Francis had prepared an OpenOffice presentation entitled 'The need for standards in KIDM representation and management.' First he said he would fill in the gaps in his introduction, and talk about the two organisations which he was representing, because it would give us some feel for how the communities that are interested in standardising these issues are debating them.

XML UK is a user group – 'a group of enthusiasts in structuring stuff'. It started off with people like himself and Conrad going to events in Amsterdam in the 1980s and hearing about the work of the Graphic Communications Association in trying to standardise mark-up, leading to the production of SGML.

Francis had felt then that something really important was going on, something that would advance the whole process of representing knowledge. He had come from a background in printing and typesetting, working for a journal typesetter. His concern was the communication of science through publishing. He could see SGML would provide something which you can't get simply with mark-up that defines how stuff should *look*; that structuring this textual material was a crucial step forward.

Joan Smith founded the SGML User's Group in the UK in the mid 1980s, and it became an international group afterwards. Then a UK chapter of that group emerged, and that is what has become XML UK. It has never been large; about 150 members, never more than 200. It holds a few meetings each year, but has kept going, and still seems to find reasons for keeping going. Its objectives are to promote knowledge and news of XML and related standards. It has an extremely static Web site, as all officers are volunteers without much time, but at least is not as bad as it was a year ago – the URL is www.xmluk.org.

IST/41 and its concerns

Francis also represents BSI Technical Committee IST/41. This committee was set up specifically to represent the UK interest in the important area of international standardisation that started with SGML (ISO 8879) and had gone through a range of other standardisation efforts. IST/41 represents the UK in the Joint Technical Committee of the ISO and IEC which is known as ISO/IEC JTC1 SC34.

Among the standards that IST/41 has dealt with have been the Document Style, Semantics and Specification Language, DSSSL (ISO 10179) – a standard approach to formatting SGML documents using a dialect of the Lisp Scheme programming language created by James Clark. HyTime (ISO 10744) has been very influential; Topic Maps (ISO 13250) followed that, and DSDL (ISO 19757), the Document Schemas Definition Languages, is very much a current concern.

XML UK fits into a community of interest which embraces users in very specific sectors and applications. In contrast, BSI technical committee IST/41 is concerned with formal standards, those extremely technical foundations which are not specific to any particular application area.

Standards-making and competition

Francis said that he could not review the standards-making process without talking about the bodies involved. It is important to know they exist, and represent communities of interest which in many respects are in competition with each other. Competition is indeed a big driver in much of this – whether between individuals with very strong personal views about how these things should be done, or between competing commercial interests, and also between competing standards organisations as they try to justify their existence.

On the one hand there is the formal standardisation process. In this field, it is represented by ISO/IEC Joint Technical Committee #1 on Information Technology, concerned primarily with general-purpose standards. In the fast-moving world of Internet development, these formal standardisation bodies have found it almost impossible to keep up with change and with demands for standardisation, and have been criticised heavily for having processes that are far too slow to meet the needs of these new communities of interest.

Yet they do perform a valuable function. Once all the dust has settled on these developments, some things are going to have to last. Formal standardisation does provide some guarantees of that. It guarantees that things will be maintained; and it guarantees that heed will be taken of the interests of minority groups – that minority languages and cultural interests, minority application interests maybe, will be taken into account.

Francis also sees ISO TC 46, which is concerned with Technical Information and Documentation, as particularly relevant in this area, in particular SC9, which is concerned with information identification and description. However these two areas of standardisation effort are very separate activities – one concerned with technology, the other with applications in specific areas of information use.

Commercial consortia and standards-making

On the other hand, as a response to the fast developments in the online world, we have commercial consortia which have sprung up to respond rapidly to the needs of evolving markets in way that the formal standardisation processes have failed to do. The World Wide Web Consortium is clearly the leader here, able to develop standards rapidly in response to technology and market developments. Some of these standards have been very successful – HTML clearly, also XML and XSL. One might argue whether RDF and OWL have been as successful, but they are certainly very relevant to what we are talking about.

Other standardisation efforts of the W3C have been more problematic. Part of the problem of this relatively informal, rapid development process is that sometimes paths are taken which are controversial – perhaps a bunch of people get together to agree something, but other experts out there disagree. In the formal standardisation process, there is much more likely to be a consensus-building approach which tries to resolve some of those arguments before things get standardised. That of course is what slows things down in the formal process, because if there really are serious disagreements between groups of experts, then perhaps no standard is achievable. In the W3C kind of situation, somebody can produce standards such as Namespaces, XML Schema or XSLT 2, despite the fact that there are communities of experts out there who are profoundly concerned and disagree that these are the right ways to approach this kind of thing.

Commercial consortia also don't cater well for minority interests or alternative viewpoints. For example in XML itself, although it is obviously a global standard, Francis is not sure that the development of the Schema Language support within XML has really addressed some of the minority positions of some of the experts involved. (This has led to the development of DSDL, and before that, RELAX.) Those could be said to be minority technical interests. He also wonders whether ultimately minority commercial and cultural interests are sufficiently taken into account. Francis said he didn't want to get into W3C politics, but there is a definite view that W3C is dominated by a few gorillas like Microsoft and the like.

Conrad commented here that many of these standards hugely privilege the English language. Francis agreed and added that they also privilege a certain technology focus, particularly with Web services. But the response to that would be, 'Well, if you're not represented by this, set up something else.'

Which is what people have done – and they have done it with OASIS – the Organization for the Advancement of Structured Information Standards (www.oasis-open.org) – which is now a kind of counterbalance to W3C for certain kinds of standards-making, in a way that the formal standards bodies haven't been able to achieve. OASIS started out very much as an SGML applications standards organisation – indeed it started as 'SGML Open', a consortium of developers and users trying to build applications standards for SGML use. When XML appeared in 1998 they saw a role for themselves and have moved into that very strongly. OASIS can take on issues that W3C has chosen to ignore or sideline, perhaps because they don't fit with the game-plan of Microsoft or whoever: issues such as OpenDocument and DocBook, DITA (the Darwin Information Typing Architecture), and other things which slightly compete with the technologies already adopted by W3C.

It's also interesting that OASIS have decided that there is a value in formal standardisation to ensure maintenance of standards for the long term, and so they have established formal relationships with ISO and other formal standards bodies in order to try to bring about some long-term stability – OpenDocument being a good example of that, having been nurtured by OASIS, but now going down the formal standardisation path.

It is also interesting to see the W3C community also moving in that direction; perhaps they see a threat, or they recognise that an engagement with formal standards is something which they should have thought of doing themselves. OpenType is now being standardised within ISO as a joint project between SC34 (with which IST/41 corresponds) and SC29, which is the MPEG community.

Sector-specific standards-making

There are smaller bodies which are specifically focused on particular technologies, or the needs of a particular sector. One example is the Dublin Core Metadata Initiative, responsible for developing and promoting nine basic metadata fields that can be used as a base standard for resource description and discovery.

Francis said he often gets asked why the book industry hasn't adopted Dublin Core for the description of books, and why they developed their own metadata standard ONIX instead. It partly depends on how narrowly or widely you define the context, he explained. If you define your context very widely, you can get people to agree on a relatively small number of assertions about things – about authors and publishers and dates and things like that. Whereas if you have a very close focus on supply-chain communication between publishers and booksellers and the intermediaries in between, there is a much greater range of considerably detailed assertions that you can get everyone to agree to accept, so that information flow between the parties is rich and detailed in content. Amazon, by the way, was a big driver behind ONIX.

Another example of a sector-specific standards-making body is IMS Global Learning Consortium, which develops and promotes the adoption of open technical specifications for interoperable learning technology, including learning object metadata. That is an important area of development, and it might well have implications for other areas outside of learning technology. We can learn from how they have gone about things, what they have achieved, and also what they have not achieved. The learning technology people are also moving in the direction of formal standardisation processes, and have their own ISO subcommittee.

EDItEUR (www.editeur.org) is the international body in the book and serials industry, that oversees the development internal communications standards (BIC is the UK body), and Francis had already described their ONIX project and its product metadata standards.

Why standardise? What standards are key?

So what is the point of standardisation? Well, standards are seen as providing quality, completeness and implementability; they enable us to feel we are covering the domain. They represent the considered, pooled views of experts. They offer some guarantee of reliability and durability. And they have the virtue of being open, interoperable, independent of vested interests and resting on a broad basis of representative participation.

What are the key standards in the area of knowledge and information management and representation? Francis offered us his own personal 'hot picks' short list:

- **In the field of structured data representation:** XML, XML Schema from W3C, RELAX NG from OASIS, and DSDL from ISO.
- **In the field of knowledge representation:** RDF and OWL from W3C, Topic Maps from ISO, and perhaps DITA from OASIS as a specialist application in this area.
- **In the field of metadata vocabularies:** Dublin Core is definitely the basis.

Document Schema Definition Languages

Two things which IST/41 is currently working on are DSDL and Topic Maps, and Francis told us some more about those. The Document Schema Definition Languages – note the plural – is a multi-part international-standard toolkit of schema languages which has been developed on the premise that you cannot model data successfully with one language for expressing that model. There are various aspects of how data needs to be modelled, particularly for validation purposes, and therefore different forms of expression are needed.

There is a place for grammar-based expression, which is what RELAX NG is, and which is what XML Schema also is to some extent, and what SGML DTDs do. There is also a place for a schema language that allows you to make assertions about relationships between different patterns in the material, which is what Schematron allows you to do. There is certainly a place for a language that allows you to describe the kind of data you are dealing with at the lowest level of information – things like dates and numbers and words, data-typing in the relational database and programming sense – for which there is a Datatype Library Language; and there is also a Namespace-based Validation Despatching Language, NVDL.

Francis described DSDL as a rag-bag of things that have been put together to meet a need, but he hopes that it will ultimately become a coherent whole, a toolbox of schema languages which will enable people to do all the things they need to be able to do in order to define how data should be structured and organised, but also to check that it has been structured and organised in that way.

In addition, the DSDL tool-kit includes standards for character repertoire validation; Martin Bryan is working on the Document Schema Renaming Language in particular, as well as being the editor for the whole of DSDL and the head of the working group that is responsible for it. Francis, for his part ,is trying to drag DTDs into the 21st century as part of the effort, by adding support for data-typing and namespaces, the aim being to take those data models which are constructed as legacy DTDs and add value to them.

Topic Maps

The Topic Maps standard, ISO 13250, was originally published in 2000, then revised, and the current standard is the second edition (2003). However, the work still goes on, and there has been a realisation that (unusually for an ISO standard) it had been approached too quickly, and without enough deliberation.

Now a re-evaluation of Topic Maps technology is under way which will result in this original standard being replaced by a multi-part standard, of which the seven parts are: the overview and basic concepts; a data model; an XML syntax; 'canonicalization'; a reference model; a compact syntax; and a graphical notation. A Topic Maps Constraint Language is being developed, and a Topic Maps Query Language – something which has some similarities to what is going on in the world of relational databases. It's too early to say what will shake out of this in terms of useful stuff that will help people's work.

Drivers for standardisation

What are the drivers for these voluntary efforts towards standardisation? Apart from some kind of altruistic feeling that standards are a 'Good Thing', there is also the perception that existing standards are insufficient to meet our current needs and certainly won't meet our future ones, and that there are gaps which need to be filled. Commercial players in the game are motivated in their support for standards because acquiring the kind of 'imprimatur' or stamp of approval which comes from embracing standards is good for market share. There is also the kudos which individuals and standards bodies gain from having their solution to the problem of knowledge representation accepted widely as the standard.

In conclusion, Francis said that developing standards for knowledge representation is a natural follow-on from the initial effort to standardise data representation (i.e., encodings like Unicode) and information representation (i.e., structuring it with XML). At present we can create static representations of knowledge – a particular view of it at a particular time, and in a particular context; but if we want to have true knowledge in electronic form then we have to have a way of representing things that are dynamic, that are changing over time, that reflect the way our knowledge develops, how relationships change within it, how new knowledge is added and assimilated.

Discussion following Francis' presentation

Martin Bryan added that one thing that Francis failed to stress is the fact that ISO as an organisation is multicultural. It therefore has to take into account different languages and different cultures. It cannot design a standard that is specific for one language only.

One of the things on which much time was spent when SGML and DSSSL were being devised, was dealing with character sets other than ASCII, in the days before UCS. That is why SGML had all that stuff in it about character sets and character entities. When developing DSSSL, they had to consider scripts that compose from right to left (Arabic), or vertically (traditional Japanese), not just the left-to-right that European languages use.

When you come to look at things from a multicultural viewpoint, there are a lot of different constraints, and a lot of different things that you need to be able to do. In Martin's opinion, one of the things that hasn't been looked at properly in the knowledge community is multicultural knowledge.

If you have to build consensus, said Francis – and the building of international formal standards is ultimately constrained by the need to build that consensus – you are forced to be fair, and to be rigorous in your argumentation. Standards that come out of that process, therefore, are ultimately of higher quality than those that aren't subject to this kind of rigour. As for the most successful standards that the World Wide Web Consortium has produced – HTML, XML, XSL – they are almost directly derived from formal ISO standards. XML came from SGML, and XSL is very closely modelled on DSSSL.

Terry referred to a standard he came across recently, ISO/IEC 11179; it calls itself a metadata standard. Did Francis have any comments on that? Francis said it was not one with which he was familiar – something to do with information modelling in an IT context. Keith thought it was the 'property' of SC32. He added that it has been driven by 'a load of statisticians' – indeed that SC32 has been hijacked by these statisticians. They seem to have suddenly realised that they have got it all wrong, and that they need to be thinking about relationships between things. As an information systems person himself, Keith thought it was very interesting that they had come to this realisation so late in the day.

The authors of ISO/IEC 11179 use a definition of metadata that is very, very different to the one that Judi and Francis had been using, said Keith. In fact, he thought the SC32 crowd were using the term 'metadata' in the way *he* has always thought it should be used – which wasn't how most of the people in this room were using it. This is part of the problem we have – people use terms like metadata without actually defining what it is they mean.

Francis said he had chaired a BSI workshop about metadata, at which were present representatives from different BSI technical committees. They had a huge problem because everybody felt they 'owned' the term. One lot would say 'Well, we don't recognise your right to talk about metadata – that's our term.' This was a serious problem. Some groups were worse than others about it, but there was very definitely a feeling that different groups for quite legitimate reasons had come up with a view of what metadata was that matched their particular context. Terry wondered therefore if we today would be able to concentrate on a particular view.

John suggested, perhaps there is a boundary between making standards, and playing the sort of games that Tim Berners-Lee or Bill Gates would play? That boundary is about deciding what the issues are, and that should be our job. John had been reminded by Francis' presentation of a magazine he used to publish 20 years ago. He had had endless arguments with the people who designed the page layout. They wanted an attractive page; John wanted semantic content and intellectual coherence. So his idea of graphic design was the intellectual coherence of the argument; their idea was of a certain 'look'.

This was the same sort of argument as the argument about metadata. Some people will hold that it is the 'author' concept (for example) which is the metadata element, while others will argue that it is the *population* of the 'author' concept which is the metadata element. And that points back to chunking and clumping again.

John thought Francis was absolutely right to stress the relationship between the narrowness of the specificity of the interest. From MARC we begat – through the Article Numbering Association – a population of the ISBN data straight into the ANA with absolutely no trouble at all – from MARC to ISBN to ANA there was just complete and straightforward population and inheritance. The whole thing just worked. Nobody got a Nobel Prize for it, nobody got a knighthood for it – it just worked.

Like chunking and clumping, we need the concepts of 'grid' and 'group'. If you have a very clear grid, then people will group very tightly. If you don't have a clear grid, then people will group in very unclustered ways. And if we can map chunking and clumping onto gridding and grouping, then may we be very close to having the schema that we need in order to try to explain the interests?

Standards-making personnel problems

Conrad said Keith's description of a group of statisticians as 'standards hijackers' made him aware that a problem with standards development is that the process is dependent upon a relatively small number of dedicated people. Where would DSSSL be without James Clark, for example, or Topic Maps without Steve Pepper?

Conrad had recently been nominated as the BCS representative on IST/41, and is having huge problems in catching up, studying the standards which are currently being discussed. The texts are written in a terrifying, user-unfriendly way.

Committees like IST/41 are often supposed to make some representation about draft standards at the ISO level – yet when they meet to discuss them, the documents haven't yet been written by the key person who is drafting it! The draft will come to the committee at the very last moment, at which point it is very hard to comment on this 400-page-long monster.

That makes it institutionally difficult for a body like the BCS, which should be involved in standards development, to get involved effectively – even if it felt like making that effort. Some individuals of course do make the effort – within the BCS-EPSG committee, Ann Apps in involved with the Dublin Core Metadata Initiative. Ian Horrocks, of course, is very much involved with the Web Ontology Language.

Standards-making too slow?

Genevieve said that back in the 1970s when she was in industry, she was aware of a lot of British Standards that seemed to be important. But what really had struck her was that by the time the standard got developed, the market had moved on. What standards actually seemed to be doing was holding progress back. Some of this was more obvious in the kinds of issues she was looking at then than in the kinds of issues we were looking at today; but she thought one has to be aware that one might come up with a very fine standard after all this huge process, but by the time you have come to it, we will be five years down the road.

On the other hand, said Martin, when you look at standards like SGML, DSSSL or Topic Maps, they all came into application ten years after they have been started on in ISO. So, sometimes it is the other way round.

Standards in government

Nic thought it interesting that governments are starting to attempt to wield some clout in terms of standards. Quite understandably, they object to being forced onto the commercial treadmill for no real benefit – indeed, being taken down proprietary avenues, such as the latest extensions that are not W3C-conformant, thereby making it difficult to have genuinely open procurement policies.

The OpenDocument issue is a clear illustration of this. Microsoft are trying to take a short cut and use ECMA to ratify the Microsoft Office standards, and there is a war going on. In a sense, the commercial organisations are exploiting the standards-making processes. There has been a period of explosion in functionality, though it is rather less clear what the significant marginal benefits have been, or will be from the next set of developments, and so the stability of standards is seen being of no small value.

Terry reported the experience of the Records Management Team at the National Archives, who do the extraction of records from other government departments (or 'OGDs' as they call them). Unfortunately each government department has its own idea of what is reasonable as a format in which to save and store these. So, when Nic talks about a governmental view, Terry has little confidence that there actually is such a thing. There may be departmental views – but in Britain, not a governmental one.

There was supposed to be 'eGov', wasn't there, said Conrad? – and the Government Category List taxonomy that Stella Dextre Clark among others worked on. And the Government Metadata Framework, added John. Well, unfortunately the reality is that when people gather data, said Terry, it may start off as a spreadsheet or an Access database or even something less structured, and it comes with the procurer's idea of what is in there. That procurer may sit in another room from somebody else who has a completely different idea, even in the same department of government. Whether standards are available or not, and what departments do about that, is out of his hands, Terry said.

David commented that one thing he always finds is a lack of understanding about transforming information formats. There seems to be a general feeling that once you have entered data into an Excel spreadsheet or Access database, you can't transform it to another format. As far as the National Archives is concerned, replied Terry, all data is transformed into their standard format.

Still, said David, it seems to him that there is a general lack of understanding that it can be done. Technically, it can *always* be done, said Keith. The question is indeed whether people understand this.

But an Excel spreadsheet may have been printed to media – this is a very common thing, said Nic. Well, for the

National Archives, said Terry, it is up to them to make sure that a transformation to a standard format takes place, even if it means typing the whole thing back in again. All National Archives information is held in XML.

This comes back to our obligations as educators and professionals, said John. He still sees books on database technologies referring to a comma-delimited file; he sees books teaching people how to build databases that refer to the comma-delimited file; and they use the example of library records as their case-study example. How ridiculous – *haven't these idiots discovered that book titles have commas in them*? There seems to be a complete lack of common sense. (Comma sense?)

But in order to do these kinds or transformation, Keith said, you have to model what it is that you have got. That's why some of the data modelling work that is going on is important, as well as always topical.

At this point, we broke for lunch.

Discussions after lunch: stories and use-cases, contexts and practices

During lunch, Conrad had created a couple of wallcharts and posted them in the room. One of these repeated his earlier point about four main ways of making information manageable; the other was a graphical display of a timeline-*cum*-concept map showing how our current 'KIDMM' technologies had arisen from a basis in text handling and databases, with the publishing industry as a very significant driver. For the purposes of this report, these diagrams and Conrad's explanation of them has been relegated to an Appendix (see page 33), but naturally they were referred to in the discussions which we report below.

Reactions to Conrad's charts

Miltos commented that Conrad's diagrams put a lot of emphasis on the *text*-based aspects of information. But increasingly, information objects are more and more complex – even the text-based ones – and information can be found in all sorts of media, for example in multimedia. Metadata and indexing can be related to those forms of information container as well – for example, to a section of a video clip, or to part of a discussion within an audio file.

Yes, said Martin, but you tag them using text labels. Miltos agreed, conventionally that is how you do it; yet he didn't want us to lose sight of these other media forms, because he felt that in the future more information would be carried in them.

Conrad replied that, as he saw it, metadata applies quite easily to such media forms – it is simple to hold a metadata record externally, and some file formats allow metadata to be embedded, such as EXIF data in photographs, or the descriptive metadata within MP3 audio files. Indexing also applies quite easily to these media forms (and one can supplement 'pointing' schemes to time-based media by using timecode references, such as SMPTE timecode for movies). But can one make media forms structured?[11]

David Penfold referred to a discussion programme on Radio 3, *Lebrecht Live*. In a recent edition of this on 26th February 2006, Norman Lebrecht had commented that everybody wants to see things in terms of pictures, and wondered what effect that was having on the written word.

To quote the Web page for this broadcast…

> 'Are we being overwhelmed by what we see? We live in a world dominated by image. What the eye cannot see, the mind will not absorb – or so television and the internet have led us to believe. If it doesn't make a good picture, newspaper will relegate important ideas to the dump bin. In a glut of eye candy, the aural and the verbal are being drummed out of sight. One of the reasons orchestral music is in decline is because it cannot be illustrated – except when subjected to extreme theatrical overlay. Books are judged by their covers. Television is blighted by visual stunts. What can be done to reverse the dictatorship of the visual? How can we put the word and the sound on an equal footing with the image? Have we actually destroyed the balance of our minds by pursuing a pictorial pornography?'

Obviously, said David, Lebrecht didn't think that this development was a good thing. (Ah, someone commented, that's just the sort of view you get on talk radio! And he said it using words!)

Pointing to Conrad's timeline/concept map chart, Genevieve said that Conrad had indicated ontologies at the most *recent* end of the process of development. But her recollection was that ontologies came into vogue at about the same time as relational databases – which would be the early 1980s. If you wanted to mention thesauruses as well, said David, Roget's Thesaurus comes a long time before that (created 1805, and first published 1865). But Conrad explained that he was referring to computerised systems for defining and processing ontologies, such as OWL – and they are very recent.

Taxonomies should have been added into that group, said Nic. It means something different from ontologies, and it goes quite closely with metadata.

[11] There are some simple ways within certain kinds of file format. For example named markers can be placed in the audio streams of AIFF and WAV audio files, and made to trigger events in a multimedia presentation. Simlarly, video streams can carry markers, and this is used to provide 'chaptering' in Video DVDs.

What's in a word?

What we might now usefully turn to, suggested Conrad, would be to look at the variation in meanings that can be attached to words like 'information', 'knowledge', 'data', 'metadata' – following on from John Lindsay's observation that the word 'information' is often redundant, and from David's reported experience of discovering that different people have different meanings and uses for the word 'metadata', on which Keith had also commented. Who would like to start?

Terry said he would. David had suggested (in email discussions before this meeting) that we look in some online dictionaries for definitions. He couldn't access all of those David had suggested, as some require a subscription; but he tried putting the term 'metadata' into Google, and as a result got lots of different definitions and interpretations. So, on which definition should we rely?

Martin asked if he might come up with a controversial definition of metadata, information and knowledge at the same time, and came forward to work at the flip chart. 'What is the relationship between these things?' he asked, and wrote a series of numbers on the chart, as follows:

12
365.25
8766
525960

This is some data. How to make sense of it? There appears not to be any obvious relationship between them. However, they *are* all connected, because they are all information of a particular type. And the type is dependent on the context within which these numbers are used. Who could make sense of these? As a clue, he added a single letter beside each of the numbers:

M	12
J	365.25
h	8766
m	525960

While we scratched our heads, he warned us to be aware that we were thinking in English. It was John Alexander who identified the abbreviations as denoting (in French) months, days, hours and minutes. Martin admitted that he had played that 'J' trick on us so that we would deliberately have to think about the problem of language. Now we can infer the missing piece of information – *not a piece of knowledge*, he was at pains to emphasise.

So, next question: what is the *knowledge* we can gain from this? We floundered around making all sorts of suggestions, so Martin underlined the 12 and the 365.25, which indicated that this data is specific to Planet Earth, which takes 365.25 days (i.e. rotations on its own axis) to orbit the sun. Terry emphasised the distinction between the arbitrary nature of the number of months, which is just a human convention, and the non-arbitrary length of a day and a year.

The figures have a pattern, and you derive knowledge by what you can infer from them, not what you can determine just from the relationships between them. That indicates the difference between knowledge and information. The abbreviations provide the clue to the context – and those are the metadata, the things that tell you about the type of information you are looking at.

John Alexander said, surely one brings that knowledge to those numbers? Yes, said Martin – that's the key thing. But just looking at this, said John, he was unable to take that knowledge from it, he could not infer and derive the knowledge from what Martin had written up on the flip chart. But with a different value for the number of days, and given a set of astronomical data about the solar system, you could probably work out the values for them, said Martin. True, agreed John.

David remarked that this is rather similar to a point he makes repeatedly when he lectures his students about XML: that XML doesn't do anything by itself, and any indexing system doesn't do anything by itself either – it is just a way of storing data.

But John Alexander thought that the definition of knowledge which Conrad had given in his pre-meeting paper was a better one. Conrad, however, said that he had offered a definition of *information* – and he was beginning to wonder whether he should take it back! He had been coming at the issue from the same kind of direction as Liz Orna, as someone concerned with creating information products, and therefore taking knowledge in someone's head, or in an organisation's collective 'head' (or collection of heads) and making information products so that people could transfer that knowlede to other heads. But he added that a communicative act is not the only context in which knowledge exists: you can discover it for yourself as well.

Terry said he was not happy with the model offered by Martin, because there was an implicit knowledge and understanding built into it. For example, knowing what days, months and relations such as multiplication are; that there are 60 minutes in an hour and so forth. This should all be part of the information that is part of the metadata, without making assumptions about people's prior knowledge. Saying that there are 24 hours in a day, that a day is defined as a rotation of the earth on its own axis – these are big chunks missing from what Martin had written on the flip chart.

Conrad commented that one of the interesting problems we bump into when we try to get machines on our side as helpmates is that we cannot assume that they have this prior knowledge. We have to give it to them.

Stories are better than definitions

John Lindsay said that he thought we needed a different approach. Trying to define things in terms of words and their meanings is a futile exercise. Words and meanings are referential, and you will end up running round in circles.

What we should do instead is *tell stories*. By telling stories, we will begin to get a shared understanding of what we take these things to be. He wondered whether we should try – say, before the end of the summer – to have a session where each of us would bring along some examples of the stuff we have done in this field, and put it up on the walls, or whatever other appropriate form of display, and explain the stories behind them.

If we look again at Conrad's paper, he continued, section two of that attempts some definitions. Conrad also shows that Shannon uses the word 'information' in a way that is completely incomprehensible. (Indeed, when he had read Shannon, John had tried to work out why on earth he had used the word 'information' at all, and he couldn't.)

Into that section of Conrad's paper, John could throw another nine uses of the 'information' word – the story of the British Government's Ministry of Information is a glorious example. We could have a little exhibition of objects from the Ministry of Information.

In the English Book of Common Prayer, the court of Elizabeth I uses the word 'information' in the context of the phrase 'laying an information' – essentially, informing the authorities that someone was not at Church, which has the consequence that someone gets burned at the stake, and the informer gets that person's money…

In all, John currently has nine meanings of 'information' in his collection, and we could gather some more stories of what this 'information' word means.

From this, we can develop what can be called 'use cases' – and use cases will be useful later on, because then we have a use case object. If we built up a little library of use cases for the concept 'information', we can quite easily do the same for 'knowledge'. We would begin to find that there are corpora where these things have context, and they have content. And that means they have coherence, and they have cohesion.

Exhibition cases and talking heads

John thought that the building we were in, the Davidson Building at Southampton Street, would be an ideal venue for an exhibition – the space just lends itself to that purpose. So we should set ourselves that as a task: if we planned an exhibition of the various use-cases, we could all contribute, and we could then produce a catalogue. Catalogues are beautiful things. And we could support our catalogue with a scholarly apparatus. It would really look quite nice.

If we are having a bit of fun thinking about this, said Conrad, he would add that some years ago there was an exhibition in Den Haag called 'InfoArcadia'. It was an exhibition of information design objects and artefacts, presented as an art exhibition would be. The organisers videotaped a short explanatory 'talking head' statement from each of the people who had contributed a display to the exhibition, and these had been made into QuickTime movies and presented on a collection of computers. A visitor could sit in front of one of these and view so-and-so's explanation of why he had done what he had done. One could accompany a KIDMM exhibition with something similar, and then make available a DVD of all of the presentations, as well as images from the exhibition.

Conrad returned to John Lindsay's idea about collecting use-cases in the form of stories. In the Oxford English Dictionary, you don't get a single definition of a word, but a series of usages, and even examples, such as how a word was used in the 14th century, in the 17th century, and so on. Also, the word is explained as meaning this thing in one field, and that in another. Doing something similar for the words that are causing us problems seemed to make more sense than trying to nail down a single meaning.

William Empson tried to do that, said John Lindsay, in *The Structure of Complex Words*. He actually tried to move a step further, and said that if you develop a nomenclature, then you can say, 'In this use-case, here is the evidence.' John could not understand why Empson has disappeared almost without trace, in comparison say with Ted Nelson or Vannevar Bush.

Genevieve said that we should spark each other off by telling such stories – we can do that in part when we break for tea, and it had happened at lunch-time in a big way. We can also do this by email. We need to hear each other's stories to be sparked into remembering our own.

Breakdown and analysis

Berin thought that trying to come up with definitions might be a lot of effort for no useful purpose. He would rather subdivide something, analyse it, understand what the components are, and then come up with a definition in which one could have some confidence, having some way of testing it. Whenever Berin has tried to look at managing information, he said, he has got nowhere until he has broken down that information to understand it. Is it news he is dealing with? Regulations? When you know what you dealing with, then you can make some progress.

Conrad asked Berin to clarify what he meant by 'breaking down' the information. To give an example, Berin described working with Butterworths, the legal publishers. They deal with legislation, which has been passed by Parliament, which their editors have annotated, and which other commentators have commented on. That is a pool of knowledge, and one which is moving forward all of the time, because the law is always changing. It has certain characteristics, and you can begin to try and manage this legislation if you can understand its characteristics.

There is also case law, coming out of the courts: there are accounts of how each case was handled, and verdicts, all the parties, and so on. Case law has its own characteristics – cases go through appeal processes, and this could lead to another case report, and so on further up the chain. Those too get annotated and cross-referenced, and have commentaries that go with them. If you are going to try to manage that sort of information, you can do so if you understand it.

Butterworths also have quite a lot in the press – legal commentaries, leaders, feature articles, whatever. So they really do have a lot of different types of information that they need to marshal in order to operate their information service as a legal publisher.

Suppose Butterworths come to Berin and say, 'We've got a problem; can you sort out my information?' – then that's what he has to try to do. He finds that it is the same wherever he goes: people think they have an information mess. It's only when you begin to break it down into its components that you can make any real progress.

Nic agreed that trying to pin down definitions would be an elusive goal. People have a mistaken belief that words have an absolute meaning. His idea of 'chair' would be different from someone else's, and if that can be true of straightforward inanimate physical objects, how much more true it is for more abstract concepts such as information and knowledge! His suspicion is that it would be a disproportionately difficult task, if achievable at all; rigorous definitions would elude us. But he thought that John Lindsay's 'stories' approach was a very good one.

What's data, what's metadata?

Keith said that when Judi gave her presentation, her first slide had three definitions of metadata – and he had totally agreed with those. But while following the rest of the slides, he thought: 'She's not talking about metadata, she's talking about something else… She's talking about what I call *data*.' Where Conrad had written on his first wallchart 'added metadata', that was not added metadata at all either – it was added *data*. Metadata is saying 'The added data I want is: the name of the author, the date this was created'. That is metadata, said Keith – that which describes the *structure* of the data that is required, not the added data itself.

He gave an example from his Open University teaching. The OU has a course which talks about metadata, and it annoys him that in the multimedia section of this course they give the example of adding metadata to images, so that you can search to get the right image. Suppose you were storing images of people, and you wanted to be able to find all the blue-eyed blondes, you would need a record of the colour of hair and the colour of eyes, and that would make a search possible; this was the illustration they used.

But this is turning things on its head, said Keith. In the days before it was possible to store photographs in computer databases, one of the pieces of data Keith would have wanted to know about all his employees would be colour of hair and colour of eyes. Throughout his military career, on all his ID cards, and therefore on a database somewhere, the colour of hair and colour of eyes were defined. Later on, it became possible to add photographs. But what came first, the photograph or the data? It's just data to Keith.

Martin jumped in to say that 'hair colour' is a label describing the *type* of data. John Lindsay objected, 'Hair colour can not be a type of data!' 'It is,' said Keith, 'though it can have a number of specific values…' And half a minute of chaos ensued…

Nic Holt aimed beyond the impasse with an observation that one can think of two 'bits of stuff' – to avoid any other contentious labelling – which are related to each other. Within one context, one bit of stuff could have the role of being metadata for the other. 'Meta' means 'with' – and 'A with B' has the same meaning as 'B with A' – the relationship is reflexive or symmetrical.

Earlier we had been discussing wikis; this is a facility within which one could create a bit of text about something or other, incrementally. If you generalise that idea, you could gradually build up a network of articles, comments on articles, whatever. It would not be structured in quite the same way as a wiki, but you could imagine within this network of linked texts and annotations, some of them would function as metadata for some of the others.

Keith insisted that he thought there was a real problem developing here, because if you now have two communities using the term 'metadata' in two entirely different ways, it is probably too late to recapture it for just one of those communities. Therefore almost certainly every time you use the term 'metadata' you have got to qualify your use by saying 'and I'm really using the term metadata in this context, and to mean this.'

Nic wondered, if we take Keith's example of an HR database, and consider the textual stuff on the identity card – is that the data of which the photograph is merely another attribute, or indeed a bit of illustrative metadata; or vice-versa? Keith replied that because it is in a database, and available to the user, then as far as he is concerned it is all data. Metadata, in contrast, is where you saying 'In my database, I will have colour of hair, colour of eyes, and I will have a photograph'.

But that, said Nic, assumes one has complete control of the database. In real life – and this is what a lot of the discussion about the Semantic Web is about – there are bits of data everywhere, and some of that data has the function of trying to tie other bits of data together and relate it and so on. In which case you cannot simply say 'It's all data.'

Communities of practice – contexts of use

Andrew offered a comment on Keith's statement that different communities view the idea of metadata in different ways. He thought that we were approaching this from the wrong direction. We had been discussing the meaning of terms which cannot be captured, because they mean different things in different communities, who have adopted them for their own particular ends.

Earlier, Conrad had read out a statement [from *The Science of Strong Materials*] about physicists, chemists and so on, who had been forced to work together in the process of developing composite materials. The reason why those communities had been able to arrive at a common understanding was not because they spent some time in a room discussing their particular views about terms, but because they identified a set of common problems and started tackling them. That was what led them to come to some common understanding about terms.

Tackling this problem head on would be a mistake, and maybe it would be more profitable – in order to bring the interest groups together – to find out what the common problems are, and let them collaborate on those problems, and effectively by subversion come to agreement. And get some useful outputs out of the process at the same time.

Miltos commented that the different communities who are using these particular terms are communities of practice; and what brings them together is those practices and processes.

One of the problems he could see is illustrated by an attempt he knew of to apply artificial intelligence to law, in which it had proved hard to develop the system. It had been fairly easy to get the Italians and the Argentinians working together, because they have similar legal processes; but it was hard to get either of those lot working successfully with Anglo-Saxons, who have such a different legal system.

What comes first? In Keith's example, you might record that someone has blue eyes – or, could one have written an algorithm that would look at a photo of the person and come to that conclusion? You can't easily grasp this kind of information, without the context of the process that created it, has formed it, or that wants to use it. And if there are incompatibilities between these, it is very difficult to get everybody to agree what we have.

If we had, in theory, a way of mapping between the different processes, then we could try to analyse it there and say 'what I mean by that, you probably call this'. Obviously the best way of doing this is by using a lot of examples, a lot of cases, where we know in practice that *this* is equivalent to *that*. That way, we could try to unravel it. Otherwise it is difficult – there is a lot of entropy of information.

The knowledge curve starts as tacit information in somebody's head. The moment you try to capture it, a lot of it is lost. The more you try to structure it, the more you lose part of that information. What you end up with is fragments that you have to try to put together, and unless you have a Rosetta Stone that can help you to put things together, it is very difficult to do anything with it.

Patterns in stories; primitives and constructions

John Linday said, fine, we have Keith's Human Resources story; we could deal with that quite quickly – we've got brown eyes, and bald people, and we understand that.

Earlier, John had taken a piece of the flipchart paper and jotted down 'Butterworths' and 'Lexis' and 'Oxford National Dictionary of Biography' and 'The National Archive' on it, and stuck it on the wall. We all know something about these four things too. John declared that he could see commonalities in those four cases, and certain common patterns, without any trouble at all.

But if he turns back to the Human Resources story, it seems to him to be different from those four examples on his chart. He cannot see the pattern that spans across them – though, he can see that there is a pattern, because they both have people's names in them. The author is a name, and human resources objects have names in them. 'Name' is a shared property.

We know from a long time ago that we have to separate family name from personal, given name. We know that the BCS doesn't know this – this knowledge has been lost. (*Much laughter.*) We know there are problems with double-barrelled surnames. We know there's the 'van der' problem. But Keith van Rijsbergen doesn't know this – this knowledge has been lost. And Keith is a K, which is further knowledge lost, because he is actually a 'C'.[12]

We can begin to build a model which starts with the concept of things like the hyphen and the space, the ASCII character code – because that we do know something about – and then we could begin to build what effectively would be a taxonomy.

This taxonomy would say that we understand the character level, we understand the name; now, the name has patterns. What are the other objects like names? John wasn't sure there *were* any – names might be unique primitives. How many unique primitives do we minimally need to at least get the conversation started? If we don't

[12] Professor CJ 'Keith' van Rijsbergen, leader of the Information Retrieval Group, in the Department of Computing Science at the University of Glasgow. So, Keith is not his real name, though everyone knows him thus.

need more than eight, then we could do it before we went home today.

John thought 'date' would be a problem, because we've got MM/DD and DD/MM, and we've got YYYY. So dates and times might be a problem.

We left GIS out of our original round entirely – it might be that 'places and spaces' constitute primitives that we need to take account of.

We've got the Article Numbering Association, with all their barcoding stuff – that issue is by and large cracked. There, now we have five data primitives. Not bad going for a day's work, is it?

Conrad replied that it would *not* be bad going for a day's work, and it would be worth noting as a useful day's work to do someday. However, there are other things that we might want to accomplish during what remained of the afternoon. We had been talking about communities of practice using words in different ways. One manifestation of communities of practice within BCS is that we have Specialist Groups, and the Specialist Groups themselves fit into different communities of practice. Perhaps we should talk about that.

How to establish definitions and categories

Terry offered us an example for consideration. Some people decided that there were so many musical catalogues, all in different formats, that they wanted to pull them all together on a useful Web site. Anybody could add things to it, could edit what was on there, and could define new terms. So, when Terry had pointed out to them that they had forgotten lyricists, they said, fine, no problem, we can add a category for lyricists.

But before they could use that, said Terry, it was first necessary to *define* what a lyricist is, so that anyone else coming into the catalogue would understand what it means, or has been defined to mean. This was a way of acquiring metadata:[13] it did mean that this allowed a community interested in music to figure out what they meant by a term, once it had been introduced.

John Lindsay said that he wouldn't approach the issue like that. He would say, 'here [example] is a lyricist'. Referring to Callimachus, mentioned on one of Judi's slides, he said he would ask, 'Is Callimachus a lyricist? Well, more than he is an epigrammist.' So, we could put Callimachus among the lyricists. And then you can come along with the example of Theocritus, and you see where Callimachus is, and you say 'No, I don't agree with that, Callimachus is one of *those* instead.'

You don't create categories by definitions, in other words – you create them by populating them with stuff. You go on doing that, and you divide the stuff whenever you bump into a Stuff Management Problem.

Nic said: You will of course have occasions when things fall in the great space between the two, or you have fuzzy boundaries – actually it amounts to the same thing.

Ownership of definitions

Terry asked: Who owns an address? There is an address; we all agree it looks like an address; but… is it a billing address? a delivery address? an accounts address? What kind of address is it? That is why there has to be some defined understanding of what that data actually means. If all I can say about it is that it is an address, it is no use.

John replied that addresses exist in a context – there's no place for addresses in a music catalogue, for example. The context of addresses is different from the context of a catalogue of musical scripts.

Keith thought that the whole concept of ownership has horrendous consequences. Going back to the example of a human resources database – he uses this example because it is one most people can understand – most people would say that the ownership of an employee's address in that database is the Human Resources Department. But surely the person who owns that address is the employee? It's the Post Office that owns the *address*, said John Lindsay.

Keith continued, saying that the HR Department owns, not the addresses of employees, but *the fact that they want to know the addresses of the employees*. Similarly with a person's date of birth: the HR Department owns the fact that we want to know the date of birth of our employees.

One of the problems Keith has when people talk about data ownership is – are we talking about the ownership of the *data definition*, in other words what Keith would call the metadata, or are they talking about the ownership of the data itself? A lot of people get those two confused.

Adrian said that he thought that we needed to remind ourselves that these things belong in what one might call 'multi-value fields'. For example if he wants to book himself on for attendance at an event, what is 'the address'? Is it the address where the event is going to take place? is it the address of the organisation that's running it? is it the address to which he has to send the cheque?

Or, if we talk about a name, and the subject is a married woman who was married before, there may be a maiden name, the previous husband's surname used as a married name, the name she uses now… You have to allow for all of these fields to have multiple values. And John's example, of whether someone should be classified as a lyricist or an epigrammist – the guy is quite probably both.

Coming closer to home – if Adrian visits the BCS Web site and wants to navigate to the Thought Leadership page, does he come down the academic path of the website, or what?

[13] Keith again objected that it is data, not metadata.

Classification versus navigation; dimensionality

Conrad said that he felt this was an appropriate moment to return to the issue of how classification of information is one thing, and the way it is presented for navigation is another. This might involve us in some discussion of hierarchies too.

Adrian had raised the issue of how the resources on the BCS Web site are structured and how they are made available, and it reminded Conrad also that Judi had said she was keen to know how the taxonomy development within the BCS was affecting the Web site. This is a kind of story – a story we don't all know – but one that is close to the BCS story. Conrad asked Adrian to kick that discussion off again.

However, it was Nic who first rose to the bait. He said that here one might have particular bits of taxonomy that have multiple facets: you might have subject area, or the kind of person interested in this sort of information, or whatever. You can get to the same piece of information through multiple facets, or dimensions, or whatever you care to call them. If you have gone down one branch of this classification, you don't want to have to go all the way up to the top and come down another one to get there – you need convenient ways of moving across.

Each facet – each one corresponds to a metadata field really – creates a multi-dimensional space within which the users move around. They need to be able to move around in any dimension that takes their fancy at any point in time, without being constrained by one particular metadata field or taxonomy facet they happen to be wandering along at that moment.

The human factors to do with usability and navigation are related to – but they are not the same as – rigorous classification. Sometimes when you have got down to some level of detail down *one* classification scheme, it is a good thing to have summaries of *other* dimensions so you can get into those quickly.

It's difficult to explain without drawings, concluded Nic! Judi said that she would like to see some drawings, so Nic stepped up to the flipchart. Suppose we have a subject taxonomy, he said, which is a hierarchy of some form. Then suppose we have a geography dimension, which can be similarly ordered hierarchically into continents, countries, countries, towns etc. Then we have lumps of information – bits of stuff – which can be classified according to both of those. (Martin Bryan suggested 'Italian opera' could be an example of the intersection of a topic and geography.)

What you have created now is a two-dimensional space – subject and geography – and each metadata field by which these bits of stuff can be classified effectively creates another one of these dimensions. For the user, being able to navigate around this multidimensional space easily is a really important factor in terms of usability. There are some aspects of classification that they won't be terribly interested in, so you won't want to expose the users to those aspects. It becomes too difficult.

No, that's easy, said John Lindsay. That's called Dewey. Dewey may be a way of specifying values for these things – the subject taxonomies, said Nic – but… No, for all of them, insisted John. And then you have a facet tree. But the point, said Nic, is how you make it usable. Just do it by mapping your character strings against Dewey, said John. That wasn't the point Nic was getting at, though. 'I don't care what value-space I draw this from', said Nic. 'The fact of the matter is that I have a structured value-space for each of them, and I have several dimensions, and I want to be able to wander round the knowledge-space…'

The conversation then became difficult to follow, as many people spoke at the same time.

Conrad observed that on a Web site, one effective way in which you may be allowed to constrain a search is to use each metadata dimension as a constraint. You may therefore look for something that is about 'information retrieval' (subject), that was a 'meeting' (kind of resource), that was 'within the last three years' (time dimension)…

Well, said Nic, we have to distinguish between two approaches to information seeking. There are searches, when you can use those dimensions to constrain and frame the search. But a lot of people just dive in and browse. The current BCS Web site home page offers two dimensions, one of which is organised by topic, and the other of which asks you what kind of person you are, and from the latter gives you links to what BCS thinks might be appropriate to those audiences. You might wander down each of those kinds of route.

Zinat came in with the observation that there was an Indian librarian, Ranganathan, who devised the faceted classification scheme, which took classification further than Dewey. What Nic had said about faceting – wouldn't Ranganathan's scheme take care of it?

John Lindsay disagreed – he thought that Ranganathan's Colon Classification does not really go further than Dewey. What it does is give you a 'PMEST' architecture when you have defined your taxonomies, and you can then adopt any scheme that you like. ('PMEST' is the acronym for the five basic facets of Personality, Matter or property, Energy, Space and Time, and the system is known as 'colon classification' from the use of colons to separate facets within a class number.) And this was all thought about in the 1930s, he added.

Nic said that he was not claiming that any of these methods were new, but now we have new technology for doing it; and one of the 'supremely wonderful' things about this new electronic technology, he said, is the ability to have the same 'bit of stuff' in several places at once; which had not been very practical with books.

John Alexander said that Nic's example seemed to disentangle navigation and classification. If we say that everything there is data, and none of it is metadata, then what you are doing is exposing more of the data, but then using the navigational imperative of usability to make it easier to get around. They are two different trains of thought – two different things going on.

From a technical point of view, said Nic, it is convenient if the things by which you want people to navigate are made explicit as metadata, because what you do not want to have to do is to have to have the searching process look *inside* lumps of stuff every time someone wants to navigate around them.

And socially, said Conrad, what you need to do when constructing a Web site is to gather a sample of your user community that you can test the site against, to find out if the categories that you come up with are ones that feel natural. At present, if he is looking to find out about the BCS Specialist Groups on the BCS Web site, he has to be able to associate it with 'Networking' – which is not how he naturally associates Specialist Groups; that's the BCS HQ view of what Specialist Groups are about.

And if he wants to find the 'Learned Society' content – well, he described himself as a non-academic who tries to be learned. He is not going to go down the Academics link, because that does not apply to him, but he still wants to find out about the BCS's view of itself as a Learned Society. So when the user's categories just don't match with those of the site owner, confusion is the result.

Nic said that there is another way in which a navigation structure may reflect a taxonomy, which is that it colours the way in which you as an organisation think about yourself, and also how you present yourself to the outside world. The BCS is trying – commendably – to be even-handed in presenting itself as a Learned Society and a Professional Body, but in some ways branding one side of the field as 'academic' and the other as whatever else the other term is [maybe business?], is actually counter-productive in terms of trying to bring the two different aspects of the BCS's existence, purposes and membership closer to each other.

The BCS Web site and taxonomy – a case study, a shared concern

The Taxonomist's Tale

Conrad said that one story he would like to hear about is how the BCS Knowledge Services Board initiative on taxonomy came about; and he knew that several people present had been involved in that. Could somebody tell that story?

David Penfold undertook to tell the tale. Back in 2002, the Knowledge Services Board set up a working party to look at BCS content. Basically, the BCS had spent a lot of money developing a Web site, and then suddenly realised – My God, what content are we going to put on this Web site? That is, content related to computing; for the site contains a lot of administrative stuff, about membership and so on, but there was virtually no information whatsoever about computing.

Wendy Hall, who was the President of the BCS at that time, offered to chair this working group. It had four or five meetings; and one of the ideas that arose was that the BCS Web site should provide a way of accessing information about computing and IT. This should include content in the *Computer Journal* and the *Computer Bulletin*, and anything else published by the BCS – but also, if possible, information on Specialist Group Web sites, and Branch Web sites, and also maybe some external materials that had been in a sense validated or approved by the BCS. Thus the BCS might say in effect 'We have no information on this subject, but there is a very good site over there which one of our SGs recommends, so go and look there.'

Now, one of the recommendations of the KSB working party was that in order to categorise this content, it would be necessary to have some sort of taxonomy. For the academic side of the BCS, for the content of the *Computer Journal* for example, the existing taxonomy of the ACM worked pretty well. However, for the more 'professional' side of the BCS, or perhaps one should say the non-academic side, a new taxonomy should be developed to cover it. And it was anticipated that the BCS Web site would be developed to have a new search engine that would actually take advantage of this taxonomy to be able to search the site. It might take a form similar to that of the Open Directory project (http://dmoz.org/), where you have a subject hierarchy, presented to as many levels as necessary.

Eventually, Judi Vernau was asked to do an audit of the kind of content held by the BCS, and subsequently to develop a taxonomy on the basis of that.

Judi explained at this point that the task had been to find out what content existed within the BCS, and how it had been dealt with previously. Following that, they had made recommendations about what could be done with it from that point forward – ranging from 'not very much' in some cases, to 'quite a lot' in others. Judi explained that some information content could be broken down into small pieces, into chunks that could find uses in lots of different contexts.

The next thing that happened was that they were asked to prepare a taxonomy that would at least describe the 'whole' pieces – the Journal articles, the meetings, the conferences. It became clear to them that the requirements of what they called 'the professionals' was different from the requirements of 'the academics'. For ages Judi had fought to make they two halves come together, but they just wouldn't. So, after consulting with David and Adrian, and Genevieve and Nic were involved as well, and quite a few others, they decided to use the ACM taxonomy for the 'academic' content, and to develop a new 'professional' taxonomy.

Judi thought that story now moves over to what Carl Harris could tell, because by then the taxonomy had been handed over.

The Webmaster's Tale

Carl took up the story. He explained that the taxonomy project had come to fruition at the same time as the BCS Web site's new Content Management System was being implemented. This now incorporates the new taxonomy as part of the system, along with the metadata schemes that Judi had proposed for them, based on Dublin Core.

Now, everything that goes into the new Web site CMS – whether it be a piece of content, hyperlink or any different way in which they might break something down as being a Web site component – can have all of these things assigned to it. Metadata can be assigned, and so can terms from the taxonomy.

Another kind of classification is built into the system as well: the navigation – where that piece of content is going to appear within the site, and what other relations it has. For example it may live in the Thought Leadership area, but they might want it linked into from another area because that is where people may be trying to find it. So, using the system they have got, the BCS can put all of those things together to give what is now the BCS front-end Web site.

To give more specific examples of how metadata and the taxonomy are being used, Carl talked about their use of metadata. They have a number of authors who are going to be contributing content for the web site, and they have been able to use the Dublin Core classifications so that if you are reading an article by a particular author, the system is able to put up 'here is a list of other articles by the same author'.

A current issue, as Carl had mentioned in our first round, is how they can apply this system consistently. They are already starting to find internally an inconsistent application of terms, which will instantly spoil everything they have been trying to achieve in terms of consistency and relevance.

To give an example from what in the BCS they term 'operational' content – such as information about what the BCS does and its products and services – things were going wrong because people were just ticking 'BCS' as a category, and so this stuff was being picked up everywhere. There is quite a lot of education that needs to be done internally about understanding what this whole metadata system is about.

Discussing the BCS Web site as a 'case'

Nic asked if there was some way in which the BCS could receive feedback about the Web site and classification. Clearly it could be disastrous if that the Society were overwhelmed by feedback, but on the other hand feedback could be an effective way of improving the classifications.

It would also, said Conrad, make quite an interesting case for a write-up as a case-study, to see that it is actually documented. Firstly, a lot of BCS members or even people outside the BCS would be interested to hear about this application; secondly, it might actually help in educating the people inside the BCS HQ about what it is all intended for; and it would be another useful piece of content.

David Penfold asked if there was somewhere on the BCS Web site – he had not been able to find such a place – where you could ask specifically for information about electronic publishing, or some other aspect of computing, organised by IT subject areas?

Carl replied that it was technically possible to provide such a facility – classifications and navigational elements could be placed on the site to enable that, so one could browse the taxonomy and pull out everything related to that topic – but that capability had not been implemented in practice.

Conrad said that it sounded as if the back end of the system had been built, and that structure now gave a good basis for moving forward, but that the interfaces to that system are yet to be fully developed. At the moment, many of the BCS Web site pages carry a little symbol for RDF, and when you click on it you can see the information that lies behind that page, but the interface hasn't yet been thought out that would give advanced access. Which is at least the right way round to develop it – it would be silly to have the interface and nothing behind to implement it!

Andy MacFarlane said that in terms of searching, there is a simple solution to this problem, called 'stop-word lists'. If you have a word which is going to be no use for searching at all, you put it in a stop-word list. So, asked Conrad, 'BCS' in this instance? Yes, said Andy – a classic case.

Carl explained that there was originally a term within the taxonomy called 'BCS' which it had been intended to use to tag internal content, but people have used that as one of their standard 'ticking terms', hence its uselessness in practice. The basic point, said Andy, is that if you are on the

BCS Web site, there is absolutely no point in using 'BCS' as a search term, because it will apply to every page.

One extra little thing that's needed, said John Lindsay, is an 'authority file' – this is where you capture your 'learning' (for want of a better word), which means that whenever a contention comes up, you can refer to your authority, so that every little case doesn't have to be worked out. He thought it should not be too difficult to make an FAQ out of an authority file, so that what other people call an FAQ would be the output of your authority file. That should be quite manageable.

But John said he was hearing a much more general problem, one which has been around for ever since people had started talking about these sorts of things, which is that we are dealing in one sense with specifics (and the particular case of professionals using the BCS Web site); but we are also dealing with generalities, because when the word 'computing' is used these days, it has become a generality. There is no part of human life that doesn't in some sense or other involve computing.

So, computers in transport *ought* to be a big part of the BCS's activity; even though he knows that nobody in the BCS has done anything on this topic. Therefore 'transport' belongs in the BCS taxonomy (not 'transport' as in the transport layer of TCP/IP of course!) And so does 'health' – there's a huge domain. But how we crack the problem of the general versus the particular is something that is going to come back over and over again.

Conrad thought there was also something of a BCS cultural problem to take into account. He had recently been talking to Rebecca George OBE, who is active within BCS Women, and they had agreed that it seemed as if the official BCS mind-set was that there existed 'Real Computing' on the one hand, real manly stuff to do with software development and algorithms and methods and so on; and then there were other areas of activity which weren't really computing even though they used computers – areas like publishing, or the arts, or media production.

These areas might use the most powerful computers around, for example to make animated film, but they were seen as not being *really* computing, and therefore not a BCS concern.

Solving classification problems; literary warrant

In response to a question from Zinat, John Lindsay talked about the choice between pre-coordinate and post-coordinate classification. The problem is that in a context like the BCS one is not going to have proper cataloguing – the whole thing is too big, too amorphous. One poor person cannot be the editor or the cataloguer. Regardless of whether you go for pre-coordinate or post-coordinate cataloguing, the problem will be relying on freewheeling individuals to pick terms, and that is just not going to work.

But if, Zinat replied, one created an authority file, as John had suggested, and as libraries have done for ages, that might help. It would be helpful, John agreed. If somebody puts in something, what we can do nowadays is bounce back and say, 'epigraphy is not the same as lyricism, because we have a think which is called literary warrant'.[14]

In the total universe of discourse, if you have only one item, there is no point in having huge philosophical debates about it. If you have got several hundred of them, you do have a problem. This is the weakness of most of the Google-type things so far: they can't discriminate between one and a hundred million. What you want is a refining process that works on the basis of your literary warrant. And if you've only got ten or twenty instances of this particular term sucked out of your taxonomy[15]… We need a capacity to say that the literary warrant for this is only five.

'TCP/IP' would be as useless a term as 'BCS'; but if you start putting TCP/IP into your stop-list, you are going to begin to hit a lot of trouble. So what you need is the capacity to do universal and particular, general and specific, on the basis of what your 'inputters' are doing. But because these days you can also track the search strings that your 'lookers-atters' are looking for, you can begin to do something useful.

Thesaurus Rex?

Conrad asked for a clarification. John has said that he preferred the term 'thesaurus' to 'taxonomy'. Is that because the concept of a taxonomy is inherently hierarchical, and sorts things into sub-boxes of sub-boxes?

[14] "In general, the warrant of a classification system can be thought of as the authority a classification invokes first to justify and subsequently to verify decisions about what classes/concepts to include in the system, in what order classes/concepts should appear in the schedules, what units classes/concepts are divided into, how far subdivision should proceed, how much and where synthesis is available, whether citation order are static or variable and similar questions. Warrant covers conscious or unconscious assumptions and decisions about what kinds and what units of analysis are appropriate to embody and to carry the meaning or use of a class to the classifier, who must interpret both the document and the classification system in order to classify the documents by means of syntactic devises. The semantic warrant of a system thus provides the principal authorization for supposing that some class or concept or notational device will be helpful and meaningful to classifiers and ultimately to users of documents"
Beghtol, C. (1986). 'Semantic validity: Concepts of warrant in bibliographic classification systems'.
Library Resources & Technical Services, 30(2), 110–111).

[15] John said he would prefer to call it a thesaurus, because he thinks that is what we need: 'not Roget's thesaurus but the ISO thesaurus that Francis didn't talk about'.

No, explained John. A taxonomy is a general form which organises genera, species and varieties. The 'thesaurus' – in the sense in which he was using it – is an ISO standard, implemented in software, which allows you to build what are called 'broader terms', 'narrower terms', 'related terms' and 'preferred terms', with a nomenclature which allows you to deliver interoperability mapping across all the different things.

So, in Dewey World, this item might be registered as 914.23; in Swindon World, it is XYZ.IP. And you can map Dewey World onto Swindon World, because you do have an architecture that links these things together. ('Architecture' because, said John, he is not using the word 'information' at all – except in that context of not using it!)

It was a pity, said John, that Dr. Leonard Will was not present with us on this occasion; because Leonard really is one of the world experts in this particular area; he knows it like the back of his hand.[16]

Tests of sensibleness and usability

Every three months, *The Economist magazine* does a special supplement on technology. The last one John Lindsay had looked at was about near-field communications. John went off to look at the ACM digital library, and typed in 'near field' – and back we were to the hyphen issue! Because 'near-field' has a hyphen in it in some contexts, and it doesn't in others.

John was so provoked by the ACM experience that he then went and looked at 'Semantic Web', and then at 'information systems'. And it seemed to John that the ACM system is so weak, because of the ideologies of the people who built it, that we need to run tests on it the whole time.

As the BCS system develops, we can take a term like 'near-field communications' and try a test with that. And if it were to become like a wiki, and if he had the energy, John could then type in something that says, 'Near-field communications, written up in *The Economist* of such and such a date – this is something we need to think about.' Now, if we could develop that sort of style of working – and there are several things that do work like that…

What should the BCS site contain, or index?

David Penfold jumped in to say that there is no need for the BCS to try to cover everything in computing, because they have often been written about better elsewhere. All the BCS would need to do would be to point to the sources, such as the Forrester Reports.

There are two different things here, said Nic. One is taking articles of interest, and putting them on the Web site; the other is creating a new thesaurus term. Nic thought that the difference between thesauruses and taxonomies is not very great in practical terms. David Penfold raised the idea of converting the BCS taxonomy into a thesaurus.

Conrad asked David to think back, to where he had started to explain the story of the BCS taxonomy project. He had indicated that there was the desire to be able to extend this classification to content that was held, not on the BCS Web site, but for example on the Web sites of Specialist Groups. How might that process happen? Might there be 'meta' tags that we should put in the pages, for example?

Well, said David, the idea as it was discussed in KSB was that we should ask the SGs to take the BCS taxonomy and refer to it in tagging their own information – though obviously there would have to be some mechanism of support for this within the BCS. There would basically have to be at the BCS a record of the metadata together with the URI for each such Specialist Group page so that searches and links would be possible.

It wasn't just the Web sites of Specialist Groups, added Nic. Potentially it could extend to cataloguing interesting articles somewhere else. Yes, said David, but one would ask the SGs to do their own Web sites. True, said Nic, because they could speak with some authority about how their pages should be categorised and what would be useful to catalogue.

This reminded Conrad about the discussion that had taken place about the Glossary entries on the BCS Web site. He had felt that the BCS Glossary as it stood didn't help the BCS look like a Learned Society when there were so few terms, explained so briefly.

Discussion of this had led to the idea of building a proper Glossary for the BCS, in which endeavour the Specialist Groups could help. To which David had replied that there are 37,000 entries in the Free Online Dictionary of Computing, and Conrad had also found that Wikipedia is a pretty useful reference for computing terms. He had tried the term 'PostScript' as a test case, and found a lot of useful stuff on both of those sites.

David said that the point he'd wanted to make was that valuable resources like the Free Online Dictionary of Computing are not generally known about.

[16] Dr Leonard Will has subsequently become a member of the KIDMM email discussion list. Some interesting and useful documents about information management can be found on the website of Leonard and Sheena Will — http://www.willpowerinfo.co.uk/

Outputs and outcomes from this day

Conrad promised that one outcome of this meeting would be that he would write up this day's discussion. Another outcome would be that David Penfold had already been asked to deliver a 'Metadata Update' to the forthcoming Specialist Groups Assembly, which would in part be a report from this day.

David wondered if Adrian might comment on whether, if we decided that we wanted to suggest to the BCS that we ought to evolve into a BCS Forum – or some sort of BCS organisation – there was some way of going about this? Something that would not be a Specialist Group, but within which SGs could take part?

Adrian replied that there is now more than one model of what a BCS Forum is. The best model, he thought, would be that exemplified by the Health Informatics Forum, which grew out of existing collaboration between existing Specialist Groups. That was a bottom-up evolution; whereas other Forums, especially the original three, had been brought about in a top-down way. More recently there have been created an Ethics Forum, and a Women's Forum. There is also a Security Forum, which developed out of a BCS Expert Panel.

There is therefore a mechanism for creating more Forums, but there is a danger there that if the BCS just cranks them out, their worth will be devalued. We would have to think carefully before getting into that, and as Andy MacFarlane commented, we would have to demonstrate some prior working.

Let's just get stuck in…

Surely, said John Lindsay, the Specialist Groups should just do what Specialist Groups have always done, which is to say 'This is a piece of work which needs doing'; set up a working group; and get a report written. Once we have something on the table, we can decide where to go after that. He thought we had two more steps to do, and with those two more steps done, we would have a piece of work finished.

The steps would be – we need to have an exhibition, which basically would mean that we would have some stories around the walls, having got some stuff together – and then we would need a working party that would tackle the rather boring task of going through a few hundred lines of text and say, 'Yes, this makes some sort of sense, this sounds like a professional and academically rigorous approach and we can live with this.' And we could then go to the other interested bodies such as the people who run librarianship – even the statisticians! – and say 'Here is what we think the work activity should be in order to achieve this.'

This rigour is important, continued John, because Tim Berners-Lee has been quite clear about what he understands the Semantic Web to be – it is about machine to machine communication; which has nothing to do with people. It is hard to argue with Tim Berners-Lee these days. The first time John had an argument with him, it was just between the two of them. But now it's as if he has become canonized; the argument is not on a flat playing-field any more.

It's rather like the Belew business in 'Finding Out About',[17] who has remarked that it has never been easy to introduce new technology to Information Retrieval. 'Why?' asked John. 'Because of Keith van Rijsbergen.'

So John thinks we should move quite quickly, and produce something which will set a debate going. We could call them KIDMMs; John thought that how that is pronounced has some nice little ironies built into it.

The real achievement of the BCS Taxonomy working party, and here, is that we have got practitioners – people who are involved in making things that work in the real market; we have academics who teach about it; and we have other practitioners involved in building the things on the ground. If we have those three communities – the industrial community, the IT practitioner community and the academic community (and actually we also have some representatives of the victims here as well) – we are quite well placed to do something, before the end of the year, with a very short deadline and clarity about what we've got to deliver.

We *could* have an exhibition before the end of the summer, and we could get a report out – Requests For Comments, they were called in the IETF – so we could get an RFC out before the end of the year, which basically says 'This is KIDMM as we understand it'.

What infrastructure & support is needed?

Conrad asked, What infrastructure is needed 'to KIDMM properly'?

John thought, rather than go for an email list, which is a bit accidental, we could go for a blog or a wiki. But nothing very complicated, just very simple. He has a JISCmail list which people could use if they wanted to – it is called 'The Hyphen Society'![18] (John also runs a Metadata list on JISCmail, but that is about the teaching of metadata. That itself is an interesting topic, because it seems there is almost *no* teaching about metadata.)

[17] *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW* by Richard K. Belew; Cambridge University Press 2001.

Or we could start our own KIDMM list, suggested David. Or we have one already? Conrad clarified that we have had a list called 'KIMtec' on Yahoo.

Andy MacFarlane had a question. What was the relationship between Specialist Groups and this forum? Any SG is going to want to continue to serve its current constituency. Is this intended as a federation of SGs? It isn't really anything, replied David and Conrad almost simultaneously. David added that he thought that John was right: if we could get something out, we might get the BCS to support the initiative, or decide that there should be some sort of official existence within the BCS for this structure.

Conrad said that one could use the phrase descriptive of those old 17th century organisations that were put together maritime trading and colonisation – KIDMM is a 'partnership at a venture' between participating SGs and others. Nothing more formal than that. And yes, to take it further than that we would need to show some evidence of prior working. It doesn't make sense at all to merge groups.

No, said Andy – he hadn't thought there was any idea of merging groups, but the question still remains of the allocation of resources. If he were to report back to the IRSG committee, their Treasurer would straightaway ask what would be the implications for budgets. Well, suggested David, we might even go back to SGEC and ask for some resources to support this.

How much money would it take? asked Adrian. David said he didn't think it would require any money for now, though SGEC had given some money to help with the organisation of this day. Said Adrian: Forums get budgets. Conrad said we wouldn't need money to set up the email list. We won't need money to start writing to each other. We would need money if we wanted to organise an exhibition and conference, though, said David. Exactly, said Adrian.

Time for closing thoughts

Conrad thought that it was now about time to bring the day to a close, and couldn't think of a better way to do this than to go round the room once again, asking each person in turn to offer their closing thoughts or reactions – or those kinds of comments that a person might have been waiting for the right opportunity to say, but the right moment hadn't come along. This was now the chance.

David Penfold said he had felt encouraged by the day. He had been interested how people's definitions vary.

Judi Vernau mused on how things can *have* metadata, or they can *be* metadata. (At which Keith came in to say that there are fewer words in the English language than there are concepts which we want to define, and therefore people will be using the same term to mean slightly different things; we shouldn't be amazed at that and we have to live with it.)

Judi replied that she had found *that* part of the discussion really interesting. Nic's contributions had also given her many things to think about, which she had been writing down, about retrieval and jumping between taxonomies.

Nic Holt said he now had a broader view on some concepts which he had thought were relatively solid – until today! The 'metadata' word is an example of that. If we try and pin these things down, we might find that they just disappear. A scary thought! – though he was sure that in practice that wouldn't be the case.

Terry Freedman asked himself if he would be going away with any answers. Not really… but he would be tempted to come back from more. He felt a strengthened confidence in what he himself thinks on these issues, however, and would like to work with others to find out some of the answers.

Andy MacFarlane felt that if we did want to develop a history of doing things together, there is no time to start like the present. Even activities which took place between pairs of Specialist Groups might help. The Information Retrieval Specialist Group from time to time runs free-standing workshops, and had already spoken to the Artificial Intelligence SG about doing so. If we wanted to build some momentum towards partnership, we could do so by getting people together and doing things.

Andy felt sure IRSG would welcome ideas from other SGs about joint workshops; perhaps each SG here should think about what other SGs they could talk to.

Andrew Tuson said that the idea that had struck him more than any other was that of the importance of context. Perhaps we could investigate that a little bit more, because it is an extremely important issue; one that is becoming more and more significant in information retrieval. Peter Ingwersen and Kalervo Järvelin have recently published a book on exactly that subject of context in IR.[19] He felt that it is something that will affect not only our group but all other groups as well. Conrad commented that one of the things that had impressed him about what he had read

[18] (John called it that because it seemed to him that the hyphen was at the atomic level at which machine-to-machine communication begins to fail. He had thought that a comma might be simpler – but then remembered the comma-delimited file. He thought maybe a space – but if he put the space in, no-one would know if it was there!) See http://www.free-conversant.com/irweblog/678.

[19] The Turn: integration of information-seeking and retreival in context. By Peter Ingwersen and Kalervo Järvelin. September 2005, Springer Verlag, ISBN 140203850X.

about topic maps was the ability to define the context, the scope within which a term applies – and Martin agreed that this was the fundamental advantage that Topic Maps had over just about every other previous approach: to be able to scope concepts and names, and say – this is what this means in this community.

Nic added that there are a few search engines that use similar methods of semantic disambiguation, so that terms that have multiple meanings can be disambiguated by looking at other terms that appear in the same document.

Miltos Petrides felt this had been a very interesting discussion to report back to the Artificial Intelligence Specialist Group. He'd enjoyed it. He doubted we'd get to a definition of what 'information' is – it could take many years – or forever. But this had been a good opportunity to put together viewpoints from different areas of computing – the practitioners, theoreticians, researchers and so on. As an AI practitioner and researcher, he felt it was a great challenge for computing: to improve the intelligence of systems, these were the issues that needed to be tackled. He was excited about the day, and saw it as the beginning of an ongoing discussion.

Robin Kyd appreciated the opportunity to come together with people from other Specialist Groups in a way that hadn't happened before. It was a good development, and he looked forward to the next stage.

John Lindsay said he had ended up with twelve pages of notes – several hundred words written down – and would be fascinated to see how they would map onto concepts. He would not agree with Keith that we are short of words – 'English has an amazing number of words which haven't been used at all!' We can have as many words as we like, as Alice in Wonderland discovered: the problem is how you join them together.

He thought we should give more time to the Oxford Dictionary of National Biography, to which he only recently gained electronic access.[20] If you go to a library which has the paper-based version, and then you go to the electronic version, you see that each of them has its strengths and its weaknesses.

No sooner had he discovered the ODNB, than he found an article[21] by a woman who had been through the National Biography, and the Oxford English Dictionary, and had mapped out how many words had been made up by men, and how many by women; and, how women appeared in the Dictionary of National Biography.

What becomes clear is that because men had the power to make the words, men also had the power to make what went into the Dictionary of 'National' Biography, and women had been largely written out of history as a result. This focuses our attention on the beginnings of what we can have proper arguments about.

You can then look at things like JSTOR and Copac[22] – and Metafind, which no-one had mentioned today. John thought that Metafind really was a step forward – but to use it at the moment, you have to be in Senate House – again, because of the intellectual property rights régime.

The capacity of people who have never been able to define their own worlds before is now on the agenda. And people who have been written out of history can begin to discover a voice. And, starting with women – well, it seems a very good place to start. He could envisage a new taxonomy coming into birth here.

Martin Bryan said he had noted down a couple of things that might be worth thinking about. The first is, 'What do we mean by KIDMM?' Knowledge, then that word we are not allowed to use, data *or* metadata. Not *and* metadata, it would seem. Then he said he'd put a very big question-mark after it.

The other thing which had struck him as being missing today was the other half of 'IT'. We'd talked a lot about information, but very little about the T. And we need to think seriously about what the T can provide, to help us with KIDMM.

Nic said that it was however important to figure out first what we were trying to do before wielding technology all over the place. Martin replied that sometimes you can define what you want to do *by the means of* technology.

Zinat Bennett wanted to mention two things. Today had reminded her that she should refresh her own librarianship skills, because they have a real value and usefulness. Secondly, she had also been hoping to hear a little bit more from the technical experts present about her problem with applying metadata. Could not all these pattern-matching algorithms and so on be of some help?

[20] The Oxford Dictionary of National Biography, edited by H. C. G. Matthew and Brian Harrison, 60 volumes, 61,440 pages, $13,000. Access to the online edition is $295 yearly for individual users.

[21] Probably John was referring to 'Gender in the Archive: Women in the Oxford Dictionary of National Biography and the Oxford English Dictionary' by Elizabeth Baigent, Charlotte Brewer and Vivienne Larminie, in *Archives*, the Journal of the British Records Association – Vol XXX, No 113, October 2005. A pre-print electronic version of the paper is available at http://oed.hertford.ox.ac.uk/main/images/stories/articles/baigent2005.pdf

[22] COPAC® is a union catalogue which provides free access to the merged online catalogues of 24 university libraries plus the British Library, the National Library of Scotland, and the National Library of Wales/Llyfrgell Genedlaethol Cymru. It is a MIMAS service funded by JISC, and is produced at Manchester Computing. See http://copac.ac.uk/

It's very dangerous stuff, said Nic. But what about technologies like Autonomy? asked Zinat. Are they just hype? It is all hype, believe me, said Andy MacFarlane.

It is interesting that we hadn't talked about technology today, said John Lindsay. Because we do by and large know about it, and yet it is just not significant at this stage.

Within small worlds and well defined problems, the technologies do work, said Miltos. But when you turn to trying to look at the wider world in common-sense terms, the technologies are eons away from delivering useful results.

'You go ahead and brush up your librarianship skills,' John Lindsay advised Zinat – 'Much more important!'

Ken Moore said that from a practical viewpoint, some issues had been raised today that he would not even have dreamed of. The potential for co-operation between the various groups represented here is really great, and he was really looking forward to the establishment of our discussion list.

Adrian Walmsley touched on a couple of administrative issues. If this group was thinking of putting on events, there is a BCS budget for events next year for the fiftieth anniversary of the Society. We would probably be reminded of that at the SG Assembly coming up in April. A BCS staff member has been put onto the programme for the Jubilee full-time.

The second thing is that if the group were to decide that a wiki would be more appropriate in some ways than a discussion list, he has one that he could make available for our purposes. There was a murmur of interest from around the room, and John Lindsay said that it was important to try out these experiments and see what would happen. Meanwhile, said Conrad, he would maintain and develop the KIDMM section of the EPSG Web site. We can use that as a repository.

Berin Gowan thought that the day had shown the value of interconnections and of tossing ideas around between the Specialist Groups. That was very encouraging. Berin does value face-to-face communication, so whatever happens electronically, he thought that if we could agree to meet in a year's time it would be good.

As a practitioner, he would much prefer if we could take a specific topic – transport, for example – and look at how information retrieval, artificial intelligence and so on can relate to that. Around that sort of forum, we could explore a topic together and get a lot out of it.

Conrad commented that it seems as if nothing quite fires up on-line discussion as the occasional experience of meeting face to face.

Genevieve Hibbs noted that Conrad had mentioned 'indexation' on one of his wall-charts. Four or five years ago she had been the first speaker at a conference on indexes, and the word 'indexation' was in the title of her talk. The Society of Indexers were there, and there was a lady there who had a big bee in her bonnet and slated her for using the word 'indexation' incorrectly. At that time she was feeling low, she was in the process of being made redundant, and nobody came to her defence… it had been a very negative experience.

In contrast, today Genevieve had felt that she had been able to contribute, and it had also affirmed some other things she had done. She had found it interesting, and she was glad we were doing it, and it was great to be involved.

John Alexander said he really looked forward to seeing how this would develop. As Judi said earlier, this is an area where at one minute you think you understand it, and the next minute you feel that you don't. At the moment, he *thought* he had got it!

Carl Harris thought that the day had been very useful. He thought it a fortunate co-incidence that this discussion came along just as the BCS Web site was coming together. He would like to see these discussions continue – both electronically and face-to-face as well. And what the BCS is trying to do with the Web site could be a real test of these ideas and principles. Yes, said Conrad; and perhaps the KIDMM group, when it gets together electronically, could also serve as a kind of helpful reference group for Carl.

Keith Gordon said that he had only heard about this event just two weeks ago, when Tony Jenkins suggested that he take part on behalf of the Data Management Specialist Group, and that he might find it interesting. And he had – and hoped he had contributed.

Winding up

Conrad then brought the day to a close by thanking everyone for taking part, and said that he hoped people would take him up on his offer to add other suggested resources to the KIDMM Web site. The account of this meeting would go there, as would the write-up he had made of Ian Horrocks' Needham Lecture. As for other links, if people could identify documents and sites worth linking to, and write a couple of sentences as a synopsis, it would be very helpful.

David Penfold said he would report back to KSB about KIDMM; and he thanked Conrad for running the day.

Appendix: two diagrams by Conrad

Before the discussion got started again in the afternoon, Conrad explained a couple of posters he had stuck on the wall during lunchtime.

Four methods

The first diagram was a table (see foot of page), in which Conrad was trying to remind the group that although metadata is an important way of making data or information can be made more manageable, it is not the only way. He identified four approaches – they are not mutually exclusive, because they can be combined.

The oldest way of organising data within a computer system is in a fielded database, which could simply be an array of text and numerical strings separated by some sort of delimiter: not a comma! – for reasons John Lindsay had pointed out before lunch. (Unless, David Penfold said, you used some sort of escape character to distinguish between commas used as field delimiters, and commas which were intended to function as real commas. Often a backslash (\) is used for this function.)

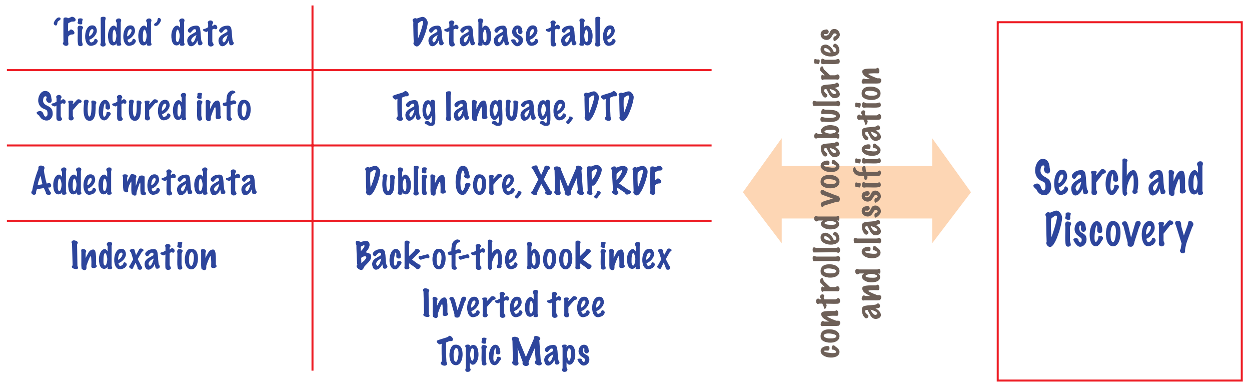
The invention of generic mark-up has meant that we can create a tag-language, based on SGML or XML principles, by means of which we can ensure that the content of a tag within a document is recognisably something – the surname of an author, a date, a price or whatever. This allows us to attach computer-processable 'handles' to elements within running text; we escape the confines and limitations of the database table.

Metadata is another approach, which as Conrad saw it was about sticking labels on things, for which there are a number of approaches; and these can apply to things like images, or physically printed books. This approach harks back to the library or museum catalogue.

The fourth approach is indexing, which has its antecedents in the back-of-the-book index. Indexing does not need one to insert anything into the information being indexed, so long as one has an adequate means of pointing to it (a page, chapter and verse, a URI). The Google type of index, an inverted-tree index, is one approach to that, one which makes free text search possible but has unsatisfactory aspects too. If one can augment such indexes with catalogues of synonyms, thesauruses, one could achieve something more satisfactory, and Topic Maps would seem to be an elaborate version of this approach.

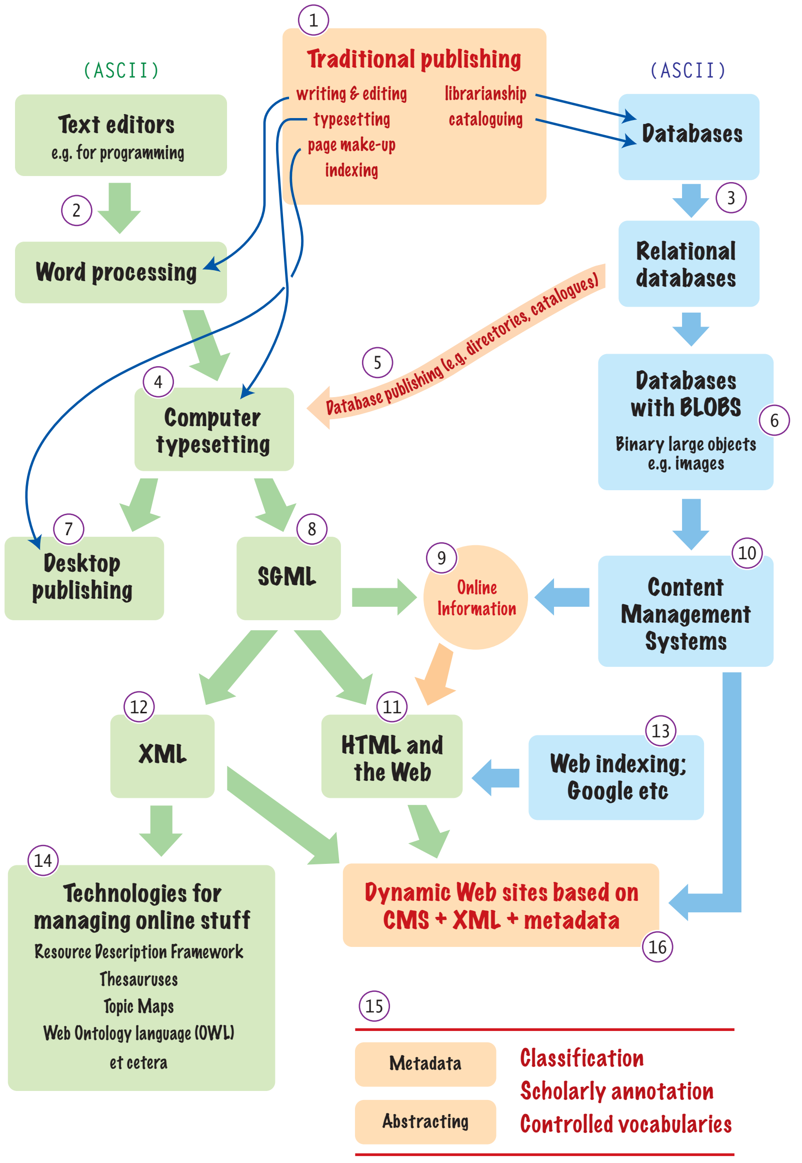
Against that, Conrad had written 'Processes of search and discovery', which involve algorithms, the application of machine intelligence, and so on. This allows queries to be directed against the data which has been structured by one or more of the methods listed – fielded databases can be interrogated with SQL, but there are also query languages for other methods of organisation, and the Topic Maps Query Language would be an example of that. (Later, John Lindsay amended the diagram by writing a column also spanning all four methods, with the words 'controlled vocabularies' and 'classification' in it.)

Four methods of organising data/information for search and discovery



'Fielded' data	Database table
Structured info	Tag language, DTD
Added metadata	Dublin Core, XMP, RDF
Indexation	Back-of-the book index Inverted tree Topic Maps

A timeline/concept map of the application of computer technologies to the publishing and management of data and information



The diagram shown left is a re-working of the larger of the two posters which Conrad created during the KIDMM discusssion day.

Whereas the original was constructed horizontally, the flow here is vertically, from top to bottom.

Also, some extra annotations suggested by other discussion participants have been incorporated here.

In constructing this version, I have used green boxes linked with a trail of arrows to mark a development stream of technologies primarily for handling text. Text editing laid the basis for word-processing, on which computer typesetting built further – *et cetera*.

On the right side of the diagram, a series of blue boxes similarly shows a development stream for database technologies as they have related to publishing and the management of information.

The light brown objects and red text captions relate not to technologies, but more to publishing and information management *processes and disciplines*.

The thin blue arrow-headed lines show how some of these processes which had previously been conducted without the help of computers took advantage of the new technologies once they had evolved to a point where they offered clear benefits over what had gone before.

Timeline/concept map

Conrad introduced his second, larger diagram, which he described as 'a publishing person's biased view of where it all came from'. In the very early years of computing, these new technologies had nothing to offer the world of publishing, and publishers continues to work at creating publications, getting them typeset etc. without computers. Librarians had no use for early computers either. ①

He identified two parallel and unrelated development streams in the earlier history of computing. Text editors initially were developed out of the need to program in languages rather than assembler code. From text editors, word processing evolved. ②

At about the same time, computer databases were invented, and were able to store numerical and textual information. ③ Of course, one precondition for both of these applications was the development of character encoding for the letters, numerals, spaces, punctuation marks and certain other functional markers such as tabs and 'carriage returns'. Enter the 7-bit character set, and ASCII.

Publishers, who up to this point had been sending typescripts off to be keyboarded at hot-metal composing machines thought: 'We could use some of this technology' – and so computer phototypesetting was born. ④ This required the concept of the escape character, to distinguish between text and instructions, and proprietary mark-up languages were developed, the function of which was to give commands to the machine that printed out the results, to make type larger and smaller, changing fonts and so on.

At this point, with ASCII text as the common means of information interchange, bright people start making links between databases and typesetting systems, and brought about the advent of database publishing systems, ⑤ used for the production of catalogues, telephone directories, bibliographies and many other kinds of database-derived publication.

Databases of course continued to develop. As more memory became available and affordable, databases started to be able to manage 'binary large objects' ('BLOBs') such as images and audio files and documents. ⑥ This laid the basis for document management systems in large-scale technical documentation projects, such as for military and aerospace technology.

In the 1980s, computer typesetting languages evolved into generic mark-up, especially SGML ⑧ – which not only was used for print publishing systems, but was quickly harnessed to produce well-structured online repositories of documentation ⑨ – Conrad wondered if anyone in the room remembered EBT Dynatext (and some did).

(It should of course be noted that only part of the publishing world was attracted to generic mark-up. The bulk of publishing projects migrated from the old style typesetting systems to Desktop Publishing systems such as PageMaker and QuarkXPress ⑦ — systems which made it easy to lay out pages on screen interactively in situations where document structure was less importance than visual appeal.)

SGML begat HTML, which rolled up the advantages of hypertext, client-server architecture and the Internet to bring us the World Wide Web ⑪ – and before we knew where we were, a paucity of online information had turned into a raging maelstrom.

Some within the Web community realised that the way HTML had developed as a pseudo-typesetting language was not ideal for working with structured documents, but saw SGML as being needlessly complex, and this led to the development of XML. ⑫

Database systems now entered the world of the Web in two major ways – as well as the obvious application of transaction processing, as is required for on-line shopping. Content Management Systems ⑩ had evolved to the point where it was possible to have the major part of a Web site's content stored in the database, and converted on the fly to HTML code for delivery to the end user.

From the point of view of the reader searching for information on the Web, however, the more significant use of databases was to hold searchable inverted-tree indexes of the contents of Web pages, the pre-eminent example being Google. ⑬

At the most recent end of this story of five decades, we get the development of technologies and methods and standards such as Topic Maps and Dublin Core, RDF and OWL, which are intended to deal with the contents within these processes, label and organise them and make them more manageable. ⑭ Significantly, these almost without exception have been built using XML.

However, these technologies would be nothing, and would never had been developed, were it not for the insights of publishers and librarians, scholars and other information professionals applying their insights about how to abstract and summarise the contents of data and information products; how to catalogue, classify and annotate them; how to develop controlled vocabularies and thesauruses to assist labelling, indexing and cataloguing. ⑮

The BCS Web site is an example of a second-generation site that has its contents encoded in XML and stored within a Content Management System, together with metadata that is organised in part from the Dublin Core set and in part from its own taxonomy. ⑯

Conrad felt it was worth re-telling this story to regain a historical perspective, and recounted an exchange of correspondence with Ian Horrocks while collaborating on a write-up of the Ian's Needham Lecture of December 2005. Though Ian works with XML every day, as the meta-language from which OWL is constructed, he had had no idea that the origins of generic mark-up lay in the world of typesetting and publishing; which Conrad had thought interesting.

Here and there along this timeline, different BCS Specialist Groups have become interested in different aspects of this web of developments, and Conrad said that our discussion day could be conceived as an opportunity to put our insights together.

Nic Holt said that another part of the big picture would include things like X-Link and X-Pointer, things which maintain relationships between information objects, which is quite important; and RDF relies on these kinds of mechanism. Doubtless the diagram shown here is not a complete one and it might be an interesting project to amend it, extend it and annotate it further.