When you think about the world, you rarely think about an isolated thing. You think of pieces of a whole, whether a department in a company or a street in a town, a cup with a saucer, a client with an address and an account history. Sometimes one thing is a part of another; sometimes they are complementary. Sometimes one could replace the other – the same thing under different names. Sometimes one piece of information describes or specifies another. Things exist in context: the economy in a time of war, one company acquiring another, a manager in a department selling to a client in another industry.

All these examples put things in relationships to one another: some of them explicit, some implicit. Things change over time and as circumstances change: You may need to change the terms you use if you start buying from a different supplier, if regulations change, if a particular vendor stops operating…or if your own needs change.

You can effect some of these changes through database transactions, but databases are a weak tool for representing the complexity and interdependencies of the real world. (More often they represent the specific data structures that were convenient for a developer – who often is no longer around.) Some of that complexity is present in computer applications, but it is often inexplicit: hard to find and even harder to manipulate or generalize. How can you represent the effects of a change in suppliers, a corporate merger or even the intersection of two separate companies’ product catalogues? Call in the programmers!

Or call in the taxonomists and ontologists. They are the people who define and represent the relationships between things and their contexts.
We originally wanted to call this issue “Mankind’s Quest for Meaning,” but the Kool-Aid wore off. The task for machines, at least, is not to find meaning, but simply to find structure – enough regularity to find things and to manipulate them. A machine needs precise data structures and rules to do anything useful.

This broad topic divides almost naturally into two parts: The first concerns finding structure in and around mostly unstructured data, for use by humans. The second concerns defining and modeling the interactions and dependencies of structured data, for use by machines. (It’s an IT thing: Such structure exists both in IT systems themselves, and in models of complex domains that (ideally) can be made explicit for use by IT systems.)

Reflexively speaking, the first goal of this and every issue of Release 1.0 is to make distinctions and classify the entities in a specific domain; that’s what the current issue is about. The second goal is to show how things work together and depend on one another; that will be the goal and the topic of our February issue.

**Meaning and structure**

Putting things into categories and defining them is a basic form of intelligence. Remember those test-your-IQ games: Which object doesn’t belong in this picture? Other quizzes test a user’s ability to derive rules from examples: What numbers come after 1, 4, 9, 16? People can also grasp analogies or metaphors, often the best way of explaining complex relationships. (This paragraph offers a metaphor for what a computer must do in finding structure….)

That ability to think abstractly – about categories or groups instead of specific instances – and to manipulate those abstractions is a key capability that distinguishes us in the larger taxonomy of things. How can we get computers to (appear to) do the same? Basically, by figuring out how to represent abstractions as concrete, well-structured data that computers can manipulate because the data is defined in ways that are meaningful to the computers and the precise instructions they follow (a.k.a. applications).
None of this is very difficult in principle or in bite-size examples – the level of five animals and a toy truck in a picture, or a short sequence of numbers – but it becomes tremendously difficult as you scale up and move beyond the relatively small set of concrete, discrete data items that fit neatly into fields in a database. The number of items increases, as does the number of relationships among them, and the number of possible different views of the data. It’s easier to make generalizations than to describe specific procedures for each instance. Computers allow us to make such complex generalizations on a massive scale…if we can get the categories and structures right. This is starting to happen as people who post information pay more attention to enhancing it with metadata.

The structure of this issue
In the issue below, we discuss search vs. structure, with illustrations from Yahoo! and Google and others, along with a peek at a data visualization tool from Groxis and an interesting new metadirectory from Microsoft. Next we look at one of the most common uses for taxonomies – shopping guides and catalogues – and then at a medical application where it helps to have a model of the user rather than just of the data she’s looking for. (Note that we don’t cover tools to create directories and taxonomies automatically; we take that as given. They include products from companies such as Verity, Autonomy, Semio and Stratify.)

In the February issue, we will focus on how enterprises can use structuring tools and modeling to deal with active data rather than mostly unstructured “content.”

Taxonomies and Ontologies
As noted in the box on the following page, a taxonomy is a subset of an ontology, which generally includes a representation of a variety of different kinds of relationships among its elements.

Analogy is another kind of relationship and can be a facet of an ontology…and it is also useful for explaining some ontological relationships. For example, to explain ontology/taxonomy in biological terms: A taxonomy defines what kind of animal you are; an ontology incorporates that taxonomy and also defines what kind of animal you eat.
In this issue, we deal primarily with taxonomies. Taxonomies underlie everything from Yahoo!’s directory structure to most shopping sites, as well as most corporate knowledge management systems and libraries. Taxonomies are a great way to organize and find things and to describe content, but they lack the complexity or procedurality needed to represent applications. But not all hierarchies or directories are taxonomies: For example, an employee directory could be hierarchical by location rather than by a taxonomy of employee types: first the whole company, then individual campuses, then the individual buildings in the campus, then the individual rooms (see page 14).

**Structure and Search**

To start, let’s just consider how the Web’s unstructured information can be organized. The two leading approaches are exemplified by Yahoo! and Google. Yahoo! has created a single, very broad taxonomy; although it has not in fact organized everything (!), it offers a directory (taxonomy) structure that in theory should be able to classify any content that shows up. By contrast, Google organizes the Web dynamically: Tell us what you want, and we’ll put it at the center of the world and find you the surrounding information. With search you can find passing references to a topic – such as the name of the person you just met – regardless of what the overall document is about, and whether or not a relevant category has been created. Or you could say that it creates the category on the fly.
XML AND THE SEMANTIC WEB

XML stands for eXtensible Markup Language, a W3C standard. It is used primarily for tagging data with meta-tags, and is very handy for defining data elements in structured and semi-structured documents. However, it is a standard for representing structure, rather than a set of standard meanings. That is, there is a proliferation of would-be standard-setters developing XML tags of various meanings, rather than a small set with consistent meanings within each industry and domain.

The pervasive use of XML (structure of data), RDF (for Resource Description Framework, semantics of data) and OWL (for Ontology Web Language, reasoning about data) is ultimately supposed to lead to the “Semantic Web” – or basically an ontology that assigns meaning to everything on the Web. But don’t hold your breath. For now, these techniques can be very useful in the small...and may lead to something as exciting and unexpected as what HTTP has already produced in the current World Wide Web.

Yahoo!: Living and linking

Diametrically opposed in philosophy, Yahoo! and Google are converging in practice: Yahoo! recently spurned Google to purchase Inktomi and its search engine. Clearly, Yahoo! now wants control of both mechanisms that are key to its business.

Yahoo! began its life as a topic directory for the Net. Srinija Srinivasan, Yahoo!’s fifth employee and a fellow student of the founders at Stanford, joined the company in 1995 from a role as an ontological engineer at Cycorp (a company dedicated to building an ontology of the world; see next issue). The title she chose for herself at Yahoo! was Ontological Yahoo!; she is now Yahoo!’s editor-in-chief. “Directories make most sense when you are browsing, when you want to discover something,” she says. “Whereas you use search when you know what you are looking for.”

Currently, as more and more Yahoo! users are experienced, they are more likely to use search – on purpose or simply because a particular item is not listed in the directory. “We can’t possibly manage the entire range of what people might be looking for,” says Srinivasan. “The directory was never intended to cover every word of every page out there. In the context of search overall, even though we don’t get comprehensiveness in the directory alone, we use the human element in other ways to improve search.”

For example, she cites how editorial expertise can enhance a query on a search-results page. This can be an offer of more – you asked for X; would you also like to see Y and Z – or it could be disambiguation. If you search for “bonds,” the category matches prompt disambiguation/refinement: An editor knew there were different senses – the athlete Barry, financial bonds, and bail bonds – and captured that in the directory structure. From that point you can either browse those categories, or you can issue a new query that narrows your search.

Nonetheless there’s a trade-off between depth and breadth; the directory offers fine-grained, carefully vetted material, while the search engine offers access to everything else. When a query produces an entry in the Yahoo! directory, it will be edited by a
human, whereas the automatically generated “page descriptions” in the search results (like Google’s) lack the human touch and are often cryptic or confusing.

Srinivasan now spends more of her time on policy questions, such as privacy and content control, but that doesn’t mean that taxonomies don’t sometimes lead to controversy. Though part of the value of a computer system as opposed to a card catalogue is that you can easily classify movies under both entertainment (US) and Culture [sic!] (France), there is still the issue of, as she puts it, “where it lives and where it links.” Yahoo! shows the path to any particular item in its directory, which reveals its “true” location in the taxonomy. “The very existence of a path can be controversial,” she says. For example, she still remembers the fuss in 1995 over the listing for Messianic Jews (Jews who believe that Christ was the Messiah). Should they be classified as Jews or as Christians? (Most Jews and most Christians believe they belong to the other category.) Eventually they got their own category, alongside Jews and Christians.

“This helped shape our thinking,” she says, as other issues came up. “Some people question the very existence of the path, or prefer their own labels that assign a particular interpretation or point of view (e.g., anti-choice, anti-life – or holocaust denial). All else being equal, we strive to call things what they call themselves, while adhering to canonical/accessible terms and phrases. Let the information speak for itself, give the consumer rich context, and let her form her own interpretations.”

**Google: Just-in-time categories**

The Google ethos is discovery (rather than creation) of structure, dynamically and locally. Call Yahoo! Cartesian and Google Euclidean.

While Yahoo! began as a directory of cool sites, Google arose from the perspective a few years later (circa 1998) that the Web is simply too vast for anyone to define or structure it properly: Best to let each query define its own neighborhood, and to start each search from the query outwards, rather than from some mythical top down, to where the answer lives. Google doesn’t attempt any kind of global structure; it takes each search term set, finds the subset of pages that matches, looks at how those pages are linked to one another, and ranks them. The ranking is not precisely “relevance,” and the process is a lot more complex and obscure – the result of years of tuning and refining.
The main focus of everything at Google is still page rank and the hundred-odd other algorithms that make up Google’s secret query sauce (only the chef in the Googlateria knows for sure). And in the end, whatever goes into it, ranking is a one-dimensional measure.

But even as Yahoo! is beginning to pay more attention to search, Google is paying more attention to structure. It has offered the AOL Open Directory Project (you can see the link to ODP at the bottom of its Directory Web page; see page 9) since March 2000. The company continually tries new tricks, some of which eventually emerge from its labs into public services such as Google News and Google Catalogs.

In this context, Froogle, Google’s new version of online shopping, is the most interesting. It’s an attempt to help users find specific, purchasable items and to present them in a useful format. “We knew we didn’t give good results for shoppers. We were good at reviews and discussions, but not so good on where to buy actual products,” says Craig Nevill-Manning, senior research scientist and engineering manager for Froogle. Froogle is a separate service, though it uses the Google engines in its initial search and orders the results by page rank.

Froogle uses parsers, some of them customized for particular vendors’ pages, to produce more meaningful results: photo, product name, price, merchant, the merchant’s classification of the product in its taxonomy, and a little bit of description with key words highlighted. A user can pick these out and, increasingly, so can software. The same information can be used to filter a search by price or category. That is, by looking inside the structure of pages, it can also create categories around the pages (or the items they display). One advantage of this kind of search vs. a directory is that you generally get directly to the product you are looking for, rather than to the home page of the merchant. (Of course, directories are getting better at this, too; no trick stays unmatched for long.)

To help in this task, Google encourages and takes structured feeds from any merchant who offers, reducing the challenge of parsing the information properly and ensuring that the information is up-to-date – especially vis-à-vis prices and even availability. “It’s open to all comers; we’ll take a feed from anyone,” says Nevill-Manning, pointing out that Froogle is a good way for a smaller merchant to get visibility if it sells goods that are unusual or unique. Froogle could also easily parse smbmeta.xml files to identify a vendor’s location and other useful filtering information (see page 6). Stay tuned.
Froogle does not charge for the service, although it does run related ads alongside the results, and it is trying to keep the whole thing low-cost. Nevill-Manning’s team is, he says, “frugal” in size, but it is already thinking about B2B as well as consumer sites, and other extensions. Nevill-Manning had known Larry Page and Sergey Brin, Google’s co-founders, when he was a postdoc at Stanford. He returned from Rutgers to join the company as a senior research scientist, he says, because it was a better research environment than he could find at any university: “We get to play with the whole Web.”

**Math vs. meaning**

Google’s other Craig, Craig Silverstein, is director of technology at Google, and has been with the company since it formally changed its name from BackRub (cf. backlinks) in 1998. He states the overall attitude clearly: “When you fool around with taxonomies and try to classify things automatically, you discover just how good people are at it. The problem, of course, is that people don’t scale.” You can automate search; it’s tougher to automate classification effectively. Thus, as with Froogle and Google News, much of Google’s research concerns producing better listings for each page rather than showing the relationships among them. In Froogle, he points out a little enviously, the site owners are motivated to include metadata: “There’s a long history of taxonomies in commerce. And the same with news – where dates are clear, and stories are generally classified by editors.”

His focus is on improving quality of results, as measured by how frequently users click on the first few results presented. Though the company doesn’t track user behavior to modify search rankings or results for specific pages (which could easily be used by third parties to game the system), it does notice whether users click on the results and uses that feedback to figure out which queries aren’t working and what kinds of refinements could help.

For example, Silverstein and director of search quality Peter Norvig want to get better at reverse-engineering metadata, such as for dates. They are also looking at some other embellishments, such as clustering results for different individuals (Michael Jackson the singer or Michael Jackson the British software entrepreneur?), or terms such as tables (furniture) vs. tables (spreadsheets) vs. tables (flat mountaintops). But as the user base becomes more experienced, most people figure out how to do that anyway, using a disambiguating search term.
**DMOZ: The Open Directory Project**

The Open Directory Project, now housed at AOL(!), began life in June 1998 as GnuHoo!, an open-content take-off on Yahoo! with a taxonomy based on Usenet. The company was founded by five engineers from Sun and other companies; the software to manage the directory and the machines it ran on were proprietary, but the content was open-source – developed by volunteers, and freely available to anyone to reuse (but with attribution and a link to the ODP).

After complaints from the Gnu folk (due to its proprietary software), it changed its name to NewHoo. It was acquired in November 1998 by Netscape, who thought it would make a fine anchor for Netscape’s Netcenter. (Remember, that was in the days before investors asked for business models.) Netscape changed the name again, to DMOZ: Open Directory Project (for Directory Mozilla), and AOL kept it when it acquired Netscape shortly thereafter.

DMOZ is now the underlying directory not only for all the AOL properties (AOL itself, Netscape and CompuServe), but also for Google, Lycos, AskJeeves, AT&T Worldnet and hundreds of other search engines and portals. The whole effort comprises two paid AOL staffers – managing editor Bob Keating (a librarian at Harvard before joining Netscape/ODP four years ago) and chief software engineer Autumn Looijen, involved with the ODP since its early days – plus 12,000 to 15,000 volunteer editors active at any given time.

They work by regular open-source principles, heavy on peer review and community discussion. The main task of an editor is to find and classify sites and to place them in the directory (using tools developed by the DMOZ founders and maintained and enhanced by Looijen and the more technical editors themselves). The editorial guidelines are on the site, and anyone is invited to join the fun. Acceptance as an editor depends on acceptance by existing editors in the particular category you want to join; as editors prove their mettle, they can gain privileges in other categories or higher up in the topic hierarchy. There are over 100 meta-editors with privileges throughout the directory. “I don’t supervise,” says Keating. “I just coordinate.”

When someone wants to create a new topic, they open a discussion, create a test area, and then ultimately add the topic by consensus of all interested editors. Overlaps or poly-hierarchy (multiple-parent) issues are resolved by cross-links or “related” links. The most frequent form of abuse is commercial – when someone creates a host of subcategories all linking to his own site, unfairly promotes or lists his own site, leaves
out competitors or detractors, or otherwise tries to game the system in his own favor. Those people are ejected.

“There’s an impression is that because it’s a volunteer effort it must be easy to spam and abuse because people aren’t paying attention,” says Keating, “but it’s the opposite. People care a lot. And unlike commercial directories, we are not aiming to please the Webmasters who are listed; we’re aiming to please the users.”

Obviously, one has to wonder what such a property – which generates no direct revenue whatsoever – means for a company such as AOL. Gerry Campbell, director of search & navigation for AOL, says, “We aren’t looking to make revenue [from ODP]. It simply makes the whole [AOL] member experience better.” It does that for Google, too, we must note. He adds, “There hasn’t been a lot of new development, but we are now looking at additional ways to evolve AOL’s usage of the directory into something in concert with the value we bring to the AOL service going forward.” For AOL as for the other users, ODP is just the seed of a directory onto which third parties can add their own proprietary extensions.

**Entopia: The spreading directory**

Directories can also enrich your standard knowledge management tool. Unfortunately, most such systems are hermetically sealed: They manage their own knowledge, in their own structures. Entopia, based in Belmont, CA, is making the leap to help its corporate customers build a company-wide directory of topics and allow their users to maintain and extend it, much like an internal “open directory.”

Each Entopia customer can use its own directory as a structure to classify almost everything in the company – and even a few outside sources, such as designated Web pages. When a user wants to store a document, Entopia’s Quantum suggests a pick-list of where it can be stored. It also notices whether the document has already been stored – though it still has a way to go on version control and avoiding redundancy (which is a small but technically challenging problem).

The value is to make the information retrieval system comprise all the information in the company, rather than just items explicitly stored in the “knowledge” bucket. In a more or less automated way it makes it easy to maintain and build the taxonomy, and then to apply the taxonomy to all a company’s documents. To use a formulation from Yahoo!’s Srinivasan, the documents “live” where they were created, but they all “link” into the corporate directory.
Quantum also includes features that hide individual data but still provide the links, so that (for example) you could find the corporate expert about X based on his e-mails or any other document he created, but you could not necessarily see the e-mails themselves. Beyond that, a user can also declare any document entirely confidential, so that it is not even indexed by the system. (Of course, the user needs to trust the software, but such permissions are set by the individual, not by administrators, who can only set defaults).

Currently the software works with Microsoft Office, Lotus Notes, Websites and fileshares. Support is coming soon for document management systems (Documentum, OpenText) and certain front-office applications (Siebel and Salesforce.com). In the case of applications, Quantum’s directory will point to appropriate records in the applications. (We hope to see Eudora support soon!)

**Searching and Mapping**

As Yahoo!’s Srinivasan notes, users have generally become more familiar with the Web, and they have (broadly) turned from browsing directories to searching, from exploring to going after a precise result. But search is not the end-all.

First of all, individuals’ information styles and needs vary, by person or occasion. Are you trying to research a space (Microsoft’s role in the industry)? Or are you trying to answer a single question (when was Bill Gates born)? Sometimes, when you’re doing a search, you want to know all about a particular item. Almost everything containing the search term is of interest (although at some point you’re likely to get tired of the redundancy). On other occasions, you want to see the context: What other similar items might be worth checking out? Is there an alternative? Or perhaps an alternative context? Rather than comparing the French movie to other movies, compare it to other forms of French Culture. . . .

Second, the tools are getting better, so that your choice is not between a simple taxonomy and a specific list of results. Instead, the best tools combine more relevant classifications and filters – not just a single taxonomy – with better visualization of the concept space they define.
As we discussed in the September 2002 issue of Release 1.0, visualization tools often depend on underlying data structures, metadata and categories from companies such as Semio (discussed in that issue) or Northern Light and Teoma (search tools that specialize in categorization and restricted result sets, well beyond the date and format limits offered by Google’s advanced search). They allow you to see (analogy watch!) a map rather than just a ranked list of addresses. The simplest ones simply show proximity and several degrees of links, like many we covered in September: “This item links to that item. Here are some neighbors...”

But you can do more... Some categories are flat: There’s simply a set of buckets that things fit into – for example, which university someone attended. But most classification schemes are hierarchical or can be made so with filters: There are smaller buckets inside each bucket – for example, the year of graduation, or the person’s major. These hierarchies comprise a variety of ontologies; that is, the up-and-down (and side-to-side) relationships vary.

For instance, there’s the inheritance relationship of a taxonomy, in which the top item is very general and each layer below it is more specialized and varied, but the lower levels all inherit characteristics from the layer above. Then there are org charts, in which all the “objects” are people and the lines between them represent reporting relationships rather than “is-a-kind-of,” but the lower-level people don’t always know how to do all the tasks their bosses can do (or is it the other way around?). There are political and geographical hierarchies, in which the lower levels are subdivisions rather than specializations of the objects “above” them: i.e. a city is not a particular kind of state or country, but rather a part of one – and the cities in one state are the same “species” as those of another. There are quantitative classifications, in which you can subset, say, age or price categories into more numerous, smaller categories.

Visualizing such cross-cutting classifications is a challenge. As the world gets better metadata and better data sources overall, we expect to see a proliferation of tools rising to that challenge. Two recent projects have succeeded, in our opinion, and the results are stunning. (They look far better on the screen – both have movement – than on our pale paper pages. Come see them at PC Forum; see Calendar, page 34.)

**Grokker: A bubbling Petri dish of information**

First is Grokker from Groxis in Sausalito, co-founded by author Paul Hawken, RJ Pittman and Jean-Michel Decombe, all of whom co-founded and led (separately) several other software companies. Grokker, soft-launched last fall, depends on the
structure and clustering of the data sets it is displaying. It has interfaces for Northern Light, Teoma, Open Directory, Amazon.com and most pc file systems, among others. It could be a real improvement over Windows Explorer, in our opinion!

You use Grokker by selecting a (supported) data source to query. A click on the “Grok” button produces a colorful graph of the results, shown as shiny, expanding bubbles, that builds continuously in real time. The graph represents the underlying structure of the data source, whether it be a list, a hierarchy, a poly-hierarchy (or directed acyclic graph), or just a set of document clusters. You can then explore the structure and its document contents by clicking and zooming around, for as many levels as the data allows, down to individual data items.

Or you can stop at any level and apply various filters/ranges to enrich the data display, using cues such as color, size, shape and luminosity to represent orthogonal qualities. Those qualities can be other discrete categories, such as language or data type or vendor or gender or the presence or absence of a specific word, or quantitative measures, such as age or price.

While traditional statistical analysis depends mostly on quantitative data that can be summed or divided, the sweet spot for Grokker is the metadata and non-quantitative attributes. (Although of course people are always trying to define qualities in ways that can be quantified, hence Google page rankings, Amazon sales rankings, search relevance (aka conceptual closeness) and just about every survey that ever gets graphed for USA Today.)

The overall effect is life-like: a fractal overview of the data space, and a curiously seductive lava-lamp-like effect as each cluster of the unfolding hierarchy swells and bursts to reveal the next layer.

Just imagine using Grokker to wade through an industry conference directory, seeing clusters of VCs, start-ups, CIOs, bankers, journalists and other types of people,
filtering on location or product type or job title, and then finding the individual names and data at the leaves of the tree.

Grokker could also work nicely with other directories and metadata such as those we describe below, as well as with Yahoo! and the like. The product, which costs $99, comes as a client tool that can display content from the sources listed above. Over time, Groxis (and licensed third parties) will add plug-ins for a variety of data sources, some of them free, and some designed to be packaged and paid for as part of various online services (news, stocks, demographics and all kinds of proprietary information services).

In addition, Groxis will offer an enterprise version with some of its own data-management tools, where it will add value to rich data structures and most current knowledge management systems. All this, of course, depends on the metadata being available. We suspect Grokker and other such tools will gain currency more for specialized data sets than for the Web, where the data are sketchy. . . . On the other hand, Grokker shows the potential of the Semantic Web (see page 5) and may help spur its development.

Microsoft’s Polyarchy: Proliferating parentage
While Grokker effectively combines a single hierarchy/taxonomy with cross-cutting quantitative classifications and filters, Microsoft’s forthcoming Polyarchy is best suited to help people see two or more orthogonal hierarchies. Better than that, it helps them see what they are seeing: Polyarchy is one of those oh-now-I-get-it! tools. It gives a 3D feeling through movement rather than by showing a static 3D structure.

Polyarchy is the new component of Microsoft’s Metadirectory (of users) from directory architect Kim Cameron. Cameron joined Microsoft in 1999 when it acquired his company ZOOMIT, which became the basis of Microsoft’s Metadirectory Services (see Release 1.0, June 2002).

As shown in the illustration on the following page, you can see the hierarchy change from one with Bill at the top to one with Redmond at the top. And you can see the link between Kim and George change from a path through the org chart, to a path up and down a different hierarchy of the same people, showing the location of their respective offices in different buildings on the Redmond campus. “We animated it so you get a sense of transitioning from one hierarchy to another,” says Cameron.
First, Bill is at the top of the hierarchy. ...and then a miracle occurs... ...and now everything flows from Redmond: same people, different branches
**What’s happening here?**

Most people know what jkhasdfha hierarchy is without thinking about it much: org charts, outlines, and the ubiquitous Explorers from Microsoft. They know that the same two people often end up being related in several ways – or in some ways but not others. (Think of six paths of connection rather than six degrees of separation. . . .) These relationships are very powerful – for understanding and for programming – once they are made explicit.

The idea for Polyarchy started when Cameron got involved with others from Microsoft in standards discussions about UDDI: “That was the first place someone ever talked to me about the organizational structure of a company being like a taxonomy,” he says. “But it was more complex; there were a bunch of ‘taxonomy’ dimensions that didn’t necessarily interact. We’d go out into the [user] namespace and people would say, ‘well, what kind of hierarchy should we have here?’

“Eventually we realized that they shouldn’t go with any single hierarchy. A hierarchy is just one view. The hierarchy that was appropriate depended on the question you were asking” – with answers based on information from the employee directory, though one can imagine that concept applies to other data sources.

Cameron continues: “So I said, if you have a completely elastic poly-hierarchy, what does it mean? And what does it look like?”

For the answer, he called on George Robertson, a senior researcher at Microsoft Research and formerly at Xerox PARC, and together they built the new Polyarchy viewer (not yet formally named) that will be available as a technology preview when Microsoft Metadirectory Services 2003 is launched this spring. “People should be able to make their own hierarchies, defining the kinds of links they want to see” – whether common buildings, common projects, or whatever else links some people and separates others, he adds. “One of the things this tool points to is shared and private hierarchies. You can publish your hierarchy like an address book, or you can simply share the framework (what attributes define the hierarchy) like a query and people can populate it with their own data.

“We wanted software that could discover the hierarchies from the relationships in the data and then display them. Take the list of people attending a meeting. You don’t know who they are, but you can see them in a variety of contexts: Are they in a building or a workgroup with someone you know? Or selling the same product but in a different country?”
Structure and Shopping

Single taxonomies are challenging enough; but most vendors of products and services (to say nothing of political parties, religions and countries) have their own views of the world and the things in it. In order to reach customers and compete with other vendors, suppliers need to align their product directories with one another’s. The big challenges are reconciling a variety of worldviews, maintaining the taxonomy as things evolve over time, and translating from one to another. Products often belong in several categories.¹

So it’s no surprise: Aside from portals such as Yahoo! (which are anyway paying increasing attention to shopping), the big markets for public directories are services designed to expose products and services to those who might want them, whether consumers or other businesses. Google’s Froogle (above) is one case in point.

It’s worth noting that such endeavors can meet resistance. Buyers generally like transparency; sellers, especially big ones, generally don’t. So the best place to start is usually by serving the buyers; eventually the vendors come in when they realize that it’s better to be transparent than invisible. That’s why so many of the B2B exchanges didn’t get very far: They were going after the sellers, who had no interest in creating a commodity market where they used to have a private club and “friendly” relations with buyers.

But even buyers don’t always want transparency, warns Mark Walsh, former ceo of VerticalNet, from bitter experience. “Yes, the sellers want stupid buyers,” he says. “But sometimes the buyers want to stay uninformed. They like those long lunches, the gifts at Christmas, the trust they build with a distributor who tells them what to buy. . . . They don’t want transparency interfering with those nice relationships.” Of course, the companies the buyers work for feel differently, as do individuals and small businesses buying for themselves.

We tried to cover one such B2B exchange (not VerticalNet!) and its data directories for this issue, but somehow we just couldn’t get a good fix on the ontology of that particular market. Here are some case studies that we could figure out:

¹That doesn’t matter so much when you’re looking for things to show to a human; it’s more challenging when you are trying to perform statistical summations and avoid double-counting. That’s the challenge faced by Sector Data, which serves Wall Street and business analysts with rich data sets supported by its own taxonomies of products and services, focused on fast-changing sectors such as healthcare, technology, media and finance, with energy and consumer products to come.
CNET Channel: Bag ‘em and tag ‘em

CNET is the well-known computer trade and consumer news site – but it also has another side, a product directory-based business called CNET Channel, that is not so visible to the public. “One of our challenges is how to leverage the editorial knowledge we have, so that it applies not just in the specific editorial coverage, but also across all products,” says ceo Shelby Bonnie. To do that, CNET needs a strong taxonomy, so that what editors say about A could be applied to A1, A11, A111 and so on. By quantifying and making explicit as many attributes of each product as possible, CNET can model customer needs and then provide customized ratings for different kinds of uses/users. It can also help computer resellers intelligently manage the mounds of product information they get every day.

The story really begins elsewhere, in Switzerland. We first met Albert de Heer back in the 90s when he and his brother Rudolf were running a Macintosh dealership in Russia from their base near Lausanne. By 1997, they had sold the Mac business and were developing software to help computer resellers manage their operations. “The software business was nice,” says de Heer, “but the content about the products was driving the value. We wanted to build a B2B exchange, similar to PCOrder, but we didn’t have the funding. So we thought we would do the dirty work and develop the data, while everyone was building software for exchanges.” It turns out they had the right idea – and their product literature still proudly compares their data-centric approach to other e-commerce vendors’ tool-centric, data-empty vision.

Their company, Global Data Trade SA, developed a substantial database of computer products and started selling it to resellers. “When we started five years ago, we knew the manufacturers wouldn’t want to work with us [because they had little interest in fostering a more transparent market], so we started down the food chain. We knew this was what consumers and resellers needed,” says de Heer.

By 1999, they had a solid little business, providing standardized data on hundreds of thousands of technology products for resellers’ catalogues and online sales channels. Says longtime svp of data operations Consta Zabrodine: “We weren’t thinking of being acquired, but more about an IPO. We already had a small but strong customer base, our data production factory and cash in the bank from a second round of VC funding.”

De Heer went to call on CNET, to see if the publisher could use his data. After two hours, both men recall, the CNET team decided to buy the whole company instead. Says Bonnie: “The main driver of our purchase was our recognition from our own
efforts of two things: how critical normalized data is to producing a useful product and a competitive edge, and...how hard and messy it is to produce.”

The company is now called CNET Channel, with de Heer as president. It is a key part of CNET’s editorial platform as well as a separate data service for third parties, reportedly accounting for more than 5 percent of CNET’s revenues. The business proposition – the use and integration of what it calls Transactive Product Data to support effective marketing and sales – is eloquently displayed in CNET’s own materials: catalogues and white papers for resellers full of product information, reseller testimonials, product comparisons and clear diagrams about how the service works, and even “community content” – rankings of products by volume of RFQs (14 of the 15 most-requested printers are from HP!).

CNET Channel produces data on over 1 million unique SKUS comprising about 50 million attributes, all carefully classified, cross-referenced and ready for slicing and dicing in over 40 markets and 18 languages. (De Heer himself speaks French, English, German and Dutch.) As vendors’ offerings change, CNET builds new SKUs and sends out updated data feeds to its commercial customers, so that they can update their own sites and catalogues in turn.

Naturally, in addition to offering effective sales materials, resellers are also interested in being able to compare prices and terms for the same vendor’s products from different distributors, who sell the same products with different SKU numbers. CNET’s analysts figure out, with the help of tools and with their own expertise for challenging cases, which product codes match which others across resellers. More interesting yet, CNET’s detailed and expert product descriptions allow resellers and their customers to compare equivalent products from different manufacturers (“though we don’t pierce the veil back to which contract manufacturers actually make them for the brand owners,” says de Heer).

Customers now include brand-name vendors such as Dell, HP and Gateway, distributors such as Ingram Micro, Tech Data and Synnex, resellers such as CDW, Insight, Microwarehouse and Buy.com, and portals such as Half.com, Epinions and Yahoo!.
where CNET data is integrated into Yahoo! Shopping. Most manufacturers are also providing direct data feeds, rather than forcing CNET to scrounge around for possibly inaccurate data.

In addition, CNET Channel now has a substantial ASP business (ChannelOnline) providing e-commerce platforms for about 400 computer VARs, including many of Tech Data’s network of resellers. It has fulfilled de Heer’s earlier dream of a B2B exchange, handling about $350 million in B2B transactions last year.

CNET Channel employs about 400 people (out of CNET’s 1500-odd). About 225 of them are programmers and data analysts who work in its Moscow “factory,” managing the painstaking process (industrial data normalization, cleansing, capture, syndication and distribution) of taking manufacturers’ and resellers’ data and properly fitting everything into CNET’s product data model. “We have a six-sigma process for data quality,” says de Heer. “We build SKUs like cars.”

**WAND Inc.: Everyman’s catalogue**

Although most wholesale markets will end up with their own taxonomies, such as CNET’s for computers, there’s also a need for a grand unified taxonomy that can handle almost anything that can be sold outside a closed inner circle.

The existing Universal Product Code system (UCC.net) is a sort of bottom-up approach where everyone lists their own products; because each registered participant has a specific prefix, there’s no duplication (much like the Domain Name System). But the UPC isn’t very helpful in searching for things; it’s simply a registration system for unique products. Any taxonomy or structure to the numbers is the company’s own.

One of the more ambitious efforts in tackling this problem is WAND, Inc. (for World Access Network of Directories), which is gaining some traction with organizations as varied as Bell Canada, IBM and Lycos.

Ross Leher was a practical-minded camera-store owner in Denver when he first came up with the idea for WAND – or it came to him. As he was a well-known local expert in cameras, insurance adjusters started calling on him to value lost or stolen cameras – based only on the claimants’ descriptions. Better yet, he says, “We could match the features to come up with an equivalent product, so that the adjusters could offer a replacement instead of a check. They really loved that, and they started
coming to us with other kinds of products. It become too much information for our salespeople, so we built our own database – Sony tvs, silverware, whatever.” Eventually he began to charge for what used to be for free, and started a data service for customers such as Farmer’s Insurance and Allstate.

By 1998, he modestly decided to develop a directory and catalogue service that would include all products and services from every industry, and sold out his $50-million, 17-store brick-and-mortar business to concentrate on WAND.

WAND’s focus is very clear; helping people to find things – specific things. “I can tell you right now,” he says over the phone, “I bet you’re wearing a textile item of cotton. . .but how useful is that? If you say, ‘I’m wearing a cotton shirt,’ is it a dress shirt, a casual shirt, or. . .? We add the attributes of size, color, style – all the things any real buyer would want to know” – well beyond what you could find in a standard government classification scheme.

People often try to use directories and abandon them, he says; they usually do so because they don’t find what they want – not just what is sold in principle, but what’s in stock. There’s no point in writing to 50 suppliers to find out if they have what you want – even with free e-mail – because the suppliers don’t like replying to those 50 e-mails if you don’t look like a hot prospect. So the good suppliers tend not to congregate where they are likely to be bombarded by e-mails. . .

The WAND taxonomy, he continues, is hand-generated, and the translations are also done by hand. (His favorite error by a competitor: Venetian blinds to “blind people from Venice.”) The WAND taxonomy now contains 65,000-plus items, in a hierarchy as many as 10 levels deep in some sectors.

He points out that much of his competition comes from paper directories or companies with a paper-directory heritage. In that world, companies employ self-defeating strategies: They might list themselves under a broad category, such as stationery, and then get asked for thousands of items they don’t have. But if they list under specific items, they can often get missed. (It’s the same problem companies have been found by search engines: they can either be lumped in with too many competitors in a large category, or miss being listed in a small one.)
The directory is poly-hierarchical. For example, you could find sweaters under textile items of wool, ladies’ tops, and perhaps Christmas gifts. The structure is not just a simple hierarchy, but in fact a complex set of links, synonyms and related terms. You can search by product classification, and also by a large number of attributes, many of them specific to just a few categories (most sweaters don’t have an F-stop number, for example).

Half the challenge is to classify everything; the other half is to keep those categories evolving as products change – last season’s remote-controlled miniature cars, for example. WAND has a feedback system for new categories or subsets, says Leher: “We accept the feedback, and we turn around new product types or synonyms in a week or two.” He looks into his records and cites a recent example: the term “tropical fish flakes,” which a company called LA Pet Connection asked to have added on December 11 of last year. LA Pet Connection got its answer a week later: The term “fish food” already exists, and so the taxonomist added “flakes” as an attribute of fish food. Anyone searching for tropical fish flakes will end up in the “fish food – flakes” subcategory. “We get a about 100 new ones a month,” Leher says. “And of course we keep all the old ones. They don’t retire; they just get resold on eBay.”

WAND’s basic taxonomy is for all commercial goods, anywhere, and is now the power behind more than 15 commercial Websites, including BellZinc, Bell Canada’s portal for small and medium businesses. Its business model varies from customer to customer, but it charges directory publishers initial activation and annual license fees, a revenue share for listing enhancements – the online equivalent of boldface ads – and hosting fees for smaller customers who use WAND’s servers.

WAND also licenses the taxonomy to third parties such as BellZinc, which makes full use of WAND’s taxonomy, with 65,000 categories of products of all kinds. BellZinc has attracted free listings from about 600,000 Canadian companies, and 1.5 million international listings from merchants wanting to reach its mostly Canadian buyer base. “We wanted a simple Internet product for the SME [small/medium enterprise] market,” says Stephane Marceau, vp e-channels at Bell Canada. “We tried online CRM, messaging, hosted Websites and the like, and met with varying degrees of success. SMEs have zero tolerance for complexity. What they really wanted was simple: products and customers. The directory has been a great entry point to doing business on the Net for Canadian SMEs. Next to Net access, it’s the best Net product on
market. And there’s no maintenance or programming. You can even just list your phone or fax number if you have no Website or e-mail. There’s relatively easy uptake, and the traffic continues to grow. The BellZinc site gets about 400,000 unique visitors per month, and about 7,000 of those have already graduated from browsing the listings to more extensive – and paid – services from BellZinc.

Although WAND’s sweet spot is product data, it is now working closely with IBM on IBM’s public UDDI service. Though IBM is focused on “Web services,” those services come from regular companies, some of them offering regular products in addition to software. For example, why not use an UDDI directory to find a component supplier that can understand your mechanical CAD drawings and ship you back a product spec to match your precisely stated needs – not in numbers or prices, but in software-specified design attributes?

That’s the basic idea, anyway. Like everyone in the field, IBM approached Web services fresh a few years ago, and quickly discovered some realities. “When we first announced our Universal Business Registry two years ago,” says Bob Sutor, director, Web services technology at IBM, “we didn’t have it quite right. Anyone can go put their stuff in, and there was no real quality control [on how things were described or their truth]. On the Net, people aren’t sure whom they can trust.” (see RELEASE 1.0, DECEMBER 2002.)

For its public directory, IBM has implemented a WAND directory not within the UDDI directory but as a complementary Web service that can classify companies by the products they offer. That information slips back into IBM’s UDDI directory using SOAP. The “service” is basically a query tool that returns relevant companies and their UDDI listings.

**SMBmeta: Build it and they will come**

Dan Bricklin, co-creator of VisiCalc, is about to make waves in the directory space, not with a hot new product but with a plain little spec: SMBmeta (for Small and Medium Business metadata). He wants small/medium businesses (SMBs) to use it to tag their sites with a small amount of metadata – address, service area, language spoken and the like – that will help them to be found or filtered more effectively.

It’s not an entirely new idea – just like .vcf files or so many other useful things people simply fail to get around to. Nonetheless, Bricklin has a good chance of kicking off a movement: In addition to being Dan Bricklin and very active in the small-developer
community, he is now (by virtue of selling his company) CTO of the largest website-hosting company for SMBs in the world, Interland.

“The spreadsheet was obvious,” says Bricklin. “A lot of people thought of it, but no one put it together until us.” He sees the same no-brainer vision in his latest idea. Indeed, SMBmeta is very simple, and it takes the Open Directory Project one step further: Instead of everyone putting sites into a directory, each site-owner can simply classify his own site, using the SMBmeta format, for any third party directory or search engine to use. This data is placed in a small XML file, with a known name (e.g. smbmeta.xml) and easily findable (by a computer) at the root of the Website.

As shown in figure 3, it is human-readable in that format, but hardly pretty. The idea is for plug-ins and search engines or directories to grab the information to use for filtering or classifying the content, and for displaying the results in more-readable form. “What gets a lot of people excited, though, is that SMBmeta seems so fertile for innovation, and we all just feel that some smart people out there will come up with stuff we’ve never dreamed about to do with it,” says Bricklin. Indeed, he hopes that SMBmeta will encourage more directories, since it will be easier for them to aggregate data.

That sort of thing is happening now with RSS and various news-feed/blog tags. To see the same kind of innovation within a more commercial community would be inspiring – and Bricklin’s aim is to make the tools simple enough to allow this to happen. “Being encouraged to fill out the SMBmeta information makes it more likely that you’ll actually put that information in nice text on a Webpage somewhere,” he adds.

One can immediately think of all the objections, of course: People already use metatags...yes, but mainly by including words like “SEX” to drive traffic, rather than metadata for serious prospective customers. But where’s the quality control? Won’t people lie? Perhaps, but the data is more concerned with specific facts – more like an ID than a resume. (And businesses would be silly to promote products they don’t carry.)
Finally, where’s the directory structure: What classifying scheme should they use? If this all works, sites will be listed in a great variety of directories with different taxonomies; the SMBmeta tags will simply make it easier to search by location, language spoken, availability of transport, opening hours and other straightforward criteria. There are a couple of required fields, including <type> for the US NAICS (for North American Industry Classification System), and also a <description> for a full-text description (no specific length, but you risk being truncated by a search engine). Bricklin figures that NAICS is the best taxonomy for now, but we assume there will be lots of them, including WAND’s (above) and many other more specialized ones. (The NAICS list was last updated last September.)

Bricklin’s vision has two strong advantages. First, it is simple: just a few data items in a structured metatag, simple enough (he claims) for any site-owner to do himself. (And if not, all the search engine consultants and Web-hosting services and domain-name registrars will be happy to help for a small additional fee.)

The second edge that Bricklin has is a channel to the community of small businesses, whose Websites generally aren’t designed to be easily findable. His company Trellix, which makes software tools for Website publishing, promotion, and commerce for individuals and SMBs, was recently acquired by Interland (née Micron), a former pc maker that morphed and rolled-up into the country’s largest Website-hoster for small and medium-size businesses. It hosts more than 200,000 Websites, and Trellix brings many thousands more (though not all commercial). SMBmeta is Bricklin’s first salvo in his mission to dream up new value for Interland customers.

.museum: A gallery of names
Generally, the Domain Name System was not intended as and is not a directory or a search engine, though people often use it that way. It was originally designed merely as a convenient way to list and identify Internet servers. We won’t go into Internet history here, but suffice it to say that the use of “real words” in the DNS is the source of most of its problems, which have to do with the political and commercial (trademark) value of the names.

Nonetheless, there’s one interesting initiative within the formal policy regime of ICANN (Internet Corporation for Assigned Names and Numbers (DISCLOSURE: ESTHER DYSON, FORMERLY CHAIRMAN OF ICANN, IS NOW A MEMBER OF ITS AT-LARGE ADVISORY COMMITTEE) that experiments with a modified form of meaningful naming and some bottom-up self-classification. It’s a welcome sign of the incipient diversity in Top-
Level Domain (TLD) Name policies since the introduction of seven new TLDs by ICANN two years ago; it would be great to see more TLDs, and especially to see more diversity among them.

Currently, most TLDs have flat namespaces, with the names allocated first-come/first-served except for trademarks (which often have too much sway, especially in the case of anti-X sites). There are some two- or more-level hierarchies among “country-code” TLDs, with second-level domains for companies, universities and the like (.co.uk and ac.uk for academic), and others organized geographically (for example, .sf.ca.us or .msk.su).

But you can get more interesting than that, and .museum shows how. .museum amounts to a loose and self-organizing taxonomy that’s not so much hierarchical as cross-tabbed: Each museum gets two identifiers/classifiers, one semi-generic and controlled by .museum, and one unique (in that category, at least) to itself.

The taxonomists classify themselves

In a sense, the museum community is uniquely suited for a structured namespace. Its members’ function in life is to define and catalogue things, so you could consider the .museum namespace just one large virtual museum in which all museums need to be classified and findable. “The museum community has reacted from two different perspectives,” says Cary Karp, who runs .museum from Stockholm. In his day jobs he is director of Internet strategy and technology at the Swedish Museum of Natural History and serves in the same capacity for the International Council of Museums (ICOM). “Some think it’s a cool idea, while others are wondering why we’re imposing control.”

The basic approach is to have a three-level structure, with two levels below .museum: Each museum comes up with a generic word for itself, such as “naturalhistory,” or a location specifier, perhaps a city or a region, and adds something specific so that the combination is unique. For instance, www.historyofscience.oxford.museum. A museum is also free to register multiple names, using as many such combinations as it feels useful.

“We applied our own museological nomenclature skills,” says Karp. “We started as rigid as we felt we could usefully be, and now we’re relaxing constraints. Generally, we don’t allow words such as “the,” but in some contexts they may actually be specifiers, such as ‘the.british.museum.’”
This approach may work because it is a combination of tight rules and freedom within them, for a relatively finite community. It’s also a model for how some light control, melded with choice, can work in a community dominated by good will rather than self-interest.

“Most museums are confused [by .museum],“ says Karp. “They can get .org for almost nothing, so we need to give them something of additional value if .museum is going to establish its long-term viability.” The fees include a one-time qualification fee at a cost of about $100, plus about $80 per domain name. About 2,000 museums have actively responded and they are at various stages in the verification and registration process. The .museum zone file currently contains about 1,700 names and as many again are somewhere in the queue.

The basic value is findability, plus some indication that they meet the criteria defined by ICOM. “Prospective registrants already familiar with domain-naming easily manage the dialogue needed to get what they want” in terms of names, says Karp. “And it’s good for small museums who are now buried through hosting services or part of some other broader institution” (such as www.lcsd.gov.hk/CE/Museum/History/english/index.html, now more easily findable as www.hk.history.museum, or www.christusrex.org/www1/vatican0-Musei.html, still un-.museum’d.

“We’ve had very few frivolous requests,” Karp notes. “Occasionally there’s some agency responsible for local tourism that wants to sell the region on the basis of having museums. We go back to those museums and ask them to tell us it’s their authorized voice, in which case we’d say fine. Beyond that, there is any number of marvelous resources that individuals (and non-museum institutions) have placed on the Web and decided to identify as museums. Very often the people doing so have no understanding of the professional museum community’s notions of what constitutes a museum. Since we can’t give a .museum name solely on the basis of someone claiming to have a virtual museum, we this as a marvelous opportunity to broaden the scope of the museum presence on the Internet, by explaining to folks what they need to do in order to qualify.

“We very rarely say no to an application for inclusion in .museum. We either say yes outright, or say, ‘Here’s what’d you’d need to do to qualify and we’ll keep your application open for a year to give you the opportunity to do it.’” Not everybody is happy
with this, though, and the discussion with the virtual museum community has not been 100 percent friction-free.

“We’re trying to prove that a well-defined concept and reasonable definitions can run smoothly, without a lot of contention.” However, he notes with a smile, “It’s uncertain how the lawyers will react if it doesn’t generate any litigation.”

Overall, .museum recognizes that there’s not a neat taxonomy of kinds of museums: For starters, geography and subject matter are almost orthogonal, and different museums define themselves according to different criteria. But in the real world, where museums simply want to establish their identity and be findable by normal people, this two-part system seems to be working fine.

**Structure and Ontologies**

To end this issue and to point to the next one, we introduce a company that illustrates the overlap or transition from taxonomies to ontologies. The point is that content structure may be in the eyes of the beholder: What you want depends on who you are. This brings us back to that notion of multiple worldviews, fighting with the brittleness of computer systems and the limits of individual understanding. How can you make a system that can be understood by many people — telling each of them something different yet telling all of them the truth?

**Clinician Support Technology: From taxonomy to ontology**

Clinician Support Technology in Newton, MA, is a sequel to Decision Support Technology and Management Support Technology. All three companies were co-founded by Larry Meador, a long-time researcher and consultant on “the strategic use of technology for competitive advantage” (see RELEASE 1.0, OCTOBER 1987). But Clinician Support Technology, founded in 2001, was primarily the outgrowth of work done by Dr. Charles Safran, who shared an office with Meador at MIT in the early 1970s and was a primary care physician and medical informatician at the Center for Clinical Computing at Harvard Medical School. Adept at finding interesting, fundable problems, he had moved on from AIDS to the problems of premature babies in the late 1990s.
Meanwhile, he had stopped practicing as a physician, partly because the average
interaction with a patient had dropped from about half an hour to seven minutes, he
says. “I looked out of my window at [Boston’s] Beth Israel Hospital, with its 550
patients inside and a budget of $600 million, and I started thinking about the 250
million people outside,” he recalls. His goal was to shift care out of the hospital
without compromising quality.

“In health care in general, there are two big issues,” he says. “One is information
about the patient – the patient record. That’s a big and interesting problem. But
equally interesting is the information the patient wants or needs.” In the old days, the
all-seeing, all-knowing doctor knew when a red nose meant a cold
and when it meant too much alcohol. Without much conscious
thought, he (rarely she) could give the right information in the right
doses to the patient. Now, as medical information becomes more
available outside the doctor’s office, patients have no way to apply
expert context to that information.

CST began with a nice $2.6 million grant from the National Library
of Medicine to the Beth Israel Deaconess Medical Center for one of
the more challenging medical specialties – premature infants. It’s an
area where the parents are almost by definition unprepared: The
birth happens early, and the baby is often surrounded by machines.
During a baby’s stay in the neonatal intensive care unit (NICU) the
doctors and nurses get an average of six calls per day from the anxi-
ous parents of each child. The first step the team took was pretty
simple, says Howard Goldberg, vp of product development: Simply
posting each child’s weight each morning on Baby CareLink (password-protected
for each set of parents/caregivers) eliminated a surprising number of the calls.

The overall project was quite successful, but it had a government-grant flavor. The
team provided computers and video hookups to the infants’ families, and the entire
population served was about 200 families.

It’s a long story, but because “the hospital did not see an immediate opportunity for
commercializing technology to support the educational and emotional needs of
patients,” says Safran, it assisted him in transferring the intellectual property to a
newly formed company. While the hospital continues to use Baby CareLink for free
as part of this agreement, it also licenses another CST product to support cancer care.
“Our current baby product has 1,200 topics [times many items] in Spanish and English. Our premise is that if someone is going to read only one thing, they should read the right thing. When you unleash patients on pure searches, they often get more scared. We want to start cascading the information, staging them through the information when they’re prepared to handle it.” Currently, it is a “plain” old text-search-based system, where the user types in a phrase and gets out a list of recommended, ranked articles. There’s also an industry-standard taxonomy of common clinical conditions, diagnostic testing, and therapies that can be used to navigate.

Now the CST team wants to extend its technology to make this complex information simpler to handle. Having discovered how much information there is available, and how ill-equipped many of the caregivers are to use it, they want to become better at deducing what information those parents and caregivers might need at the time that they need it. That’s everything from deducing an infant’s durable equipment needs from its current clinical condition, to knowing the long-term implications of a proposed therapy in leukemia treatment. Yes, perhaps this is something the doctor should be telling them, but he doesn’t have time.

*It’s not about the data; it’s about the patient*

Goldberg’s insight was that the issue was not understanding the data; it was understanding the patient. So instead of a better tool for structuring the data, he looked for something that could better describe the patient and use it to query the educational content base. Creating this machine-readable vignette of the patient requires not a taxonomy, but something closer to an ontology (or the traditional expert system so beloved of researchers and so despised by doctors, who consider themselves the experts) – not what *is* what, but what *follows from* what. The company is now developing an ontology covering neonatal care, including interactions among everything from clinical conditions, diagnostics, and therapies, to weights, timelines, durable equipment, developmental milestones, and outcomes. To do this, the company is working with the Cerebra ontological engine from Network Inference (next issue), designing a tool that will discover just what the parent needs to know.

“Neonatal care is very high-tech,” says Goldberg. “To be an effective parent, you need a PhD in intensive-care medicine, and you need to get that very fast. Our challenge is to help grow a better parent in this very high-tech environment. But ironically, as we add more information and enlarge the search space, it makes the right information harder to find. We want more information, but less noise. But ‘noise’ depends on who’s listening. For example, parents of a 29-week-old infant born at 1000 grams 7
days ago, now on a ventilator, and fed through an intravenous line, should begin to be educated regarding the course of treatment, milestones to look forward to, and risks to the baby, such as the risk of infection. At this stage, it would be inappropriate to broach issues such as bathing or picking a car seat. Subsequently, knowing that the parents had been educated regarding the expected course of events, we can drill down on specific topics in greater detail as the baby progresses or suffers setbacks. In this way, we can enhance parents’ interactions with the health system, even under difficult circumstances,” says Goldberg.

Is this blatant paternalism? Or is it clever progressive disclosure, tailoring information to the patient’s needs?

You can argue about these issues forever, but in the meantime, Baby CareLink results in cost savings, according to CST. It’s a small amount – two fewer days in the NICU plus fewer readmissions – roughly a 10 percent savings overall. That’s a small absolute savings on a base of only 70 infants, but if you got the same ratio of savings against the $18 billion that is spent in the US each year on the care of premature infants, you’d have real money. Furthermore, a peer-reviewed study of the project found that parents and caregivers reported 75 percent fewer problems than normal during the study. . . . There are some things that just can’t be measured in money!

**Conclusion(s)**

Even as search and structure tools get better, there remains a delicate tension between searcher and searchee. The searcher wants the best – even if he can’t define it precisely – or at least he wants to see things in accurate context. By contrast, the searchee wants to be found, and to appear in a context in which it is the best. Transparency makes it easier to compare equivalent things, and to see that they are equivalent, while money sometimes makes it easier for things to be found, whether with ads, enhanced listings or even undisclosed paid listings. All these factors mean that search will never be a simple thing as long as people and companies are jostling to protect their interests in a competitive world. But the tools and services we have described above help to put commodities in their place, and to let all the diverse things in the world find their proper places in it.
Indeed, the companies and initiatives described in this newsletter are an epitome of a directory: a small, pale subset of the full reality, but enhanced (we hope) with an editor’s insight and clarifications. Some people might look at all the different efforts from all the different directions – up, down, sideways – and wonder how we’ll make sense of it all. Can the Semantic Web really emerge from all this confusion?

It will have to, because people are not simply going to give up their own views of the world for someone else’s. (Our favorite illustration of this point is the flower known in England as Sweet William, after William, Duke of Cumberland. North of the border, in Scotland, it is called Stinkin’ Billy.)

Just as the World Wide Web emerged from people making their own links, the Semantic Web will emerge from people making their own taxonomies and ontologies – and other people making tools to reconcile them. The fact that there are multiple taxonomies and ontologies and classification schemes, such as those described here, does not mean that most of them are wrong and we need to wait for the right one to emerge and win. Their proliferation may be a problem in terms of making the world easy to represent, but it is merely a reflection of the reality that the world does not have a single center, any more than one particular language is “true.”

*The Language Question*

Why have we listed the language skills of the founders and executives in the companies we covered in this issue? We have this little theory that we are testing: that people who speak more than one language are more likely than most to get the notion that there are many ways to understand the world. Language itself imposes a point of view, specifies categories and defines the world – something you may not notice unless you are multi-lingual. Of course, being the child of a divorce may be equally useful for creating such awareness. . .but no, we didn’t ask!
Resources & Contact Information

Gerry Campbell, **AOL**, campbellg01@aol.com
Stephane Marceau, **Bell Canada/BellZinc**, 1 (514) 933-3920; stephane.marceau@bell.ca
Charlie Safran, **Clinician Support Technology**, 1 (617) 614-2600 x123; fax, 1 (617) 614-2525; csafran@cstlink.com
Howard Goldberg, **Clinician Support Technology**, 1 (617) 614-2600 x104; fax, 1 (617) 614-2525; hgoldberg@cstlink.com
Albert de Heer, **CNET Channel**, +41 (21) 943-0356; fax, +41 (21) 943-0369; albert@cnet.com
Shelby Bonnie, **CNET Networks**, 1 (415) 344-2486; fax, 1 (415) 344-1234; shelby@cnet.com
Bob Keating, **DMOZ: Open Directory Project**, keating@netscape.com
Lionel Baraban, Peter Katz, **Entopia**, 1 (650) 632-0101; fax, 1 (650) 802-6709; pkatz@entopia.com
Craig Nevill-Manning, Peter Norvig, Craig Silverstein, **Google**, craigm@google.com; pnorvig@google.com; csilverstein@google.com
Paul Hawken, **Groxis**, 1 (415) 332-6990; fax, (415) 331-0556; paul@groxis.com
RJ Pittman, **Groxis**, 1 (415) 331-0555; fax, (415) 331-0556; rj@groxis.com
Bob Sutor, **IBM**, 1 (585) 243-2445; fax, 1 (914) 766-1834; sutor@us.ibm.com
Dan Bricklin, **Interland/Trellix**, (978) 318-7201; fax, (978) 318-7296; danb@trellix.com
Kim Cameron, **Microsoft**, kcameron@microsoft.com
Kevin O'Brien, **Sector Data/Gradience**, kobrien@sectordata.com
Ross Leher, **WAND Inc.**, 1 (303) 623-7716; fax, 1 (303) 893-1574; rleher@wand.com
Srinija Srinivasan, **Yahoo!**, 1 (408) 349-3322; fax, (408) 349-5101; srinija@yahoo-inc.com

For further reading:

NAICS taxonomy at http://www.census.gov/epcd/naics02/naicod02.htm
SMBmeta tag, at http://www.softwaregarden.com/smbmeta.xml
For a list second-level names currently used in .museum, visit http://anythingyoudarnplease.museum
See also www.w3.org/2001/sw (despite the URL, this is up to date)

Correction:

In last month's issue, we neglected to include contact information for the co-authors, **John Hagel** (john@john-hagel.com) and **John Seely Brown** (jsb@parc.com). You can also find a rich collection of Hagel's and Brown's writings on John Hagel's website, www.johnhagel.com.

We regret the error!
Calendar of High-Tech Events

2003

FEBRUARY 16-18  Demo 2003 – Phoenix, AZ. Come play with the latest gadgets from the newest startups. Register online or contact Lavayne Harris at 1 (800) 633-4312 or 1 (650) 577-2700 (outside the US), or via email at registrar@idgexecforums.com. www.idgexecforums.com/demo/

FEBRUARY 17-21  3GSM World Congress – Cannes, France. The world’s biggest mobile communications show. Featuring Esther Dyson as a keynote speaker. Register online or contact Tamara James, 44 (1932) 893-853; fax, 44 (1932) 893-894; 3gsm@informa.com; www.3gsmworldcongress.com

FEBRUARY 24-MARCH 1  TED2003: "Rebirth" – Monterey, CA. This will be the first TED not run by Richard Saul Wurman, but the new organizers promise the same experience: "Electricity for the Brain... Tools for the Mind... Food for the Soul.” Contact Chris Anderson, 1 (650) 851-6464; fax, 1 (650) 851-9172; chris@ted.com. www.ted.com

MARCH 7-11  SXSW Interactive – Austin, TX. In its tenth year, this offshoot of the South-by-Southwest independent film festival draws cutting-edge new media developers and content creators. Register online, or for more information call 1 (512) 467-7979; fax, 1 (512) 451-0754; email, inter@sxsw.com. sxsw.com/interactive/

MARCH 12-19  CeBIT – Hannover, Germany. The biggest technology event in Europe, officially dubbed the World Business Fair for Office Automation, Information Technology and Telecommunications. Call +49 (511) 89-0 or visit the website at cebit.de. www.cebit.de/homepage_e?channel=1

MARCH 17-19  CTIA Wireless 2003 – New Orleans, LA. This year’s keynote speakers at this annual wireless convention include Ted Turner, Paul Otellini and Michael Powell. Register online or call 1 (301) 694-5243; fax, 1 (301) 694-5124. www.wireless2003.com

MARCH 23-25  PC Forum – Scottsdale, AZ. EDventure’s premier conference...25 years and running! Our theme this year is "Who? What? Where? Data comes alive." Confirmed speakers include this issue’s Shelby Bonnie (CNET), Sergey Brin (Google), Larry Ellison (Oracle), Jonathan Miller (America Online), John McKinley (Merrill Lynch) and Kevin Turner (Sam’s Club/Wal-Mart). Registration is now open on our Website. For more information contact Daphne Kis, 1 (212) 924-8800; fax, 1 (212) 924-0240; email daphne@edventure.com. www.edventure.com/pcforum/

---

E  Events Esther plans to attend.
D  Events Daphne plans to attend.

Lack of a symbol is no indication of lack of merit. The full, current calendar is available on our Website, www.edventure.com. Please contact Christina Koukkos (christina@edventure.com) to let us know about other events we should include.
MARCH 31-APRIL 2  InfoWorld CTO Forum – Boston, MA. Truly a technology conference! Speakers include Ray Ozzie, Greg Papadopoulos and Bob Sutor. To request an invitation, visit the website: ctoforum.infoworld.com

APRIL 1-4 CFP2003 – New York, NY. The Thirteenth Annual Conference on Computers, Freedom & Privacy. This year, the focus will be on the freedom to move, think and speak. Register online or by fax, (407) 366-4138. www.cfp2003.org

APRIL 27-MAY 2 NetWorld + Interop – Las Vegas, NV. Network with the networking community. For more information, call 1-888-886-4057; fax, (650) 372-7000. www.interop.com/lasvegas2003/

MAY 2 Good Experience Live – New York, NY. Organized by Mark Hurst, founder of customer experience consulting firm Creative Good. GEL will gather a diverse set of speakers to explore what it means to create a good, meaningful, or authentic experience. Register online or email updates@goodexperience.com. www.goodexperience.com/gel

MAY 4-7 CIO Forum Financial Services – New York, NY. Strategic IT forum for the US financial services industry. Presented by Richmond Events. This year’s event takes place on board P&O’s newest ocean liner Adonia, sailing from New York City. For information, visit the website or call 1 (212) 651-8700; fax, 1 (212) 651-8701. www.cioforum.com

MAY 13-16 GigaWorld IT Forum – Phoenix, AZ. Giga Information Group’s flagship event, addressing the issues facing managers of technology. Register online or call 1 (781) 792-2669; events@gigaweb.com. www.gigaworldus.com

MAY 15-16 TV Meets the Web – Amsterdam, The Netherlands. This year’s theme is "Digital Media: The Path to Profitability" and will cover video on demand, SMS TV, content billing, DRM, and more. Contact Vanessa Vigar, vanessa@vandusseldorp.com or +31 (20) 535-6979. www.tvmeetstheweb.com/2003/

MAY 18-20 Vortex 2003 – Dana Point, CA. An invitation-only event where executives from the telecom, Internet and data-networking industries gather to discuss the future of networking. Request an invitation online at www.idgexecforums.com/vortex/register.html. www.idgexecforums.com/vortex/

MAY 20-24 Twelfth Annual World Wide Web Conference – Budapest, Hungary. Discuss the latest developments in web technology and the issues and challenges facing the web community. For more information, visit the website or email info@www2003.org. www2003.org

JUNE 11-13 TedMed3 – Philadelphia, PA. Discover how technology can help you achieve a healthier life. Imagine! Register online or call (401) 848-2299; email, wurmanrs@aol.com. www.tedmed.com
Release 1.0 Subscription Form

Complete this form and join the other industry executives who regularly rely on Release 1.0 to stay ahead of the headlines. Or if you wish, you can also subscribe online at www.release1-0.com.

Your annual Release 1.0 subscription costs $795 per year ($850 outside the US, Canada and Mexico), and includes both the print and electronic versions of 11 monthly issues; 25% off the cover price when you order from our online archives; a Release 1.0 binder; the bound transcript of this year’s PC Forum (a $300 value) and an invitation to next year’s PC Forum.

NAME __________________________________________
TITLE ___________________________ COMPANY ________________________
ADDRESS ________________________________________
CITY __________________________ STATE ______ ZIP ______ COUNTRY ___________
TELEPHONE __________________________ FAX __________________________
EMAIL* ___________________________________________ URL _______________________

*personal email address required for electronic access.

☐ My colleagues should read Release 1.0, too!
Send me information about multiple copy subscriptions and electronic site licenses.

☐ Check enclosed   ☐ Charge my (circle one): AMERICAN EXPRESS MASTER CARD VISA

CARD NUMBER ___________________________ EXPIRATION DATE ___________
NAME AND BILLING ADDRESS ______________________________________
SIGNATURE ______________________________________

Please fax this form to Natasha Felshman at 1 (212) 924-0240.

Payment must be included with this form. Your satisfaction is guaranteed or your money back.

If you wish to pay by check, please mail this form with payment to: EDventure Holdings, 104 Fifth Avenue, 20th Floor, New York, NY 10011, USA. If you have any questions, please call us at 1 (212) 924-8800; email us@edventure.com; www.edventure.com.