Arise, librarians: Today the world is yours

Ian H. Witten
Department of Computer Science, University of Waikato, New Zealand
September 2005

In the mid 1950s the field of linguistics suddenly shot into the academic limelight. In the 19th and early 20th centuries it was a dusty corner of the humanities, entered by would-be scholars through a rite of passage that involved journeying to a far-flung corner of the known universe, locating an obscure and heretofore unknown language, conquering it, and returning with a detailed record of its characteristics. Though they respected these endeavours, no one in other disciplines cared much about them from an intellectual point of view. But in 1957 everything changed with the publication of a book entitled *Syntactic Structures* by Noam Chomsky, a young man still in his twenties. He postulated a universal basis for language (called deep structure) and a series of generative rules (called transformations) that act on the deep structure to produce the complex sentences that we use for everyday communication. All languages have the same deep structure, which is innate in human beings. Languages differ from one another in surface structure because of the application of different rules for transformation, pronunciation, and word insertion. This revolutionary and completely unexpected hypothesis promised to unlock the secrets of the mind. It catapulted linguistics from a sort of academic marginalia into the very centre of the study of the brain, cognition, and the mind-body problem. It was as if the sun had suddenly and unexpectedly risen over the field.

Today the sun is rising for librarians.

The woes of the profession

But first, the downside. We stand at the epicentre of a revolution in how our society creates, organizes, locates, presents, and preserves information. Librarians are without doubt among those whose working lives are most radically affected by the tremendous explosion in networked information sparked by the Internet. Advances in information technology generate shockwave after shockwave of opportunities and problems that seem like a fierce and long-sustained onslaught on librarians' own self-image. How are you supposed to react when a technology juggernaut like Google suddenly declares that its mission is something that you thought society had entrusted to you alone, and which you had been doing well for centuries ("to organize the world's information and make it universally accessible and useful")?

In the brave new world of digital resources, what are librarians to do? How can they cope with massive changes in a time of shrinking budgets? How can they swim against a current of popular opinion that questions why on earth, in today's culture of information overload, we need libraries at all?

To take just one example of change, history will view the move from owning information to renting it—a trend that strikes at the heart of the librarian's customary way of doing business—as a tragedy on a par with the burning of the great library of Alexandria. Today, publishers rarely sell digital copies of scholarly research—they license them instead. The Digital Millennium Copyright Act in the US, inspired by a draconian interpretation of the World Intellectual Property Copyright Treaty, enshrines in law a stark and sometimes brutal distinction between licensing and ownership. The effect on libraries—and society—is immense. A distasteful sea change in the librarian's profession is underway, from information-enabler to restriction-enforcer. Many brave souls are standing up and speaking out in favour of revoking this pact with the devil by transforming digital content licenses from lease to sale.

Into the dark world of commercially restricted information the open access movement shines a warm and beckoning light. That scholarly research so jealously guarded by publishers was donated to them by publicly funded academics! Why not cut out the middleman? But open access (which is not necessarily the same as cost-free access) has its problems too: notably the difficulty of encouraging harried academics to electronically self-publish their work in appropriate and standard ways (the current buzzword is "interoperable"). Here is yet another new job for overstretched librarians.

Then there is the challenge of combining access to traditional library material (owned), on-line journals (licensed), and information on the Web (freely available) in a one-stop shop that recognizes the different legal status of the material. Many difficult issues arise, such as how to reconcile site licensing with walk-in library patrons—which incidentally raises the spectre of what will happen to today's students when they graduate and find the library's doors metaphorically closed by licensing restrictions. Some libraries even aspire to service their patrons' e-books and personal digital assistants, which raises legal questions of delivery models and technical ones of expansion-card compatibility. Countless new roles are being forced upon our hapless librarian.

Who'd be a librarian today? It could only be someone who yearns to be where the intellectual action is and relishes the impossible challenge of dealing with an onslaught of novelty. But we will argue that tomorrow's world belongs to those capable of organizing information, who understand the importance of accurate metadata and mechanisms like classification systems, authority files, controlled vocabularies, subject headings, and thesaurus structures—in other words, to librarians.

My laptop as a library

I am writing this essay on my laptop. I have just checked: the file I am writing is one of 150,000 others. These are only my personal files—the total number of files on the disk, including system files and those of other users, is far larger. My computer, now nearly three years old, is not a particularly powerful one. In fact, I periodically back it up on my iPod personal music player (which I bought a few months ago when my child showed me hers: I was envious). As well as all 150,000 files, my iPod contains many thousands of tunes, and many thousands of photographs—and it's only one-third full.

In terms of sheer quantity, 150,000 documents constitute a fair-sized library, far larger than all but the most immense personal libraries of yore. In the 17th century, the world's largest collection was Duke August's of Wolfenbüttel, Germany: it reached 135,000 imprints by his death in 1666 and was acclaimed as the eighth wonder of the world. Even the great library of Alexandria was not very much larger than my laptop "library." At its zenith it boasted 700,000 volumes, and more than 2000 years would elapse before any other library attained this size—notwithstanding technological innovations such as the printing press. However, if digitised, it would fit on my portable music player.

Unfortunately, the standard of organization of my files falls woefully short of the standard set by even the most primitive of libraries. Indeed, the word “organization” is a misnomer. Although I am a fairly tidy person and take pains to keep my workspace clean and usable, my file space is haphazard. The only trace of organization is the folder structure. Along with my files are 10,000 folders into which they are placed. Folders are hierarchical, and most are buried deep within the file structure. The top-level folders are like rooms of a great mansion. Superficially the house looks tidy, but every door opens to reveal a disgusting mess. Let’s peek into a few.

One room holds all my email. It is organized into folders that correspond approximately to topics, most of which have subfolders. Compared to the other rooms it is relatively well structured, because I work with email every day. Nevertheless, there are various subsubfolders called *misc* containing messages that I can’t be bothered to classify, and others called *old* that are probably obsolete but I can’t risk deleting. Many of my emails contain vital attachments. Most relate to projects that have entrails in other parts of my file structure, but there are no explicit links—I have to try to remember where things are.

Another room (I’m not sure it’s a folder, it may be a single file) contains my web bookmarks. Unlike many of my colleagues I have few of these, and they are organized as a flat list, not hierarchically. Again, most relate to projects stored elsewhere, though there are no links except those in my head.

Several other rooms represent top-level folders in my file structure. Some are covered in cobwebs: they haven’t been used for years. Others are used only sporadically. On opening the door to some of these rooms it would take considerable effort for me to recall what they contain, and what they are for. Many are lingering vestiges of my academic past. Any given project has information in several widely scattered rooms: some file folders, some related emails, and some related bookmarks.

My desktop is like another room. Mostly it contains links into other parts of my file structure. However, on it are several folders in the “too hard to file” category. Really I should clean up the desktop and file these. But some have been there for months and the fact is that I haven’t got round to dealing with them. Perhaps I never will.

My laptop is three years old and most of my files are younger than this. However, there is an embarrassing room called *legacy files*. When I get a new laptop, I will not sort through my files, I’ll just copy them, lock stock and barrel, into a “legacy” folder on my new computer. That’s what I did when I bought this computer, and my legacy folder here contains within it a legacy folder for my previous computer. I have no idea what many of these files contain: it would be a major effort for me to look into them and clean them up. I have no intention of doing so.

The web as a library

My laptop spends most of its life connected to the Internet. In my mind I regard the Internet as a massive extension of my readable file space. However, in practice I am forced to use completely different tools to access it, and I have to worry about explicitly and painstakingly downloading some parts that I may need now or in the future.

The level of “organization”—if you can call it that—of the web is appalling: not so different from my laptop. The web divides into sites that resemble my top-level folder—except that there are hundreds of millions of them. Each site has a hierarchical structure of web pages that is evident from the URLs within it. And, of course, unlike my files, there are countless explicit hyperlinks between different web pages.

How big is the web? At the time of writing, Google, the predominant web search engine, claims to index 8 billion documents, totalling 25–30 terabytes. Big as it is, Google by no means covers the whole web, and it would be reasonable to increase this by a factor of say three, approaching 100 terabytes. Four years ago (half Google’s life, or an age in Internet time) the Internet Archive's collection of online data contained about 10 billion web pages—over 100 terabytes of data and growing. But like an iceberg most of the web lurks beneath the surface. The “hidden web” can only be retrieved through information retrieval services rather than by clicking hyperlinks. Some parts are open; others are password protected. It’s anybody’s guess how much is hidden there, but it’s certainly many times the open web.

To visualize these gargantuan figures, let’s compare the web with the world’s treasury of literature, in terms of size—we are not talking about quality here. The US Library of Congress has 30 million books. A smallish book contains 100,000 words, or nearly 700,000 characters—700,000 bytes of text in uncompressed textual form (compression might reduce it to 25% of that size without sacrificing any accuracy). Illustrations increase this substantially, depending on how they are stored, but let’s consider the words alone. Suppose the average book in the Library of Congress is just under a megabyte—say 2/3 MB. That makes a total of 20 terabytes for the textual content of the Library of Congress—or 20,000 copies of the Encyclopaedia Britannica. Of course, the Library of Congress certainly doesn’t contain everything. It was estimated in 1975 that some 50 million books had been published up to that time. Perhaps we should multiply this by three to account for the books published since?—that’s five times the Library of Congress, 100 terabytes. About the size of the web.

What about all the information produced on computers everywhere, not just the web? It took two centuries to fill the Library of Congress, but it has been estimated that today’s world takes about 15 minutes to churn out an equivalent amount of new digital information, stored on print, film, magnetic and optical media. Over 90% of this is stored on ordinary hard disks. And the amount doubles every three years.

You might take issue with all these individual figures. They’re rough and ready, probably out by a large factor. But in today’s exponentially growing world large factors are overcome very quickly. The library of Alexandria, with 700,000 items, can fit on a teenager’s portable digital music player. Wait five years until storage has improved by a factor of 30, and the Library of Congress will fit there. Wait another couple to store the world’s entire literature. What’s a factor of 10, or 100, compared with exponential growth? The web has reached the point where it dwarfs its entire ancestry.

What are the prospects for actually putting all our literature on the web? Well, storage is not a problem. And neither is cost, really. To show page images, backed up by the kind of low-accuracy searchable text that automatic OCR can produce, might cost $10 per book. The 30 million books in Library of Congress would cost $300 million. That’s only half the Library's annual budget.

The problem, of course, is copyright. There are three broad classes of material: works that are in the public domain, commercially viable works currently in print and being sold by publishers, and works still under copyright that are not being commercially exploited. Current projects are digitising material in all three of these areas. But be warned: things are moving very quickly. Many radical new developments will have occurred by the time you read these pages.

Finding things in these libraries

When people access the web, they invariably start with a search engine. Search engines are the Internet’s “killer app.” A recent survey found that 73% of Internet users had used one in the previous four weeks, compared with 68% for email. For me (and most of my colleagues), Google automatically appears when my web browser is invoked.

Full-text search is an embodiment of the classical printed concordance, with the advantage that, being fully computerized, it works for all documents, no matter how banal—not just sacred texts and outstanding works of literature. Multiword queries are answered by combining concordance entries and ranking the results, automatically according rare words more value than commonplace ones. Web search services augment full-text search with the notion of the *authority* of a source, estimating this mechanically on the basis of the number of web pages citing that source, and *their* authority—in effect weighting popular works highly. Efficient and effective ways of searching through immense tracts of text is one of the most striking technical advances of the last decade or two. And today’s search engines do it for free.

Search engines developed extraordinarily rapidly. The web began in 1993 and the first full-text search facility appeared in 1994. The idea caught on like wildfire. By the next year I (along with most other academics in my field) was using a search engine regularly. A decade later the entire world was doing the same thing. The web became so huge so quickly that search is the only way of dealing with its immensity.

Surprisingly, this is not yet the predominant means of finding things in one’s file space. Though searching has been built into popular operating systems for years, it’s slow and clunky. Paradoxically, it’s easier to find something on the web than on my computer, and when I know that the information I need is in both places I seek it amongst the billions of documents on the web instead of on my own desktop!

After a long period of totally inadequate searching tools for personal computers, efficient and effective search at last became available with the development of Google’s “desktop search”: they advertise it as making it as easy to search your own file space as it is to search the web. Faced with leadership from a third party, major computer vendors—Windows, Macintosh—are rapidly following suit and incorporating comprehensive search facilities into the latest versions of their operating systems.

Historical development of file-space organization

Until you can search your file space efficiently, you must rely on your ability to organize items into the hierarchical folder structure and remember where they are. With my 10,000 folders, this is patently impossible. But even this rudimentary facility has not always been in place. Here is a broad-brush history of information processing, in 17-year chunks.

The first stored-program computer was unveiled around 1948, although most scientists did not have practical access to computers until the 1960s. In those days, users did not employ a file store at all. They submitted their job on punched cards or paper tape, and could not leave anything on the computer from one run to the next. As a humble graduate student in the late 1960s I had no files on my university's computers; neither did most faculty. Those lucky enough to warrant their own disk quota were given a directory in which they could retain a few files. You could list the files and look at them, but there was no hierarchy and no tools for organizing your file space.

Seventeen years after the first computer, around 1965, the hierarchical file store was born. It was implemented in a specialized computer research lab at MIT on an advanced timesharing system called Multics; just one of several revolutionary innovations that project produced. Of course, only a select few researchers knew about it. Eventually Unix, a mini version of Multics, copied the idea of a file hierarchy. Unix came out in the first half of the 1970s and was widely adopted. I myself used it for the first time in 1976. Like others I was captivated by the fact that directories could contain other directories, and so on ad infinitum—though I had to list their contents and painstakingly explore the hierarchy using arcane typed commands. To me this seemed like the ultimate tool for organizing information. Who could ask for anything more?

Seventeen years later, around 1982, researchers at Xerox PARC, the Mecca of interactive computing, invented the metaphor of folders. They developed the Xerox Star computer, the first to have a modern desktop-style graphical user interface. Again it was confined to the research lab and only a few lucky cognoscenti had access. Years later, after waiting in vain for Xerox to come out with a viable product, the desktop and folder metaphors were popularized by early Macintosh computers. I first experienced this in the late 1980s. While the file structure resembled Unix's hierarchy, you could now click around it instead of issuing typed commands. The idea that files could be placed into folders, which could be put into other folders, seemed like a great advance. It was almost as good as a filing cabinet!

After another seventeen years, around 1999, the World-Wide Web was in full swing and full-text search had established itself as the way to find things. Full-text search was practically unknown until the early 1990s, because it stretched the limits of available computer power. Today it is ubiquitous on the web, and is finally beginning to find its way onto our computer desktops too. When Google introduced their e-mail system in 2004, they dispensed with the traditional mail folder metaphor. Instead they let you tag emails with arbitrary labels, giving a non-hierarchical organization that can associate several different categories with a single object. And, of course, you can search: the main way of finding things. We belong to the era of full-text search.

How will we be finding information seventeen years later, in 2016? By then librarians will have entered the scene and helped us make sense of it.

Search is not enough

Search is the only practical way of finding things on the web. But with commercialization, search is falling apart at the seams. Access to information that is in principle freely available is in practice mediated by centralized, quasi-monopolistic search engines that evaluate web pages to determine what to present. Contextual evaluation is based on the court of public opinion, which encourages spamming by artificial communities that construct networks expressly designed to promote certain products. This means that although the broad principles are public, the precise mechanisms that search engines use to evaluate information are necessarily kept secret. This has led to a speculative "bubble" of web visibility, and a great deal of speculation about the operational details of search engines.

We have little idea how the tools we use to find information work; how they select information for us to see. And not just because it's a commercial secret—though it is. Even if a search company were willing to tell us exactly how its service worked, making that information public would open it up to limitless spam.

These issues will soon transcend the web as we know it today. Academic libraries are engaged in putting information such as their own special collections on the web, and, in the spirit of librarianship, giving as much free access to them as copyright law permits. Initiatives such as the Gutenberg Project and the US, China and India Million Book Project are striving to create open libraries of out-of-copyright material. Commercial web giants such as Amazon present actual pages from books offered by publishers—though access is restricted according to the "fair use" principle of copyright law and subject to a publisher-approved page-viewing limit. Google has partnered with major libraries to digitise their collections and make them searchable. All these initiatives are in rapid flux, and new ones will emerge from the turmoil, but there is no doubt that we are witnessing a radical convergence of online and print information, and of commercial and non-commercial sources.

Traditional libraries cannot provide full-text search of their holdings, because most of what they have is on paper. But any librarian will tell you that even if they could, it would certainly not replace the traditional method of access via the library catalogue and other library finding aids. Search is simply not enough.

Foraging for metadata

Librarians know that the way to find information quickly, reliably and accurately is through carefully prepared metadata. Libraries have put vast human resources into the production of metadata. There are reported to be a billion MARC records in existence worldwide, each one taking an average of two hours to produce. Two billion hours is over one million person-years, a stupendous investment.

The content of my laptop's file system will never receive such loving attention—nor does it deserve to. However, there is much metadata that could be extracted automatically and used to organize the file space, providing an alternative view to manually assigned folders on the one hand and automatic full-text search on the other. All files have a name, type, position in the file hierarchy, creation date, last modified date, and last-read date. Filenames are often indicative of content (particularly for pictures). Titles can be extracted from most text files, more or less accurately depending on the type of the file—and as a last resort the first few characters can serve as a surrogate title. Many common file types contain their own metadata.

Documents in Word format have “properties” that include title, subject, author, keywords—though these are rarely used in practice simply because there is no point: their values are ignored by current software. Some picture files (TIFF) and audio files (MP3) contain metadata. E-mails include sender, receiver, subject, and date fields.

At the other extreme, web pages at large will never receive carefully assigned metadata either—although efforts such as the Semantic Web may go some way towards this. But as with personal file systems, some metadata can be derived automatically. HTML documents have titles, dates, filenames, URLs—and some express more extensive metadata in HTML Meta tags. Other file types also contain metadata, as noted above.

Metadata can also be extracted from the raw textual content of files. Heuristics can be used to identify the title, author, date of writing, author’s affiliation. There are reliable ways of detecting the language in which the document is written. “Entity extraction” techniques can be used to estimate its geographical and temporal coverage. Statistical and linguistic techniques can be deployed to extract key terms that describe the document, or produce a structured list of all its salient phrases. Summaries or abstracts can be generated automatically. Automatic techniques are nowhere near as reliable as manual metadata extraction, but they are far cheaper and can be applied on a gargantuan scale.

Web communities

Metadata gleaned by foraging existing sources and extracting features automatically from text has important applications in helping to organize information. But however hard we try, it will never compete with high-quality manually assigned metadata. The next generation of solutions to the problem of organizing and finding information will arise from a layer of latent organization situated somewhere between my laptop and the web as a whole: the community.

Communities already play a formative role in mediating access to information. Take the university, one of many professional communities of which I am a member. Through its library, my institution subscribes to several online publication databases. When my laptop is connected through the university network, I automatically gain access to the documents in these databases. Mechanisms that used to be awkward five or ten years ago are gradually becoming refined to the point where they operate invisibly. I locate a document on the web that is not open access, and, if I am working through my institution, it is automatically available to me without any further layer of authentication.

The production of metadata is a community affair. High-quality metadata can only be assigned by individuals, but the amount that any one person can do is severely limited. It is communities that organize and collate the work of individual people into a comprehensive and worthwhile body of metadata. Communities are of many kinds: academic, social, recreational, professional, family, artistic. Only community members care enough about the documents that fall within their purview to gather together the resources to index and catalogue them properly. Some communities involve professional librarians; others utilize dedicated amateurs.

We noted earlier that full-text search is beginning to fall apart through the artificial manipulation of search results—spam. There can be no universal solution to the problem of spam, for it requires identifying socially undesirable behaviour, which is only defined relative to social norms. This in itself will force people to form communities and restrict their searches to within their boundaries. Community membership schemes will be devised that exclude spammers, and quickly identify and eliminate any infiltrators that find ways to join. Search will become a community affair.

The sunrise

Information access in the world today is broken. From my laptop to the web at large, the mechanisms that are available—hierarchical folder organization and full-text search—are woefully inadequate to cope with what we ask them to do. The result is frustration and loss of information. The fantastic power of computers and networks has been used to create an unselective, disorganized, and unmaintainable world in which all information is available in some theoretical sense. In principle it is egalitarian; in practice it is almost unusable if high-quality information is what you seek.

Selection, organization, and maintenance are precisely what librarians are trained to do. If "data" is characterized as recorded facts, then "information" is the set of patterns, or expectations, that underlie the data. You could go on to define "knowledge" as the accumulation of your set of expectations, and "wisdom" as the value attached to knowledge. Information is not created equal; it is wisdom that librarians put into the library by selecting material, organizing it, and maintaining that organization as the collection grows.

Today we desperately need librarians to help us put the wisdom back into information collections. We need them to show the way. To do so, they must be open-minded and able to work comfortably and effectively in the new information environment. Here are ten concrete suggestions.

1. In what ways can automatically extracted metadata be used most effectively?
2. How can its quality be improved by using existing web resources as authority files?
3. Are there new kinds of metadata that can be gleaned automatically?
4. Can web communities be organized so that trust is maintained within them?
5. How can communities be recruited to produce high-quality manual metadata?
6. How can ordinary people be trained to undertake metadata assignment?
7. Can free sharing of metadata be reconciled with quality control?
8. How can the benefits of institutional access to copyrighted information be assimilated into a community-based information support system?
9. Can information from minority groups receive appropriate visibility in a world where the mass market dominates?
10. How can information quality be reconciled with an egalitarian approach to information availability?

These are tough questions. There are no standard answers—perhaps no entirely satisfactory answers at all—and ingenious and imaginative thinking is needed to address them. The people most likely to make progress with them are those trained to think clearly about information and access. Librarians, arise: the world needs you.