Legal information and the Internet

– Experiences and challenges

Content

LEGAL INFORMATION AND THE INTERNET

Introduction: Legal Information and the Internet

Peter Seipel, The IT Law Observatory:

30 years ago the so called "System for legislation and case law" had just begun its services in Sweden. It was only available to a limited number of public authorities, mainly the ones active in the criminal law sector. The thought that electronic legal information could ever become a major resource for legal work was distant, to say the least. There were even many people who ridiculed it.

How different the situation today. Certainly, legal text databases had developed over the years but the growth of the usage was slow rather than steep. The suddenly there was the great Internet leap. It meant three things that have become increasingly visible:

– Widespread access to the technology
– Simplicity and ease of use
– A change of attitude towards electronic information retrieval. Today, it is obvious that the Internet is on its way to becoming something of an information backbone for legal work and the administration of justice.

But there are problems and uncertainties. Many of them have existed for a long time. Let me mention some:

– To what extent should legal information be made available free of charge to the citizens?
– What tasks ought to be performed by the public sector and what tasks by the private sector?
– What kind of co-ordination and central steering are desirable and for what ends?

In Sweden these issues have been discussed since the development of electronic information resources first begun in the late 1960s. In recent years the Government ICT Commission has involved itself in the issues. Among other things it has arranged conferences and offered advice on policies etc. Since the early 1980s the Foundation for Legal Information has focused all its work on the issues. Among other things, the Foundation has taken an initiative to develop basic legal data quality requirements. In academia the Swedish Law and Informatics Research Institute has been actively involved in matters of electronic legal information since its creation in 1968.

Now these three institutions have joined forces to arrange yet another conference, this time on legal information and the Internet.

There are three aims behind the initiative.

Firstly to open the minds of both developers and users of electronic legal information. For this reason, experts from the outside have been invited to give their perspective on ongoing developments and main issues.

We welcome, Tom Bruce from the US who will participate via videoconferenc. We welcome also Andrew Mowbray from Australia and Henric Stjernquist from the European Union. All three are recognised experts in the field, experts who also have substantial practical experience.

Secondly, the aim is to inform about ongoing work in Sweden, not least the important project titled "the Legal citation", in Swedish the "Lagrummet". Mr Christian Levander will assist us and describe the principles that guide the design of an important part of electronic legal source materials in Sweden.

Last but not least, the conference will leave plenty of room for discussion. We hope that this discussion will be inspired by the viewpoints and thoughts of our invited speakers and we hope that it will be lively. There aren't so many occasions of this kind and I urge the participants to be active this afternoon and not let the panel decide what is to be discussed and what answers accepted. It is essential that the viewpoints of both the developers and the users of electronic legal information are put forward. Use the occasion.

The whole session is a highly open event and, as I take it you have noted, our deliberations are broadcast on the Internet. We will also store the recording of the conference in an electronic archive from where it can be accessed, viewed and listened to in the future. How I wish that something similar had been possible at some of the early conferences on electronic legal information in the 1970s.

We also aim at some kind of traditional, written documentation and, why not, also a set of recommendations to be considered by the ICT Commission.

So, I conclude this introduction by wishing us all a successful, active and thoughtful conference. Again, use the opportunity, put forward your viewpoints, pose questions. We are all involved in restructuring important parts of the legal order.

Legal information and the Internet in USA

Tom Bruce, Co-director of the Legal Information Institute of Cornell University Law School

> *You should be able to read a building. It should be what it does.*
> *Richard Rogers (1990)*

> *Well building hath three conditions.*
> *Commodity, firmness, and delight.*
> *Henry Wotton (1624)*

I'd like to begin today with quotes from two architects, because architecture is in some sense what we are about here, and because what I have to offer are in some sense architectural observations.

I suspect that like a lot of architectural tours this one will suffer a bit from uneven focus. I may linger a little too long over an ornament or a special technique or to spend too little time on an important archway or colonnade or some other feature of a building. And it is certainly true that I only imperfectly understand the relationship of purpose to design in the legal information environment. But there is more going on here than that simple analogy.

Wotton, in the quote above, speaks of "commodity, firmness and delight". At least two of those words have multiple resonances for us today. As I take it, Wotton's "commodity" is what we would think of as functionality, or in Heidegger's wonderful phrase, "readiness to hand". But there is another sense to the word "commodity" in the legal information world of today – that of something for sale in a market economy. "Firmness", too, has dual meanings. We might speak of the firmness that makes a technical infrastructure or an information collection "robust" or "durable" or even "up to date". But we might also speak of the firmness or certainty of information that is accurate and backed by authority.

But what are we to make of this word "delight"? I don't have any idea what Wotton meant. But I can guess that he meant something that goes beyond functionality and good engineering, beyond simple "how-to" prescriptions. Something more than either of those things, something that fills the gap between engineering in itself and practical art.

In structuring this talk I found that whatever that "something else" is, it's both complex and frustrating. I have observations to make. They will be observations about the setting in which an architecture of public legal information is located, about the technical means used to realize that architecture, about those who commission and inhabit the building that results, and about the industry that builds it. And those observations don't lead us inexorably to one particular place, but instead

they inform design in a much less deterministic way.

No collection of observations leads inexorably and scientifically to an ideal design. We can observe what we will about the audience for legal information, the technology with which legal information is delivered, the purposes to which such information is put, and the market forces at work … but that falls far short of a blueprint or a how-to manual. The factors involved are so complex and bound up with normative concerns and the needs of national legal cultures, concepts of good citizenship, and variations in local economics and practices that rigid, one-size-fits-all prescriptions are impossible.

Architecture begins with solid design that accounts for people and purpose. But it doesn't end there. In talking to you today I find myself caught in the gap between observation and plan. And for those reasons there is a sharp divide in my talk between (on the one hand) observations based on my experience of legal publication and (on the other) some prescriptions for legal publication systems. These latter are certainly strongly based in conclusions drawn from my observations but have nothing of the inevitable about them. There is no strong train of syllogistic inevitability between observation and prescription -- just design informed by experience. There is more than one way to do things, and a host of factors to be considered in the doing.

My plan is to give you a few brief weather reports, some scattered observations on the state of things as we found them in the electronic information cybersphere. I want to talk first about technology and some of its implications, then about the American scene and what it can tell us, and move on to some more theoretical observations and finally a prescription or two.

Let me start first with a description of my own vantage point, which is that of an institute for applied research, one that studies the practice and theory of legal information.

I

In 1992 Peter Martin and I started the Legal Information Institute – LII – simply, as we then thought, to put legal information on the Internet. We thought of ourselves, and we still do, as applied researchers. We wanted to make things, try them out, see where they led. At that time, the field in the US was absolute empty of all but the large commercial publishers like LEXIS and Westlaw. Those two (along with a very few others) had established both commercial and intellectual dominance over legal information, and there had been little serious investigation of alternatives for more than ten years. The new capabilities of the Internet broke that open, and we were the first there with legal information.

There were more LII firsts. Not only were we the first legal information site, we were first to deliver information intended for professionals other than those in the field of high-energy physics. We developed the first web browser for Microsoft Windows, the first legal Web site, the first site offering a United States code: we worked as consultants on the first-ever portal site and so on and so forth. In the beginning we were excited at getting a hundred hits per day on our US code collection, which then consisted only of Title 17, the Copyright Act. Now we take nine milli-

on web hits a week and we run mail-based current-awareness services that reach tens of thousands. And on admittedly rare occasions we service as many as five thousand hits per minute. We are the largest and best-known wholly non-commercial legal information site in the world. We have a very small full-time staff of five, a part time administrator and a part time editor. Two of the five are programmers. One is a systems administrator with a law degree. Two are law faculty members, one of whom (me) does not have a law degree. Four of the six of us live in Ithaca; two do not, but telecommute from as far away as Boston. It's rare that more than three of us are in the building at the same time. We are as much a product of cyber-space as we are a product in cyber-space.

Some early decisions continue to define us today. We chose a small number of high impact collections, in particular the decisions of the US Supreme Court and the United States Code, as our flagship efforts. We didn't attempt to be comprehensive. In the complex environment of the United States, that would have led to a delusion of effort bordering on the foolhardy. Instead, we wanted to make deeper investigations of a limited number of important collections of legal texts. We wanted to seek out the place of a graduate law school in developing architecture, commentary and secondary sources that would promote intellectual as well as technical access to the law. We wanted to draw on the broad resources of a large university to inform our work – on computer scientists and social scientists in particular. In other words, we were doing what it is now fashionable to call "information science" in the tricky domain of law. We were doing it, not just talking about how it might be done.

We have no formal ties to government, nor do we have special or exclusive arrangements regarding content. We acquire our raw texts from public sources that are, theoretically, available to anyone, whether private citizen, entrepreneur, or large commercial publisher. Indeed, we have generally refused to take on projects where public domain legal text is not available. We have a relationship with government agencies and legal publishers that is perhaps best described by the phrase "leadership by example" or more often "leadership through shame". We try to do things well and let others follow on. We have sometimes done commissioned or bespoke work for government and for private legal publishers. More recently we have begun to form communities of interest with particular government technical groups; my e-mail this morning contained an exchange between one of our analysts and one of the US House of Representatives technical staff discussing parsing techniques for identifying subsection labels in the United States Code, and some lexical-analysis tools we've developed to help us do that. In short, we are more or less what you expect from an academic effort that is interested both in experimentation and which maintains the kind of intellectual nimbleness that permits us to put our efforts where we imagine they will do the most good.

II

Now a few things about technology.

I have little to say about the impact of Internet technology on the distribution of legal information that is new. But there are some foundational matters that it would be well to keep in mind. First of all, Internet technology implies new market structures by changing the loci where value is added to legal text. Second, the Internet brings an accessibility that brings dramatic change in the audience for legal information. Third, technology affects costs, and the relative importance of different kinds of costs, in ways that makes some new things possible and some old things undesirable.

Implied market structures

First, let us consider these implied market structures. In contrast to print, which demands that the publisher add all the value that will be added to a text before it finds its way into a book or printed document, electronic hypertext permits value to be added by many people in different ways at different times. One publisher may put simple text on-line and another may add indexes and commentary to it. All of these different levels of value can exist simultaneously in different versions for different markets. My colleague Peter Martin says by contrast with electronics, print is both static and lumpy – hard to change, and inextricably trapped between the covers of a single, lumpy book. Hypertext is neither of those things.

Hyperlinking, and with it the ability to construct portals and other resources that point to resources, makes it possible to build collections that are federated rather than concentrated. A collection of the decisions of the US Circuit Courts of Appeal might in fact be searching separate collections, one for each court, linked together with hypertext references. Or one might, as we have, build a collection of commentary on professional practice rules in a number of states, with common indexing and cross-referencing, aiding navigation and comparison. An obvious but very important corollary to this notion is that effort and expense can be divided among multiple actors and institutions.

Continuity of process

The idea of continuity of process is perhaps best illustrated when we think about legislation. In the US, as practically everywhere else, federal legislation begins with a drafting process and continues through a number of varyingly-public stages, ultimately resulting in passage by a legislature. At that point it is codified, incorporated into a larger, topically organized body of law. The practicalities of print lead us to think of these happenings as a series of discrete processes, each demanding different printings and distributions of the material as it passes through the various stages of its life-cycle. But the fluidity of electronic text, taken along with sophisticated mark-up technology such as XML, might lead us to a more process oriented view. In reality, legislation is a continuous process that begins with a draftsman seated at a word processor, and ends up in the ears and eyes of the public. In an electronic world, this is a process that cries out to be re-engineered. Surprisingly, this is not often been done, and at this point so far as I know only Tasmania has a system of electronic publishing that embeds the concept of legislation as a kind of continuous process.

Audience effects

Internet technology explodes the meaning of public. It brings not merely a linear or even a geometric expansion of the number of people who can access law texts, but instead a combinatorial explosion of niche-markets for particular types of legal con-

tent. It is not simply that greater numbers of people have access to law -- it is that what they are seeking and the reasons why they are seeking it are hugely diverse and complex.

A legal information product is in fact a cross-product, the intersection of a particular primary source with a particular audience. In turn it is not so simple a matter as creating secondary sources for niche audiences as lawyers do when they produce client newsletters, practice guides or legal education materials offered as refresher courses to practicing attorneys. Marc Galanter writes that "law usually works not by exercise of force but by information transfer, by communication of what is expected, what forbidden, what allowed or whatever consequences of acting in certain ways" . Like any other content transmitted through a communication system, primary legal sources can be rendered more or less understandable, locatable, and hence effective by structuring and presenting them differently for different audiences. And secondary sources must of course be constructed for a particular market, audience, or level of understanding.

Publishing for niche-markets like this is an activity that is both profitable for the private sector and a matter of necessity for public bodies that have an explicit or implicit mandate to serve audiences with special needs and perspectives. Unsubtle examples of such activities include language translation, specialized presentation for the physically disabled, or materials for the illiterate. Other, less obvious needs exist as well. Agencies in the United States routinely rearrange statutory material into structures and sequences optimized for use and reference by field workers or program administrators. Professional associations, trade groups and lobbying organizations build legal guides intended for niche-markets in a particular industry or demographic slice such as the elderly. In the private sector, the fact that publishers must provide products for a range of media and audiences – law text bundled in different packages – is often cited as a major competitive reason to adopt mark-up technologies such as XML, that facilitate re-aggregation of legal information into new publications. Private sector publishers are not the only organization with reasons to do this.

The legal information market is really no longer conceivable as bipolar – it can no longer be seen as a question of lawyers on the one hand versus a largely legally ignorant everyone else on the other. In the old days we imagined that legal information systems were for the most part professional research systems used by lawyers who were lawyering. And we also believed that that condition implied its opposite, that non-lawyers using such systems were for the most part a slightly foolish fringe of pro se patrons unwisely representing themselves in matters of more or less consequence. Within a year of going into operation we learned that the Internet audience for legal information is different. Internet-based legal information systems are used by many cases and conditions of people for many different reasons. We shall look at the more prominent ones later. Probably the most interesting group is the one I refer to as non-lawyer professionals. These are people whose interest in law is vital, ongoing, and professional rather than either being casual and hobby-like or sporadic and trauma-driven. These are hospital administrators interested in public benefits law, police officers, people in regulated industries with a need to know the regulations. Often they are a species of lawyer-by-osmosis, someone who is deeply familiar with a particular area of law that is tightly coupled to her work. In that respect they are not "average citizens", but they are not lawyers either.

Such new and diverse audiences require new and diverse legal information architectures. They will want specialized collections of law of particular relevance to them. They will want those collections organized and presented in ways that reflect their profession or their situation, in ways that collections organized according to the legal abstractions and legal terms in use by lawyers do not. They are concerned with situations and fact-patterns rather than theories, doctrines, and concepts. They are, in short, a very intelligent and exciting type of lay users, and a potentially enormous audience.

Costs

Let's turn now to costs.

By contrast with earlier technologies of print or mainframe-based computing, web-technology requires comparatively small capital concentrations, if any. The distribution infrastructure, the Internet, is shared. Computing resources needed are (by contrast with mainframes and printing plants) very inexpensive. You can, in short, get into the legal publication game with very little money. Whether you can stay there or not is another question. The long-term cost of editorial maintenance of collections has not decreased, although as we saw earlier the new technology allows them to be distributed insofar as collections can be aggregates of smaller collections with different sponsors.

It is a commonplace that web distribution of primary legal text has a marginal cost near zero. Given the existence of text in electronic form the additional cost of putting it on the web in raw form is relatively small. On the other hand, initial data-conversion and editorial costs may be quite high.

All this adds up to the notion that the first copy cost predominate, just as they do in other kinds of publication. In thinking about public policy, some would end the discussion right there, saying that first-copy costs have been entirely borne by the tax-payer as a matter of course, and that therefore the government has, in fact, no right to withhold or charge extra for electronic versions. This is an argument for which I have a lot of sympathy, but at the same time we have to recognize two things. First, electronic distribution is very inexpensive but it is not completely costless. Second, we have to ask what level of value we expect in this inexpensive first copy. There may well be things we want that we have not, in fact, already paid for with our taxes.

It may therefore be useful to look at first-copy costs in a little more detail.

First-copy costs

It's not my intention here to consider all possible first-copy costs, but rather to look at a shorter list that I believe to be significant in this environment. The first of these cost items is format conversion. This is both a matter of getting raw materials – legal texts as they come from the creating agency, court, or legislature – from one electronic file format to another, and of getting them from one editorial format to another. It is expensive insofar as automated methods of moving (say) word processing documents to HTML for web publication tend to be very sensitive to the original data

format and are therefore both corpus-specific and subject to change and drift in editorial format over time, as well. This is illustrated for us by our experiences with the New York Court of Appeals, where it seems that we are re-writing transformation software every five minutes because of changes in the ways the court is publishing its information.

The second important cost is the cost of linkage – the business of extracting cross-references and adding linkage to other relevant legal documents. It is costly, because the problem of recognizing cross-references is non-trivial and also because even once those cross-references are recognized there is no guarantee that the document being pointed to by the cross-reference is actually on line somewhere. That suggests in turn that there may be a minimal critical mass of documents that needs to be available on-line before a law collection is truly useful.

A third category of cost is metadata extraction and addition. I mention both extraction and addition to suggest that there is more than one general class of metadata we might be interested in. Some metadata actually exists as text within the four corners of a document -- for example the name of the author of a judicial opinion, or the effective date of a piece of legislation. Other interesting metadata doesn't exist within the document itself, but is externally constructed by editors -- such is the case with headnoting and with various forms of data that provide formal classification for the document. And there are some fuzzy types that could go either way. Such metadata might be deduced from the document indirectly or it might be imposed from outside – for instance, with some kinds of abstracting or intellectual indexing, which might be done either in software or by an editor, depending on the nature of the texts themselves and what sort of abstract is wanted.

Finally there is the cyclical cost of currency – that is, of keeping everything up to date. Obviously this cost is much higher for legislation than for judicial decisions, because legislation changes frequently and must be updated and judicial opinions generally do not.

III

With those foundations in place, I'd like turn now, if I may, to the scene in the United States.

There are three notable features of the American landscape. First, courts, legislatures, and agencies are numerous and their relationships are complex; second, there are few restrictions on who may publish legal information; and third, a private-sector oligopoly has long had a stranglehold on the market.

There are many, many courts and legislatures in the United States. In fact there are 103 federal courts (by my count), 50 states that have anywhere from three to nine courts apiece, and countless local jurisdictions. Second, we have a long tradition of judicial independence that some would say amounts to judicial stubbornness or orneryness. That independence is deeply reflected in a lack of standardization of practices and procedures from court to court. There is no set of uniform technical standards for electronic publishing (or any other kind). In fact the thirteen US Circuit Courts of Appeal share almost nothing in the way of publication standards or technical apparatus, save for three or four that use extensions to a small and very

outmoded case-management system as their means of putting case information on the net. A lot of case-law is simply not available to public search engines such as Yahoo or Google.

Creation of law is so dispersed that government becomes both a consumer and a redistributor of law. In our experience, it is often people inside government that are most grateful for the ability to access legal information from public sources. This seems odd. But it is easily explained by a combination of two factors. First, the agencies or law-making bodies that turn their output over the third parties for publication, whether these parties are inside or outside government, have the same difficulties of access that the general public has. In effect, they've donated their material to a system that is opaque to them. Second, those agencies or law-making bodies are much more strongly motivated by a sense of obligation to their particular constituencies than centralized publishing offices, such as our Government Printing Office, are. For example, our Social Security Administration has a well-known audience, one that it has learned how to serve with a degree of customization and sensitivity that the Government Printing Office simply can not and will not match because that organization serves a much broader and more diverse group of constituencies and hence must go wide rather than deep. An agency like SSA, at its best, knows how to provide customer service, and to the extent that it has to republish law in order to do that, it will do so.

Obviously, a setup like this is absolutely ripe for a distributed approach to publishing legal information, in which each lawmaking body becomes its own publisher. Equally obviously, there are enormous problems of harmonization and standardization that must be solved if such a distributed approach is to work. Attempts to do so in the US have largely failed for the same reasons that the government sponsored OSI networking standards did. The standards-creation process has a habit of making the perfect into the enemy of the good – that is, of making the standardization process a lengthy and ponderous one that seeks perfection at the expense of practicality and (most importantly) timeliness. It is also true that such long, drawn-out processes often fall victim to the internal politics of standards-making bodies. It would be better if we had a standard that, however imperfect it might be, is quickly arrived at, easily implemented and iteratively refined in ways suggested by actual experience with real-world examples, rather than ponderously debated in the abstract.

A second characteristic of the American scene is the absence of legal constraints on legal information and its publication. There is no copyright restriction on US government works. The counter-balance to that has been expansive private-sector claims to intellectual property rights in the apparatus of citation, particularly those of the West Publishing Company. Fortunately, that battle is largely over, and the good guys won; those claims are now seen as relatively meritless. But for us there was some value in being forced to confront the issue; in doing so, we were forced to focus on the problem of vendor-neutral citation, and in the process of doing that, in turn, we focused also on the problem of media-neutral citation, with more or less satisfactory results. In my own view our solutions to the citation problem are myopically national in character and orientation and need to be extended beyond our own borders.

Mention of the West Publishing Company puts one in mind of what is probably

the single most remarkable aspect of the American scene, and the one that seems strangest to people abroad – real dominance of a few private sector actors in what has been essentially an oligopoly market. It may well be that the American experience in this respect can be generalized globally, because it seems to me likely that any nascent legal information market will have very few sellers and hence will tend to exhibit oligopoly characteristics. I have no idea, by the way, if that is true in Sweden or not, but perhaps someone will tell me in the question and answer period.

What exactly are the characteristics of such oligopoly markets?

First of all, oligopolies are unstable, and very behavioristic in nature, with all-out price warfare as their degenerate case. In markets with few sellers everyone is eyeing one another with the intent of divining what their competitors' future pricing policies are to be. They are waiting for the first competitor to try to capture larger market share by discounting. And if anyone does make a move, and discounts too heavily, prices will spiral down toward marginal cost as competitors try to cut each others' throats.

Oligopolies, then, value stability, and do what they can to ensure it. Obviously, with price wars threatening to break out at any moment they want to keep the market as quiet and stable as they can. One consequence is that it is very difficult for oligopolists to stray outside their high-profit core markets. This is so because if they go outside high-profit core market they must often discount in order to attract new kinds of buyers and once they begin discounting a price war is hard to avoid. So they tend to stay near their home turf where they can maintain high market share.

Thirdly, oligopolies may innovate, but they tend to do so very reactively, and they tend to do so in a way that focuses on what their competitors are doing rather than on the utility of the innovation per se. For years LEXIS and Westlaw, the two largest American providers, have been trading punches by introducing technical features on their systems that are aimed more at the other guy than they are at utility for the market. In a sense they build a lot of bells and whistles that the market is not interested in, simply because their competitors have done so.

What are the general implications of this oligopoly market structure for the American legal information scene? In general, American legal publishers have only wanted to serve their core markets in high-end law firms and other customers with deep pockets, like government. If they were to stray into more price sensitive markets with higher demand elasticity they would have to begin discounting. And if they were to do that they would end up in a price cutting war ending at marginal cost. Or so they feel. To be sure, there are exceptions. Educational discounts are an obvious example of that, although it is significant that while such discounts are universal in American law schools they are not generally offered to international markets. And Internet-based products themselves have the potential to be used as a discounted second line of goods, or second brand, which is what Lexis appears to be doing with the web-based Lexis One service, and Westlaw has done with their acquisition of FindLaw. But it is worth noting that it has traditionally been, and will probably continue to be, extremely difficult to get oligopolists to create systems of cross-subsidy that result in free or heavily discounted public information. Such a regime is simply too threatening to their unstable market positions in markets with a few sellers.

And now for some surprises. Or perhaps they are not so surprising – these are things that are only surprises if you happen to be, as I am, involved in the culture of an American law school. These are things that surprised us initially, because we had been conditioned to expect otherwise. In retrospect they seem obvious. But they serve as cautionary examples of how one's thinking can be colored by familiar (if odd) institutional structures, and by institutional priorities and ways of thinking.

We have discovered that the priorities of a public legal information system don't necessarily reflect hierarchies of legal prestige. That's a fancy way of saying that in some sense our thinking about what needs to be published first is in some sense inverted in its priorities. Most "law people", and that phrase takes in most legal publishers, believe that things that are legally or technically important or prestigious should come first. Thus, the decisions of the highest appellate courts are often the first things that are made available online. That is not, however, what most people need on a daily basis. Their municipal code is more important to them for day to day purposes than the United States Code; either is more important than most appellate-court decisions. When we start talking about law that actually affects people on a daily basis, we find that we are talking about things that are much harder to access electronically. Consider for example the following rather comical report from a friend who is a lawyer in the nearby city of Binghamton, New York, who was sent to get a copy of the municipal dog law.

"Believe it or not," she says, "the city clerk said that no complete copy of the Binghamton code is available to the public anywhere, even in the public library. The only way to get an up-to-date version is to go to or call the clerk's office. I know they sold copies of zoning and building codes for USD10 each, and when I called the clerk faxed me an up-to-date dog law at my request. He knows I am a lawyer and we have spoken before. I don't know what would happen if a citizen needed to see part of a code that was not available for sale. I am pretty sure there is not even a complete version of the code in the clerk's office. I think he'd have to tell someone which section you were interested in and have it printed out."

Well, dogs or no dogs, the fact is that the needs of lawyers usually don't accurately reflect the needs of the public. Nor is that grounds for assuming that the needs of the public are unsophisticated and limited to so-called "dog law" – the somewhat patronizing term that American lawyers often use to describe the material most wanted by what they perceive as the great unwashed.

If the view a lawyer has of the system is a biased one, a view from inside a graduate law school of the United States is even more distorted. Graduate legal study in the United States largely focuses on important judicial decisions made in the highest appellate courts. It becomes easy to forget that, for the most part, legislatures speak much more powerfully than courts do. Courts speak to surprisingly few people. Indeed, in some US appellate courts the proportion of unpublished decisions, those decisions that by design speak to no-one beyond the parties immediately involved, is around 80 percent. By contrast, legislatures and agencies speak powerfully in what Richard Susskind has described as the "hyperregulated state" – still more so in places like the European Union where unity is cemented by layer upon layer of regulation meant to ensure uniformity and equity.

These things seem obvious once we assume a bit of an outsider perspective. The

problem is that we don't often subject ourselves to the discipline of thinking like outsiders. As a result, when we talk of public access to law, we too often talk about it as if the public's priorities and needs were those of lawyers. But they are not.

There is a need to get beyond platitudes. When we talk about these systems (which we generally do only at Law Society luncheons) we talk in the ringing tones that usually disguise a fairly empty rhetoric. And in common law jurisdictions, we go on and on about "ignorance of law is no excuse", and "the rule of law" and other sorts of high-sounding things. The needs that most people have for a legal text are rather more pragmatic than that. At the same time we need to keep in mind their needs are not the same as what lawyers want from an on-line legal resource. What, generally speaking, is this broader audience trying to do?

Well, first of all it is worth noting, as I did a moment ago, that law only functions by communication. It is not just that ignorance of law is no excuse. It is that law only works when it is known. At its simplest, law is as Galanter said a way of communicating what's expected and what's forbidden. And so we would like above all else for law to communicate, and we would like it to do so as clearly as possible. True, there are limits to clarity. Technical concepts and abstractions often rear their ugly heads in ways that are unavoidable, and hard for the uninitiated to understand at a glance. But generally speaking law cannot work unless it is communicated, and that means not just one-way, broadcast communication-by-edict but communication that results in genuine understanding.

Second, a huge number of people are interested in law as a simple matter of civic involvement and understanding. Two unusual cases from recent LII experience illustrate this. The first example is the abnormally high hit rate we experienced after the decision in the Florida presidential election case. Requests for "Bush v. Palm Beach County Commissioners" spiked at about 5.000 per minute, roughly ten minutes after the Supreme Court handed the decision down, and they flooded in at that rate for about fourteen hours. The second is the huge spate of requests we received for information about the display and handling of the American flag after the events of September 11, both in the form of web hits and e-mail. To be sure, these are both increases propelled by national trauma or political and social uncertainty. But they illustrate a much more general phenomenon that is sometimes hard to see because it is so pervasive – the fact that citizens really do care about the law, about government, and about the civil order in which they live, and that when they feel it to be threatened or uncertain they turn to law as a kind of grounding information or guidepost. Their concerns are very often more direct and self-interested. Often there is a need for legal information to "level the field" between parties in a dispute. Beginning in 1998 we undertook some work for the New York Court of Claims, a small (and in the greater scheme of things, a very unimportant) court in the New York State. Unimportant, that is, unless your car is hit by a snow-plow on a state highway or you break your leg in a state park. For the Court of Claims is where one recovers claims against New York State. What goes on in that court is a classic example of a mismatch between plaintiffs who, along with their attorneys, will appear before the court once in a lifetime, and defendants of whom there is only one, the State of New York, represented by the Attorney General's office. On the one hand, a one shot

plaintiff, on the other, an experienced repeat player. Experience is very much on the side of the defendant, of the State, as is information. Until recently the judgements of the court were not published generally, but simply given to the parties in single copies. Thus, the Attorney General's office was able to compile an archive unavailable to individual scattered plaintiffs, an archive that then contained, in effect, a great deal of institutional experience and savvy in practicing in front of that particular court. We built a publishing system for the court, one which made this information equally available to everyone, including the public at large. At the dinner celebrating its launch an assistant attorney general came up to me and said: "You have taken away our advantage". And so we had. And the international scene provides even more dramatic examples. Consider, for instance, the state of things in Kenya – a common-law jurisdiction where until quite recently no caselaw had been published since 1968. It is hard to see how a system based on precedent could survive such a thing.

But if we stay with the more mundane example of the Court of Claims, we quickly arrive at the more general idea of legal information as a means of risk management. Citizens want legal information because they want to gauge the implications of some future action or determine the most advantageous way to proceed. If they can do that without engaging an attorney they will. And one powerful way to do it is simply to read the law and understand it as best one may. In the US this has sometimes taken on the label of "preventive law", a clear reference to the medical profession and to the value that profession increasingly places on prevention. Interestingly it is this preventive orientation that business clients seem to value when they choose multi-disciplinary practices (MDPs) or international consultancies like Ernst and Young or McKinsey in preference to large law firms. To quote a knowledge manager at one such MDP : "The client wants to manage risk, wants a business solution, and she doesn't give a damn whether or how a lawyer is involved". Clearly in many cases the availability of public legal information is an important foundation for assessing and managing that risk.

"Transparency", as the term that is now used in the context of globalization, is a derivative of this more general notion of risk management that has special economic implications. Consider the simple case of a global investor wanting to do business in a developing country. Minimally that investor wants to know first of all that there are rules pertaining to foreign investment and secondly what that rules are. Or there might be important differences in very basic concepts of ownership or property rights in operation, or other cultural differences that an outsider would not be aware of. Such examples make it clear that legal information is important part of the international business and investment climate, and thus it is no accident that investors like George Soros and institutions like the agent development banks are funding legal information projects as part of broader efforts to stimulate global trade and investment.

My larger point is that only when we think about what we want to enable people to do can we get beyond platitudes and start talking about real priorities and operational criteria for legal information systems. There are plenty of practical reasons why people other than lawyers want legal information, and the reasons why they

want it must inform our designs very strongly.

Authority

I'd like to take a second and briefly consider the issue of authority in legal information systems. As a question it is both fundamental and broad, and probably more than I want to take on here. But knowing that you will be spending at least some part of your day today in considering authority, I thought I should nod in that general direction.

Whether or not a particular law text is authoritative or not seems to me to be a matter of decree by an issuing body, which is then followed up by reasonable technical guarantees that the text has remained unaltered since the decree was given. Making such technical guarantees is certainly within the reach of present-day cryptographic technology. Whether or not anybody believes those guarantees or not is another matter.

As Clifford Lynch points out in his wonderful paper on authenticity in the digital environment, which I recommend to all of you, authenticity and authority are as much a matter of trust and social construction as they are anything else. In a very real sense electronic documents would be authentic enough for our purposes even now if we just trusted them to be so. Yet there is a feeling, not by any means limited to legal information, that electronic information is inherently untrustworthy compared to print. No matter what we do technically, it will be quite a long time before people feel any other way, simply because trust as a concept rests at least in part on a sense of the familiar, and electronic legal documents are new.

Secondly, an American has other reasons to be wary of grants of official status or authority. In the US, such grants of authority have become the basis of a sort of barter, lucrative for both the body making the grant and for the publisher who is its exclusive receiver. Under such arrangements, official status goes hand in hand with an exclusive publishing contract. The organization whose work product is being published is heavily compensated by the publisher in return for what amounts to a monopoly over the "authoritative" version of the law. From a public-policy perspective, I believe that we want authenticity and authority in electronic law texts, but we do not want exclusivity in the arrangements that surround their distribution. For that reason we want to find ways of insuring authority while being sure that it does not become a tradable commodity of the sort that leads to exclusive arrangements, exclusive in both senses of the word.

At this point I want to leave quite abruptly from observation of what is to questions about what ought to be. The question I am attempting to address is who should publish the law and how, and what structures we should use to accomplish these things. It seems to me that there are three axes of choice that we may want to consider. Public sector versus private sector, as Peter Seipel said, centralized versus decentralized, and government versus non-government. What sort of structures do we want to use to publish the law?

As to this last axis, the question of government versus non-government, it would seem at first blush that it ought not to be separated. To a degree the government versus non-government question substantially overlaps the others; in fact it is contained in them. But it is worth pondering separately for a moment because of the extent to

which (in the US at least) the notion of government legal publication is weighed down by a public perception of historical faults, the belief that the system is broken. In the US, government publication is believed above all else to be slow. This is actually not the case. What is the case is that like all government bureaucracies the publishing system is responsive to political considerations and to political pressures when they are present, and unresponsive to anything whatsoever when they are not. In reality, the government does a quicker job of publishing than most daily newspapers when someone wants it to badly enough. Publication of the Starr Report during the Monica Lewinsky scandal is proof that when lurid scandal and political pressure are involved, publication can take place quite quickly.

In reality, some branches of the American government do an excellent job of competing with private-sector counterparts long believed by the public to be superior to government – witness the competition of the US postal service with the United Parcel Service, with the postal service largely winning by virtues of an ability to streamline its operations and to create innovative new products. The government legal publishing could easily do the same, particularly if such a plan were sponsored by those who have the most immediate reason to want laws to be published, namely those who make them.

That in turn brings us to the question of the publication by some centralized entity versus self-publishing by creators of information.

In favor of centralization we have a few arguments. The first of these is the idea that law is somehow too important to let just anybody publish it. That is essentially the position of those who traditionally believes in strong government control over the publishing of law. The need for accuracy is so great, they feel, that not just anybody can do it. In fact, they say, you need a specially qualified professional to do it in a place called a Government Printing Bureau or, perhaps, in some exclusive and special part of the private sector.

In the past, as we have seen, there were other arguments for this kind of centralization. These were particularly strong when law publishing involved the concentrations of capital needed for facilities demanded by older technologies like printing presses, privately-built specialized telecommunications infrastructure, or mainframe computers. One such argument for centralization still does have some teeth, the idea of concentration of expertise. Electronic publishing requires technical knowledge that is expensive or duplicative to house in multiple places. It should be borne in mind, however, that this only implies a need for a central consulting service or a locus of expertise, not for a central full-service publishing operation.

Finally, the most compelling argument for putting all one's eggs in one basket in this way is simply that centralization results in standardization without the need for ponderous consensus-driven processes of standards development. In such a scenario, the standard is whatever the central body says it is. This has a lot of appeal, especially to anyone who has ever served on a standards committee.

Conversely, here are a few arguments in favor of decentralization. The first of these and the most compelling in my mind is scalability. Very simply, effort is divided so the more hands become available as more work is undertaken. In a place like the United States, where there are a large number of jurisdictions, each with its own courts, legislatures, and agencies, scalability is probably the most compelling argu-

ment for a decentralized system. Nothing else will work, because the effort involved is simply too large and too complex for any single entity. Unless the work is parceled out it will not get done.

Secondly, there is the question of locus of control, particularly quality control. As I mentioned a moment ago there are few who are more motivated to see law correctly published than those who make it. In a distributed system based on self-publishing by law creators, errors are inevitably caught and corrected early in the process.

Finally, content creators tend to believe that it is less expensive for them to let someone else do their publishing, but in fact this is not the case. Ultimately the idea of a centralized, outsourced legal publishing operation has problems. Here is why:

Arguments in favor of centralized publishing regimes tend to rest on questions of standardization, reliability and efficiency. Those three lines of argument share the presumption that a central publishing entity could use a set of common internal tools and standards for information handling and storage, ultimately resulting in presentation of the information to the public or to consumers in some uniform standardized way. Thus, when we talk about centralization we imagine that a centralized third-party publisher can wrestle a diversity of incoming information streams, each with its own idiosyncratic format, into a kind of standardized product that is available to an audience that wants the benefits of common appearance and functionality. This activity is imagined to be efficient because the group involved is able to spread the costs of technology and of developing technological expertise across information streams produced by many creators, whereas without them each creator would have to develop publishing expertise on its own. Other presumed advantages includes standardization and reliability.

Standardization is a simple label for what is in fact a complex set of very practical problems for publishers generally and for electronic publishers in particular. Different information creators use different electronic file formats to house their information, such as word processing programs, ASCII, HTML, or a database format. Some formats are less readily converted than others.

Transformation of a creator's work into something that conforms to a set of standards or a common format is not merely a matter of electronically transforming one file format into another, however. There are issues of editorial conformance involved that go beyond typography to the structure of documents. For example to reliably convert judicial opinions and format them for online distribution we have to be able to reliably (and preferably automatically) determine what the names of the parties are, what the date of decision is, who the author is, where headings, major and minor sectional divisions occur, and so on. We do this not only for typographic purposes but so that we can appropriately tag metadata in our standardized version, extract text features for special treatment by search engines or indexing software, create links to other materials, cite the case and so on. Since the different courts and legislators vary in the way they format materials, our text conversion and conformance software must vary from court to court, legislature to legislature and agency to agency. The underlying techniques on which such software depends are fundamentally matters of sophisticated pattern recognition, not unlike a very sophisticated version of a word processor's search-and-replace function. Because it is so dependent on recognizing patterns in the text that are known in advance, conversion software tends to be extre-

mely difficult to generalize beyond the work of a single court or legislature.

Basically the effect of this is to drive up the cost of conversion, and cost in turn affects the scale at which centralized third-party publisher can operate. Every new, non-standard corpus added to the collection published by the centralized publisher represents a significant new short and long term conversion cost. How tempting it is then to suggest that all newcomers need to conform to some sort of input data standard that makes all conversion job both easier and less expensive – indeed, which means that your conversion software can be reused from corpus to corpus -- particularly if you are under budget pressure. In fact, this is what some organizations do, and I envy them their ability to do it. In the United States, with our tradition of judicial independence, we can not compel that sort of cooperation.

So we can see that there are intense practical reasons why a centralized third-party approach doesn't scale well. Because it is expensive to convert diverse input streams into a common format, a centralized approach works better when the number of creators being serviced centrally is more limited, where a one-size-fits-all approach is suitable, or where a better funded private sector is at work. More importantly, any third party publishing operation will have real limitations; it can only scale up to a certain point before conversion problems force it to compel a structure that is fundamentally indistinguishable from self-publishing by creators. This is so, because third party publishers have to insist on increasing levels of standardization in the data in order to maintain operations as the volume and complexity of their input streams increase out of proportion to resources. Ultimately the extent to which the input data sent by creators so closely reflects the practical needs of the output product that the creators might just as well be publishing the material directly.

This isn't bad in itself. There is a useful time dimension at work . Even if things reach the stage I just described, the creators will have bought some time and probably some opportunities for technological development and transfer, by using the third party publisher as a kind of buffer against change. Even so however there are real questions about how well a central operation can work.

Let's turn now to the question of public versus private.

One powerful argument in favor of the private sector is that it is the only apparatus that can really expand legal information service to the vast number of possible niche markets. You remember that we talked about combinatorial explosion of audience. There are many markets out there, and it would be impossible for government to service them all.

Secondly, and again fairly compelling, there is the idea that the government really should not have exclusive control over knowledge of its operations. You want third party publishers in order to guarantee that the government operations are being correctly and fully reported to the public.

Finally there is the argument that the government should simply not interfere in areas where the private sector is already active, that is, it should not take away business from those who already have it. This has been a particularly pervasive argument in the United States, where as I said the private sector is firmly in control of the situation.

In favor of the public sector is the idea that dissemination of public information

is a legitimate government function. Indeed it may be more than a legitimate function – it may be an obligation of government if the public has substantially paid the first copy cost we talked about already. Second, one can argue that the private sector inevitably underservices unprofitable markets. They go where the money is, and often leave unprofitable markets unserviced. Finally, as we pointed out, the private sector is inevitably heir to the problem of oligopoly which mean in effect that outlying or nontraditional markets again do not get served.

IV

Finally some prescriptions.

How are we to resolve these dilemmas? I think we can take it as given that there needs to be co-existence of public and private sector publishers. Optimally, to my way of thinking, this would take the form of free-to-air self-publication by government bodies, augmented by as many private sector publishers as the diversity of general and niche markets will sustain. The struggle, if there is one, is really over questions of how much value-added metadata is to be put in at source. If free-to-air systems add too much value, the market for some type of niche providers will simply be unprofitable for the private sector while remaining unserviced by government. If, on the other hand, public systems add too little value they are essentially unusable. Thus it seems to me that the crucial question is how we think about how much value we are going to add in free-to-air systems.

Adding both linkage and formatting information seems to me essential. Also essential (but frequently neglected) is the need to have law accessible via public search engines such as (in the US) Google and Yahoo. Increasingly those services are kind of catalog to the public library, and hence it's important that law be included in them, since they are where the public expects to find things.

Often the line between free and fee services is defined by two different versions of the same information, such as the (free) time-delayed version of (fee service) stock quotes. Some of these version methods simply seem inappropriate in the context of public legal information. For instance, some US courts currently use a variation of time-delay versioning, a kind of a sunset rule to limit the amount of time the judicial opinions are available for free once they are handed down. The approach strikes me as nothing short of perverse.

Other types of metadata, such as commentary that has to be added editorially are a little more up for grabs. Some should clearly be left to the private sector, and indeed it is probably only the private sector that can afford to do them properly. Some kinds of head-noting and commentary fall into that category; so do some kinds of intellectual indexing. It is not clear whether some other types of value-added service are in fact something to which the public has a right or not. What about commentary or instructional material that in effect provides intellectual access to the public, allowing them to understand as well as read the law? It may be that clear-cut rules are impossible here, leaving the public sector in a kind of friendly competition with the private sector, one in which expectations are continually rising because the public sector simply does as much as it can and assumes that there always are things left over for the private sector to do at a profit.

As for centralized versus distributed systems, as many of you know my sympathies are firmly with the distributed model, largely because in my own setting it is the only one that will work. And it is the only one that makes fitting use of Internet technology. I am not blind to the problems that it poses. The question of how to develop standards is particularly crucial, and all the more exasperating because standards committees are the sort of debating societies that lawyers love.

As for a final point, I think that in general we are at once both too pragmatic and too idealistic about the kind of systems we are designing. We are too pragmatic because we believe that all needs can be served as our own needs are served, that legal information systems that serve lawyers serve everyone. Too often it is the case that we simply design these things in the way it seems quite useful to us but may not be useful to anyone else. And at the same time I think we are too idealistic because our thinking about the needs of others is often too tied to high-sounding phrases like "rule of law" and too removed from direct observation of what real people actually do.

For to have commodity, as Wotton talked about, we have to think about the needs of others. To have firmness, we have to use distributed technology as best we can. And as for delight – in this case, it should come from serving the needs of the public.

Peter Seipel:
Thank you very much for this lucid and very comprehensive overview, and I also thank you for leaving us with these dilemmas we will have to discuss during the afternoon. I have asked Andrew Mowbray to use this opportunity to make a comment on your viewpoints or pose a question. Please, Andrew.

Andrew Mowbray:
I was interested, of course, in your comments about decentralization and a distributed system, given that we have followed a different model. What sorts of things do you think is necessary in order to achieve the same distribution of resources and how are they different to the sorts of standards that you would need in order to maintain a scalable centralized resource?

Tom Bruce:
Primarily I see this as a metadata issue. The primary problem in operating distributed systems is that people come into them unaware of how to find things and very aware of the fact that how you find things may differ from system to system. Minimally you need the same metadata contained in all documents furnished by everybody – some agreement on what metadata would be present for example in judicial opinions. Are we going to have author data, decision title etc etc. That standard can be kept reasonably simple, and it can be made the basis for reasonably simple bibliographic-style search that will serve most people.

At the same time, in the application of metadata, you need ways of relating corpuses in different categories to one another – of linking a case to the relevant statute and of linking each of these to contextual or explanatory information. The situation I often imagine is one in which a hypothetical lay user with a problem – a pro-

blem with a small business or with the school district or what have you – walks up to an internet search engine and throws some terms at it, and by some miracle gets back a list of stuff. And in that list, figuring quite prominently, is the US Supreme Court decision. And they look at that decision, sitting there in the list, and because they had high school civics they say: That's dispositive, that's something important. But when they click on the link in the hit list and read that Supreme Court decision, they suddenly realize that they have fallen down into a very deep hole, they don't understand any of this – what does this mean? They want to have a button they can press at this point that says: Explain this to me, please, or: Show me more stuff like this, or: Show me the statute this relates to.

And any ability we might have to help them with this is ultimately based technically on being able to say: This case, this statute, this regulation here shares metadata with this other one over there and therefore the two are in some way conceptually related.

Both intellectually and technically this kind of interoperability depends largely on standardized metadata being present in the components of a distributed system. The difference between that and a centralized system of course is that in a centralized system you are completely free to impose that sort of standard yourself, without even having to consult anyone. When you try to get numbers of courts or public bodies to do that you are essentially trying to herd cats. It has proven possible to do that with things like court filing standards, or with bibliographic metadata. I see no reason why we can't do it with on-line material if we are talking about simple metadata and bibliographic search. If you get off the four corners of of the document and start talk about intellectual indexing or conceptual classification it gets much, much harder.

But basically it is a metadata issue.

Andrew Mowbray:
Would you be envisage some sort of restricted web-search engine, because obviously the problem with metadata in the general world is that people have abused it to the extent it becomes fairly unworkable. Hence Google search engine which basically ignores metadata. So would you be suggesting that we have a targeted sort of spot that only including sites which are cooperating properly?

Tom Bruce:
As you say, I think this is probably the only practical system that works. You could also have it driven by a system of harvesting engines. I am getting quite interested in RDF and some of the more interesting stuff that has been done by RSS channel as a way of doing this kind of thing. There are a lot of work going on in digital libraries that squarely rest on this. A thing that they seem to have done with stuff like the open archives initiative is to have constructed schemes for metadata harvesting in which compliance with the harvesting scheme can basically be taken as evidence of good faith. So I suppose it is restrictive in the sense you describe, but you get a centralized harvester looking at a number of sites that conform to the standard pulling metadata out of those. In essence I am talking about the same things that you are. Some kind of targeted search engine or some kind of targeted metadata harvesting.

An international perspective on delivering free access to public legal information.

Andrew Mowbray, AustLII

1. Introduction

The AustLII web service went live in July of 1995, when it provided Commonwealth Consolidated Acts, decisions of the High Court of Australia and a small index of Internet legal materials [1]. This was made possible through the combined efforts of David Grainger and his staff at the Commonwealth Attorney-General's Department's SCALE together with AustLII's initial staff [2].

Over the past four years, the system has expanded considerably to the point where it currently contains the full text of the legislation of all 9 Australian jurisdictions, all of the decision of superior courts and many other primary and secondary materials databases. AustLII has been consistently ranked in the top 100 Australian Internet sites [3], regularly getting approximately 200,000 'page hits' per day. In addition there exist a wide variety of projects that promise to expand not only AustLII's content but also its user base and profile.

This growth has occurred very quickly, and is being spurred on by a number of factors, including:

– User demand for a national law collection;
– Pressure from stake-holders to increase both the number and scope of primary materials databases;
– A desire to meet the needs of 'lay users' through better secondary materials;
– Funded project aims, such as Project DIAL and the World Law Index.

In 1998, AustLII conducted its first user survey [4] to assess user's perceptions of AustLII's performance. The overall tone of survey responses was positive in terms of both AustLII's public policy agenda and its technical delivery of materials. However users also took the opportunity to voice concerns about currency, accessibility and other issues. These are discussed in some depth below.

Technical and management problems are also starting to arise. While reliable at current levels, the work practices that were used to build the original AustLII system could become difficult to use due to the sheer size that AustLII is growing to. Each database uses a unique and seldom reused set of tools, written by a single database maintainer, whose idiosyncrasies are known only to them. AustLII is also finding itself up against the familiar technical limitations related to CPU capacity, memory usage and disk storage, as well as limitations imposed by the operating systems.

In summary, the pace of change has introduced not only problems of strategy and policy, but has also raised important issues related to scalability. The crux of the issue is that solutions developed for a particular problem do not necessarily continue to be

appropriate as the scale of the problem grows.

This paper discusses the current state of the AustLII system from a technical perspective and focuses on the scalability of the Web resources that we are using. It includes an overview of the technical setup of the AustLII system, including a detailed description of the tools and technologies involved.

The current system is then critically analysed in the context of assessing its scalability. Finally, the new technologies currently under development are outlined.

2. AustLII's Current Technology

2.1 Current System Dimensions and Configuration

AustLII operates what has become one of the largest legal database collections in the world and is certainly the largest non-commercial system. The current dimensions of the system are 86 major text databases, over 7 GB of searchable raw text containing over 1 million documents cross-referenced with over 22 million hypertext links.

The current hardware configuration which directly supports the live system consists of a number of Sun Microsystems servers [5]. Primary disk storage consists of two RAID arrays with over 210 GB of space.

The machines are linked to the UTS network via 100Mb/s fibre connections, and from there to the rest of the world via the NSW Regional Network.

2.2 Systems Software Overview

From the outset of the project, most of the software that has been used on AustLII has been written in-house. Some of the more major software systems that have been developed include:

– sino[6] – a free text search engine which is capable of delivering fast retrieval times over large text databases and which provides a flexible user search language and software interface;
– the hypertext markup software [7] – a suite of programs which facilitate massively automated hypertext markup of cases, legislation and other materials;
– feathers [8] – a system which allows for indexing and presentation of links to Web materials;
– gromit [9] – a targeted Web spider which is used to provide search facilities to materials on sites other than AustLII, as well as providing updates to some AustLII legislation databases;
– wysh [10] – a Web based expert systems interface which allows expert systems based on the ysh [11] inferencing engine to be made available on a distributed basis

Apart from these programs, the AustLII production system also uses a number of other pieces of commercial and public-domain software. The operating system on all of the AustLII machines is Sun Microsystems' Solaris 2.6. The web server is Apache [12], with some local modifications. The GNU C Compiler [13] and Perl [14] are used for compiling and running locally developed code.

The interaction of these pieces of software is complex. The following diagram represents an overview and each element is described in further detail in following sections.

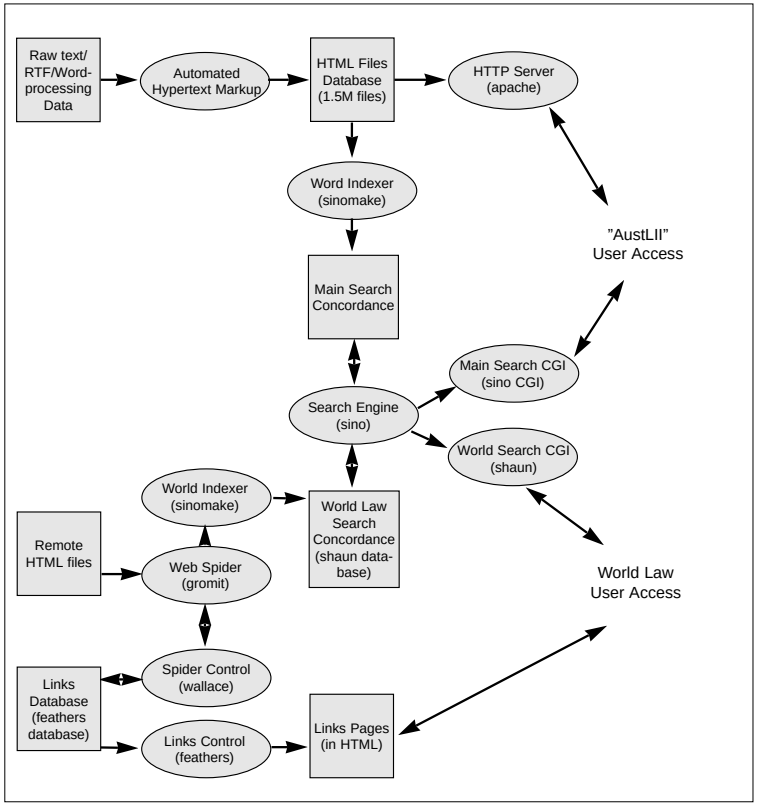


Figure 1. AustLII Technical Overview

2.3 Hypertext Markup

From the user perspective, one of the most obvious features that distinguishes AustLII from other large text databases is the extent of the hypertext markup (currently with over 22 million links). AustLII employs no editorial staff to assist with hypertext markup. All hypertext links on AustLII's databases are inserted on an automated basis with no editorial input.

The hypertext markup is achieved via a number of programs and scripts that employ similar approaches. Some of these are implemented in C and others are written in Perl.

2.3.1 General Approach

The general nature of the markup scripts is highly heuristic and is designed to identify a number of salient text features. Some of the things that are currently processed include:

– references to Act names:

– references to sections of Acts (both internally and externally);
– references to other structural legislation elements (parts, schedules etc);
– references to legislatively defined terms:
– references to case citations

Although some of these can be dealt with without reference to any contextual matters, a lot of these items are highly context sensitive.

For the most part, all markup is done ahead of time. Dynamic markup is kept to an absolute minimum in order to maximise system performance. The major exception to this is in relation to the Noteup functions which are included for all legislative documents and some cases. The noteup function allows users to conduct canned sino searches which are based upon stored URL addresses.

The effect of noteups is to perform a 'reverse hypertext lookup' – thereby returning related documents which refer to the current document.

The main aims of the overall hypertext markup approach are that:

– the markup should be as rich is as possible;
– it should minimise the number of erroneous links; and
– it should be as simple as possible (both for speed and maintainability).

Unfortunately, these aims tend to be contradictory. Rich hypertext markup involves complexity and so challenges maintainability and speed of execution. Similarly, the more ambitious that the markup programs become in terms of identifying unusual text patterns, the greater the risk of error and so forth.

The current markup approaches represent a set of design compromises that seek a balance between the constraints. This has been achieved over a number of years through experience with legal data.

As further discussed below in the section on document management, the markup tools rely heavily on the use of file organisation for document management. Currently, there are no separate document control databases. Partly because of this and partly for reasons of markup efficiency, all hypertext links on the system can be mapped on a 'one way' basis. The central idea is that whenever a potential link is found, it is possible to determine an appropriate destination without any database lookups (other than a check to make sure that the target HTML file actually exists).

2.3.2 Contextual Sparse Natural Language Parsing

In previous work on the DataLex Project, a fairly simplistic approach was initially adopted to identify hypertext links that relied upon simple pattern matching. Whilst this is quite acceptable for identifying obvious textual features (such as references to Act names), it is very restrictive otherwise.

The approach adopted by most of the current markup programs is one of contextual sparse natural language parsing. This is a methodology that takes into account the context of where a potential hypertext link appears (including information that can be gleaned from previous and subsequent hypertext) as well as being capable of doing sophisticated parsing of disjointed pieces of textual material within a document.

Context is very important for a lot of the markup and operates at various different levels. The scripts take into account simple things (such as, what type of document is being processed and the relevant jurisdiction), but often also need to take

into account where words occur within a document and what has come before and what comes afterwards.

The approach of sparse natural language parsing is similarly important. Unlike pattern matching (via regular expressions for example), parsers have the potential to operate more subtly and also to take account of the context under which they are operating.

The hypertext markup programs and scripts vary dramatically in their level of sophistication, but mostly work along similar lines. Some of these are of general application and others are more specialised.

2.3.3 Legislative References

One of the most commonly used and most easily explained programs is called findacts. This is a program that recognises references to Act and Regulation names, to references to Parts and Divisions of legislation and to section references. An interface to findacts is provided on the main page of AustLII to assist organisations that wish to add their own links to AustLII legislation.

Findacts works by first finding apparent references to legislation. Often these are simple to find (by the occurrence of the word 'Act' for example), but on other occasions these are more difficult (for example, a reference to 'the Constitution' or to the 'Corporations Law'). Once found the sparse parser operates to gather in the complete name of the piece of legislation and does a check to determine the appropriate jurisdiction (possibly with 'hints' from the calling program). It then checks to see that AustLII holds an Act or a Regulation by that name (after completing the name, by the addition of a year or converting an abbreviated form as necessary). If successful, a hypertext link is inserted for the reference found. The program then examines the surrounding text and, in the context of the Act or Regulation found, tries to determine if there are any internal references to it. This is done via a process of repetitive sparse parsing and where a reference or series or references are identified, these too are marked up.

Findacts also is a tool that is used to markup individual pieces of legislation. In this case, context is much more important. The program takes account of the Act that is being marked up and changes its assumptions (as the majority of references are likely to be internal). For example, it takes account of likely references to pieces of related legislation (for example, where the text in a Regulation appears to be referring to a section, the software is intelligent enough to assume that this is to the enacting Act).

2.3.4 Legislation

The most complex markup that AustLII currently does is in relation to legislation. In one sense, legislation is an easy target for automated hypertext markup: it has hierarchy, order and, to some extent, consistency. The difficulty is that all of these advantages are embodied in natural language, which needs to be identified and responded to.

The current approach to legislative markup is to deal with the underlying text as a set of problems in series. In some ways the most vital step is to just gather in the basic organisational information (such as simple things like where sections start and

stop). This type of information is often difficult to reliably generate from what is often just effectively ASCII text. Nevertheless, it is vital both in practical terms of dividing an Act up into its component parts for delivery purposes as well as the more difficult issue of understanding the context of words that appear within it.

Once the basic elements of a piece of legislation have been determined, the next step is to pass the text through the findacts program which is passed the context that it is marking up an Act, the Act name, any related pieces of legislation and so forth. Findacts also internally takes account of where in an Act it is and makes adjustments as necessary.

Other programs follow to reprocess the data adding links as they go. One of the more important of these is a program called finddefs. The task of finddefs is to identify and mark up references to internal definitional terms within an Act or Regulation. Definitions in legislation can be global, at Part level, or just refer to particular Divisions or even sections. Finddefs deals with these contextual issues and inserts links as appropriate.

The rest of the legislative markup ends with a call to a tool called act2html which does the final division of the legislation into individual HTML files, tidies up the Table of Contents (or builds a new one if none is present), adds legislative Notes and many other things.

2.3.5 Case Citations

Another example of a tool that performs automated markup is a program called findcases. With the advent of vendor neutral and medium neutral citation [15], referring to case law and automatically inserting links to it has become much easier. In respect of the historical material however, there still exists a difficult technical challenge.

Conventional case citation refers to cases on the basis of where they are published. A case can appear in multiple series of reports from different publishers and although for most courts there is a preferred (or 'authorised') series of reports, there is no guarantee that this citation will be present.

The role of findcases is threefold: it extracts parallel citation references from cases in the database and matches these to file names; it identifies references in text to things that appear to be citations and that appear in the citation/filename list; and it does the actual replacements of citations with hypertext links.

2.4 The Sino Search Engine

Apart from the hypertext markup software, the centrepiece of the current AustLII system software is the sino search engine. Sino is designed for simplicity and speed. The software is written in C and is very compact [16]. The major trade-off in sino-'s design was to sacrifice disk usage for speed of execution [17].

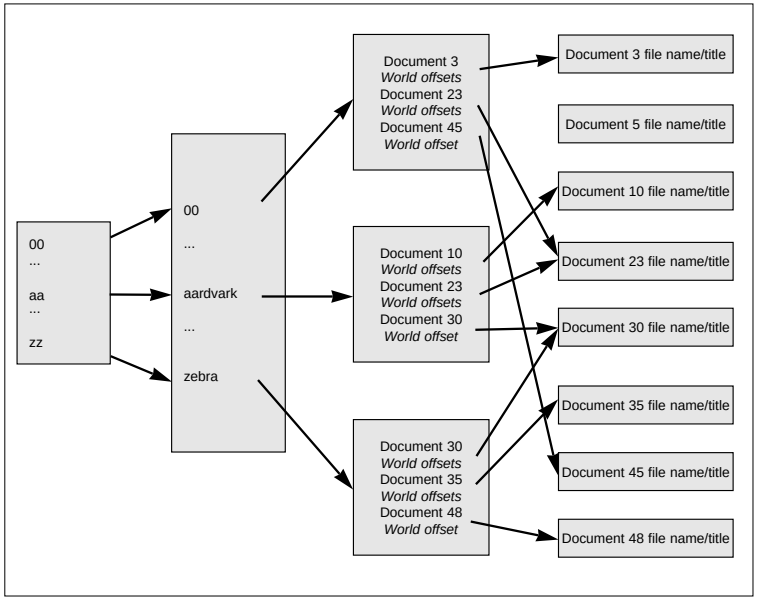


Figure 2: The Sino Concordance

2.4.1 The Sino Concordance Structure

Like most search engines, sino relies upon a word occurrence database (sometimes called a concordance or an inverted file) to speed up search times. This database consists of a dictionary of every word in the indexed text files, along with linked references to the documents and word offsets for each occurrence of each word.

The sino concordance consists of several files. The word dictionary is stored in a file called sino_words and consists of a sparse index based upon the first two characters of its contents, followed by a compressed list of the words themselves and an offset to the location of occurrence information.

The word occurrence information is stored in a file called .sino_hits. This file contains one block of information for each word. This starts with a header setting out the number of entries that exist for the word and the number of times that the word occurs. The occurrences (or hits) are stored as a series of 32-bit references to a document number followed by word offsets within each document.

The document number is a reference to a third file called .sino_docs which gives information about individual documents. This is an ASCII file, which maps document numbers to the name of the file that has been indexed, the HTML title of the document (to save individual file lookups when presenting search results), the date of the document and the size of the document.

The concordance ratio (that is the size of the text indexed versus the size of the index files) is around 40%. Although this is relatively large, the concordance is easy

to read and minimises unnecessary file input/output. Although concordance building on the current model is very memory intensive (using up to 300M of core memory), the build times are very fast. In sustained terms, the sino database creation utility (sinomake) is processing about 500M of text per hour.

2.4.2 The Sino Interface

In execution, the sino search interface uses very little memory. For Boolean searches, the amount of memory that is used per search is around 250K. For freeform ('conceptual' – now known as 'any of these words') searches, this figure increases to about 400K. The sizes of the temporary files that it generates are fairly large (up to 200M for complex searches).

From an interface perspective, sino offers a flexible set of alternative mechanisms. At the simplest level, it can be invoked in a non-interactive fashion to perform a single search and to return results. It also has an 'interactive' interface that is suitable for processing by custom written scripts that make use of pipes or Unix sockets. For this purpose, sino supports a simple command language. A typical interactive sino session follows:

```
sino> search banana
sino: total-docs: 7 message: 7 matching documents found
au/cases/cth/high_ct/173clr33.html
CALIN v THE GREATER UNION ORGANISATION PTY LIMITED (1991)
au/cases/cth/high_ct/124clr60.html
KILCOY SHIRE COUNCIL v BRISBANE CITY COUNCEL (1971)
au/cases/cth/high)ct/115clr10.html
MEYER HEINE PTY LTD v CHINA NAVIGATION CO LTD (1966)
...
sino>
```

Figure 3: Low level communication with sino

This approach is very flexible and means that sino searches can be easily shared across a number of machines. Sino also supports a full C language API and has an associated C library.

From a user perspective, the sino user search parser is very forgiving. It will accept searches in a number of standard search languages which legal researchers might be familiar with. The current search syntaxes which are recognised include Lexis, Status, Info-One (now Butterworths On-line), DiskROM (now LBC), C and agrep. The desire to handle all of these command languages mean that there have been a number of tradeoffs (eg the use of characters such as minus for a Boolean 'not' in Status). Nevertheless, the compromise is designed to work in the majority of cases and seems generally to work well.

2.4.3 Freeform Searching and Ranking Issues

Apart from conventional Boolean searches, sino also supports 'freeform' (that is, 'conceptual') searches. These sorts of searches do not involve the need for operators or other formal syntax and are designed for users who do not have experience with Boolean systems.

Freeform searches are processed as follows:

– All non-alphabetic characters are stripped and common (non-indexed) or non-occurring words are removed;
– Based on the relative infrequency of the remaining search terms, sino builds the biggest list of matching documents (that is, any document which contains at least one search term) that it can within set memory constraints;
– The system then ranks these on the basis of (a) how many search terms appear; then (b) how many 'weighted hits' appear. The weighted hits are calculated according to a formula which gives preference based on how early word 'hits' appear in a document, how commonly the word occurs and on (inversely) on the document size.

The current formula that is used to determine the relative weighting of each word occurrence (or 'hit') is:

$$(a * b + ((c * (d/ ((e+1)+1))) * 100 / ((a + 1) * b)$$

where:
a = the total number of search terms
b = the largest number of occurrences for any of these search terms
c = the no of occurrence for this word
d = a constant to reflect how early a word must occur to deserve special weighting (currently 300)
e = the document offest for this word in the current document

Figure 4: The Sino Freeform Ranking Algorithm

The effect of this ranking algorithm is to yield a percentage. A document receives 100% where it contains all of the search terms and the greatest number of ranked hits. The relative 'importance' of other documents is proportional to this figure.

As is the case with most conceptual ranking systems of this type, the correctness of the search results is best judged in terms of their usefulness from a user perspective. Whilst it is a bit difficult to gauge this with total accuracy, it appears from user feedback that the approach seems to work well. The ranking mechanism for Boolean search results works on a similar basis.

2.5 AustLII's World Law Index

Apart from the databases which are stored on the system proper, AustLII also provides a database of links to other Australian and international legal web sites. Originally this index was searchable only by searching link titles and keyword descriptions. However the index (called Feathers) has since been combined with AustLII's

web spider, and now the full text of most indexed sites can also be searched.

2.5.1 The Feathers Indexing Software

Originally, the links database was maintained manually, but it grew rapidly and contained more than 500 entries by the middle of 1996. In order to maintain this list on a more sustainable basis, Geoffrey wrote a database management system (initially called Chain, but later renamed Feathers) which was based around an SQL back end. This software provided a new user interface that allowed hierarchical browsing and used sino to provide text search facilities. It also introduced an easy to use interface for editing and maintenance of index entries.

The Feathers database has been redeveloped by Austin and has grown to about 4,000 [18] references to external web pages. From a user perspective, these can be browsed on the basis of their source (categorised by countries and jurisdictions) or by subject. There are 50 'top-level' categories and these are organised and divided in a way that is similar to the conventional paper based Australian legal subject indexes.

The database is maintained manually by AustLII secondary materials staff. Users may also contribute links, but these are edited and approved by AustLII editorial staff prior to being added to the database.

2.5.2 The Gromit Targeted Web Spider

There are a large number of generalised search engines that facilitate searching of web pages (Alta Vista, Lycos and the like). From a legal research perspective, there are two problems with these sorts of system: they generally do not index pages exhaustively and the quantity of data makes legally specific searches difficult. A recent paper estimated that general search engines do not index more than 16% of the web and can take several months to find new pages or update their databases [19].

In 1997, Austin wrote the first version of a targeted web spider that was designed to overcome these difficulties. The program is called gromit and has an associated interface and control program called wallace. The aim of the system is to index web pages and make them searchable via sino, but to be selective about what sites are indexed and to be exhaustive in respect of relevant legal material. In the current system, gromit makes use of the feathers database to select which sites it indexes.

As a major information repository, and in response to the impact of other web spiders on the AustLII system, gromit is very conservative about the loads that it places on the remotely hosted sites that it is indexing. Apart from the generally accepted compliance to the Robots Exclusion Protocol, it also ensures that no site is accessed twice in a 1-minute period. Where implemented, gromit also takes advantage of appropriate HTTP headers (such as If-Modified-Since and Last-Modified).

AustLII has recently received a major grant from the Asian Development Bank (Project DIAL) to remotely index the legislation and laws of 11 developing Asian countries. The prototype of this facility in many respects provided the impetus for the creation of gromit. As more information providers publish their own material, the significance of distributed indexing will become very important for the project.

2.6 The Wysh Distributed Inferencing Engine

As part of the work on an ARC funded research project and following on from the earlier research conducted as part of the DataLex Project, AustLII is conducting research and development into the production of scalable enhancements to the service based around artificial intelligence technologies and in particular expert systems.

The essence of this aspect of AustLII's research is to investigate how inferencing technologies can be used to 'add value' to underlying legislative data on a massive scale over the Internet.

2.6.1 The Ysh Expert Systems Shell

The current approach is to use an expert systems shell called ysh that had been previously developed. Ysh is a quasi-natural language based expert systems shell that supports simple propositional logic. The system is primarily rule based and by default all rules are both backward and forward chaining. The quasi-natural language based knowledge representation allows for rules to be written in a close paraphrase of the underlying legislation that is being modelled. All dialogues, explanations and reports which are generated by the system are constructed dynamically by parsing and manipulating the English sentences which are used in rules. Ysh also has limited support for case-based reasoning (based around Alan Tyree's pannda system) and general document generation.

2.6.2 The Wysh Web Interface

The wysh interface was written by King and Cant to facilitate the operation of ysh consultations over the Web. This operates using an inferencing server which maintains state information that is associated with each user session on the server (over separate Unix sockets) and uses a simple forms based approach at the client end. Wysh reads knowledgebases directly from HTML pages. An important feature of wysh is that knowledgebases can be distributed across different machines and pages.

The wysh interface is tightly coupled with the underlying hypertext paradigm and makes use of sino to provide search facilities. Hypertext links can be added in all consultations and reports either explicitly or automatically (using tools such as findacts). The knowledgbases themselves form part of the system and can be displayed or searched over in standard fashion.

3. Current Feedback from Users

From a technical perspective, the AustLII system is a diverse one and is constantly being changed to keep up with the growing size of the database and add increased functionality. This section discusses a few of the current issues and indicates current and likely future systems development.

In September 1998[20], AustLII ran its first user survey, designed to gather feedback on AustLII's user base and user attitudes towards AustLII's performance as a legal web site. The general tone of user's responses was quite positive, with 96% of respondants rating AustLII as being as good or better than other legal web sites, and 42% feeling that AustLII was the best legal web site they used.

There were however a number of concerns reflected in user comments. Regular users may be familiar with some of the issues identified in the survey. Among the

most common themes in the user's comments were:

– 'Simpler' and 'more accurate' searching – including confusion of the Freeform search function and how it worked;
– Improved currency (especially for Commonwealth, ACT and SA databases) and clearer notes on the currency of databases;
– Improved coverage, both in terms of the breadth of AustLII databases (more state legislation and courts) but also the depth (older court cases, missing NSW Supreme Court cases);
– Easier printing (better formatting, full legislation downloads in RTF).

These user problems stem in part from various technical problems, many of which are related to scalability. The main issues are identified below.

3.1 Concordance Size

Whilst AustLII's Sun Ultra computers are 64-bit capable, Solaris 2.6 is a 32-bit operating system. It's file system only uses 31 of those bits in its file pointers. This means that the largest file that can be stored on a Solaris 2.6 computer is 2 gigabytes [21].

In June of 1999 the AustLII search concordance exceeded this limit for the first time [22]. AustLII builds a search concordance over the entire set of AustLII databases (primary and secondary materials). The process typically takes eight hours. In this case the process could not complete, because the resulting concordance file was too big. In addition, it has become difficult to complete the World Law search concordance because the size of the dictionary (which must be kept in core memory) is prohibitively large.

For performance reasons, the temptation to split the concordance has been resisted. A temporary solution has been to expand the list of common words (which aren't indexed) however this is unacceptable for a number of reasons, including loss of accuracy in search results. The single concordance approach also means that there can be unacceptable lead times between the time that a document is added to AustLII, and the time that it can start appearing in search results.

Currently, work is focusing on a new concept in sino – that of a 'virtual concordance.' This is part of the new 'beta interface' and the embryonic Anarchivist , which is discussed below. The idea of a virtual concordance is to have multiple physical concordances linked and accessed as if they were a single concordance.

Future investigations and research will be conducted into the viability of using HTTP to distribute components of a virtual concordance over multiple machines. Although this is not an entirely new idea this does represent a natural extension of the distributed web paradigm which, with increasing bandwidth, may become a practical proposition in performance terms.

3.2 Document Management

AustLII's approach to document management has been affected by the need to produce results under time constraints and competing considerations. When a new database is to be added AustLII starts by receiving sample documents from the data provider. Generally, AustLII requests documents be provided in RTF format, however this is not always possible. If the data provider supplies another format, then it is usually converted to RTF first before being converted into HTML. For acts, an

intermediate standard format called STATUS is used. Standard HTML headers and footers are then added and the resulting document run through AustLII's automatic markup software.

As part of the negotiation process with the data provider provisions are made for continuing data feeds. This is increasingly set up as an e-mail process, where the data provider is able to e-mail new documents which are automatically received and processed (thus, High Court judgements can be available on AustLII within minutes of being sent from the court). AustLII has also begun using Gromit to fetch updated data from remote sites before converting them on AustLII. The Commonwealth, ACT and South Australian legislation databases are maintained this way, using SCALEplus [23] as the data source. The recent addition of Tasmanian legislation[24] is also due to this process.

Once established the markup and updating process becomes largely automatic. However the increasing number of databases and variety data sources and delivery mechanisms is becoming problematic in document management terms. In particular, the current approach means that each of the 86 databases currently published has its own unique 'front-end' scripts, which are controlled and understood only by their author.

A further problem results from manual editing of databases. This happens rarely and is generally avoided, but is sometimes required to quickly remove a case from publication where a suppression order has been made or to correct significant markup problems. Currently no mechanisms exist for tracking this kind of work and make it difficult to be confident that automated rebuilds do not override the sometimes important changes that have been made.

3.3 Maintenance and Other Problems

Maintenance of AustLII databases is generally handled fairly well however there are a number of occasions when manual editing is required. Manual editing is difficult to track and labour intensive. It is important to track such things as the removal of a court case that has fallen under a suppression order; updating and checking legislation; and correcting or updating documents that contained errors. Unfortunately, current methods do not allow audit trails to be developed, and make it difficult for data providers to update their own databases to check and correct errors.

The current system also does not allow for particularly sophisticated user authentication and so requires updates to be 'hand checked' by staff members before being uploaded onto the live system. AustLII is currently conducting research into digital signatures and electronic delegation of legal authority. However there must first exist a technical platform from which to test such systems.

AustLII faces many other technical issues, some of which have to do with scalability and some of which are concerned with the constant imperative to add functionality. Some of the other items that are on the technical agenda include:

– consideration of a second generation search engine to replace sino;
– further generalisation of hypertext markup approaches; and
– Research and investigation into the extent to which the existing expert systems knowledgebases can be automatically generated.

AustLII has always sought to automate as much as possible, and this has been one

reason why AustLII has been able to build such a large legal database so quickly. However the 'glue' which keeps AustLII's parts together is becoming stretched and it has become clear that it is time to start planning and designing for the next development cycle: the 'next generation' of AustLII.

4. The Anarchivist Solution

What AustLII has come to need is a sophisticated document management system designed specifically for the web. Such a system would include a common toolset and standard practices, while still remaining flexible and tailorable to the unique capabilities and data formats of data providers. This would not only help solve the problems of the production server but go some way to providing a platform for continuing AustLII's original R&D aims.

AustLII's new Anarchivist project consists of four emerging technologies, the first three of which are based on open standards. These are:

– LDAP: Lightweight Directory Access Protocol;
– WebDAV: Web Distributed Authoring and Versioning Protocol;
– XML: Extended Markup Language;
– SinoCGI / API: New SINO technologies.

4.1 LDAP

LDAP stands for Lightweight Directory Access Protocol. This is an open standard derived from X.500 – only without all the intervening OSI layers. LDAP is a hierarchical, distributeable directory service (ie a database). One advantage of LDAP over traditional relational databases is that it naturally allows us to organise AustLII's information collections in a hierarchical and potentially distributed manner. Hierarchies are how users are used to accessing complex databases, and how AustLII's maintainers are used to organizing them.

While most developers view LDAP as a way of maintaining distributed phone and e-mail directories for personnel, LDAP's chief advantage lies in the flexibility of the objects it can store. A database maintainer can create a schematic for any kind of hierarchical database they care to create, and then populate the database and enforce the schema to ensure data integrity and consistency. For AustLII, LDAP can therefore become the backbone of an extensible and distributed document management system.

What such a system would provide is a central repository of meta-data [25] on every document in the AustLII database, easily organised into a logical hierarchy. The potential of such a system, from AustLII's technical point of view, is enormous. Many of AustLII's 'blue sky' technical plans rest on such a system being in place. However it is important to emphasise that the end goal is to find ways to increase access to justice through better access to legal information and to avoid building structures and technologies that do not contribute to that goal.

4.2 WebDAV

Tim Berners-Lee's original vision of the World Wide Web differed significantly from what we have today in one important respect: the current web methodology involves publishers, who maintain control over content, and users, who are generally pas-

sive consumers of information. The original vision was that the web would be a collaborative medium – a global conversation. However, at the time that Netscape and the web gained mainstream media attention, only half the picture had been implemented. To describe most web sites as 'interactive' is to completely misunderstand the term – the true potential in that word has barely been realised.

WebDAV is a set of extensions to the HTTP standard that allows web clients to update server documents in a secure manner. A WebDAV server consists of 'collections', which are analogous to directories, and documents. WebDAV specifies a set of protocols for creating, editing, moving and deleting these documents and collections. WebDAV allows a genuinely distributed authoring environment.

An organisation like AustLII can benefit from WebDAV in a number of ways. It can be used internally as the mechanism for updating databases. It could also later be used by the courts themselves, to update and maintain their own collections. AustLII's current plan for WebDAV is to be the interface protocol to Anarchivist – the document management system.

Some important deviations from the current WebDAV standard may be required. For example, Anarchivist will be required to track not only documents local to AustLII, but also documents and document collections existing on remote servers. This will require WebDAV clients to be able to update the meta-data relating to a document, but not the document itself. There is also poor client support in the current environment, however strong client support is not required yet for Anarchivist to work well.

4.3 XML

XML is the eXtensible Markup Language. XML is a simpler version of the popular SGML. It is designed to lower the costs (in terms of time, money and expertise) associated with using SGML to represent structured documents. It is also designed specifically for use on the web, with expanded linking capacity beyond that currently offered by HTML. XML is already supported by a number of web browsers and is likely to replace HTML over the next five to ten years. XML is a crucial part of the WebDAV protocol, since it forms the basis upon which clients and servers communicate in a WebDAV environment.

The early HTML standards focused on representing data by describing what it is (eg 'this is a heading') rather than how it should be displayed (eg 'bold, 14 pt Times Roman'). However the standard was designed around representing academic papers, and allowed authors very little control over how documents were displayed. This led to browser makers extending HTML in unplanned (and occasionally bizarre) ways. While the need to represent structural information in documents is very important, it is also clear that the variety of applications upon which HTML is built requires a more flexible approach to both the structuring and display of data.

This is the realm of SGML, a 'language for writing languages' upon which HTML is built. Much work has already been done in the legal domain using SGML [26]. However SGML is a complicated language and requires significant investments of time and money before it can be put to practical use. XML is an initiative supported by the World Wide Web Consortium and has been designed as a simplified version of SGML.

AustLII's immediate concern with XML is as the language used by WebDAV servers and clients to communicate. However AustLII has already begun long term planning for switching to XML as its primary data representation standard. The ability of XML to support legacy data in a structured way will help this process.

4.4 SINO CGI & API

AustLII's current search interface is based on Perl CGI scripts which interact with the sino process using TCP/IP sockets. On the one hand this allows sino processes to be distributed among multiple machines, allowing for some form of load balancing. However there are considerable performance overheads associated with each component of the search interface. Perl is interpreted and must re-compile the interface script for each search. The CGI protocol requires a new process to be 'forked' (started) for each incoming search. And communicating with sino over TCP/IP places extra load on the local network.

AustLII's new search interface is based around two technologies:

– FastCGI [27]: An independent and open replacement for the CGI standard, where persistent search servers continually listen for incoming search requests. This avoids the per-search performance hit of CGI which requires a separate process for each incoming search;
– Sino API: A new interface to the sino search engine that allows the sino library to be embedded directly into the interface program. This avoids the TCP/IP and Perl overheads.

In addition to the interface changes, the new search interface introduces the concept of a virtual concordance. A virtual concordance is one or more physical concordances, which may be associated with zero or more mask paths. Mask paths in sino are the mechanism by which search results are restricted (used for example when a user only wants to search High Court cases). By combining concordances and mask paths into one virtual concordance architecture, AustLII has been able to introduce the simple but powerful World Law search facility [28].

The change in interface architecture from Perl/CGI to sinoAPI/FastCGI has led to performance improvements well beyond initial expectations. AustLII usually processes one search a second, however at peak times the rate may increase to two searches per second. However performance testing revealed that AustLII's maximum search performance under heavy load was just under two searches per second in ideal circumstances. This created a serious bottleneck with a major impact on system performance during peak times. While sino itself was very fast, its interface to the web was not scaling.

The table below gives performance figures obtained during development. The searches were conducted on an otherwise idle web server. Each test was conducted twice, with the results of the first test discarded. This removed the issue of caching and disk head seek times. The performance figures revealed a 'base case' sixteen fold increase in search capacity.

	CGI/Perl		Fast CGI/C API	
	Searches per Second	Transfer rate	Searches per second	Transfer rate
Single word search ('banana')	0.59 rps	14.91 Kb/s	9.48 rps	190.68 Kb/s
Single proximity search ('environment near pollution')	0.56 rps	14.59 Kb/s	4.38 rps	107.48 Kb/s

Key: Rps: Requests per second
Kb/s: Kilobytes per second

Notes:
1. Conducted using Apache Bench with CGI 'GET' URLS setup to conduct searches with similar options.
2. Apache Bench conducted searches in blocks of 10 with a concurrency level of 5

Table 1: Performance figures for sino searches

The results indicate the impact of the old interface on overall performance. In the new FastCGI model, performance is much more closely tied to the performance of sino itself (hence the more pronounced drop in searches per second when a more complicated search was done). It is important to note that the performance figures are only relative – the tests were conducted on a significantly slower machine than the current production server (whose current maximum search rate is approximately 1.5 searches per second).

4.5 Anarchivist Architecture

In the new Anarchivist, document meta-data is stored in LDAP objects. Actual file data is stored in the Unix file system, using a simple mapping between LDAP distinguished names and file system paths. Some objects will be 'remote' in that there will be no corresponding local data file.

To update the meta-data, a modified WebDAV protocol will allow updated information to be sent via an XML encoding. Where the file contents itself are to be created or modified, WebDAV will be used to store both meta-data and document body. In the medium term, HTML will be the data format for document storage, however Anarchivist should also allow for the original 'pristine' source to be stored along with the HTML version.

The contents of the Anarchivist document repository should be 'mirrored' to the local file system, for 'static page' serving by the web server. Standard headers and footers would be added at this point, along with any output from the automated markup scripts such as findacts. This is done mainly to speed user access to the data and to provide clean source for sino to index.

Anarchivist will also be able to store compressed versions of files (particularly case

law), which can be served directly in compressed form to the latest browsers, or decrypted on the fly at the server end for older browsers. This saves both space on the server end and time on the client end for those clients which support streaming decompression.

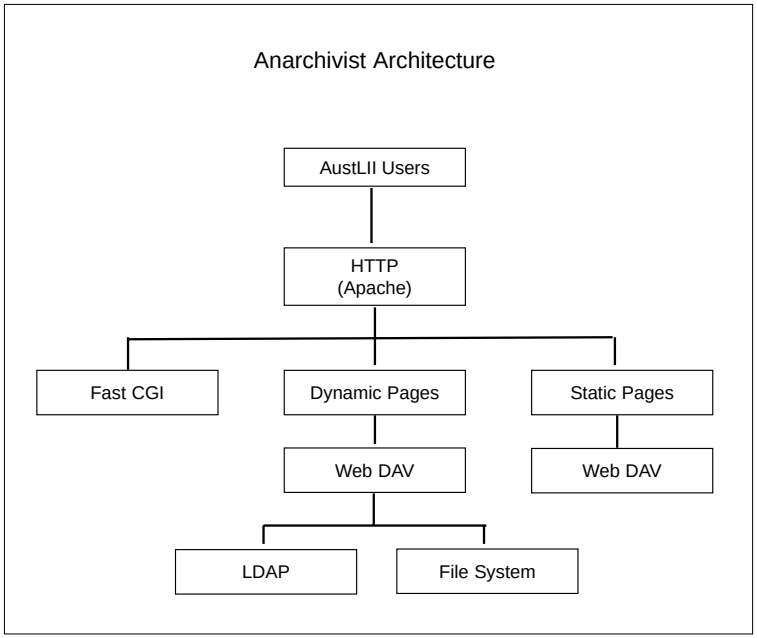


Figure 7: Simplified Anarchivist Architecture

5. Conclusion and Future Directions

From a technical perspective, the AustLII project has involved the development of a number of new approaches to dealing with legal information. The current system reflects a mixture of the practicality necessary for a production service with a large user base and ongoing research based experimental systems which attempt to expand the expectations that users can reasonably expect.

AustLII hopes that over the next year, the emerging Anarchivist platform will serve as the basis for new research projects, including:

– A new system for courts to send and update new cases directly, using industry standard WebDAV clients, authenticating themselves via digital signatures;
– A parallel citation database containing vendor and medium neutral citations, as well as vendor standard citations;
– A second generation sino search engine, supporting multiple document type indexing, concordance distribution and parallel processing.

All of the above must be tempered with AustLII's usual pragmatic approach to technology and focus on broad public policy goals.

Footnotes

1. Greenleaf G et al, 'Introduction to the AustLII Papers', Background papers for presentations by AustLII staff at the 'Law via the Internet 97' Conference, 25-27 July 1997.
2. Andrew Mowbray, Graham Greenleaf, Geoffrey King and Peter van Dijk.
3. Source: 'Where did we go in Australia?', <http://usrwww.mpx.com.au/~ianw/> (as at 1 July 1999).
4. Austin D, 'AustLII User Survey 1998', < http://www.austlii.edu.au/austlii/survey/> (as at 30 September 1998).
5. An Enterprise 3000 with 2 x 400MHz CPUs with 1.3G of memory called 'bar' and two Sparc Ultra 170s with 256M of memory called 'bronte' and 'bondi'.
6. Mowbray 1995-. See Mowbray A 'Sino User Manual' AustLII 1996 <http://www.austlii.edu.au/austlii/sino_full.html> and Greenleaf G, Mowbray A and King G `Public legal information via Internet: AustLII's first six months' Law Technology Journal, CTI Law Technology Centre, Vol 4 No 2 November 1995, 5-10, ISSN 0961-6902 (also in Australian Law Librarian, Vol 3, No 4-5, 1995, 144-153, ISSN 1039-6616).
7. See Greenleaf G , Mowbray A, Tyree A (1992) 'The DataLex Legal Workstation – Integrating tools for lawyers' Vol 3 No 2 Journal of Law and Information Science (1992) 219 -240 (also in Proc. Third Int. Conf. on Artificial Intelligence and Law, ACM Press, 1991).
8. King and Austin 1996-. See Greenleaf G, Mowbray A and King G (1997) `New directions in law via the internet - The AustLII Papers' Journal of Information, Law and Technology (JILT), Issue 2, 1997, University of Warwick Faculty of Law, (electronic journal) located at <http://elj.warwick.ac.uk/jilt/issue/1997_2>, 30,000 words (also published as `The AustLII Papers' in Proceedings of the Law via the Internet `97 Conference, AustLII, UTS/UNSW Faculties of Law, June 1997).
9. ibid.
10. wysh was originally written by Geoff King and Simon Cant in 1996. See Greenleaf G, Mowbray A, King G, Cant S and Chung P (1997) 'More than wyshful thinking: AustLII's legal inferencing via the World Wide Web', Proc. 6th International Conference on Artificial Intelligence and Law (Melbourne 1997), ACM Press, Association of Computing Machinery, New York, 1997, 9 pages.
11. Mowbray 1993. See Greenleaf G, Mowbray A and van Dijk P (1995) 'Representing and using legal knowledge in integrated decision support systems – DataLex Work Stations', Artificial Intelligence and Law, Kluwer, Vol 3, Nos 1-2, 1995, 97-124.
12. Apache Software Foundation, <http://www.apache.org/>.
13. Free Software Foundation, <http://www.fsf.org/>.
14. Wall, L and the Perl Software Consortium, <http://www.perl.com/>.
15. Vendor and medium neutral citation provide a system of court assigned unique descriptors for cases based upon a court designator, a year, a decision number and, if necessary a paragraph number.
16. Sino currently consists of less than 8,000 lines of C code.
17. And hence the name Sino – 'Size is no Object'. Apart from being a reaction to the very slow retrieval times of glimpse vs the very good concordance ratios that it was achieving, the name was also meant to reflect the fact that sino could handle very large text databases.
18. 3,969 as at 19 July 1999.
19. Lawrence & Giles, 'Accessibility of information on the web', Nature (Vol 400), 8 July 1999.
20. Austin, 'AustLII 1998 Survey Results', <http://www.austlii.edu.au/austlii/survey/> (as at

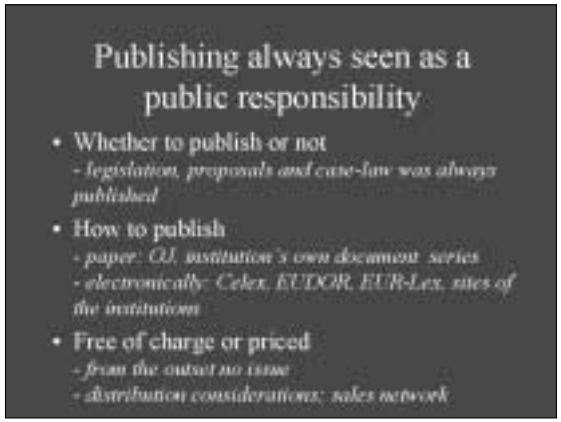
30 September 1998).
21. More accurately, 231 or 2,147,483,648 bytes.
22. See The Sino Search Engine, above for an explanation of a 'concordance file.'
23. See <http://scaleplus.law.gov.au/>.
24. See <http://www.thelaw.tas.gov.au/>.
25. For AustLII's purposes, such things as document title, data source, publication date and version history.
26. Poulin, Lavoie & Huard, 'Supreme Court of Canada's cases on the Internet via SGML', Law via the Internet 1997 Conference Proceedings, 25-27 July 1997.
27. See <http://www.fastcgi.com/>
28. See <http://beta.austlii.edu.au/links/World/>

Internet and legal information in the EU administration

Henric Stjernquist, Publications Office of the European Communities

I have worked for three years at the European Commission, or more precisely at the Publications Office in Luxembourg, where I am responsible for the production and management of the contents of our legal information services Celex and EUR-Lex. Today I have been asked to make some comments from a European point of view. With European in this context we mean the institutions of the European union. I would, however, like to stress that what I say today does not necessarily reflect the official position of the European Union, but should rather be seen as my private opinions.

I will start with a background description of the public information policy of the European Union in general and how it has developed in the last few years. In particular I will focus on the new Public Access to Documents Regulation, which is applicable from 3 December this year, and how this will affect the availability of legal texts for example. Then we will take a look at the users, what we believe that our users want and who could and should produce the value-added information they need. Finally I will also discuss the costs of producing and distributing legal information compared to the revenues they generate. Should the users or the European taxpayers pay for the value-added information and is it public responsibility to produce it? In the end you will hopefully make some conclusions. From me there will probably be more “concluding” questions than real answers.



I think it is correct to say that publishing of legal documents has always been seen as a public responsibility by the European institutions. You have some founding member states where legal information is more or less a private business, like in the Netherlands, whereas in France, for example, publishing is seen as a very important public task. So the question for the founders of the European Communities was probably not who should publish, but rather what to publish. And we can see that already from the outset not only legislation but also proposals and case law have always been published by the institutions themselves, via the Publications Office.

There also questions, of course, on how to publish. From the beginning the institutions set up their own Official Journal and the institutions also created their own document series. But the European Communities also entered into electronic publishing quite early. The Celex database in 1970 was one of the first legal databases in the world, and it has continued with other interinstitutional electronic information services and, more lately, various Internet sites run by the institutions themselves.

Whether legal information in general or the legal documents themselves should be free of charge or not was not really a question anybody asked in the beginning. It was considered more or less self-evident that this kind of information was something you had to pay for, and this view has been predominant up until now. The rationale behind this was hardly to cover all the production and distribution costs, I don't think that the revenues have ever exceeded the costs, but rather to make it possible to build up a distribution network. It became clear at very early stage that it was not possible to distribute and have a contact with all the subscribers centrally from Luxembourg. Therefore a network of national sales agents was set up, not only in the members states but all around the world. And such a sales network can of course not exist if they have nothing to sell. This is still a main concern even if the attitude towards free of charge distribution of legal information is changing.

And indeed things have changed. The movement towards increased openness and transparency, together with the emergence of the Internet has both contributed to this change. For the EU it started with the opening of the EUR-Lex service in 1998. EUR-Lex, which is operated by the Publications Office, was at that time a sort

of free Official Journal on-line, intended for the citizens, but is now developing into a portal site for access to EU law. Following EUR-Lex a whole range of free web sites has been opened by the institutions themselves. In spite of all these free sites the general price policy has more or less remained the same. But now, it is not only the sales network that must be protected. We also have to discuss the value-added we provide and how to protect this. We will come back to that later.

We will now take a closer look at the new Public Access to Documents Regulation, which was adopted during the Swedish presidency and will be applicable from 3 December this year. Formally, it covers access to European Parliament, Council and Commission documents, but it is clearly said that it should as far as possible be applicable to the other institutions.

The aim of this Regulation is to give access to all kinds of documents, very much like the public access you know here in Sweden. But the Regulation has some rules that are of importance directly for the access to legal documents. If we look at Point 6 of the recitals:

> "*Wider access should be granted to documents in cases where the institutions are acting in their legislative capacity, including under delegated powers, while at the same time preserving the effectiveness of the institutions' decision-making process. Such documents should be made directly accessible to the greatest possible extent.*"

What does then 'directly accessible' mean? We can look in Article 12, 'Direct access in electronic form through a register'. The idea is that all institutions should create such registers and the aim is to make them available on-line. Article 12 says in Point 2:

> "*In particular, legislative documents, that is to say, documents drawn up or received in the course of procedures for the adoption of acts which are legally binding in or for the Member States, should, subject to the Articles 4 and 9, be made directly accessible.*"

Point 3 is also worth mentioning which refers to, for instance, white papers and green papers of the commission:

> "*Where possible, other documents, notably documents relating to the development of policy or strategy, should be made directly accessible.*"

Which means directly accessible in electronic form on the Internet.

In Article 10 there is an interesting line in the end of Point 1, concerning consultation of these registers, saying:

> "*Consultation on the spot, copies of less than 20 A4 pages and direct access in electronic form or through the register shall be free of charge.*"

This doesn't, of course, mean that we will not be able to continue to provide legal information services against payment, but the conclusion we can draw is that the texts as such should be obtained free of charge.

There is another article I would like to draw your attention to, saying which acts should be published in the Official Journal. We can see that, apart from legal acts in the strict sense, Commission proposals, common positions, framework decisions, conventions established by the Council, conventions signed between the Member States and international agreements concluded by the Community should be published there. This is actually the first time that it has been clearly stated that Commission proposals must be published in the Official Journal. They always were, but the interesting thing is that even though it is said that they must be published in the Official Journal this doesn't mean that it has to be done in the printed version. A year ago we created an electronic Annex to the Official Journal. Commission proposals were published in the C series – you know there are two series, the L and the C series of the Journal – and the Commission proposals are now published in the C E, which is this electronic annex to the Official Journal. The interpretation of this is that we have to continue to publish them in the Official Journal, but only electronically.

The same article presents also a number of documents that "as far as possible" should be published the Official Journal: initiatives presented to the Council by a Member State, common positions referred to in Article 34 of the EU treaty, and certain directives and decisions not mentioned above.

So, as I already said, one conclusion that can be drawn from this Regulation is that legal texts should be available free of charge, at least if they are in an electronic form. And that is also why of the Publication Office decided this year that legal documents will be available free of charge on-line from January 1st 2002 – irrespective which format they are in, whether it is html, pdf, tiff or ASCII.

At the same time the management committee stressed the importance of having a clear strategy concerning the value-added. What is the value-added we will provide to paying clients? We have tried to identify a number of elements that will be parts of this continued value added service. This is very much focused on the service itself and not the value-added you may put into the documents themselves.

- First of all a *fast and reliable access* is something the paying clients will be expecting. Of course we all want a fast and reliable access, but if you pay you can rightly expect to have the service running whenever you wish to consult it.
- Certain *advanced search functions* you don't find in the free of charge interface.
- *Analytical data*. Some of the meta data that cost us a lot of money to produce could be restricted for certain search and display purposes that are only available for paying clients.
- Access to various *export facilities* enabling users to download documents in certain formats for further processing such as creation of databases etc.
- *Help, training and documentation* is of course essential elements of a value-added service.
- *Newsletters, profiling services and alert services*, meaning that you will get a message when a new document meeting your predefined this profile has been loaded, could be considered as value-added.
- So could access to *interactive communication* between the users, with the helpdesk and so on.
- It could also be that some parts of the information would *only be for paying clients*, such as certain references. One example is an internal database of the Court of Justice with references to national case law interpreting EU law. This information could perhaps be available only in the pay service.
- Finally, *certified authenticity* of the documents would indeed be a considerable value-added.

These elements of value-added information that I have just mentioned are already or

will in the near future be available for paying clients in our Celex service, whereas free of charge access will be available in EUR-Lex. Whether it is the responsibility of a public organisation to produce and provide users with this kind of information at all is another question. Whether it should be available free of charge for all users and not restricted to paying clients yet another. I will come back to these questions later.

Just a few words about the users, who they are and what they possibly want and also, to come back to the question before, who is the best producer of added value.

It has for a long time been assumed that there are citizens that are never prepared to pay for anything and professional users that are always prepared to pay. We can, however see that when we open a free of charge service, like EUR-Lex, that also a lot of professional users go to this service instead of the paying service if they can find the information they want there. So I think the reality is more complex. Perhaps a better way of distinguishing between different groups of users could be legal professional users, non-legal professional users, and non-professional users. Whether they are prepared to pay for value-added information or not is still difficult to say. I think features like high accessibility, good service functions etc. are of interest for all. Whether it is worth paying for or not depends of the quality of the basic free of charge service. But if you go a bit further. Aren't compilations and interpretations of the documents even more interesting? And if you look at the citizens' need of information, I wonder if it is not really this kind of information that is most useful, rather than just providing raw documents without further explanations and additional information. But this kind of information, that indeed possesses a high degree of value-added would be difficult to charge for from ordinary citizens whereas professional users in most cases gladly would pay for it. But beyond the question of pricing, what is the public responsibility in all this?

First of all you can distinguish between a centralised and a decentralised solution. When I talk about a decentralised solution in the European Union I talk about a handful of institutions. It is not like in Sweden with hundreds of courts and government agencies. But we can still see the difficulty. The way we work when we produce some of the value-added in our legal information system is to make a legal and documentary analysis to add descriptors, links between documents and so on. And already with five or six information providers this is a huge problem. So I think that if you opt for a decentralised solution you will probably have to create a system with not too much of meta data and analytical information. Otherwise it will be very difficult to maintain coherence and a common methodology.

But then you can of course ask yourself whether there is a market for all this information. Perhaps we exaggerate and produce information that is not really needed at least in proportion to the costs involved. Or, if that is not the case, perhaps

we should leave to the private market to produce and distribute more of this information. Maybe that would be more cost-effective? I will come back with some examples of the costs we have for the production of value-added information. But first, let's have a look at some other interesting figures.

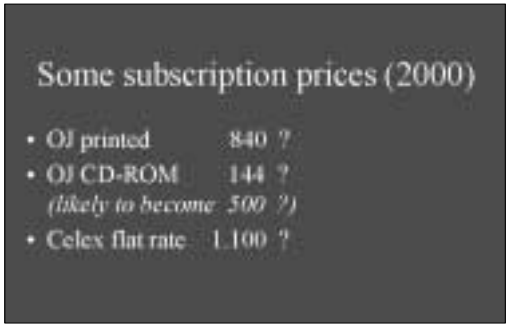
Official Journal, C and L series

	Paper		CD ROM	
	1996	2000	1998	2000
ES	1988	1320	74	222
DA	505	260	36	51
DE	2231	1654	91	328
EL	318	169	11	64
EN	2832	1682	168	387
FR	2951	2017	124	433
IT	2359	1604	151	551
NL	1025	681	51	85
PT	615	419	46	179
FI	245	116	4	20
SV	261	137	13	22
TOTAL	15330	10059	769	2342

This is the number of subscriptions to the Official Journal. It is the distribution that goes outside the institutions of the European Union, which of course, are important users of this publication. But on the other hand, all officials of the institutions that want to receive a paper copy have to justify that they really need it on paper. Otherwise they can consult it on the Intranet of the institutions. In the table we can see the figures for 1996 and 2000. If we look at Sweden, for example, there were 261 subscriptions to the paper edition in 1996 and 137 in 2000. You can imagine the cost per copy. If you look at the English and French the number is not very impressing either. What you can se is a decrease from almost 3000 to 2000 in four years. In 1998 we started to produce a monthly CD-ROM containing the Official Journal including the electronic annex I talked about earlier. In 2000 there were 22 subscriptions in Swedish and some 500 in French and Italian, so that doesn't explain the whole decrease in the printed edition.

On the other hand, on the EUR-Lex free of charge Official Journal site we have the contrast to these statistics. We are able to distinguish 17.000-18.000 individual users per day and they display 120.000-180.000 pages per day. At the same time in the paying service Celex we count around 1200 users/subscribers and the numbers of documents displayed has increased from 20.000 in 1996 to some 80.000 in 2000.

If we then take a look at the prices of the various products we find that a subscription to the printed version of Official Journal costs 840 euro and the CD-ROM 144 euro. Next year we will produce a CD-ROM with some more value-added information and better search functions and the price is likely to be about 500 Euro. If you want to access Celex there is a flat rate meaning that one simultaneous user can access the service as much as he likes in a year for 1.100 Euro.



These are some approximate figures for what the production of the value-added service Celex costs. The budget for Celex, which is now also used as a common catalogue for the free of charge portal and not only for the paying subscribers, is about 2.000.000 euro, excluding the distribution costs. Out of that 700.000 euro is spent on the legal and documentary analysis. Just as an example, indexing with Eurovoc – a thesaurus specially developed for the European institutions used among other things to index everything we publish in the Official Journal – costs us 12.5 euro per document. And with 10.000 documents a year indexed that is of course a significant cost.

You can of course ask yourself if the possibility to make a search using Eurovoc descriptors is worth this cost. But another problem is that, since we have a partly decentralised system, certain documents from e.g. the Court of Justice do not contain Eurovoc descriptors at all. If you make a search using Eurovoc you will not find anything from the Court of Justice. If the indexing is not used for all documents the value of it can definitely be questioned.

Finally a look at the revenues of Celex tells us that we get less than 45% back of what we spend, or 842.000 Euro. 400.000 Euro, or about 50% of the total revenues, come from 28 license holders. More than 50% of the rest, or 284.000 Euro, come from 400 flat rate subscribers. So you can definitely say that it is a few big users that stand for the largest part of the revenues.

To conclude all this I only have a number of questions.

As you have already heard the texts will now be available free of charge but what about the value added? Should we do more as a public body? Or should we do less? Is explaining the law a public responsibility? When we create sites for non-professional users, is it sufficient just to give them access to the text, or should there be at least some explanations?

As you have seen, we have few paying users, we have quite high production costs. Could the market do this better or is this still a public responsibility?

A fact that I didn't mention is that for the Celex database we don't have that many paying clients, but the by far biggest user group is the institutions themselves, which use the service for various purposes. And probably we would have to do a lot of the production and development work for them in any case.

Last but not least there is the question of multilinguism. What would happen if we left all this to the market? What would happen to the minority languages? You saw the number of subscribers to the OJ in various languages. I find it difficult to believe that the market would develop the same value-added services in Swedish, as they would perhaps do in English, French and German.

Principles of the construction of the Swedish legal information system

Christian Levander, The Government Offices for Administrative Affairs

I will now tell you something short about the Swedish official legal information system, what we call Lagrummet – what we have done, why we have done it, and what we want to do in the future.

Some historical background: We have in Sweden, as most of you probably know, since 1980 a legal data system. I won't tell you the details about that since that was before my time, but it was not then as big as today and, since it was 1980, it was not at the Internet. In 1996 the IT Commission wrote a paper to the Government proposing a wider responsibility for the State of publishing basic legal information on the Internet. The Government thought it was a good idea and appointed a working group within the Government office. The group presented its findings in a report in the beginning of 1998 (Ds 1998:10). The report proposed a new legal information system. The report was referred for consideration to over 70 agencies, and generally it had a very positive reception. With that ground the Government informed the Riksdag about its ambitions with the system (skr 1998/99:17) and decided on a regulation, called *Rättsinformationsförordningen.* These are two documents that are very important for us that are working with the system.

According to the regulation, July 1st 2000 we set up a website, www.lagrummet.gov.se, or "the Legal Citation" as it was translated earlier today. It is a portal with links to all legal information the system consists of.

The government proposed that we should have a legal information system, and the main reasons for that were as follows:

– Firstly, there is a legal reason, a principle called in Latin "Ignorantia Juris Nocet", which means that every citizen has to be familiar with the contents of the legislation. You can't say, when you have committed a crime, "I didn't know it was forbidden", or at least you wouldn't be listened to. To make it possible for the citizen to have a theoretical possibility to know, the legislation should be provided free on the Internet.
– Secondly, a democratic reason, the principle of every citizen's right to free access to information about the agencies and the society. It is similar to the principle behind the 24-hour, seven-days-a week government ("24/7") agency.
– Thirdly, there is the efficiency reason. The Government thinks that an important reason for building the system is that access to laws and regulations makes it easier and more efficient to work for the public sector.

The main target groups, or target audience, are the citizens and the public sector. That is stipulated in the regulation.

The Government's demands on the system consists of two parts.

– Firstly, there are the demands in the short run, as posed in the regulation. These demands were fulfilled in the first of July 2000.
– The Government's paper to the Parliament also contains demands in the long run. You can say that these are the Governmental ambitions of the system. There is no time limit for these.

The demands in the short run are

– The State is responsible for basic legal information at the Internet. I think the most important word is "basic". The State is not responsible for anything else than this. The meaning of the word "basic" is an important discussion, and it also concern the issue of the co-existence of the public and the private sector in the legal information area
– The system comprises a mandatory content, that has to be in the system, Preparatory works. Government bills, laws and regulations, including regulations of Government and agencies under the Government, Swedish case law, including some decisions of agencies and other legal information. What other legal information to be included is grounded in agreements between Government offices and the agency that wants to publish it. One important agreement is with the Parliament.
– The information is free of charge.
– The responsibility is decentralised, which means that the public organisation that issues the information is responsible for the publishing on the Internet, and that the information is correct and up to date. This means that building the system is a co-operation between the Government offices, the Parliament, the higher courts and about 80 agencies. We are all part of the system. It also means that the information is not localised to one place. It is in several databases all around Sweden.
– The Government Offices is responsible for the co-ordination of the system. Within the office there is a Council, with representatives of the agencies on a high level: the Government offices, the Parliament, the National Courts Administration, the Agency for Administrative Development, and some other agencies. The council discusses strategic issues and can also make some decisions.

The demands in the long run, what we can call our visions, are:

– A uniform presentation. The Government has said that you should recognise when you are in the legal information system.
– The Government also want to make it possible to make structural searches in all the information at the same time. This means that searches in all the different databases take place in different places. The Government also talks about links between documents.

I will now mention something of our experiences in technical issues.

In the regulation it is stated that the Government Offices can decide on the use of mark-up standards. The Government Office has delegated this to the Council. One decision we have made is the use of XML, Extensible Mark-up Language. I think most of you know what XML is; it is a language which is not depending on any private company or privately owned software product. It is a recommendation

from the organisation W3C, consisting of all the big software producers of the world. Most experts say that XML is probably an important standard on the Internet in the future, and probably also for the legal information system.

Our experience is that there are problems with XML today for our types of documents, at least problems we haven't solved yet.

– XML editors is a tool where you write in documents in XML. We think that these editors are too expensive and too difficult to use for ordinary users. That means we have to rely on MS Word or other similar programmes most of the public sector use today.
– The conversion from MS Word to XML is not 100% safe according to our experience. It requires manual handling, and manual handling often leads to errors. The quality of the information system is very important to us. If the user can't trust the information he has to go to the printed version.
– We have also problems with searches in technically different databases.

These are problems we haven't solved yet.

Finally something about how we think about the future. We have an official inquiry, called Sverige Direkt. Sverige Direkt is a portal to public information in Sweden. Now, the Government wants to investigate how this portal and other portals within the public sector, including the legal information system, will be organised in the future. Sverige Direkt has in a report in June 2001 proposed the start of a new agency responsible for these kinds of questions. That means that the responsibility for the co-ordination for the legal information system will be transferred to this new agency. We are of course very curious about this inquiry. But in the meantime we will work with the following:

– We think it's very important to show quality, to make sure that the information of today regarding technique and searching possibilities are correct and up to date. The information in the system has to be trusted.
– We want to perform a uniform presentation. You should know when you are in the system. You should recognise yourself.
– Concerning searching we have tried earlier to mark up all information in XML and tried to make it possible to search in all text. We can use much of this work in the future, but we have still a few problems left. Meanwhile, when we wait for the technical development, we want to think of some smaller steps, for example to search for titles. The use of metadata is one possible solution to this.
– For the visions, what the Government wishes to see in the long run: We will of course consider what the Government has said in the paper to the Parliament, but also follow the current situation. What is the situation in other parts of the world? What do the private market and the academic world thinks? We want to consider the technical development and wait for the tools that are good enough for us. Most important, I think, is the demand of the target groups. What kind of system is best for the citizens? Is it important for them to make qualified searches? Or it is more important to write easier texts about difficult law and regulation texts?

Panel discussion

Peter Seipel:
With the assistance of the panel, let us now try to single out some of the issues that have been brought forward during the previous parts of the conference. This is not an easy task. For one thing, we have been discussing many of the issues for many years now, and the preconditions for this discussion have been changing all the time. The changes that have taken place with the advent of the Internet and the World Wide Web are certainly dramatic. They have made things possible that used to be regarded as more or less a product of fantasy. Take for example the development that Andrew Mowbray described, an academic institution with limited resources establishing itself as one of the leading institutions in the field of web-based legal information. Evidently, the web has played an important role for making such developments possible.

You may recall the questions I tried to formulate in my introduction. The first one is often asked, viz. to what extent should legal information be made available free of charge for the citizens? We have not yet answered this question. Some of you have commented on it, but let us attempt once more to try to define precisely what we are talking about and whether it is possible at all to find a general solution to this particular issue.

The second issue: What tasks ought to be performed by the public sector and what tasks by the public sector? This has nothing to do with the free of charge issue, because you can very well imagine a situation where tax money is used to pay for "free" information services to the public provided by a commercial operator. That is but one example of how things could be arranged.

The third question I posed concerned what kind of co-ordination and central steering is desired and for what ends. This question is also independent of the two other ones. You can think of many combinations of solutions and strategies.

With these short introductory reminders, I leave the floor open for comments.

Henric Stjernquist:
I more or less have the same questions. If you say free of charge, what kind of information do you mean that you want to make available free of charge? It has now been decided to make all texts available free of charge, but if we don't provide the tools to find and interpret all this information what is this free of charge access worth? It has also been decided that more advanced search facilities should not be available free of charge. But isn't that something that is of importance also for the citizens? What we do now is to give them a free of charge portal. All the documents are there, but it is very difficult to search. You may end up with search results of thousands of documents you have to go through. Is that perhaps more of a way to give yourself a good

conscience that you have done something for the citizens?

Peter Seipel:
I can think of a new legal principle here: "Information polluter pays".

Andrew Mowbray:
I think I will concentrate on the first question – to what extent things should be free. Firstly I think it is internationally agreed that the public has a right to public legal information, and that it should be free. This is not exactly as easy as it sounds; what do you provide as free? One thing we say in the paper I distributed is that as the very least the responsibility of the government is to make the basic materials available. If a search facility can be added, this is an extra thing you can do that is very valuable. One of the things that struck me when Henric Stjernquist was talking was that if the raw EU data is freely available, that some international organisation like AustLII will put it up on the Web.

Peter Seipel:
Would you say that the problems of charging are almost marginal then?

Andrew Mowbray:
One of the big things with the Internet, and one thing that is very different from the old-style dialog service, is that the way you can make money out of it is quite different. Charging is something which will always have the effect of dramatically reducing the audience, especially in the modern environment.

I think that if you do need to raise revenue to do what you are doing you need to think more laterally rather than to trying to charge individuals that happen to use the service.

Peter Seipel:
In your lecture you mentioned the Canadian solution, where every solicitor pays 7 CND per year to obtain the service. This seemed like an interesting option. Would you like to comment on this?

Andrew Mowbray:
The type of audience that use public legal information systems are predominantly not lawyers. As Tom Bruce said in his speech, professionals that are not necessarily lawyers. They are people interested in occupation health and safety, privacy concerns for their firms and whatever it is. If we can extract some money from the lawyers to fund everything else then this is a way of achieving social good.

Peter Seipel:
I would like to invite the conference participants to also join in the discussion.

Björn Westberg, Jönköping International Business School:
It is almost an understatement to say that fast and reliable access to basic legal sources is essential, and also that it should be free. It has been underlined by all repre-

sentatives of the European Union and of individual States, by experiences and also by the presentation before coffee-break when we really had the reference to a governmental bill and a parliamentary bill, saying that there would be free access to Swedish case-law, and that it was supposed to be treated by what was called High Courts, by which I suppose was meant the Supreme Courts and the Courts of Appeal. My question is simple: When? For it has actually been a question since long. I, like many others was very positive with the presentation of the governmental legal site. And my question is: When will there not only be examples of case-law? When will there be real case-law available? That's the reason why I ask: When will it be available?
I have also a small question to Henric Stjernquist. It relates to the same basics. He mentioned as examples of added value "fast and reliable access". My question is: Is not that the basic features, at least if I with "fast" not mean in a technical meaning but "updated information".

Peter Seipel:
The poor man's version: slow and unreliable.

Bengt Nordqvist:
I think a full text version is something for the future. It has been under discussion for years, and today is only a short notice. I don't think it is possible today to go further.

Peter Seipel:
One argument that is quite often put forward in Sweden is that it is not possible to add a full text version of case law on the Internet because this would violate the Personal Data Protection Act, and it would be too expensive to purge the documents from all personal data. A committee is looking into this problem right now in a more general way. I understood from your description of the materials in AustLII that you actually put the full text of the sentences on the net. Doesn't that cause any problems with regard to personal data protection?

Andrew Mowbray:
I think that the attitudes to privacy might be somewhat different between Sweden and some of the other jurisdictions. Essentially in the common law jurisdictions we have a principle of openness of justice. There are very few exceptions to this. If you look at traditional law reports it is very unusual that names are ever removed or anonymized.

Peter Seipel:
It would be interesting to hear some views from the audience about the possible anonymization of court decisions. Quite often when we have discussions about this issue in the Foundation for Legal Information the views collide – the foundation, by the way, has been set up by a number of bodies, the Administrative Department of the Swedish Parliament, the Prime Minister's Office, the National Courts Administration, Norstedts Juridik Publishers, LO-TCO Legal Protection, the Swedish Bar Association, etc, and its sole purpose is to discuss the future of electronic legal information in Sweden. The question raised by Björn Westberg has kept coming up, and quite often the representatives of the public authorities point at the difficulties of protecting personal data, but they also quite often say that there is no need for case law of this sort. You need only the official reports from the Supreme Court and the Supreme Administrative Court etc. Other parties maintain that this is not correct. Practicing lawyers maintain that they also need case law from lower level courts because this will inform them about issues coming up, about the kinds of arguments that can be used in certain legal situations etc. So for various reasons they want to have a broad access to case law. It would be interesting to listen to some arguments and comments on this.

I see an unlucky, possible split here, where we may have one official, free of charge system, where you can only find short descriptions of possibly important cases. If you wish to perform more thorough searches of case law you will have to pay may be quite a lot to have access to the full text material. And some material will perhaps never be made available, like material from the lower level courts, because nobody will systematise and arrange it for distribution. But it is not at all self-evident that Sweden will find itself in this situation. In many other countries – in Australia, in Norway, you name them – it has been possible to arrange for comprehensive case law information retrieval systems on-line. So the present Swedish solution is not self-evident. I think there is much of history in this. We have been used to handle things in a certain way. So, Björn Westberg, I don't think you can get a precise answer to the question "when?", but you can get a number of answers to other questions such as "what is the present line of argument?" and "why does it take time?"

But please, it would be interesting to listen to comments regarding the protection of personal data. Are we exaggerating the problems, can we handle them in a different way? And is it a general interest to have wide-ranging access to case law, not only from the higher courts but also from the lower level courts?

Henric Stjernquist:
I should perhaps answer the question about fast and reliable access. I fully agree that this is a basic requirement for everybody. What I meant is that if you really want to ensure that a system is available 99 percent of the time, this is associated with a considerable cost. We are a public body, and in fact it is the Commission's data centre that is responsible for the distribution of the Celex and EUR-Lex services. They don't perhaps always act like a private company, and it happens that the system for technical reasons is down for several hours. As a value added I think you could provide a better accessibility and reliability, but only to an additional cost. Then it is a question if you do this to all users or just for those who pay. Obviously, if you pay 1.100 Euro you probably expect the service to be available if you want to consult it 9 o'clock in the morning, when if you are a free user maybe you can accept that you have to wait an hour. I don't really have an answer; it's a matter of costs and how you prefer to pay them.

Andrew Mowbray:
I think in terms of availability, one thing we have always said that even if we provi-

de a free service we aim to run it on a professional basis. I am not sure that if we were charging money that we could do it any better than we do it now, in terms of reliability.

Peter Seipel:
These are of course complicated matters. What you are saying is that technical bells and whistles don't count anymore. You can't get paid for adding an extra hyperlink. It would also be interesting to listen to comments on this. The question is also: how bad can a citizens' information system be, without being too bad?

As for so-called added value, we also have the issue of state organs beginning to explain and comment on their own normative materials. This is not only a question of charging or not. It is also a question of whether it is sound practice that lawmakers also comment and hyperlink and explain how texts should be interpreted. Of course such a practice has a strong tradition in Sweden – I am referring to the "preparatory materials", the "förarbeten" (travaux préparatoires), where the head of a ministry comments on the interpretation of a particular statute in a Government bill. In the Internet environment such practices can be an extremely sensitive matter. Briefly, where is the borderline or limit to such explanative activities performed by state and municipal authorities?

Liselott Söderlund, Stockholm University
I thought that if the government has problems in gathering the complete reference of case-law, is it unthinkable to get it paid by ads on the Internet? Just like the Underground is partly paid by commercial ads? Would that imply problems with the credibility of the material?

Bengt Nordqvist:
We have discussed that matter. I think it is a bit far away from the point when we have ads in "Lagrummet".

Andrew Mowbray:
I am not sure it is entirely just a matter of money. To do this kind of stuff is not terribly expensive

Henric Stjernquist:
2 million Euro is a lot of money but it is not very much in the budget of the European Union, so there is no real problem of funding.

Peter Seipel:
I think the comment you now make deserves to be underlined: we are talking small money, absolutely nothing compared with agricultural support or regional development aid or the like. But the informational environment is invisible and nobody suffers or gets hurt, therefore things tend to take time.

Cecilia Magnusson Sjöberg, Stockholm University:
I would like to raise a question with regard to co-ordination and public/private ini-

tiatives and how to avoid reinventing the wheel over and over again… More precisely, my question has to do with the EC Commission's work concerning collection and dissemination of legal information. I am just curious: to what extent if at all and also if you find it worthwhile, you take advantage of earlier experiences from projects within the field of "added value" with regard to modern means of supplying legal information on the Internet? Earlier projects, actually financed by the Commission – I am thinking, for instance, of the so called EULEIS project, co-ordinated by Finland, including parties from Belgium and Sweden, and to some extent also the Elvil project. In these attempts one has addressed questions such as how to set up united document structures? How to approach demands for multilinguism?

Henric Stjernquist:
Of course we try to look at what is happening around us, and indeed if there are projects financed by the European Union we try to look at them in particular. But the problem is that while we can use these projects for our visions about the future, we are very much stuck in the daily production, and there we have of course also questions about multilinguism and the way of treating texts. We take most of the information, apart from what we get directly from the institutions, from the Official Journal. The Official Journal is more or less the basis for what we are doing. A big part of the value-added information also comes from there – the Eurovoc indexation for instance. The format they use is a very complicated SGML that sometimes poses problems even for the printers. It may look like a very simple operation to publish legal information on the web, and it can be if you choose simple solutions. But unfortunately they didn't always do that in the past. So, even if we try to look at simple projects like AustLII or various interesting development projects funded by EU or elsewhere it is unfortunately not the reality we are living in for the time being.

Peter Seipel:
History is a burden, isn't it?

Andrew Mowbray:
The aim is always to create something which appears to be simple, to create the illusion that all these beautiful data come out in a consistent format. I don't know the detail that is involved in doing that particular set of data.

Cecilia Magnusson Sjöberg:
I did not try to convey the opinion that these tasks in anyway would be simple. Not at all. Perhaps there is a need for a global community approach in order to cross-fertilise results from similar experiences. In this context mention could be made of growing networks for XML and law in US (Legal XML) as well as in Europe (LEXML)

Anders Jonasson, Norstedts Juridik:
You mentioned, Andrew Mowbray, that in court cases you had started to make numbers on paragraphs in case-law. Is that an example of how digital publishing can improve general publication of case-law. We have a long tradition in publishing case-law on paper. In Sweden it is page-oriented, and it is difficult to make links to the

exact place. I think you showed an interesting example of how electronic publication can enhance the accessibility in the way of how the document is constructed – the architecture of legal documents, even on paper. There is an interesting synergy in that, between paper and electronic publishing. Are there others? Is it possible to explore that, in the way Cecilia Magnusson Sjöberg talked about? Making standardisations, for example?

Andrew Mowbray:
Paragraph numbering is pretty much universally recognised as a fairly simple solution to a complex problem, which is essentially to achieve so-called media .

Peter Seipel:
Talking about paper and electronic information – why aren't we prepared to leave print information as the primary source of legal data? Couldn't we switch over and decide that the electronic version is the original. We could then concentrate our efforts on standardisation and on making the information manageable. You may remember that this morning Tom Bruce mentioned different needs of different categories of users. Henric Stjernquist also went into this when he commented on the three categories of legal professionals, non-legal professionals and non-professionals. If you take non-typical legal professionals like customs officers or policemen, what Tom Bruce said was that these people need legal information presented in such a way that they recognise themselves and their questions. That is something I believe we have to work on. Since a rapidly increasing amount of legal text material exists in electronic source format, one of the tasks is to make it manageable so that it can be presented for different purposes and different people. How far are we from that situation? Is it only a dream or is it now on the drawing boards?

Andrew Mowbray:
All the big publishers now do precisely what you say. The practice of the moment is that you publish the material, generally in HTML, and you basically can from that generate a web page, a CD-ROM, or a book.

Peter Seipel:
This is a rather crude version of the dream of adjusting the normative materials to the needs of the end users. What publishers now do is to produce different products. They produce a CD version, a text handbook version etc. But I think of even more radical ways of handling the electronic information. Some of you may remember a doctoral thesis that was defended many years ago by Britt-Louise Gunnarsson who is now professor of Nordic languages at the Stockholm University. Its title was The language of the Labour Co-Decision Act. What Gunnarsson did was to analyse the language of the statute and to suggest a reformulation so that employees could easily read and understand and see their problems in the text of the statute. Of course this was a very sensitive operation from the traditional legal point of view, because what she did was actually to draft a new statute, written from the point of view of the employees. She did it all manually, of course, and she used the instruments of manual, traditional text processing. Today we can do things like this more rapidly

and at least partially by automated means. We can do it quite openly and explain that this is the version intended for the police, this is the version intended for the prosecutors, this is the court version, and so on. If you want to see the original version you can go to the electronic source, but perhaps studying the one that has been adapted to your particular need may facilitate the task of understanding.

Is it to go too far and to change our thinking about what law is and how legal norms ought to be handled? I think the situation in which we find ourselves makes it necessary for us to discuss things like this. They are no longer theoretical.

Andrew Mowbray:
I think we currently see in some of legislation examples of this. The trouble is that as soon as you put it into a statute it becomes something that is open to interpretation.

Peter Seipel:
Let me pose a question to you: Is Sweden lagging behind? We have heard a description of the US situation where a few big companies dominate the picture. We have heard about what goes on in Australia. Sweden used to be an advanced country in the field of electronic legal information retrieval in the 70s and set an example in many ways. What about its position today?

Andrew Mowbray:
I know much less about it than probably anyone else in this room. From what I have seen the Government's involvement is laudable. Put the data in a format so that people can use it and somebody is going to come along and roll up a system. If nothing else is going to help to push things along.

Karin Jönsson, Lund University Law Library:
I would like to return to the question on primary sources, authority and authenticity. How far are we from any electronic legal text-version to be regarded as primary source? Is that a question of a technical format to be approved? We have the example of EurLex where we have the PDF version for the latest 45 days; then they are removed and we have to rely on the HTML format from CELEX. So I wonder where the problem is?

Bengt Nordqvist:
First of course it is a technical question. Today we have the technique, maybe not well-tested but we have it. The next thing is to go from a secondary source to a primary source. I think we have to change our way of thinking and that we are more than a couple of years away from that.

Henric Stjernquist:
I think that it is not so much a technical problem. You mentioned the PDF files – from 2002 you will have them not only for 45 days but forever. I mentioned this Official Journal CE, the electronic version, where we have documents that are only published electronically. They are not binding legal acts, only preparatory docu-

ments and parliamentary questions. But the experiences I think are very good. The thing is that these are produced exactly as a printed Official Journal, in HTML, in PDF, we even make TIFF files that are facsimile images, the only difference is that they are not printed. We may have quality problems, not least because we have to handle eleven languages at the same time. Already in the printed version we have probably more printing errors in the EU Official Journal than in the various national gazettes, but technically speaking I can see no problem. It is just that the Treaty says that the text as printed in the Official Journal is the only authentic one.

Bengt Nordqvist:
I think it is the ambition of a lot of people working with it, that the primary source should be the electronic version. It has been discussed a lot. But my experience is that it will take more than a couple of years.

Andrew Mowbray:
A few practical things about it: First, if we are going to say that one version would be the authorised one, it would probably have to be the original one. So if it is stored in HTML, I guess you would have to say was the absolutely correct. At a technical level you then just apply a digital signature and this then probably has much greater confidence that what you are looking at was absolutely right than you can now with paper. It would be a much better solution than with the paper solution we currently have. .

Peter Seipel:
I think Tom Bruce also made the point that we have to distinguish between formal status and practical acceptance. My guess is that we will see a drift into the electronic world. There are many reasons for this, one has to do with costs. The paper version of the Government bills to the Parliament weighs several kilos. Many subscribers probably throw away what they find irrelevant. This is a waste of trees and a waste of users' time. The reasons for still sticking to paper is a question of attitude of people who find it difficult to switch to electronic material, but they also have to do with some of the still existing weaknesses of electronic materials. For example, it is not unheard of that it takes longer time to get the electronic version of a document than to get the printed version. But this will probably change. The final sign will be a formal decision that electronic source materials have been lifted up to the level of authentic original texts. It is hard to tell how long time this will take – a couple of years at least, I believe.

I have one final question that has to do with the users. My question to Andrew Mowbray is: How do you cater for the needs of the users? Are they a part of the organisation of your system? And I would also like to put the same question to Bengt Nordqvist: You mentioned the council, but it consists, I gather, only of a few public authorities. Are there no user representatives in the council?

Andrew Mowbray:
I don't know. AustLII is not a company – we are a joint venture of two universities. We have a management committee, which is the most formal thing we have, and it

consists of our two deans, myself, Graham Greenleaf and AustLII's executive director Philip Chung.

Bengt Nordqvist:
It is correct that there are no users in the council. We are supposed to have meetings for the users at least once a year, but I can say that up to today we haven,t held any big user meeting.

Peter Seipel:
Since there appears to be no further questions or comments, I wish to thank you all for participating. In particular I wish to thank Andrew Mowbray for coming the long way from Australia to share his experiences and thinking with us. It has been inspiring indeed and there is no doubt that we have a lot of things to learn. Let me also thank Tom Bruce both for his valuable comments and for taking the trouble to appear on our computer screens. Thank you all.