BCS

Knowledge, Information, Data and Metadata Management

METAKnowledge Mash·up 2007

Conrad Taylor reports on the 17 September 2007 one-day conference organised by the KIDMM knowledge community (Knowledge, Information, Data and Metadata Management), based in and supported by the British Computer Society. This report and other outputs of the event can be found at http://www.epsg.org.uk/KIDMM/mashup2007/outputs.html

Background: about KIDMM

The British Computer Society [1], which in 2007 is celebrating 50 years of existence, has a self-image that is much to do with engineering, software, and systems design and implementation. However, within the BCS there are over 50 Specialist Groups (SGs). Among these, some have a major focus on what we may call 'informatics', or the *content* of information systems.

Until quite recently, those SGs with a focus on digital content lived their lives in fairly discrete intellectual silos. True, the BCS Electronic Publishing Specialist Group (EPSG) has for 20+ years has strong participation from academics, librarians, designers and commercial publishers, especially in scientific, technical, medical and legal publishing. But EPSG did not talk much to other groupings within the BCS.

In 2001, the BCS held a gathering of representatives of its Specialist Groups; the SG Assembly is now a twice-yearly event. Through these meetings, SGs got to know each other better, but tended to compare notes about organisational matters, rather than about their specialist topics of interest. Then at the Spring 2003 SG Assembly, Dr David Penfold, then Chair of EPSG, gave a talk about metadata. Several of the SGs present declared an interest in metadata.

At a BCS SG Assembly in 2005, one workshop discussed shared-interest topics around which SGs could collaborate. *Knowledge, information and data management* was identified as a candidate. Grassroots discussion among BCS activists led to a workshop in March 2006 [2], from which emerged the KIDMM group (Knowledge, Information, Data and Metadata Management).

KIDMM is an extremely informal entity – essentially, a community gathered around a JISCmail discussion list, which at the time of writing has 66 members. About a third of subscribers are from communities outside the BCS [3].

During 2006, a Synergy Working Party of the BCS made a study of prospects for closer collaboration between SGs and the BCS Forums, which were more recently established. The resulting paper, discussed at the Autumn 2006 SG Assembly, put forth the idea of 'knowledge communities'. While no clear idea emerged of what a KC might be, many delegates thought that KIDMM looked more like a nascent KC than anything else around – as a result of which, KIDMM received BCS backing and funding in 2007 to run a conference appealing to a larger audience, and to construct an exhibition about this aspect of computer use.

Mash-up day

The MetaKnowledge Mash-up conference was billed as a 'sharing and thinking day'. In creating the event's name, Ian Herbert and Conrad Taylor reflected an aspiration to put together, in new ways (i.e. 'mash up') the knowledge that different communities maintain about ways of managing knowledge, data and information.

Over ninety people attended, from diverse backgrounds – ten BCS SGs or forums, three university libraries, several university departments of computing or information science, the British Library and the National Archive, two museums, JISC and UKOLN, ISKO-UK, the Metropolitan Police, a few national and local government departments, two large consulting firms, and many independent consultants.

Introduction: 'Handles and labels'

The day was chaired and introduced by Conrad Taylor, who acts as co-ordinator for KIDMM. Conrad identified three ways in which we add 'handles and labels' to information within computer systems, to make it easier to manage.

In a database, data obtains meaning from its container: a number is made meaningful by being placed in a 'date of birth' field, or a 'unique customer ID' field, etc. Explicitly documenting the meaning of fields and the relationships between them becomes important when databases need to be merged or queried – as Christopher Marsden of the Victoria & Albert Museum would later demonstrate.

Since the 1970s, the term *metadata* has been used within the data management community to refer to information that describes the allowed content types, purposes *&c.* of database fields and look-up relationships.

To hear this may surprise librarians and information scientists, who started using 'metadata' to mean something different in the early nineties (Lorcan Dempsey of UKOLN may have started the trend?). KIDMM now recognises this, and as a result, any conversation about metadata within KIDMM has to include disambiguation about which of these meanings is intended [4].

Information in text form is more difficult to manage, as it tends to lack useful 'handles'. There are exceptions: mark-up languages such as SGML and XML allow entities within text to be given machine-readable semantic contexts. Military systems documentation has used SGML/XML structure in support of information retrieval for a couple of decades.

Failing that, we have to rely on free text search, with all its frustrations – as Tony Rose would illustrate.

The third 'handle-adding' approach which Conrad described is the attaching of cataloguing data – labels, or 'metadata' in the librarians' use of the word. This may live outside the information resource, as in a library catalogue, or be embedded in the resources, as Exif or IPTC data is embedded in digital photographs.

Dublin Core [5] is a well-known starter-kit of metadata fields. One troublesome Dublin Core field is the Subject field: here we bump into problems of classification. Conrad called on Leonard Will to give the conference a briefing on what the issues are – which later in the day he did. Conrad also referred to the practice of using controlled vocabularies, an example of which is SNOMED–CT for health records, about which Ian Herbert would later speak.

Information Retrieval today: an overview of issues and methods

TONY ROSE, Vice-Chair of the BCS Information Retrieval SG, offered Wikipedia's definition of information retrieval: *search for information in documents, or for the documents themselves, or for metadata which describes documents*. The task faced by IR is formidable, given that 800 MB of new information is added each year for each person on the planet, and about 80% of corporate information is unstructured.

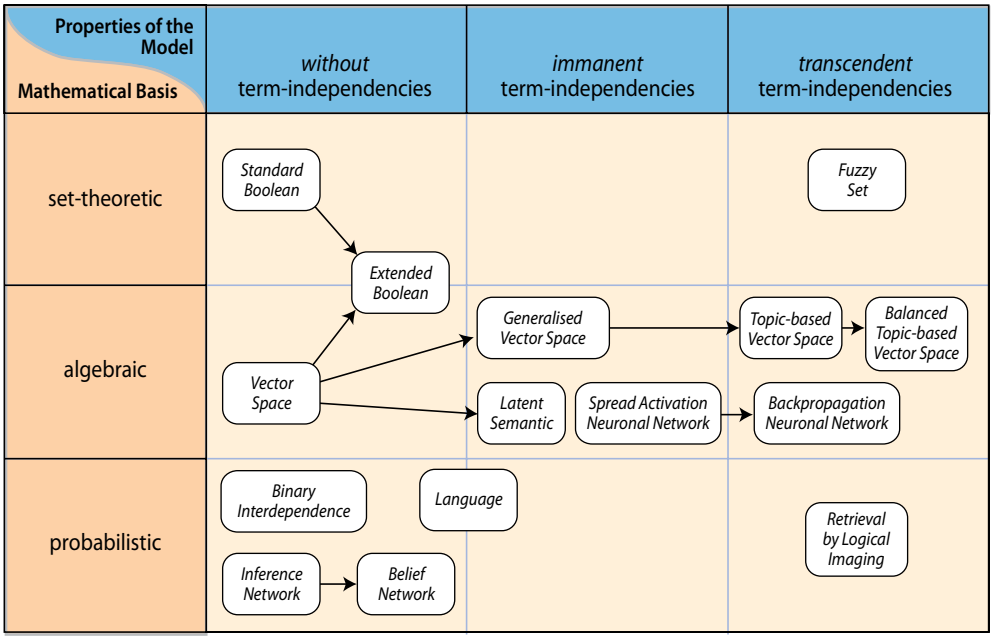


Fig 1 – An array of contemporary approaches to Information Retrieval.

Naïve view, technical view

Most people's experience of search is of Web search, a field that is dominated by Google, Yahoo! and Microsoft Live Search in that order. There are also other search engines that operate within internal or specialist document collections.

The 'naïve science' view of how search engines function is that they understand what users want to find. However, all they really do is count words and apply simple equations; they measure the 'conceptual distance' between a user's query and each document in their database. Authors of documents express concepts in words – as do searchers, in the search terms they choose. The central problem in IR is whether a good conceptual match can be found between the searcher's terms and the author's.

To process searches, a search engine must represent the search terms and author's terms internally. The usual 'bag of words' approach treats every document as a collection of disconnected words. The similarity of the contents of the 'bag' to any search terms is calculated by set theory, algebraic or probabilistic methods, the algebraic approach being the most common. Tony spared us the mathematical details! However, we reproduce as Fig. 1 the diagram with which he showed the variety of types and sub-types of approach used.

In evaluating the effectiveness of information retrieval systems, two key terms are 'precision' and 'recall'. A search is insufficiently precise if it brings back a lot of irrelevant documents, and it scores poorly on recall if there are many documents relevant to the searcher's query that are not retrieved. Unfortunately improvements in precision are usually to the detriment of recall, and vice-versa.

Why is search hard?

What makes search hard? The bag-of-words model has the weakness of treating 'venetian blind' and 'blind Venetian' as equivalents. Ideally, words shouldn't be treated as if they had independent existence in documents; word order is important; relevance is affected by many structural and discourse dependencies and other linguistic phenomena.

In many search engines, the database is reduced in size by excluding words such as *the*, *and* or *of* – the 'stop words' approach. The trouble is, these are useful function words: by excluding them, much linguistic structure is thrown away.

In specialist systems, words may be added to a stop list because every document in that collection can be expected to have them, making them useless for discriminating relevance – for example a patent database might exclude the word *filing*. But what if the patent being searched for has something to do with iron filings, or the process of filing a piece of metal?

Word-stemming may be used to expand the effectiveness of recall. The words *compute*, *computer*, *computation* are essentially about the same concept, so searching on *comput** or having the system treat the variants as equivalent may be useful. But as examples of false positives we have *organ* and *organisation*, *army* and *arm*. As for false negatives, stemming may fail to note the relationship between *cylinder* and *cylindrical*, depending on how the stem is defined.

Though stemming tends to focus on suffixes, prefixes may indicate related concepts: *de*focus, *un*structured, *dis*assembled. There are false positive here too: *dis*tress is not about cutting someone's hair off!

Then there is the problem of named entity recognition. The identity of 'New York' as a named entity is destroyed by the bag-of-words approach. One corrective solution is to add a gazetteer of geographic entities. One could do the same for people's or organisations' names.

Concept matching

Moving from syntax to semantics, matching concepts for effective recall presents greater challenges. At one time the IR community had great hopes of Natural Language Processing technologies. Though this was laid aside in favour of the bag of words approach, there is renewed interest in NLP: two companies to watch here are Powerset and Hakia, which aim to deliver meaning-based retrieval.

It would be handy if a system would recognise *car* and *automobile* as synonyms. Some systems use a thesaurus, but it can be problematic to decide at what level of abstraction to set equivalence (in some use-cases *car* and *bus* might be equivalent, in other cases they might even be antonyms, e.g. in a debate about sustainable transport policy!). Building thesauri always involves a huge amount of human editorial input; but there is currently renewed interest in the use of subject-specific taxonomies and thesauri for concept matching in specific domains.

The flip side of concept matching is the problem of disambiguation. There are at least three meanings for the word 'bank'. One should have the riverside distinguished from the financial institution and the aerobatic manœvre.

So why does it work?

If search is so hard, how come Google gets it right so often? Google's originators had the insight that on the Web there is another index of relevance: the links to document being indexed. The more links point to a page, the more highly it is rated by users.

Amongst other hot and current IR topics, Tony pointed to vertical-market search applications in very specific domains; 'rich media' search as on YouTube; personalisation of search engines to specific users; answer engines; multilingual search; and search agents.

Discussion

David Pullinger (UK Cabinet Office), who is in charge of the pan-government search solution, commented that their research shows that ordinary people searching for government documents use terms other than ones found in the government databases: governmental language and people's ordinary language are different. Ironically, Google is the only tool that discovers these documents effectively, because it picks up words that are associated with links to the documents; and the links are often written in plainer English.

Someone wondered how important it is to train users to be better at search. He's noticed that when he uses different search engines, he formulates his terms differently, because (over an admittedly long period of time) he's learned how the different engines behave.

Conrad noted that in January 2003, the BCS Developing Countries SG held a discussion workshop on 'Information Literacy' [6]. One definition of information literacy offered by SCONUL, the Society of College, National & University Libraries, includes the skills of using search engines to get the results you want. Conrad is often asked by friends to search for information on the Web because he finds it faster and more effectively than they can. But he thought that all the arguments for *user education* as the answer to efficient search should not let engine-makers off the hook: better interfaces, better usability are also vital.

Someone recalled the MSc he did 20 years ago in natural language processing and semantic analysis; what Tony had outlined seemed to point at that space. He wondered if the issue of prefixes and suffixes in morphological analysis was relevant to English, but less so to other languages. Conrad mentioned Scots Gaelic, where consonants within nouns are 'morphed' to an aspirated form in the genitive case – much more complicated to teach a computer about than suffixed inflections.

Professor Ronald Stamper, who established the MEASUR programme at the LSE for applying semiotics in information systems development, mentioned the techniques that his group had been developing for disambiguation, leading to a 'semantic normal form'. This, he said, is applicable to a whole range of problems, and captures the core aspects of meaning, thus removing ambiguity. It's an approach that has been applied in particular to legal expert systems.[a]

Another questioner wondered whether we are changing the way we write in anticipation of the needs and limitations

a. Ronald Stamper has sent us an explanation: 'The Semantic Normal Form (SNF) is a canonical schema that captures the core aspects of meaning. The SNF is based on a theory of perception that leads to a new concept of *ontological dependency* that expresses how the existence of each thing depends on the coexistence of others. An operational definition of meaning then relates one's terminology to the SNF. The SNF is not a formal construct but an empirical one; it is stable across languages and communities but can be fine-tuned to them by means of perceptual norms. As a canonical form, different analysts should arrive at the same SNF for a given application. It serves as the schema for a Semantic Temporal Data Base (STDB) that always incorporates time constraints and indicates where responsibility lies for the data, which make it almost impossible for the STDB to accept meaningless data. When used in practice, the SNF has enabled development, support and maintenance costs for an administrative Information System (IS) to be a reduced by a factor of ten. The SNF and its method of semantic analysis are products of the LEGOL/MEASUR research programme on specifying an IS as a system of social norms. It was conducted initially at LSE funded by EPSRC, ESRC, IBM and Digital, later at the University of Twente and now at Reading, Greenwich and other universities. Other products of this research include: Problem Articulation Methods for *soft systems* analysis based on the architecture of the social norms; Norm Analysis for the representation of social norms in a precise form; and software to generate a computer application for the relevant domain, derived directly from the SNF and its associated norms.'

of information retrieval systems. Tony thought it not very likely, since technology changes faster than language; but agreed that for archives of historical literature a thesaurus might have to be customised to recognise a different set of synonyms than in modern usage. (Consider the changing meanings of 'gay' and 'fantastic'.)

Returning to the issue of terminology mismatch raised by David Pullinger, Conrad drew attention to a 2003 paper on e-democracy by Danny Budzak [7], reporting on a study that compared the terms that local government Web sites use to describe local government services, to those chosen by users.

(This paper was just one of a collection of PDF documents and MP3 audio recordings of various lectures and interviews, which KIDMM wrote to a CD-ROM and handed to everyone attending the MetaKnowledge Mash-up day. All of these resources are also available online.)

Data Mining, Text Mining & the Predictive Enterprise

Tom Khazaba of SPSS was the next to speak, focusing on data mining issues in commercial and government contexts. Tom had been invited as a speaker by Tony Jenkins, who represents the BCS Data Management SG within KIDMM: and data mining is one of the varied subjects that DMSG addresses through its regular programme of meetings.

What is predictive analytics? One definition by Gareth Herschel is that it 'helps connect data to effective action by drawing reliable conclusions about current conditions and future events'.

Most large organisations may derive insights of practical use by applying an analytical process to the data they keep. 'Data mining' refers to the core of that analytical process. 'Predictive analytics' is a term encompassing the broader context of use: it 'helps connect data to effective action by drawing reliable conclusions about current conditions and future events'.

In business, a common goal of predictive analytics is to get customers more efficiently and cheaply; also to identify services to sell to existing customers, and incidentally to avoid trying to sell them things they don't want as well. By analysing why customers close accounts and go elsewhere, for example, banks could learn how to retain customers.

The next largest group of business applications of predictive analytics is in risk detection and analysis, *e.g.* detecting fraud and suspicious behaviour.

Governments use predictive analysis to prioritise tax actions in the field of tax collection, especially with large taxpayers such as corporations. Very few tax errors are due to fraud; most are simply mistakes, and taxation authorities can use predictive analytics to guide their checking of tax records. Getting this right, said Tom, can be worth billions to a government.

In policing, data analysis leads to more efficient resource allocation, e.g. by identifying crime hot-spots. In detective work, computer-assisted analytics can detect patterns across huge volumes of crime reports – for example, to catch a serial offender. These methods can speed up records analysis so dramatically in comparison to manual methods that such searches, which could previously be justified only for very serious crime, can be used more routinely and to solve a wider range of cases.

Analytical technologies

Describing the technology, Tom introduced an algorithm in common use called a 'decision tree'. In data mining, decision trees breaks a population into subsets with particular properties, and can do so repeatedly, sifting data to map from known observations about items, to target categorisations with predictive use.

Neural networks are relatively simple algorithms plugged together to experiment with stated relationships between different elements in a dataset, which may reveal hitherto unnoticed patterns in data. Neural networks are trained against a sample dataset.

A third class of algorithm used in predictive analytics is clustering algorithms, which find natural groups within data. They are often used in statistical analysis; but in data mining they are commonly used to find anomalies. In the detection of fraud, spotting an anomaly is not a proof, but it can draw attention to instances that merit investigation.

Association discovery sorts out 'what goes with what' – an example might be 'customers who buy charcoal and beer are also likely to buy lamb chops or steaks'. Associations may be sequential – 'a lot of people visited this Web page, then this, and ended here'. Tom pointed out that such analysis of Web trails can help identify points of navigational confusion that cause users to abandon a site.

In summary, data mining algorithms discover patterns and relationships in the data, and from this they generate insight – things one had not noticed before, which can lead to productive change. They also have predictive capability.

Using predictive analytics effectively

There is an increasingly used standard data mining process model called CRISP-DM – the CRoss Industry Standard Process for Data Mining [8]. A benefit of this model is that it gives experts who work in this field a common terminology, and guides organisations in their thinking about how to solve problems with the aid of these tools and algorithms. The CRISP-DM standard 1.0 is about eight years old, and the consortium is currently working on version 2.0.

The really big returns from predictive analytics come when it is highly integrated with an organisation's main business processes. Every business process is a chain, with key decision points. The more decision points an organisation can support with predictive analytics, and the more kinds of decisions that can be improved at each point, the more benefit can be derived.

SPSS has noted a number of factors that make predictive analytics work effectively for organisations. Firstly, it should be easy to do. SPSS's Clementine software made its debut a

decade ago as the first visual data mining workbench, making it possible for non-technologists to understand it. Sophisticated analyses can be set up in diagrammatic form. Also it is important to have tool support for the whole CRISP-DM process, not just the data mining algorithms.

Tom admitted he had focused on data – it's his area of expertise – but reminded the audience that 80% of the information in an organisation is typically found in text form, and this is worth 'mining' too. Customers' Web-browsing behaviour can be a rich information source; and results from surveys are another.

Discussion

Mentioning SPSS's roots in statistical analysis in the social sciences, Conrad asked Tom about applications in health and social services, where there are needs to mine epidemiological and demographic data and discover patterns such as poverty, teenage pregnancy and low birth weight. Did Tom know of organisations using Clementine tools in these fields? Yes indeed, he replied; in fact, about half the user base is in the government area, split roughly equally between national and local government. If fact it is easier to get examples of successful use of predictive analytics from the public sector because there aren't the same issues of secrecy about which analytical methods give companies competitive advantage.

Tony Jenkins commented that analysis of data for business intelligence (BI) is not new, but in the past it relied a lot on the prior analytical task of knowing what to ask from the data. What is new here? Is it that interesting facts are 'pushed out' by data mining processes that would otherwise not have been looked for?

Tom said that this was both correct and incorrect: the difference between predictive analytics and traditional BI is that in the latter one does start from an understanding of where the patterns are to be found, whereas in predictive analytics one starts from a position of *not always* knowing in advance. However, you do have to have a starting point, such as 'what pattern might indicate fraud?' Serendipitous discoveries *are* possible – but successful implementations of predictive analytics always start from a business goal; it's not just random rumination.

Jan Wylie asked if this sort of software was used for predicting credit-worthiness; and if so, where had things gone wrong in the financial meltdowns caused by the 2007 collapse of the sub-prime mortage lending sector?

Tom replied that these tools aren't magic – they deliver good results only when good business knowledge is applied in framing the investigation, and in evaluating the results. These companies may not have understood the risks they were taking by lending to such customers. Also, because prediction is always based solidly on historical data, there is no way that *unprecedented* events can be predicted.

SNOMED Clinical Terms – the language for healthcare

Ian Herbert is Vice Chair of the BCS Health Informatics Forum, and currently Acting Chair; and is on the committee of the Primary Health Care SG. HIF, though it was the 4th BCS body to get 'forum' status, in fact is the forum with the oldest roots, having existed for many years as the Health Informatics Committee – a collaboration of several healthcare-related BCS Specialist Groups.

Ian defined SNOMED–CT as 'a terminological resource that can be inplemented in software applications to represent clinically relevant information reliably and reproducibly.' It has been adopted by the NHS as a standard controlled vocabulary for use in the Care Records Service.

The purpose of SNOMED–CT is to enable consistent representation of clinical information, leading to consistent retrieval. In direct patient care, clinical information is what documents people's health state and treatment, expressed as entries in electronic patient records. There is also decision support information – this assists health professionals in deciding what action is appropriate for the care of a patient. Secondary uses of health information are in epidemiology, medical research, resource allocation, administration and management.

Because people move between different health providers (GP, hospital, specialist clinic), there is a need for semantic interoperability between the information systems of all the providers: meaning must transfer accurately from system to system. A controlled terminology will help deliver this.

Nobody believes that all clinical information can be structured and coded – there will always be a role in medical records for diagrams, drawings, speech recordings, movie clips and of course free text. However, SNOMED–CT enables consistent 'tagging' of information about individual patients, plus knowledge sources such as drug formularies.

If the clinical record tells us 'this person has diabetes', we should be able to link into knowledge sources that describe diabetes and how to deal with it.

Also important is automation of decision support, for example to ensure that prescribed drugs will not interact adversely with other medication the patient is receiving. (Ian described this as a 'killer application' for semantic interoperability, because unsafe prescribing currently kills a lot of people each year – certainly more than the 3,000 or so who are killed annually on the roads.)

Under the hood

SNOMED–CT stands for 'Systematized Nomenclature for Medicine – Clinical Terms', and is the result of the 2002 merger of the American SNOMED Reference Terminology (RT) with the UK National Health Service Clinical Terms, Version 3. The latter was developed from the 'Read Codes' developed and made commercially available by Dr. James Read, and purchased by the NHS in 1998.

Why has the NHS chosen SNOMED–CT instead of some other standard such as the International Classification of

Diseases (ICD-10, which has been adopted by the EC), or Operations and Procedures Classification System (OPCS)? These function well as classifications, but are not detailed enough for unambigious use in medical records, to guide treatment. They also lack the facility to combine expressions to clarify meaning (whereas SNOMED–CT lets one combine e.g. 'emergency' and 'thoracotomy'; or even 'recurrent' + 'ingrowing toenail' + 'left' + 'great toenail').

SNOMED–CT is updated more frequently than the alternative classifications, so it is more responsive to changes in medical knowledge and practice. Remember, thirty years ago, there was no classification for HIV/AIDS. Individual specialists often ask for new terms to be added; and administrative terminology changes all the time.

Many GPs use a standard terminology – the old Read Codes, generally version 2 but occasionally even version 1. This terminology is much simpler than SNOMED–CT.

As for free text, of course it is valuable in medical records and there will always be a role for it, but it has two major drawbacks. For one thing, meaning may be ambiguous – does 'fit' mean a seizure or a state of health? For another, meaning in free text is not available for computation. It cannot be analysed automatically for auditing and to direct payments by results; it cannot direct care pathways; and it cannot trigger automated warnings, such as about allergic reactions to medications or interactions between them.

Concepts and descriptions

Any SNOMED–CT term is a class descriptor: 'jawbone' as a term describes all jawbones. The boundary between a terminology and a classification is slim – it depends what you use it for. In linking to knowledge sources, one is not talking about the fracture of this particular patient's jawbone (as one would in a medical record), but a *class* of such fractures.

SNOMED–CT consists of hundreds of thousands of concepts, each given a unique 'ConceptID' code. A concept may have many names, each with a unique 'DescriptionID' code. For example, the concept with ConceptID 22298006 has as its Fully Specified Name: *Myocardial infarction (disorder)* – for which the DescriptionID is 751689013. This particular Fully Specified Name is not likely to be found in a medical record. A preferred term is offered: *Myocardial infarction* – DescriptionID 37436014. There are also several synonyms such as *Cardiac infarction* or *Heart attack*, each with its own DescriptionID. (Note that in practice, the users of health information systems should never have to be confronted with these codes directly.)

This formality can resolve the ambiguities of free text. To a neurologist, 'cord compression' means compression of the spinal cord; but to a midwife, the umbilical cord is implied. In each context, the text term makes perfect sense – but as machines don't do context well, we need to have two distinct concepts in SNOMED–CT, with distinct ConceptIDs.

Does using a controlled terminology like SNOMED–CT limit what can be said in a medical record? Not much. With about 400,000 health care concepts, a million clinical terms and 1.5 million semantic relationships, SNOMED–CT provides a very long pick-list. Terms can also be further qualified by contextual modification such as 'family history of' – 'planned' – 'refused'. (The ability to record a 'family history of' something is important for example in the case of diabetes.)

The extent of the SNOMED–CT vocabulary allows specification of conditions and procedures in a great deal of detail: a *femoral-femoral crossover angiogram* is a kind of *peripheral graft arteriogram*, which is a special *peripheral angiography* procedure within the more general framework of *peripheral angiography*.

SNOMED–CT is organised as several hierarchies, based on multiple top-level concepts, an example of which is 'body structure'. Below that, *index finger* is 'a kind of' *finger*, which is 'a kind of' *hand part*. (Most structure in each hierarchy is based on 'kind of' relationships.) A single concept can belong to more than one hierarchy: tuberculosis is a 'kind of' *disorder of the chest*, but is also part of the hierarchy of 'infectious diseases'. In other words, SNOMED–CT is a polyhierarchy, not a taxonomy. Each concept may have permitted qualifiers, too: 'pain' is a concept, and can have a 'severity' qualifier.

Ian briefly showed a table which gives demonstrates some of the hierarchies in SNOMED–CT, plus examples. This has been reproduced as Table 1 at the top of the following page.

SNOMED–CT terms are used first to define an item, then the definition may be qualified. In making the definition, concepts can be combined with attribute–value pairs. For example a procedure can be represented thus:

method	=	excision
site	=	both tonsils
using	=	laser device

This is a *post-coordinated* representation of a clinical procedure. Its pre-coordinated equivalent would be 'bilateral laser tonsillectomy'. In the medical record, a practitioner may employ either the array of post-coordinated terms, or the pre-coordinated form. Most practitioners prefer to use pre-coordinated terms, but within such a large terminology finding a single pre-coordinated term is difficult: and this is one of the tensions within SNOMED–CT.

Terms can also be added which are not actually defining characteristics; this allows the specification of some richer relationships. In the case of our tonsillectomy, there could be a *qualifier* attribute so that outcome = success; or one could define a skin rash as *caused by* something. Defining characteristics of course should always be there; as for how many qualifiers and relationships are described in the medical record, this depends in part on how diligent one is in authoring the record, and on how significant the relationships are in the case in question. The classics are fundamentals such as *reason for* and *caused by* – others may not have yet been added systematically throughout SNOMED–CT.

Using SNOMED–CT in practice

Finding the right term to use in any particular case can be a challenge. You can search for a word or phrase, in which case you'll be presented with a list of all terms using that word or phrase. In a large terminology, that can be daunting! You can

Selected SNOMED–CT Topic Hierarchies	Examples
Clinical Finding: – Contains the sub-hierarchies of Finding and Disease. – Is important for documenting clinical disorders and examination findings.	Finding: Swelling of arm Disease: Pneumonia
Procedure/Intervention: Concepts that represent the purposeful activities performed in the provision of health care.	Biopsy of lung Diagnostic endoscopy Foetal manipulation
Observable entity: Concepts represent a question or procedure which, when combined with a result, constitute a finding.	Gender Tumour size Ability to balance
Body structure: Concepts include both normal and abnormal anatomical structures. Abnormal structures are represented in a sub-hierarchy as morphological abnormalities.	Lingual thyroid (body structure) Neoplasm (morphological abnormality)
Organism: Ocoverage includes animals, fungi, bacteria and plants. Necessary for public health reporting, and used in evidence-based infectious disease protocols.	Hepatitis C virus *Streptococcus pyogenes* *Acer rubrum* (Red Maple) *Felis sylvestris* (cat)
Substance: Covers a wide range of biological and chemical substances. Includes foods, nutrients, allergens and materials. Used to record the active chemical consituents of all drug products.	Dust Oestrogen Haemoglobin antibody Methane Codeine phosphate
Physical object: Concepts include natural and man-made objects. Focus is on concepts required for medical injuries.	Prosthesis Artificial organs Vena cava filter Colostomy bag
Physical force: Includes motion, friction, electricity, sound, radiation, thermal forces and air pressure. Other categories are directed at mechanisms of injury.	Fire Gravity Pressure change

Table 1 – a selection of SNOMED–CT topic hierarchies and examples of what they include.

try a search for the pre-coordinated term, but there may not always be one available in SNOMED–CT. Or, you could start from a point you know, and browse the term hierarchies. For example you could start with the name of a procedure like 'excision', find your way to a site definition, and so on. These approaches can easily be combined.

The ideal situation for input is where the record-making system is designed for a well-defined area of practice. On a data input screen for recording blood pressure, the template can limit the choice of terms to only those that are relevant, on a drop-down menu.

It may be possible to have text automatically encoded as it is being entered. Such intelligent systems would be highly desirable, but at present they are unreliable, and the codes they generate must be reviewed and approved before being committed to the medical record.

One problem that is being encoundered in implementing SNOMED–CT is detecting equivalence between expressions of a concept in pre-coordinated and post-coordinated representations. Most pre-coordinated terms in SNOMED–CT were taken in *en masse* from the Read Codes terminologies, and can represent quite complex conditions. An example is Colles' fracture – a pre-coordinated term named after the 19th century surgeon Abraham Colles, who described a particular kind of fracture of the wrist end of the radius bone, in which the lower fragment of bone is displaced backwards, leading to a characteristic deformity. Achieving a canonical post-coordinated description of Colles' fracture involves several fundamental concepts, the relationships between them, and qualifiers.

Representing pre-coordinated identifiers canonically through relationships between fundamental concepts is not always yet available in SNOMED–CT. This is a significant problem, because unless equivalence is detected 100%, all data retrieval has a grey margin – with the risk of possibly insufficient recall of all records to match the query.

SNOMED–CT has problems expressing negation, which is critical in medical records – diabetes *excluded*, appendectomy *not performed*, no pain in right leg, and of course 'NAD' – nothing abnormal detected. The problem is, there are so many ways of expressing negation in the English language.

As the NHS starts to use SNOMED–CT – e.g. to build data entry templates – errors are being discovered in it. Concepts are being found to be in the wrong hierarchies, or the wrong position in them. And there's a huge problem of enabling accurate, speedy use of SNOMED–CT in unconstrained situations, such as when taking a patient history, in which circumstance almost anything could come up. 'GPs aren't data collection clerks,' warned Ian, and in the complex three-cornered clinical situation with GP, patient and computer, the computer system must be easy to use.

Read Codes have worked surprisingly well in general practice. But SNOMED–CT is going to be a tougher proposition. The benefits it can bring will be enormous, but the gain will cost some pain. Until now, hospitals have not tackled the

serious encoding of data about patients as it is collected, and this is going to be the first time that many clinicians have met controlled terminology.

Human beings are lazy, and good at inference. As a result, patient records are typically full of short cuts. 'BP 140/80' means 'blood pressure was taken; the systolic pressure was observed to be 140 mm of mercury; diastolic pressure was observed to be 80 mm of mercury.' Also, any human reading this will assume that the blood pressure reading is that of the patient in whose record it is written, and that it was taken within a particular encounter with the patient.

Computers in contrast are pernickety, and very bad at inference, and want everything made unambiguous. And SNOMED–CT is likewise: there are numerous codes for a blood pressure reading. Ian recalled an occasion when the question was asked which code for a blood pressure reading should be used in which situation, and the SNOMED–CT experts had to leave the room to think about it! GPs and clinicians are not going to do that.

Users want the biggest return possible per keystroke or mouse-click. For this reason, neither post-coordinated input methods nor unconstrained searching through labyrinths of terms are going to be popular.

Despite SNOMED–CT's great scope, it is not sufficient. For one thing, it deals with concepts, and concepts identify types. It does not deal with numeric values, e.g. 'weight = 70 kg'. It does not identify individual objects in the world, such as people. Therefore SNOMED–CT terminology has to be used within an external syntax that binds instances of concepts to their context, and this will include records of

- who the record is about – the subject, typically a patient;
- when and where the action/event occurred, or the observation was made;
- who performed the action or made the observation.

Where are we now?

SNOMED–CT started as an American system, then was developed as an Anglo-American collaboration. Since April 2007, an independent international body has been in charge of it (IHTSDO – the International Healthcare Terminology Standards Development Organisation). Several countries have adopted it, led by Spain and Germany, with more on the way. Nor can it be said to have significant global rivals.

It may have been adopted by the NHS, but there is still little practical experience of using it for keeping patient records, and virtually no experience of using it in real time, during the patient encounter. It's a huge experiment – some may say, a huge gamble, with a lot at stake.

How will we know when SNOMED–CT is successful? Ian closed with a quote from Alan Rector of the Medical Informatics Group at the University of Manchester:

> We will know we have succeeded when clinical terminologies in software are used and re-used, and when multiple independently developed medical records, decision support and clinical information retrieval systems sharing the same information using the same terminology are in routine use. [9]

Discussion

Someone asked how SNOMED–CT relates to the MeSH (Medical Subject Headings), National Library of Medicine and National Institutes for Health classifications and so on. If you have a set of medical literature that is indexed according to MeSH, NLM or whatever, how does this connect?

Ian asked Janette Bennett of BT whether there has been a mapping exercise between SNOMED–CT and MeSH; she thought so, which matched Ian's own perception. It is highly likely that they can be mapped at the concept level, though how far such mapping has been done is unclear. There has definitely been mapping from MeSH to ICD10.

Tony Rose acknowledged the wealth of experience being gained in constructing SNOMED–CT, and wondered how much of that experience is translatable and has been used to inform other projects. Ian replied that in talking recently to some Australians at an international conference, he got the impression that quite a lot of learning was being transferred.

When you build a terminology like SNOMED–CT, said Ian, you construct many little theories. For example, there is a theory around *Action* – that it involves a performer, a recipient, a lifecycle etc. SNOMED–CT likewise is realising the need to embed little models in various places to make the thing work, and this will help it relate to other terminologies outside – providing they share similar models of what an action is, of course.

The maintainance issue, and input paradigms, are things other people should think about and learn from. Consider: SNOMED–CT is a compositional terminology. You can use individual concepts where they relate, and you can put them together in post-coordinated fashion. To Ian's knowledge, this is the first time that a compositional terminology has been constructed on such a massive scale; and the lessons have been frankly quite traumatic. So yes, there is much that has been learned which should be shared.

Someone mentioned the long lifetime that's intended for SNOMED–CT, and asked us to recall the state of medicine 80 years ago. Had SNOMED–CT been built 80 years ago, it might have fitted into a small book. What is being done to make sure SNOMED–CT will still be valid in 80 years' time?

Ian explained that no concept in SNOMED–CT will ever disappears: once created, 'it's immortal'. A concept may, however, be deprecated for current use. There are concepts in SNOMED–CT that came from James Read. We know them to be ambiguous; they remain in SNOMED–CT, but marked as deprecated.

Imagine SNOMED–CT had started in the 16th century, the concept of malaria as 'ague', based on the theory that the illness was caused by breathing bad air near swampy ground, would still be there – but deprecated for current use. Over time, new science will cause concepts to move from one hierarchy to another, or new hierarchies may be introduced. Genomics, for example, will cause many changes.

Mark Phillips (Dept of Children, Schools and Families) said that his department is developing a similar kind of project. He asked about strategies for improving take-up of SNOMED–CT. Ian replied with regret that various pilots of

the use of SNOMED–CT have been postponed. This could be disastrous: we have no idea how things will work in practice.

All major systems procured under NHS Connecting for Health will use SNOMED–CT – it's a huge experiment, and Ian wouldn't have wanted it done this way. But… sooner or later you have to bite the bullet in using standardised terminologies. Use in real-time patient care will be the acid test. If it is too cumbersome to use, there will be major issues. This must be exercising the brains of all of the manufacturers and suppliers. Ideally, use of SNOMED–CT would be broken into constrained contexts where the number of terms used is actually quite small and manageable. But there are irreduceable areas where you can't do that – taking a patient history is a typical example, certainly in general practice.

Conrad noted there seems to be a Information Design or Human–Computer Interaction issue here. However elegant the terminological system, if when you implement it in software it is unusable, it will be a failure. (Alan Rector describes this as the test of 'clinical pragmatics'.)

Robin Clark of the National Cancer Research Institute asked about governance of the process for expanding the terminology and guaranteeing its consistency. Ian replied that one of the most difficult aspects anyone has in looking after SNOMED–CT is deciding whether a proposed new term represents a new concept, or is a synonym for something already there.

There is still a lot to be learned about the authoring process. If you understand what a new concept means – and most of the authors are expert clinicians – then you know to what hierarchy it belongs. Example: this is a surgical procedure, so you search the surgical procedure hierarchy. You would know what qualifiers are available, what defining characteristics are important. Therefore, if you are an expert in your field and you understand the principles on which SNOMED–CT is based, you have a fighting chance of doing a good job. Sure, better tools are required – but there is always going to be a big element of human skill and knowledge in authoring any such terminological resource.

Geospatial information and its applications

Dan Rickman of the BCS Geospatial Specialist Group spoke to introduce this subject. Geospatial SG is one of the BCS's most recently founded Specialist Groups.

Dan noted that geospatial data had already featured in previous talks: mash-ups with Google maps in the case of Tony's talk, police identification of hot-spots in the case of data mining. He promised to give us a good introduction to geospatial data, and share some of what the SG has learned in its meetings so far.

Geospatial Information Systems allow the 'capture and modelling, storage and retrieval, sharing, manipulation and analysis of geographically referenced data.' Other terms that are used for GIS are geographic information systems, spatial information systems, desktop mapping, etc.

At the heart of any GIS is a database, containing objects with attribute data. It is true that in the past, GIS has been 'sold' with a more pictorial view, putting the map at its heart. The more recent trend is to enphasise the data instead, and say that any map is a spatial representation of what's in the database. This is a paradigm shift in thinking about GIS, which its practitioners are encouraging. The presentation of geospatial information doesn't even have to be map-based, anyway. If you want to know which stores are nearest, search of the database may be spatially calculated, but information delivery could be a list of names and addresses.

Spatial information retrieval is important in any system which must reflect the fact that Aberdeen is nowhere near Aberystwyth – except in alphabetical terms.

What on earth…

Geospatial data is spatial data defined in relation to the surface of the earth. This relies on the obscure subject called geodesy, from which models of the earth are derived.

The system of latitude and longitude is based on the earth modeled as a sphere; a more sophisticated model recognises the earth's shape as an ellipsoid; on the basis of this we have a co-ordinate reference system called WGS84 (World Geodetic Survey 84). This is used by the Global Positioning System of satellite platforms, and therefore by any SatNav system you might care to buy.

Dan confessed that the GIS community is still a Flat Earth Society in many respects. The national mapping agency in Great Britain (but not Northern Ireland) is the Ordnance Survey, which uses a two-dimensional representation called the National Grid – based on Eastings and Northings. Any altitude data is quite secondary, regarded as not fundamental to the co-ordinate system.

Spatial relationships are based around a concept of neighbourhood and this relates to two 'laws' of geography:

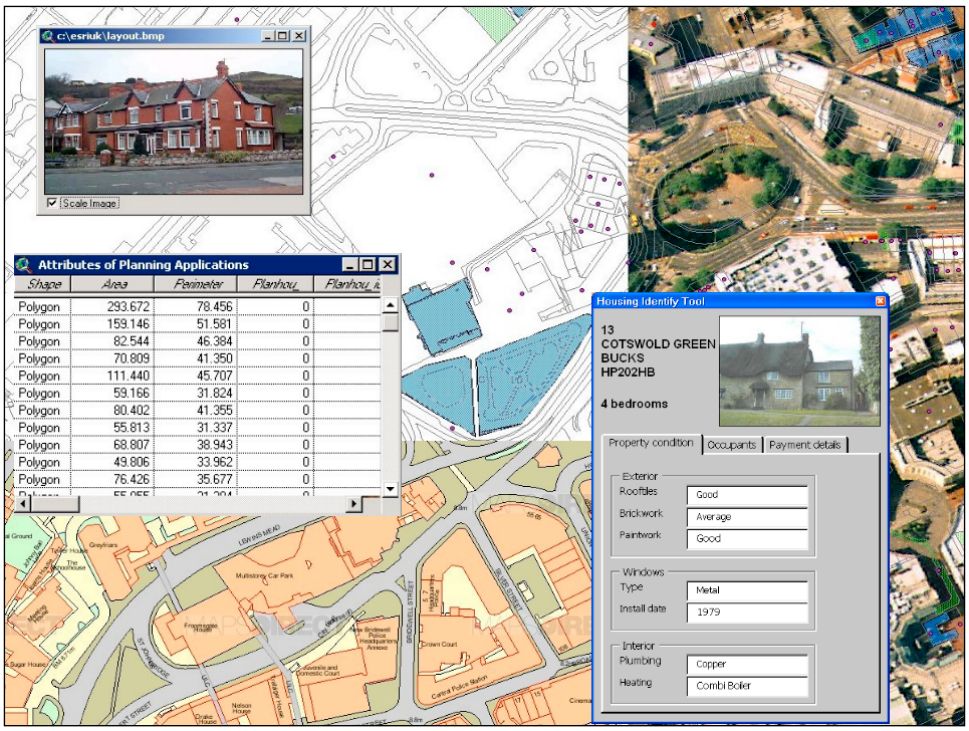
- 'most things influence most other things in some way' (this raised a laugh)
- 'nearby things are usually more similar than things that are far apart'.

The quality and richness of geospatial data has been improving over the years. Formerly, a simple join-the-dots model was used – Dan referred to it as 'spaghetti data'. These days we use concepts of topology, structure, objects etc. GIS requires metadata, as Dan supposed all data systems do: at its simplest, GIS data requires a co-ordinate system, without which the co-ordinates would be as useless as a financial database that didn't define its currency.

Geospatial information becomes more complex if we must also accommodate temporal modelling. Consider a system handling subsidy payments for farmers, such as the former Rural Payments Agency, handling the payments for 'set-aside'. But over time, have boundaries changed for the area set aside?

There are different ways of modelling geospatial information. For some kinds of information you can use an object

basis – a building, a reservoir. Other phenomena are field-based, such as rainfall – this is continuously variable across the landscape. However, as far as Dan is aware, almost all GIS applications are based on an object model: if you want to store rainfall data, it would have to be as an averaged value for each object, e.g. for the roof of your building.



Map components © Crown copyright Ordnance Survey. All rights reserved.
Fig 2 – a montage of various types of geospatial data and imagery.

Use cases

Who uses geospatial data? Users include central and national government agencies, public utilities, insurance companies (consider flooding! very topical in 2007) – and public health (especially the field of epidemiology). Travel is a very large and obvious field of application: SatNav of course, also multimodal route planning and optimisation of delivery routes.

Addresses are interesting. Although where people live or work is a spatial issue, many systems deal with address data in a stubbornly alphanumeric fashion. But mail-order companies and utilities don't just have customers with addresses; they also have routes, networks & infrastructure for delivery (vehicles, pipes, cables). If you deal with that information spatially, linking your CRM liaison servies to the service delivery section and to engineering, you will get a better management overview of how your business works.

How do we take a territory and split it up for spatial data analysis, such as demographic and epidemiological statistics? The question is, which boundaries we use. Some boundaries are physical, some are administrative, and many of the latter are quite arbitrary. In discussions of ontologies for geospatial data, we often distinguish more physically-grounded boundaries from *fiat* boundaries which are social constructs.

Historically, much geospatial data has been sidelined in proprietary formats, sitting outside the database, not readily accessible by other systems. There is now a shift towards moving spatial data into mainstream database technology, examples of which include Oracle Spatial, or Oracle Locator. These allow you to model spatial data as a complex data type inside your database. There are also Open Source solutions based on MySQL or PostgreSQL – an interesting development.

Much of the time, we are interested in the connections in information; in the case of geospatial data, this includes networks and topology. Showing two examples – a SatNav unit in a car, and a live train map for Cambridge based on Google Maps, Dan commented on the large amount of information that has to be maintained in such systems, for example about which routes are one-way, which turns are restricted.

Three-dimensional geospatial applications include those for calculating line-of-sight, radio propagation, prediction of flooding, and water pressure calculation. These days it is also possible to design 3D structures in CAD or architectural applications and load them into Google Earth; Yahoo! and Microsoft have also made big investments in this direction. You can place your building design against satellite imagery for a preview of your design *in situ*. There was recently a bit of a fuss about Google Streets, in which 360° images were taken of streets buy driving down them with cameras. Not everyone whose image was captured wanted to have their presence and activities broadcast across the Internet!

Temporal analysis of geospatial data is another specialised area, in which changes are plotted over time, such as extent of urbanisation, changes in population density, or land use.

Data types in GIS

Dan showed three kinds of GIS representation applied to an urban area: old-style Ordnance Survey Landline 'join the dots' mapping; the object-oriented approach, in which each building is a database object with a 'TOID' (Topological Identifier); and aerial photography, which is rich in features but hard to interpret. This screen image (see Fig 2 above) also included a window of Planning Application attribute data, which relates well to the object-oriented GIS approach.

GIS databases now support more structure than in the past. Ordnance Survey, for example, can provide you their MasterMap product – a structured map-base onto which you can build your own data. But how do you do that? And what are you actually doing in practice? To some extent it boils down to creating a set of 'foreign keys' for your data. This brings up issues of referential integrity.

There are also issues of data management, and lifecycles. When Ordnance Survey update their data, is that change relevant to your organisation, and how you use geospatial information? The transformation may be complex, because the Ordnance Survey quite naturally looks at things from a national mapping agency point of view, not with respect to your business reason for using information.

There is an initiative called the Digital National Framework which encourages organisations to structure data with reference to objects, rather than to recapture and duplicate data, though it must be said that progress with the DNF is slow. The principles behind it are:

- Capture information once, use it many times (taking benefit from third-party maintenance of data, and data management based on Change Of Use). This can then be exploited in updating the non-spatial business object data.
- Capture the highest resolution possible, avoiding the need for later re-capture, and improving the potential for data interoperability.
- Where possible, use existing proven standards.

Recently there has been a 'quest' to define the BLPU – the *Basic Land and Property Unit*. The problem is, the entities one wants to track can be complex. On the Ordnance Survey Master Map, 'St Mary's football stadium, Southampton' is a single unit. The nearby railway station is recorded as multiple objects – main building, several platforms, pedestrian bridge… As for hospitals, there are a number of buildings, and ownership may be complex. Clinics off-site are organisationally part of the hospital but not contiguous with it. How these relationships are described can also change according to requirements for data use.

What is the role of raster map data in geospatial information systems? To make sense of scanned images of maps in GIS, metadata such as the co-ordinate system and projection system used is needed. There is a trend to placing scans in databases as BLOBs (Binary Large Objects), together with such media-oriented metadata as number of bytes per pixel, what file type, what compression algorithm etc.

The benefit of holding raster data is necessarily limited, because the 'intelligence', the knowledge that is stored in a raster map requires human interpretation. In a way this is analogous to the problem of detecting semantic meaning in unstructured text, but worse. There is limited progress in map-based pattern recognition systems, though there are some semi-automated solutions from companies such as LaserScan.

An interesting field of application for geospatial information is in geospatial data mining, which is linked to the issue of data visualisation. (A classic early-Victorian example is the mapping work whereby Dr John Snow tracked down the cause of a cholera outbreak in Soho, by spotting the location of affected household in relation to the Broad Street water-pump, which had become contaminated by effluent.)

The Geospatial SG believes that geospatiality is a field whose time has come. There is greatly increased public awareness, thanks to Google Maps and SatNav. This can only increase in the future as more phones will incorporate GPS, driving a host of location-based services.

We are seeing greater use of mash-ups involving GIS, the classic original being the overlay of property availability over Google Maps. Dan showed a 2005 example in which D*f*ES achievement and attainment tables for Key Stage 2 results from schools was mashed-up against Google maps.

Through Freedom of Information as a driver, there has been a great push in the direction of electronic document and records management, and this can contain potentially significant geospatial data components. One of the enduring problems in this field is matching up address data (which is constructed with minimal spatial intelligence) with accurate co-ordinate-based location data. This usually requires the construction of electronic gazetteers, which map postcodes or street names or named locations to spatial co-ordinates. (The list of street names at the back of an *AtoZ* map is a form of gazetteer familiar to many, though it maps to book pages and a grid to help users find streets on the map.)

What does it cost to engage with geospatial data? These days, surprisingly little, thanks to Web-based resources like Google Maps, adequate of network bandwidth, and service-oriented architectures. However, the cost and complexity of data can be an issue, and it's hard to explain business benefits of imaginative data management to senior managers: data, it seems, is doomed to be regarded as 'boring'.

Data availability and data quality are bigger issues in geospatial information systems than is usually understood, and this is an area in which the Geospatial SG could work to awareness (collaborating with others in KIDMM). Similarly, there is a rather weak presence within GIS of standards for metadata and cataloguing; Dan feared this could become a brake on the market if it isn't addressed.

Discussion

Sabine McNeill asked Dan if he was aware of any mash-ups between geospatial and climate data. Conrad noted the widespread use of data, geospatially referenced and often acquired by remote observation systems such as satellites, which feeds into climate prediction models such as those used by the UK Meteorological Office at its Hadley Research Centre on Climate Change. It's also interesting to note the collaboration between the Met Office and a Japanese climate research centre, in which the climate model is being run on historical meteorological data, as a way of testing the model. Ian Herbert added that in meteorology, the models are not simply two-dimensional but also take altitude into account. (Indeed the depth, temperature, direction, speed and salinity of ocean currents are also relevant here.)

David Pullinger asked if lack of standardised geography is holding back the use of GIS; if so, what standard would Dan opt for? In clarification he said was thinking of postcodes or Office of National Statistics (ONS) output areas, for example. Dan agreed that the postcode is a powerful concept, but the area definition is basically a set of delivery points. Any geometry put around those points will be arbitrary.

Also, the postcode system was devised for delivering the mail, but many organisations (including the ONS) use it to split the country into areas. Indeed the phrase 'postcode lottery' reflects the way that the postcode gets treated as a *de facto* geographical unit.

In drawing the session to a close, Conrad recalled that a few weeks before, he and Aida Slavic had been talking about the problems with these entities called *countries*. In classifying geographical locations, one imagines a monohierarchy, in which a town is situated within a country. But over time, boundaries change. Conrad has a Jewish friend whose knew her grandmother was born in Romania. In trying to trace her family history, she approached Romanian authorities with no success. In fact grandmother was born in or near Chernitvtsi – the city was in Romania from the break-up of the Austro-Hungarian Empire until 1940, but since then it has been part of Ukraine. This problem comes up repeatedly when you try to attach geographical and historical data together; it's as if you need a 'spatio-temporal gazetteer' to track such changes.

Integrating museum systems at the Victoria & Albert Museum

In introducing this topic, Conrad reminded the conference of a challenge often faced in data management when databases, independently developed, must be integrated together. The problems are not just technical; the databases which are to be integrated reflect a particular way of looking at their subjects, and organisational politics may be involved. The Core Systems Integration Project (CSIP) at the V&A can illustrate what may be involved, and is worth studying.

Christopher Marsden is an archivist and records manager at the V&A. He spoke in place of Sarah Winmill, the Museum's head of IT, who was unable to be present for family reasons. Christopher pointed out the V&A's technical partners – System Simulation Ltd (SSL) was represented by George Mallen and Mike Stapleton, who later contributed technical details.

Christopher offered to describe CSIP; to talk about the challenges which it has presented; to explore the technical and content-related issues that have arisen from taking information about the V&A's collections from a variety of data sources and delivering it via a single coherent interface, for the benefit both of staff and the general public; and to offer some lessons, from which others undertaking similar initiatives might benefit.

The challenge

The Victorian and Albert Museum contains about 1.5 million objects. These are inventoried and in some cases catalogued electronically in the V&A's Collections Information System, based on SSL's MUSIMS software [10]; this in turn conforms to the MDA's SPECTRUM standard. In addition, there is a large amount of paper-based documentation about objects in the collections.

There are also about 1.5 million items in the National Art Library, plus vast archives. The library database is a standard library system supplied by Horizon. There is no database for archives; these are catalogued in sets of XML files, following the Encoded Archival Description (EAD) standard [11].

The Museum also holds about 160,000 images of objects in the collections, in both analogue and digital form. These are documented in a Digital Asset Management database, and images are available for purchase online.

The information systems for the V&A's library, archives and collections are all large, well established, and based on different recognised standards. One might wonder, given that a library catalogue is not so different from an archive catalogue or a collections records database, why someone has not hitherto come along and knocked the heads of some librarians and curators together and said, 'Why don't you just have one standard for the lot?' But Christopher felt that the development of an overarching standard was not likely to happen in the near future: the separate standards are of long standing, they have been carefully developed, and there are justifiable reasons for the differences. But it creates problems in getting such different systems to work together.

At the V&A there have also been problems with 'copies' and 'harvests' of data. They have sprung up when people have been working on particular projects, such as a scholarly study or an exhibition for a gallery. Sometimes people have found that the main systems have not worked quite the way they wanted, so individuals or project groups have set up some mini-application or database of their own, to do their own thing with that data. Thus there has been a serious problem of people sucking out data and manipulating it in their own systems, until it no longer matches the core data.

Historically, some quite complex links have been set up between systems, but those inter-dependencies between systems have not been properly documented.

Aspirations and the status quo

In terms of use of information, the Museum staff have begun to have larger aspirations about ways in which data owned by the V&A could be used; and it may be regarded almost as a scandal that despite the huge amounts of data the Museum has, the public cannot get to quite a lot of it.

Christopher showed a diagram representing some of the data systems in the V&A. There are data systems which for our current purposes are irrelevant – finance, personnel and so on. Others are certainly relevant: the photo catalogue, which is now a proper digital asset management system; the Collections Information System; the library system; the archives; and there is also a Content Management System which lies behind the V&A web site and feeds information though to it. But it is all rather fragmented.

From this starting point, the aims and objectives of the Core Systems Integration Project were:

- to develop a system architecture whereby various applications could access information about objects in the collection through a Virtual Repository, rather than remastering object data locally; and

- to integrate the Museum's core systems, and remove the dependencies on manual data manipulation inherent in existing practice, thus improving the efficiency and accuracy of data delivery.

Hopes and deliverables

Near the start of CSIP, a list of the project's deliverables was drawn up. To encourage 'buy-in' from Museum staff, at the top of the list was something of evident value – a Gallery Services application, to help staff give customers what they wanted to know at the point of enquiry: information about collections, where they might see a particular object, where it is stored, what else in the collection relates to it &c.

However, logically speaking, before one can implement these applications, there is a need for other fundamental deliverables, in particular the 'Virtual Repository' (VR). This in effect is a central data switch through which the core data systems can relate to one another, serving as foundation and data source for any applications built on top of it.

An early ambition was to be able to link images to the National Art Library (NAL) catalogue. Unlike the collections database, the library catalogue software couldn't talk to the digital asset management system. The library is not just a collection of reference books held as text; there is a lot of artefactual material there – rare books, manuscripts, book art, a lot of early material relating to the history of the book. These are as much museum objects as a piece of furniture or ceramics in a gallery would be.

When the Museum works on a big gallery project, staff need to be able to manipulate data about the collections in order to think about how the gallery will be developed and the collections presented. In the past this had to have been done with separate, stand-alone databases. One of CSIP's aims was to be able to do this without re-mastering data.

They also wanted a publishing process to help deliver the data to the public; plus a Data Mastering Protocol – a policy statement that would clearly identify where data items are to be mastered, and how the various data items in the Museum should relate to one another.

Christopher next showed a diagram drawn up by SSL illustrating the system architecture of CSIP (see Fig. 3). Four boxes at the bottom of the diagram represented the sources: the Collections Information System, the Picture Library, the National Art Library and the Theatre Museum Archive. In the middle sits the Virtual Repository, and at the top were shown two sample applications – Gallery Services, which serves up information to the front desk, and a system which gives access to images from the National Art Library system via an Online Public-Access Catalogue (OPAC).

Potentially, many other services can be envisaged. For example, a Loans module is desirable, because whereas the Loans Service has good access to CIS data about objects in the collection, at the moment they cannot access Art Library or Archive data.

Although Christopher did not address these technicalities, the diagram shows that the gathering of information from the databases is achieved through a Z39.50 standardised access mechanism. EAD-encoded archive material is gathered through a OAI-PMH process (Open Archives Initiative's Protocol for Metadata Harvesting). Exchange of information between the Virtual Repository and end-user applications is done by sending structured requests and messages back and forth, following the SOAP protocol (Service-Oriented Application Protocol).

To illustrate the Gallery Services Application, Christopher showed a photo of staff at the museum's Information Desk, using both screen-based reference and paper reference to answer the queries of visitors. At the V&A they talk about 'the William Morris question' … How do you deal with the visitor who comes to the desk and says, 'I'm interested in William Morris – what have you got?' Given the breadth of Morris's involvements in arts and crafts and design and printing, there is no one place where you can find all that information; and it is difficult to reply 'We have this and this; it's in this place; if you go to the gallery now you'll see them; and by the way we also have some books printed by Morris, and some archive material from Morris & Co., and this is how you'll find it.' Christopher commented, 'You probably think, why is it difficult for us to answer questions

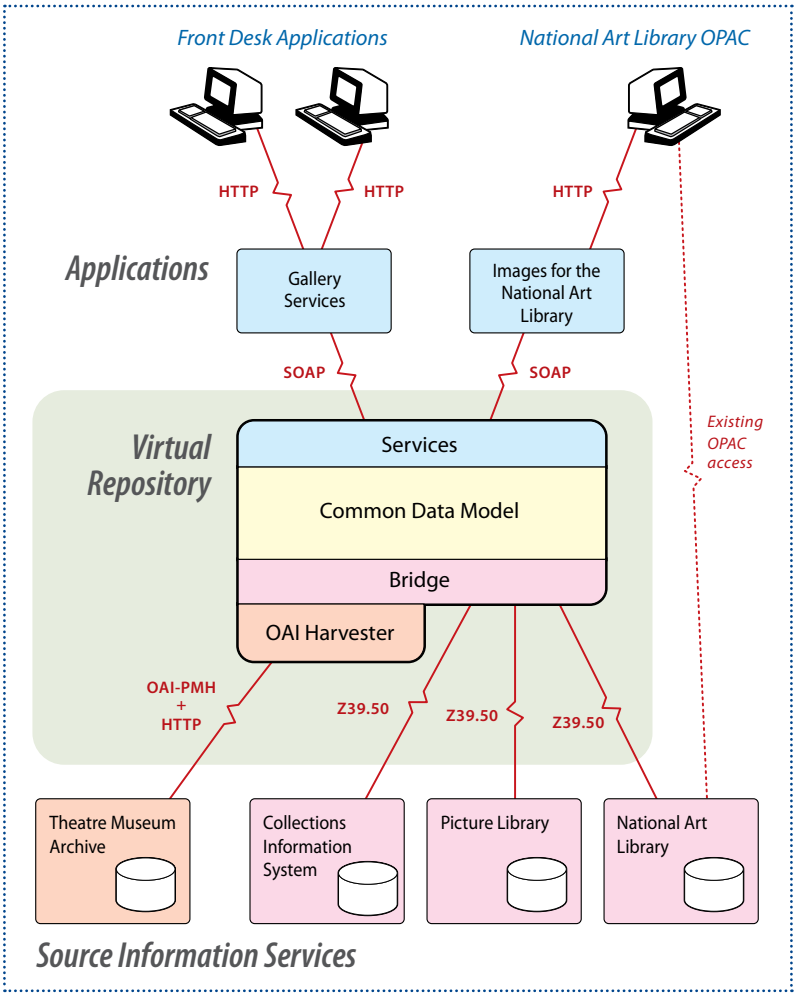


Fig 3 – The architecture of the V&A's Core System Integration Project.

like this; haven't we thought about it before now? Well, we have, but we just hadn't quite got there.'

The key requirements for the Gallery Services application are that the Museum should be able to provide access to information about objects in the collection, and say where they are currently located. Details about the object should also be reachable from 'surrounding information', such as who was the artist or maker, who is depicted in the object (e.g. a commemorative plate with an image of Lord Nelson), the period, the culture, the date and place of creation, &c. The Gallery Services application should provide quick access, and full results which are easy to understand – ideally, with illustrative images.

The Stuff in the sources

Christopher next showed some samples of the data as it is currently kept, and how it might be presented to an enquirer. First, he showed a bibliographic record from the National Art Library. The format for bibliographic data storage which the V&A uses is MARC, and this slide showed an entry as it would be seen by a cataloguer, and how the information gets presented to the public as the result of an OPAC query.

Next Christopher showed views of an object record. Again, one view was of how such a record appears in the CIS, and the other showed how a user of the 'Access to Images' Web site would see it. There is currently public access to information about some 20,000 objects through a facility called 'Search the Collections'.

The third example was of archival records, and we saw a page of an archive catalogue as it would be seen through public access, together with the EAD XML markup behind the scenes of that system.

When the V&A staff started to work on this project, they faced a number of issues. One was the hierarchical nature of information in archives and collections information. A chest of drawers for example is an object, but also has components e.g. the individual drawers. Consider conservation work: if a drawer is removed and sent for repair, the Museum needs to be able to track where it is, so each drawer has its own record and photographs. However, you don't want to bother the enquiring public with these records of components. Likewise in a library, you may have a description of a series of books as well as the books individually.

'With archives it is the same problem, but writ large', said Christopher. Consider the V&A's archives for the Habitat company. The individual document is catalogued, also the file it comes from, the series from which that file comes, and the department of Habitat it came from. Then you need records which deal with Habitat as an entity – when the firm started, how it kept records, and how those records were organised. Archival catalogues typically have layers of description, making it complicated to work out what to serve up to an enquirer.

Different museum standards provide for different levels of granularity. Take people's names: in the archive standard, a name is a single data field. In the library system, it is more broken down; and in the museums standard (SPECTRUM) it is broken into lots of bits: forename and surname, title and pre-title and post-title; dates-of, occupation and much more. Getting such differently-constructed record sets to work together and look the same is hard.

Mapping model

At first, the team thought that things could be kept simple by using the Dublin Core model, which provides fields such as Description, Subject, Identifier, Coverage etc. They tried to sketch a draft mapping of CIS data to DCMI; but was soon found to be 'too blunt an instrument'. The DCMI Subject field mapped to ten distinct library fields, and half a dozen collections information fields, and losing those distinctions would be anathema.

The team also studied the Conceptual Reference Model of CIDOC, *Le comité international pour la documentation des musées* – International Committee for Museum Documentation. The CIDOC-CRM is described as 'a formal ontology… to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information' [12]. CIDOC-CRM provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation.

The V&A team went to a seminar about CIDOC-CRM and got very excited about it: it seems to offer logical perfection. It is not impossible to use this model in practice; and the V&A team did collaborate with a project called SCULPTEUR (Semantic and content-based multimedia exploitation for European benefit), whose Web site shows CIDOC-CRM in action, indeed with a demonstrator based on objects from the V&A collections [13]. But when the team's excitement had subsided, they found that, in Christopher's words, 'It was beyond our feeble mental capacities to apply it to the whole of our collections and bibliographic and archival data.'

They settled for a compromise: an expansion of Dublin Core, adding a column with fields that they needed, mapped to the CIS, library and archival system data. And thus they derived their own Common Data Model for mapping source material into the Virtual Repository.

But surely there are thousands of museums all over the world facing similar problems; wasn't the V&A re-inventing the wheel? Apparently a few *have* made the attempt – there has been similar work done across the road at the Science Museum, on a small scale – but at the moment there seems to be no model to be followed. This is why the CSIP team at the V&A are keen to go out and talk to people about they work that they have done, so others can learn from the difficulties and successes they have encountered.

Where has CSIP got to?

The Virtual Repository is now in place – and it works. Not quite perfectly, but it works. There is a prototype Gallery Services application, serving up information and pictures in a way that could be used by Gallery staff and by the public coming in, but it isn't installed at the Information Desk yet. Progress has slowed recently, for a mixture of personnel and political reasons, but Christopher declared optimism that they will be able to push ahead and deliver applications.

Sarah Winmill, Head of IT at the V&A, has drafted a list of 'lessons learned so far', and Christopher undertook to read them out and comment on them.

- ***It is possible to integrate your data without putting it all in one place*** – this is evidently true. However, the V&A team found that putting together the Common Data Model on their own was time-consuming, and stressful, as each department has its own area to defend – librarians who have worked to produce the perfect library catalogue don't want their data to be wrecked by being stripped of some of its sophistication for other purposes. Curators and archivists feel equally protective towards their own data and systems.
- ***High-level buy-in for the project is essential*** – this may hint at why the team has not been as able to push CSIP as far forward and as fast as they would have liked.
- ***Market your project carefully; talk about benefits and deliverables, not technology*** – in an organisation like the V&A, there is often cynicism about projects of this sort unless people know what the end product will be and whether they are going to be able to use it.
- ***The major challenge is no longer the technology, but the underlying understanding of our data*** – that's easy for the V&A to say! remarked Christopher, because they have very good suppliers: museum staff can talk to SSL and they go away and do all sorts of amazing things.
- ***Don't wait for perfection!*** – True enough; perfection indeed may never arrive.

Discussion

Someone asked, since a museum is essentially about organising artefacts in different categories in a particular space, had the museum given any thought to spatial information? Only, said Christopher, to the extent that the catalogue includes a reference to room and case location. The museum is very far from being able to SatNav visitors around the galleries – nice as that would be in such a large and labyrinthine building. But such issues are beginning to be raised.

Dan Rickman asked for elaboration on the 'ownership' of data in the source databases and how the Common Data Model is used in practice in building the Virtual Repository. Mike Stapleton of SSL explained that the VR has a brokerage module. It collects information from the databases, and runs queries as required against them; and does both these things in response to queries from applications. The EAD archival data is harvested and held locally to the VR in a structured database. As for the Common Data Model, its main benefit is to enable conversations between Museum colleagues and suppliers, and clarity is its main benefit.

Conrad asked George Mallen of SSL, considering that his company has worked with many museums and galleries over the years, so has been in a position to learn something with each new project – what has SSL learned from this? George explained that in such projects, an important starting point is the knowledge organisation system (KOS) of each institution. Each has its own attitudes to its data and knowledge.

Another influence is the various standards in the field, such as CIDOC and SPECTRUM. Technology suppliers, such as SSL and its competitors, must interact with all such factors, and tailor solutions to each client's needs.

It's important to keep abreast of developments in the field, and here SSL's answer may be unusual, in that the company gets involved in European-funded research initiatives – they were not involved in SCULPTEUR, but have been involved with its successor. This is a good way to keep a company and its technology responsive. And George reflected on what Ian Herbert had said the National Health Service – perhaps the culture & heritage sector is more adaptable, and the amount of invention that is happening in the field lends resilience to cope with new developments as they come along.

Conrad referred to an event which the BCS Electronic Publishing SG ran some years ago with the MDA, at which there had been a presentation about the issue of conducting searches about information and objects in the culture and heritage domain, across the databases of several museums or galleries simultaneously. The Z39.50 standard was at the heart of that – could George say a bit more about its role? Essentially, explained George, it's a protocol for distributed access. It translates the query against each database into a form that makes sense locally.

Aida Slavic raised the issue of metadata 'crosswalks' – there are so many of them, between EAD and SPECTRUM; recently a schema came out, which is a UKOLN project. How useful are any of these things, like Dublin Core? Might there be a point to some form of metadata registry, so for example there might be a common solution for names? Ian Herbert picked up on this, saying he was not surprised that the V&A had had problems with Dublin Core, because it is so general; and if you have different forms of information architecture, they require different metadata sets. There may be a common core, but there is no way that one size fits all.

Christopher described how they might have got well on with their Common Data Model, then suddenly someone would think of some new applications the Museum might be using in the future – for example, a conservation module. But really the Model, which had been put together with some aspirations of simplicity, didn't have the level of detail to support that.

Susan Payne of De Montfort University Library said their current experience in data modelling is that people are scared about missing something out. If something's later discovered to be missing, how does it gets built in? How is agreement reached? Mike Stapleton said there comes a point when you have to draw a line and stop modelling – and yes, it is a problem. If later, while developing applications, something is found missing from the data model, it can be expensive to fix at that stage.

Susan then asked, if a change needs to be made to an existing categorisation scheme, what sort of turnaround time might one expect? All Christopher would comment is that things don't tend to happen fast at the V&A – they happen well, but not fast.

Preservation of Datasets at The National Archive

Terence Freedman was our speaker on this subject. Terry works for NDAD, the National Digital Archive of Datasets at The National Archives, and is actively involved with many BCS Specialist Groups, and with KIDMM.

The National Archives (TNA) exists for historic reasons. There was long an issue of who was responsible for looking after government records, and this became the job of the Master of the Rolls. In 1838 the Public Record Office was established for this function; and the Master of the Rolls retained responsibility for the PRO until the Public Records Act 1958 transferred it to the Lord Chancellor's Department.

In 2003 the PRO merged with the Historic Manuscripts Commission and the Family Records Centre. Together, these became The National Archives. The Office of Public Sector Information, responsible for Her Majesty's Stationery Office (HMSO) among other things, merged with TNA in 2006. Though HMSO itself has not come into TNA, and remains separate as the retail side of government publishing, it does influence the way in which TNA uses the data it has, because HMSO has an interest in public data re-use – they make it available to people by selling it. The National Archives, on the other hand, does not charge – all its data is free.

The information TNA looks after goes way back – to the Domesday Book! The Public Records Act 1958 stipulated that government data should be released after 30 years (the '30 year rule'). Later the threshold dropped to seven years; and thanks to the Freedom of Information Act, a lot of government data can now be accessed immediately. In theory. But in practice, some information is 'redacted' – edited out of the record, through a decision of the Lord Chancellor's Advisory Committee. Thus there are blanks in some released documents.

Who owns this data? Much government information has a geospatial component, for example about North Sea oil deposits, or tubercular badgers. The Ordnance Survey are happy to let people have sight of their information, and use it, but are also concerned that people might steal their data by printing out or copying maps. Some government departments have an agreement with the Ordnance Survey, but TNA does not; because of variations in mapping technique, this issue makes preservation more difficult. Indeed, Terry has brought some of his TNA colleagues to meetings of the BCS Geospatial SG to try to explore some of these issues.

From chaos and dependency to order and freedom

The National Archives has responsibility for documents, records and datasets. As for documents, they come from the Electronic Document and Records Management Systems of government departments; they have provenance, and various sorts of metadata attached – certainly sufficient attached data to make cataloguing them fairly straightforward. But the National Digital Archive of Datasets mostly deals with tabular data – and this is much more problematic.

NDAD acquires 'born digital' data, which can be up to 30 years old. That might mean punched cards, paper tape or magnetic tape. The threat of loss of data through decay is quite pressing in the case of magnetic records. Surrounding information is often missing too – what methodologies were used for collecting that data; why it was collected; who collected it; what format it was contained in and how it was encoded.

The aim is to improve access to this data by indexing it. This does work! Type 'NDAD' into Google and you will be into TNA's data in no time at all, promised Terry: type in 'NDAD badgers' or 'NDAD trees' and you will be into that particular selection of datasets. (Though Richard Millwood later noted that when he tried the latter, Google asked if he had meant 'dead trees' instead!)

To make such access possible, it is vital to move rescued data in its typical chaotic state through common processes, resulting in a standard encoding – for example, tabular data is converted to CSV files (comma-separated values) – and relationships are created between tables, such that they are maintainable from that point on. That done, the original software can be forgotten; it is no longer relevant once data has been freed from those dependencies.

Nothing in the record must change – it is inviolable. Even if the original data was wrong, the record must stand. Other agencies, such as the Department for Work and Pensions, or the Office for National Statistics, do correct this data, resulting in highly useful results. But NDAD's role different, and preservation of data is key to that role.

'Born digital' government data started to be generated from about 1967. In 1996, the Conservation Department of the PRO took responsibility for conserving magnetic media. The consulting firm of Cornwell were called in; they recommended setting up a computer-readable data archive. In 1998 the University of London Computer Centre was awarded a contract to do this, and 'NDAD' was named that same year. In 2001 the Digital Preservation Department (DPD) was set up inside The National Archives, and the contract has been managed by DPD ever since.

Recovering information from government departments can be difficult, and involves liaison with Digital Records Officers (DROs). NDAD staff acquire the datasets, together with surrounding information which might be called 'metadata'; check them for completeness and consistency; and arrange for secure transport to the University of London Computer Centre for processing. Note that NDAD doesn't take anything and everything on offer; on average, about 5% of the material that a department holds is worth preserving.

It's not necessarily the case that data recovered from a particular department is owned by that department. It may be owned by another agency such as Ordnance Survey, and trying to sort out ownership and copyright is an ongoing issue. Sometimes records have to be redacted simply while rights to make all data public are being ironed out – a map, for example, might have to be omitted.

Provenance is not usually too difficult to determine, since NDAD staff were involved in retrieving the datasets from the

government department; it gets checked, and the originating department is asked to agree the provenance record, though changes of staff can make this tricky at times.

NDAD maintains an audit trail for everything they do: the processing of the dataset, whether all the retrieved elements form a 'contractual entity', whether or not the material can be turned into something meaningful, whether there is permission to go forward, whether the Lord Chancellor's Advisory Committee has said 'stop!', whether any redaction is necessary, whether the dataset is being chased by a Freedom of Information Act request… all sorts of things need to be tracked. Incidentally, so far no government department has asked for special privileges to see an 'open' (unredacted) version of a redacted dataset (such permission would go only to an originating department or its successor).

Early in the process, materials are copied bit-for-bit from the original media onto current media; then the original data carrier is disposed of securely. A further bit-copy is made, and conversion to current encodings and standard formats proceeds from there. 'Fixity checking' is performed at frequent intervals to ensure that data has been copied faithfully, and multiple redundant back-ups are maintained via remote storage.

When NDAD acquires digital data created with obsolete software, certain assumptions are made at the start, such as ASCII encoding. There may be the problem that the record has been 'packed' in some way: it might be six-bit or eight-bit, for example. Then there is a process of testing against the bit-copy until the record is 'cracked' and becomes legible.

Adrian Walmsley (BCS Engineering & Technology Forum) asked if therefore it was part of NDAD's responsibility to maintain access to old software tools for this purpose, and perhaps surprisingly Terry stated that it was not. As the 30-year rule is now effectively dead, these days almost all the datasets being preserved by NDAD are about seven years old, and it isn't too difficult to find software for that far back.

If the originating department can also supply the software (they own the license, after all), so much the better; it will be borrowed, used and returned. And as for the future of the data, the whole point of the conversion is to shift it into a canonical representation that will free it from dependency on particular software.

Security comes in three forms. First, all the buildings involved in storage must be made secure against e.g. fire or flood. Then there is access security: all TNA staff are cleared to SC level, as are the staff of all contractors; even the cleaning staff. There are plenty of firewalls and other measures to ensure that data access is read-only: nobody is able to tunnel in and change the data. There is a system of local and remote access permissions, and of course read-write access is particularly closely controlled.

Access to NDAD data is free. On average there are 55,000 Web page accesses each month; there are 116 million records currently, spread across 400 datasets. Access to data has remained excellent – only a few hours outage per year in excess of planned down-time for maintenance.

Discussion

Aida Slavic said that her area of expertise is more to do with discovery metadata, but she was curious to know what sort of preservation metadata was in use by NDAD, for example to record the data formats, the software in which data had been originally created, etc.

Terry replied that yes, technical metadata about how the information was originally recorded is collected, plus contextual information about what was going on in the world at the time and why the information was collected. To this will be added a record of how it was transferred, and into what structure; plus a contractual entity reference code, an NDAD reference code, and TNA's own reference. All of this data is connected to TNA's catalogue at the highest level.

Someone asked about back-up storage; traditionally this was tape, now is it more likely to be a copy made to another disk? Terry said that currently back-up is done to specially formulated HP terabyte tapes; server-to-server backup has been proposed; but even then, the preference would be for a further back-up to tape.

Conrad Taylor raised the issue of the digital preservation of documents, rather than datasets, lest we forget them; and related a concern that has been raised by Adam Farqhuar at the British Library. A huge amount of electronic documents which they receive are in Microsoft Office formats, such as Word and Powerpoint. For this reason, Adam is enthusiastic about Microsoft's Office Open XML format, which should render the closed proprietary binary formats into a publicly-documented, XML-rendered encoded text form.

Ecma International has adopted Office Open XML as one of their standards (Ecma-376), but the fast-track ISO process which they then championed to make it an ISO/IEC standard (DIS 29500) has proved controversial, because Office Open XML is seen by many as a 'spoiler' launched by Microsoft against the ISO/IEC 26300:006 OpenDocument standard, derived from the OpenOffice.org XML format

(John Alexander commented that the DIS 29500 proposal had been voted down in a ballot that ended on 2 September 2007 [14].)

Terry said that TNA has an agreement with Microsoft guaranteeing access to all previous Microsoft operating systems and applications. This is a partial solution, though only for TNA, and only for documents authored in Microsoft's products – where that would leave WordPerfect or AmiPro documents is another matter.

Conrad wondered similarly about long-term accessibility of Adobe's PDF – Portable Document Format – and reported that Adobe Systems is developing an XML representation of the content and structure of PDF documents, under the title of the 'Mars Project' [15].

Issues in Classification

Leonard Will (Willpower Information) rose to the task laid on him in the morning by Conrad Taylor, to sketch out what are the main issues in building a classification system.

Leonard, formerly the Head of Library and Information Services at the Science Museum, London, is now a consultant in information management, with an interest in libraries, archives and museums; he's also a member of a BSI working party on structured vocabularies for information retrieval (BS 8723). That is worth studying; as is a talk which Leonard gave at the EPSG/KIDMM panel discussion about classification, the slides and MP3 recording of which are available [16]

Leonard told us that, for starters, we must distinguish between making *descriptions* of documents and providing *access points* to them. When you retrieve a list of documents, you should be presented with information so you can assess which documents are relevant. We should also distinguish between (a) the structured form of data that you need for consistent indexing and (b) information presented about the document once you have found it – which can be free text.

Controlled vocabularies are used to bring consistency in indexing, so there is a better chance of matching terms used by a searcher with terms applied by an indexer.

Concepts, labels and scope notes

Tony Rose had talked earlier in the day about concepts for subject access. If we are not doing automatic indexing, but human or intellectual indexing, the first step should be a subject analysis – to decide what the document or group of documents is about. Having identified the concepts represented in documents, you translate them into the controlled vocabulary for labelling those documents.

In a controlled vocabulary, we define concepts, and label them with a preferred term. We may also give them other, non-preferred terms. The preferred term is used as an indexing label, and is to some extent arbitrary, although it makes sense to use a term people will actually look for. However, it should be equally possible to achieve access using any non-preferred term.

We normally label concepts by plural nouns, not by adjectives. This is where museum people tend to get upset: they tend to think we're labelling 'this chair', but in indexing terms we are labelling a class or category of 'chairs' – and the particular item in the collection is a member of this class.

For disambiguation, the standard practice is to follow the label with parenthetical additions so that we can distinguish bank (riverside) from bank (financial institution).

The definition of the scope of a concept – what it means – is best expressed in a *scope note*. It is the scope note rather than the label that defines the concept. It should say what's included in the concept, what's excluded, and what related concepts should be looked for under another label.

Grouping and relating concepts

Concepts can be grouped in various ways in a controlled vocabulary. In a thesaurus, they can be labelled separately, with relations between them specified. The relations in a thesaurus are *paradigmatic* – they are true in any context. The term 'computers' is related to 'magnetic tapes'. Broader and narrower terms are common forms of relation, e.g. 'computers' and 'minicomputers'.

However, it is not the job of a thesaurus to link terms like 'computers' and 'banking'. Such a relationship is not inherent in the concepts themselves. Of course there are documents about computers and banking, but the relationship between the concepts is said to be *syntagmatic* – they come together in a syntax to express the compound concept.

'In a classification scheme, we build things using concepts from different parts of the thesaurus – from different *facets*, if you like,' said Leonard. 'Though I find the term facet is often mis-used because it has become fashionable. I don't like the term *taxonomy* either… it's used by people who don't know whether they are talking about a classification scheme or a thesaurus. Best leave taxonomy to the biologists.' (He added this is not an official BSI point of view!)

A facet, in the British Standard, is what is sometimes called a *fundamental facet* – a particular category such as objects, organisms, materials, actions, places, times. These are mutually exclusive: something cannot be an action and also an object, for example.

Facets can be expressed in a thesaurus; but you can have only broader–narrower, or 'is-a' relationships between terms within the same facet. But in a classification, we can combine facets together: e.g. banking as an activity and computers as objects, combining to express a compound concept.

When combining facets to create compounds in a classification, you have to do it in a consistent way. Therefore we establish a *facet citation order*. The Classification Research Group (CRG), founded in 1952, devised such a citation order, starting with things, kinds of things, actions, agents, patients (things that are operated on) and so on, ending with place and time. It's an elaboration of the facet scheme proposed by Ranganathan, the 'father' of faceted classification – his formula for the citation order was PMEST, for Personality, Matter, Energy, Space and Time.

A classification scheme doesn't bring things together just alphabetically, so it is often necessary to have some notation, perhaps a numbering scheme such as Conrad had applied to the 'BCS Taxonomy', [b] which allows you to sort and maintain

b. The 'BCS Taxonomy' was one product of a 2002 initiative of the BCS Knowledge Services Board, which set up a working party about BCS content, chaired by Wendy Hall, to see how knowledge sources created by the Society and its SGs could be collected, organised and made available more widely and systematically. Judi Vernau of Metataxis was commissioned to manage the process of devising a classification scheme that could be used to 'tag' resources. The resulting document is not strictly a taxonomy (it is a polyhierarchy) – terms appear more than once, in various contexts. To assist identification of which instance of use is meant in a particular case, for the purpose of KIDMM discussions, Conrad Taylor devised an ad-hoc hierarchical numbering scheme; it is to this that Leonard was making reference.

things systematically. Otherwise you will get Animals at the beginning and Zoos at the end; yet because they are related, we expect to find them close together.

This brought Leonard to address the difference between a thesaurus approach and classification approach. A thesaurus approach lets you do post-coordinate searching, and to find things by particular terms. In post-coordinate searching, you combine concepts after indexing, i.e. at search time. Pre-coordination means that terms are combined *at the point of indexing* into a compound term.

The value of a classification should be that you can browse it usefully. Leonard thought it a pity that in modern online library catalogue systems, we can no longer browse as used to be easy to do in a card catalogue, finding related concepts to either side of where you started looking.

Classification schemes allow (or should allow) this kind of useful browsing, and a thesaurus is useful for searching. The two are complementary, not alternatives; you don't have one or the other, and it is best if you can have both.

All too often we see a user interface with a tiny box, into which we are asked to type search terms, without guidance about what we can do. Even our advanced interfaces often give insufficient guidance about how to compile a sensible search statement.

Whereas Google might say to us 'Did you mean…?', a thesaurus could do so much more. It could say, 'I've got this, but I index it under this term; I'll do the search on that instead,' or, 'You've searched for this, would you also be interested in searching for the following related terms?'

This kind of interaction with the user is what librarians call the *reference interview*: like when somebody asks 'Where are the chemistry books?' when what they really want is to find out if the weedkiller they put down their drain is going to do any damage. (People are generally thought to ask for topics broader than they really want to find, because they are frightened to be too specific.)

Ideally a computer search interface should be able to lead the enquirer though that kind of reference interview. It can be done, in principle; but the computerised systems we have at present don't do that.

Discussion: taxonomies and tagging

One document provided for study in the Mash-up delegate pack was the BCS Subject Taxonomy (see note *b* on previous page), commissioned at the request of the BCS Knowledge Services Board as a resource for classifying BCS information products. Conrad asked, If you had to classify a Web page or a document using this taxonomy, how would you start? Wouldn't you regard the task with some foreboding? You would have to learn the general structure before you could use it with any confidence and facility.

Carl Harris, the BCS Webmaster, in March 2006 reported that there are no *technical* problems in using these classifications to annotate the XML files in the Content Management System behind the BCS Web site; but they have found that at the point of authoring the classification, people are unsure about what category, what keywords to use. This seems to have led to some resistance to doing the work of classification. Conrad thought that it is often the case that if people are called upon to add metadata, and they cannot see the benefit of doing that work from their point of view, they are reluctant to do that work 'for the good of humanity'.

Someone asked if classification isn't always contextual. In the BCS, we classify subjects a certain way because we are all generally interested in things to do with computers. He said he had just finished reading a book by Umberto Eco, who is a semiotician. Eco describes an early Chinese attempt to classify animals: those that walk, those that fly and those that swim – also, those the Emperor likes, those the Emperor doesn't like, and those the Emperor hasn't made up his mind about yet. It seems crazy – but if your life depends upon not offending the Emperor, it is a very reasonable classification.

All classifications have a purpose, agreed Ian Herbert, and their structure is defined by their purpose. When you sub-type things, you need to declare the axis or basis on which you speciate a concept. If the basis is whether the Emperor likes it or not, then fine. But the six categories described by Eco are not mutually exclusive: an animal might be able to walk, fly *and* swim (a duck?), and the Emperor might hate the creature as well.

Someone asked to be able to introduce the concept of 'point of view'. He is working currently with the Integrated Public Sector Vocabulary, IPSV, and particularly with the Communication and Information section of it. Some of the oddities one can find within that arise from different points of view: for example the view that an indexer in the public sector would take of ICT, versus the view that professionals in ICT themselves would take.

Conrad Taylor suggested we turn the discussion upside down. In systems like Flickr and Del.icio.us, people classify things with tags that they choose – 'folksonomies'. It is less effort to get people to tag things that way; is it useful?

When Conrad and Genevieve Hibbs made a 'KIDMM visit' to the Improvement and Development Agency, they found that in the I&DeA blogs and forums people tag items 'folksonomically' (as Richard Millwood later describe). This is interesting in that organisational context, given that the existance of the Integrated Public Sector Vocabulary. It would be interesting to know how effective folksonomic tagging schemes can be, and what the relationship might be between the idea of central authority, and people just tagging things as they see fit.

John Alexander described the practice of displaying a 'tag cloud' so that you know about the categories and keywords that other people favour, and two understandings of the function of a tag cloud emerged:

- The terms in the cloud can be the tags most frequently assigned; this can become a preferred, semi-normative pick list; though one can still introduce one's own.
- The cloud might instead show the terms on which most searches had been made against the site.

Richard Millwood suggested classification will be guided by the tools on offer. 'If we gave infant children tools as good as some of those out there in the Web 2.0 context, I think

children would learn to understand what taxonomies are, practically, day by day, in an enjoyable context, and they would think very hard about how they would make sense of them communally when they come to secondary school and beyond.'

Major classification schemes

We were talking about classifying stuff as it arrives online; but Aida Slavic reminded us of the huge amount of information already there, online or in libraries.

The big classification schemes – Library of Congress, Dewey, Universal Decimal Classification – aim to cover the whole of knowledge, and to do so scientifically. The structure of these classifications serves the purpose of co-locating books to match how people are going to use them: for example, all the books about heart diseases are put together. Not to index heart attacks together with poems about the heart in love, or Braveheart the movie.

A huge amount of humanity's information is organised according to this system. Do we dare call it obsolete? It has the huge advantage of being a map of knowledge, to which we can map other things. These large classifications are not stupid – they have a good facet analysis system behind them. As Aida sees it (and this is why she helps to maintain UDC), if such a classification schemes is encoded and exposed in a machine-understandable way, then perhaps you can correct and improve on it by noting how it touches upon other vocabularies you use.

For example, in building the 'BCS Taxonomy', it would be useful to build links to those large classifications. Surely among 200 million books listed the Library of Congress Catalog, you will find some computer books worth linking to. The point is to use mapping, and not just one classification but many; and to link and switch them through some kind of registry.

Conrad wound up the discussion saying that perhaps the BCS, to become an organization that shares its knowledge, should put further effort into considering how to improve the way it gathers its own resources of knowledge together and makes them accessible to people.

Enabling knowledge communities online

Richard Millwood is the Director of the not-for-profit company Core Education UK, and Reader at the Institute for Educational Cybernetics at the University of Bolton. He was formerly the Director of Ultralab at Anglia Ruskin University, where he worked on the Ultraversity project.

Richard is also a consultant to the Improvement and Development Agency (I&DeA). Introducing him, Conrad referred to he'd been impressed by the I&DeA's knowledge community support systems. Marilyn Leask of I&DeA had recommended Richard to lead a discussion of how we can build and foster similar knowledge communities online.

Ultraversity: learning community online

The Ultraversity degree is a BA (Hons) degree in Learning Technology Research (LTR). It is a three-year 'learn while you earn' course, structured as part-time but with a full time credit rating. It is not located in any particular *subject* discipline, but in a *methodological* discipline of action research, where the action is to improve the work that you already do. It has received validation twice, and quality assurance is provided by Anglia Ruskin University.

Richard played a clip from a local TV news show, *Look East*, which reported how a 46-year-old teaching assistant from Harlow, Jane Day, who had left school with just three O-levels, gained her Ultraversity LTR honours degree after studying entirely over the Internet, without meeting tutors or fellow students face to face. However, she had made friends over the Internet with several of her fellow students, and looked forward to meeting them in person at the graduation ceremony.

Richard emphasised again that these students are not studying a subject discipline, but the work in which they are already engaged. Work may be defined quite broadly: for LTR student Eve Thirkis, her focus was on using technology to help with the education of her autistic son. With the knowledge she has gained, she now plays a leadership role in the Autistic Society in Doncaster.

The Ultraversity degree combines several innovations. LTR is inquiry-based learning, using an action-research methodology. The course is personalised: for each student a study focus is negotiated with the university to create an individual learning plan. Assessment is by dissertation and presentation of an e-portfolio, assisted by a peer review process. Most relevant from KIDMM's point of view is that the Ultraversity system contextualises learning within an online community which offers support and critical feedback, with input from experts and access to an online library. (The online community is effected through the FirstClass Communications Platform from Open Text, a system also used by the Open University.) [17]

Inquiry-based learning, also known as problem-based learning, emphasises a 'double-loop' learning process. The inner loops are cycles of *action*, each cycle having *plan–do–review–reflect* phases. Around these are cycles of *learning*, in which findings from reflections within action cycles are used to reflect on the practitioner/student's norms, identifying the opportunities for improvement and re-entering modified norms back into the learning loop.

(For reading in this area, see Donald Schön, 1987, *Educating the reflective practitioner*; Chris Argyris, 1982, *Reasoning, Learning and Action*; Chris Argyris and Donald Schön, 1974, *Theory in Practice: Increasing professional effectiveness*.)

The LTR course was shaped additionally to give to its participants a set of skills, knowledge and confidence in the use of online services, equipping them for lifelong learning as online practitioners beyond the point of qualification. The online community shared by Ultraversity students provides a safe, private space as a launch pad for a public identity online.

Before constructing the Ultraversity scheme, Richard and his Ultralab colleagues had worked on prior projects with teachers such as *Talking Heads*, the purpose of which was to ensure that head teachers keep up with each others' learning and skills. The Ultralab team discovered that as these online communities developed in strength – creating spaces in which participants could trust and believe in each other – it raised the quality of debate and led to deeper learning [18].

Learning in Ultraversity arises through students expressing ideas, and then evaluating these expressions. For this to happen, people must feel safe from ridicule. Students' prior experience of schooling means that they often do not trust the concept of learning from other 'ordinary' people like themselves; one student wanted all information volunteered online to be validated for her by her Learning Facilitator. Over time, she realised her fellow students 'were extraordinary people, with a wealth of knowledge and especially experience among them.'

Richard insisted that we should not ignore questions of *affect* – feelings. One quotation offered a fascinating insight into how students, who never met face to face, felt about each other, speaking of 'the very deep bond our cohort has forged'. It was written in response to an article in *The Guardian*, the title of which – 'The University where Everyone is a Stranger' – alleged that students didn't know each other [19].

In the Ultraversity project, organisational improvement of the learning process is assured through a constant cycle of evaluation and adjustment, whereby the content of the curriculum, the needs of the student and the needs of the employer are constantly revisited. This is in contrast to traditional work-based learning curricula, where university and employer organisation together agree an one-size-fits-all curriculum in advance, which the university then delivers to the student.

The question of Authority came up in the Ultraversity situation: the issue here was the balance of contributions from Higher Education staff, and from the practitioners (both the action-research students, and their workplace colleagues). The former were able to offer authority regarding process and overview, while the practitioners were the main source of authority regarding practice and context.

Richard showed some more video of LTR students at the graduation ceremony, sharing the curious experience of meeting 'for the first time' people they already knew well. They spoke of the importance of critical feedback within their online community – which raises interesting questions. How do you give critical feedback to someone you've never met? How do you trust the person giving the feedback, that they are not trying to get at you or show you up for a fool, but really trying to help you to learn? It is the job of the staff running the degree, as facilitators, to build this confidence between people.

I&DeA Knowledge

In January 2007, Richard went to work with Marilyn Leask's team at the Improvement and Development Agency. I&DeA Knowledge is a 'one-stop shop' Web site giving rapid access to background knowledge required to do the job of local government effectively. It pools national, regional and local perspectives; it provides a means of communicating with experts in other local authorities; and it is a source of useful, authoritative and reliable information.

This last point is of interest in the light of the previous point about Authority. If you work in a local authority, who would you trust for knowledge about tackling climate change at local level: central government? or the local authority that's been winning prizes for its climate change policy?

The I&DeA team set about the task of trying to explain to local authorities how the agency could help them learn from best practice. For example, on the I&DeA Knowledge Web site, they pulled together knowledge about climate change sourced from local authorities particularly forward-looking about the issue. Richard displayed a Web page on the site with links to relevant discussions: encouraging sustainable travel behaviour, calculating CO_2 emissions, carbon and ecological footprinting for local authorities, and sustainability training materials.

Such discussions are not, strictly speaking, online communities – they are bulletin boards hosted on the system, where people make contributions. These are not contexts in which people form relationships, by and large. Richard repeated the point to hammer it home: *Online Community does not equal discussion forums* – a community is a collection of people, not a software facility.

Some resources have been centrally produced: toolkits to support a range of actions, and documents to support learning and staff development. But for KIDMM, the I&DeA Communities of Practice part of the site would be of greater interest, Richard thought. There are about 50 of these CoPs, and membership is open to people who work within a local authority or are elected representatives. Joining a CoP helps people to obtain practical solutions to issues quite quickly; to access cutting-edge practices and thinking; and to network with others to share knowledge and experience.

Richard, as a consultant to I&DeA, is a member of two CoPs. He logged in live to his account, and showed us an internal CoP about CoPs. There are discussions there about what we can learn from online community software. In this space, Richard can share bookmarks, post up slides from a meeting, and otherwise communicate with colleagues at the I&DeA in a form that will stay 'on the record' – the thoughts and documents and slide-sets continue to stay online as a resource for the members.

The I&DeA system also has a wiki functionality, used for developing shared documents; and Richard believes there is a huge role in the future for the shared, collaboratively-authored document.

Richard showed us an event of which he had posted a notice, and pointed out that he had tagged it with keywords. Many different kinds of posted content can be tagged. Those who do the tagging can choose their own keywords; but the system also displays a list of the most commonly used, and through these means the tagging terms with which people feel most comfortable are rising to the top of the heap.

Richard admitted that the knowledge management team at I&DeA are probably not all that experienced in knowing how to build classification schemes; but thanks to these tools, and by actually doing tagging – in a meaningful context, and in meaningful relationships with other people, to get meaningful work done – they are learning.

Richard wondered if in making the I&DeA pitch to local authorities, they had perhaps underemphasised the ability to build professional relationships that will endure. If you are the one person in a local authority in Cornwall with a particular problem, and the nearest person who has faced such a problem is in Devon – that may be the county next door, but it's still too far to have a chat over coffee. You have to get online to have that sort of discussion.

Social networking services

Recently, Richard has been drawing the attention of I&DeA colleagues to phenomena such as Del.icio.us, Flikr, Frappr, Twitter and Facebook.

Flickr lets you share your photographs, Del.icio.us your bookmarks, Frappr your location, and Twitter what you are doing at the moment. Those are things about you as an individual, shared with other people. Some of these interactions seem trivial – perhaps none more so than Twitter messages like 'I'm thinking of watching Eastenders', but Richard argues that such chit-chat all builds deeper bonds. Some people will use a mobile phone to tell a friend, 'I'm at the supermarket, I'll see you later.' No substantial content, but the relationship is cemented and improved.

As for Frappr, it les you mark a map to say where you are. In the Ultraversity project, it was used so that students could declare their location. Once everyone had done that, any one of them could see the spread of locations for the students on the course. Sometimes people took advantage of the revealed proximity to arrange to meet up.

Facebook is more significant: it is pulling many of those services together in one space, and adds what Richard calls 'automated gossip'. As people join groups, add applications, ask questions and have them answered, become friends with each other… no one person writes all the happenings you can observe on your Facebook page. All those messages are generated by the actions of users on the system, and it helps to maintain community.

The I&DeA team is paying particular attention to Facebook, as situations are arising like the one at Kent County Council, which in August 2007 banned its 32,000 employees from using Facebook at work. The Council explained its motivation as being about keeping its systems secure, which is clearly nonsense. In truth, managers fear that employees will waste a lot of time online – and they are prepared to disrupt access to social networking services, discipline staff using them, and ultimately dismiss them for it. Facebook use is also banned in the Metropolitan Police and Transport for London. It was also blocked for a while at I&DeA, until their own workers persuaded the agency that they needed it for their professional purposes.

Conrad mentioned that there is a BCS Members group on Facebook, started by a member in the USA; a few people active in the BCS in the UK joined it recently. That group has over 260 members – but almost nothing happens there. Similarly, Sue Black of BCSWomen started a CSWomen group, which now has 88 members, again with very little activity. These things seem common: a group is started; lots of people join; it stagnates soon after. Establishing a group is not enough; something else is needed.

True, said Richard, and there are many frivolous groups on Facebook. But the RSA London group to which Richard belongs has significant discussions going on, supplemented by face-to-face meetings. This example shows how Facebook is becoming seriously useful in Richard's professional life.

As for the chitter-chatter aspects of Facebook, it helps to keep the people in Richard's life present when he is at the computer. Facebook has good profiling tools, good support for interpersonal relationships, and great integration with email, used for notification.

Some thoughts

Beyond the technology focus, Richard wanted to look more broadly at human facilitation online. Many Web 2.0 tools such as MySpace are about *online personality* – it's all Me, telling Me to the world. But the *interpersonal* has now developed, especially on Facebook. Maybe the *communal* will soon emerge – where there is shared purpose, an intention to form proper relationships of trust and productivity, around something we aim to achieve together.

Richard had been reflecting during the day that problems with things (data and information management) seem to be the same ones he faced 20 years ago. Progress may have been made on the theoretical front – but how much change has there been, practically speaking? Perhaps that is because theoreticians haven't focused enough on people.

People are central to knowledge creation, knowledge management and knowledge relationships. Richard also commented that we won't get away from diversity; we won't get away from neologisms; and we won't get away from dissent. We need systems that allow us to accept and deal with that, mash up the various inputs, and put things together.

A serving of Mash-up

Richard quickly showed us a mash-up experiment he had made during the day: a spreadsheet in Google Docs that references information on the Web about the population of cities. For London, it is stated to be 7.5 million; that data is retrieved by a formula that googles off and gets the data. He filled in the same formula for other cities, and it pulled back results before our very eyes in the same way. Shall we add another location? 'Swindon', was proposed, and back came the answer '161,000'. Whether that is true or not he could not tell. What he did know was that without much effort, in five minutes, he had created an information retrieval tool.

'It makes plagiarism look stupid. Wouldn't you want your secondary school children to be able to use these tools

effectively, to build knowledge for themselves, and for others in their class, for the community they live in, for companies they go to do work experience in. Wouldn't you want that? How is that going to happen in the climate we've got now?'

Discussion

Someone asked how much knowledge Richard expected users to have, not just to be able to interpret information, but also to be able to tell whether it looks vaguely correct. Conrad said, this is another point that people have made about information literacy. In the Developing Countries SG workshop on Information Literacy referred to previously, the SCONUL document deemed the ability to *critically evaluate* the information one has retrieved as being as important a skill as the ability to go out and find stuff in the first place. When it comes to school students, it's often that critical evaluation that is missing.

Richard said he would go further; he'd say that a lack of critical evaluation is not new, and has always been in short supply; except in the best schools, and with the best teachers. The problem is that the model of schooling which children absorb is: 'You are going to tell me what the truth is, and so I don't need to think. All I need is the ability to remember it and write it down when it comes to the exam, the way you told me it should be written down in the exam.' That is the challenge we face in education.

It isn't just schoolchildren who fail critically to evaluate, someone remarked; it's in government as well – e.g. a huge problem of people being uncritical about statistical data.

A woman in the audience noted that when we heard the Ultraversity students appreciating the value of community, and how nice it was to be able to post problems and have people rally round, she thought: is this a particularly female sentiment? She said, *she* wasn't such a nice female (!) – *that* wouldn't be the reason she would want to use a community; *she* wouldn't want to help others sort their problems. From this she wondered two things. Firstly, did these people stay as closely bonded after they had actually met, and got to know more about each others' foibles? And were there any use-cases involving men?

Richard said that the course population was 80% women, and the average age was 40. That is not uncommon among part-time learners in the caring services – some of them were teaching assistants, some were health professionals. But there were men, and these did give similar stories.

Richard noted that some Ultraversity students dropped out after 6 months on the course, with the riposte, 'How could you possibly start a course like this without quality materials such as the Open University provides?' They had not accepted the idea that they might construct knowledge themselves, in conjunction with others filling in the gaps in their expertise. For some, taking responsibility for their own learning was particularly hard.

Jan Wylie recalled that when years ago we started to participate in email discussion lists and bulletin boards, there was a big problem with 'flaming'. Has that gone away? Have we become more mature? Or did Richard find that if people felt strongly about something there was still some of that behaviour? It can lead to some very nasty outcomes.

Oh yes, said Richard; it still happens. One way to improve things is not always to do all the communication in one big forum. A large community sharing a single communication channel such as an email listserver is quite vulnerable to flaming. But where many subcommunities operate, like the learning sets in Ultraversity which consisted of just a few people each, one can have experience of community in smaller spaces where trouble is less likely to start.

The challenge for Ultraversity staff, as academics, was how to deal with such situations, given that the ethos of academic freedom considers the right to express opinions as being paramount. In particular the facilitators suffered from real angst about how to deal with people's destructive behaviours. Conrad said that from his years of experience in email-list communities, he feels one does need people around who are able to be the carbon rods that are inserted between fissile material when reactions start to run a bit hot.

It was asked if the software on which I&DeA Knowledge applications run is available to other communities. Richard said it has been developed with a company who created a bespoke solution. They do permit any government body to use it as a foundation for their own online community; but you apply to use that company's service, rather than taking the software away and running it yourself.

To wind up the day's formal procedings, Conrad wondered how this conversation might continue. He hoped the participants present had enjoyed the conversation and would continue discussion informally over drinks (and they did, for several hours!) He expressed great pleasure at the great variety of people who had wanted to join the event; it felt like a real vindication of the ambitions within KIDMM to break out of the silos of specialism, firstly between the BCS Specialist Groups (and we had heard from several during the day), and then breaking out beyond those.

Conclusions and aftermath

Outputs from the Mash-up are being made available on the KIDMM Web site, including slides, some audio from talks, and this Report [20]. Several people joined the discussion e-list after the Mash-up [3], and an e-newsletter service has been set up so that people who'd like to keep an eye on activities around KIDMM can be informed on a fortnightly basis [21].

Inspired by Richard Millwood's advocacy of online community, the new domain *www.kidmm.org* has been registered, and the aim is to develop there an electronic discussion space eventually open to wider participation. The other major current KIDMM project is the construction of a portable consciousness-raising exhibition, *Issues in Informatics*, two prototype panels of which were on display at the event.

References & Links

1. BCS Web site http://www.bcs.org
2. 6 March workshop report http://www.epsg.org.uk/KIDMM/workshop.html
3. BCS-KIDMM@JISCmail.ac.uk – membership list with affiliations is at http://www.epsg.org.uk/KIDMM/email-list.html
4. See Conrad Taylor, 2007, *Metadata's many meanings and uses* – a personal briefing paper obtainable from http://www.ideography.co.uk/briefings
5. Dublin Core Metadata Initiative Web site: http://dublincore.org/
6. BCS-DCSG workshop report *Information Literacy, the Information Society and international development* is accessible from http://www.epsg.org.uk/wsis-focus/meeting/ 21jan2003report.html
7. Danny Budzak, Metadata, e-government and the language of democracy. http://www.epsg.org.uk/dcsg/docs/Metademocracy.pdf
8. CRISP-DM – see http://www.crisp-dm.org/
9. Alan Rector. 2000. 'Clinical Terminology: Why is it so hard?' in *Methods of Information in Medicine* 38(4): 239–252; also available as a PDF from http://www.cs.man.ac.uk/~rector/papers/Why-is-terminology-hard-single-r2.pdf
10. See comprehensive description of MUSIMS at the MDA Software Survey: http://www.mda.org.uk/musims.htm
11. Official EAD site at the Library of Congress: http://www.loc.gov/ead/
    See also Wikipedia description at http://en.wikipedia.org/wiki/Encoded_Archival_Description
12. See CIDOC Web site: http://cidoc.mediahost.org/standard_crm(en)(E1).xml
13. SCULPTEUR – http://www.sculpteurweb.org. Includes 'a web-based demonstrator for navigating, searching and retrieving 2D images, 3D models and textual metadata from the Victoria and Albert Museum. The demonstrator combines traditional metadata based searching with 2D and 3D content based searching, and also includes a graphical ontology browser so that users unfamiliar to the museum collection can visualise, understand and explore this rich cultural heritage information space.'
14. See results table for the 2 September 2007 ballot at http://en.wikipedia.org/wiki/Office_Open_XML_Ballot_Results
15. See Adobe Labs Web site: http://labs.adobe.com/technologies/mars/
16. To retrieve all the materials from that meeting, including MP3 recordings of panel members' presentations and the discussions in full, go to http://www.epsg.org.uk/meetings/classification2006
17. For an introduction to FirstClass, see http://www.firstclass.com/AboutFC/
18. Carole Chapman, Leonie Ramondt, Glenn Smiley, 'Strong community, deep learning: exploring the link'. In *Innovations in Education & Teaching International*, Vol 42, No. 3, August 2005.
19. Stephen Hoare, 'The University Where Everyone is a Stranger'. In *The Guardian*, Tuesday 20 June 2006.
20. Outputs are being posted at http://www.epsg.org.uk/KIDMM/mashup2007/outputs.html
21. The KIDMM News e-newsletter can be subscribed to at http://lists.topica.com/lists/KIDMM_news