Building the Virtual Health Sciences in the Eastern Mediterranean

Progress in Health Digital Libraries

Global and Regional Perspectives

Chris Zielinski, Consultant

18 Monks Orchard, Petersfield, Hants. GU32 2JD, United Kingdom

Tel:+44-1730-301297, Fax: +44-1730-265 398, e-mail: zielinski2002@cs.com

Major libraries are going digital.- the US Library of Congress met its goal of putting 5 million historical items online by the year 2000, and a unified virtual European Library has just been announced. What are the options in the Eastern Mediterranean Region?

This paper aims to summarise key concepts in the broad and growing field of digital libraries in the health sector. To an extent, these concepts are generic to digital libraries, although there are a number of health-specific developments that will be reviewed as well.

After a presentation of basic concepts, we will provide an overview of the numerous open-access initiatives that are expanding possibilities, particularly in information for developing countries. A brief introduction to the key networking activities being proposed is provided. To conclude, some of the key technological issues are reviewed, relating to search technology, metadata and locator technology.

1. What is a Digital Library?

As summarized by the World Intellectual Property Organization,

> Digital libraries are electronic equivalents of paper collections of records… the concept of a Digital Library … [is] an organized collection of electronic information disseminated to a designated community through network technologies providing easy access to data. Provided that a global secure network can be established, digital libraries hold the potential of vastly simplifying the process of providing access to timely and complete collections… Digital libraries, accordingly, present an attractive alternative to the paper-based collections maintained today.

Each of the key concepts in the notion of a digital library are identified here: it is electronic, it is an organized collection, it is disseminated to a designated community, it simplifies access, collections are timely and complete, and they are an alternative to paper-based collections.

This, it must be acknowledged, is an ideal picture. What we have seen to date are digital libraries that are often incomplete, ill-focussed on any community, with unchanging old materials, and at best supplementary to paper-based collections. Nevertheless, there are a number of significant currents in contemporary discourse that offer remedies to these problems, and we will touch on these below.

In general, it is being found that providing training and technologies to libraries is one of the most important ways to bridge the so-called Digital Divide. "…putting computers in libraries is increasing access to information for those adversely affected by the Digital Divide and is having a positive impact on libraries and their patrons…. Libraries, with their long history of providing free access to information, are a natural vehicle for leading this effort to expand public access to technology."

An interesting observation in this regard is that, where libraries have been strengthened in this way, library users who have had little prior technological access to information – primarily students and low-income residents – had the heaviest usage on the new computers. Studies also show a positive shift in staff morale in libraries that receive computers and training, as well as increased library traffic.

2. Basic Concepts

2.1 Information Overload and Loss

There has been a fundamental shift in the world of information in the last century. The information age was built on the rise of mass literacy, mass education, and mass communications. These social changes have been abetted with inventions and applied technology. Writing was invented by the Sumerians around 5,000 years ago, paper around 2,000 years ago, printing 500 years ago – and then, in a mad run of marvels tumbling from the magician's sleeves in the last century or so, the electric light, telephone, radio, television, computer, CD-ROM, Internet. Each of these factors and all of these technologies – even the light bulb – have been about information.

Allied to these social and technological developments has been the rise of business models that have favoured the spread of information, and that have given it a recognized commercial value even when it cannot be measured. Such models have lead to the development of highly sophisticated and complex technologies that are nevertheless designed to be easy to use – a paradox that can only be explained by the commercial valorization of information, and the recognition that there is a mass market for it. By and large, information has become a low-margin/high-volume commodity and everybody wants some. There is a parallel high-margin/low-volume market for information about patented medicines and immediate-impact commercial information, and it might be argued that much of the literature in academic journals is priced beyond the reach of many people who could use it.

This outburst of content is unparalleled in human history. There is so much redundancy in the information age that, according to one article there are over 75 *names* for it – information revolution, age of access, communications age...

Richard Feynmann, estimated that there were some 24 million different books in the world, and that, if atoms were used for digits in digital storage, all of them could fit into a grain of sand. Soon all those 24 million books will be available in electronic format – perhaps not in a Feynmann grain of sand, but certainly on the Internet. Another estimate suggests that 10 million documents and 75,000 books are published annually (20 times as much as in 1910).

Journal literature is in a similar state. For example, Japanese medical researchers publish 109,700 scientific papers a year – about one "scientific paper every minute of the working day". Other nations show similar industry in producing journal articles. Overall, there are 100,000 journals publishing 7,000 journal articles every day.

With all this productivity, it is sobering to record the familiar statistics that 10% of the journals in any given library account for 80% of usage, and at least 50% of the journals are never even consulted once.

The problems of information overload are familiar, and it is clear that we have faithfully transplanted the problem from the analogue world to the digital one, increasing our problems a thousandfold. Right now on the Internet, you can use any one of a hundred search engines to find information contained in over 2.1 billion unique, publicly accessible pages. This cache of information is growing at a rate of over 7 million pages a day.

In addition to these readily accessible pages, information is contained deep within web-based databases. Most search engines will not search within databases, or as Elinor Abreu calls it, "the Deep Web." A recent survey suggests that there may be 550 billion documents in the Deep Web.

Not only do we produce too much information, and not only is most of what we have put onto the Web currently unfindable, but even what is accessible is hopelessly approximate – to the point that we are not even able to cope with indexed results.

Boolean algebra is, formally, a mathematically complete system – but Boolean searches on keywords are not efficient in rapidly winnowing down the mass of plausible information into the specific information that you want. It is not successful information provision when you launch a search for a particular piece of information to have 1,000 articles identified and presented with only the smallest concern for relevance. And this situation will get worse, the more information that is available for electronic searching. What will happen when every search using Boolean operators and standard search techniques results in 10,000, or 100,000 "hits"?

As a result of these three factors – the geometric explosion of information production, the digitisation of everything, and the absence of robust and sensitive searching tools – the potential rewards of an information economy are being stymied. With so much information being produced, so much is being lost. In the "attention economy", the market belongs to those who shout the loudest, and grey literature never reaches grey matter. Research lies entombed in local journals and so the wheel of basic research is reinvented with depressing frequency. Information equity – where all information is evaluated according to the same standards and with the same chances in the marketplace of ideas – is a dream.

To counter these negative tendencies, it is a relief to be able to report that technology is riding to the rescue, with the rise of metadata systems that are consistent descriptors of digital content of all kinds, or searching technology based on concepts, and not just on keywords, and ion locator technology that enables the location of content in a persistent, reliable way. These will be describe further in the following pages.

2.2 Information Specificity and Information Equity

The key vision in the concept of digital libraries is of a world where everyone will be able to find the specific information they want quickly and easily, and every piece of information has an equitable chance of being identified and accessed – without overloading the human capacity for assimilation.

This vision of information specificity and information equity would represent a genuine quantum leap in knowledge management capability, and would be one of the keys to what Tim Berners-Lee calls "the Semantic Web" (see below).

- ***Information Specificity:*** The concept of information specificity is that everyone will be able to specify and find exactly the information they want from every available source, whether formal or non-formal, to a new level of accuracy and relevance. The perspective is from the person searching for information. The stress is on automating the individual personalisation of the searching and finding process: it has to feel natural, respect privacy and be easy to learn and administer.

- ***Information Equity:*** Looking at the information world from the perspective of the information producer, the vision is that every piece of information that is available online or offline can be found by those looking for it, if an electronic version of the full text, or even a synopsis or abstract exists. Moreover, no information should be lost despite the overwhelming "noise" competing for the attention of individuals in the electronic age.

These are crucial concepts in information for development and yet, when one considers the statistics for languages of publication and origin of scientific content, the picture leaves much to be desired:

Languages of publication of internationally indexed articles

60% in English

10% in German

8% in Russian

7% in French

15% for the rest of the world's languages

Articles published in scientific journals of international coverage

82% are from OECD countries

7% from Eastern Europe

5% from Asia

4% from Latin America, and

2% from continental Africa

Even if we should presume that the preponderance of the world's valid, important biomedical information originates in the West – and there is evidence to suggest that this is not the case – the 2 to 5% participation in international scientific discourse allowed by Western indexing services is simply too little to account for the scientific output of 80% of the world. Digital libraries with input from all countries can act against this trend.

In general, information from all sources should be accorded equal access, equal economic value and equal rights. This simple credo does not insist that the balance of information flow should be 50-50, but rather asserts the principle of equity. We should also recognise the mutual interdependence of our information needs.

2.3 Information and Developing Countries

As summarized by the present author in an editorial in the *British Medical Journal* in 1975,

> The "Third World" of developing countries is far from homogeneous. Nevertheless, as consumers of information, the countries have a stark regularity of features that allows for convenient grouping: most of their medical libraries subscribe to fewer than 50 journals, less than one library in ten has a computer or CD-ROM player, and budgets for new books, software or online charges are tiny or non-existent. Telephone and telecommunications systems are sparse, unreliable and expensive, so network use is rare. Where network access does exist, it is mainly used for simple communications, often in non-interactive batch modes, rather than to scan health literature.

Although things have improved somewhat in the last six years, this remains the picture for many health libraries today.

A number of international and non-governmental organisations have been helping developing countries acquire health literature and contemporary technology. However, although such projects are generally launched for laudable reasons, they almost never include the provision of local journals and other formal information. Whatever the donors’ intentions, Western information assistance usually serves as a vehicle for opening up developing-country markets to Western information product providers. The implicit assumption is that the information superhighway, whenever it comes into sight in the developing countries, is a one-way street from the First World to the Third.

One reason for the lack of developing country literature in global databases is the simple fact that it is very difficult to find. International services such as MEDLINE or the Science Citation Index typically index some 3,000 journals—95% from the First World and only 5% from the Third World. This is a starting point for the vicious cycle affecting Third World literature: journals that are not indexed are rarely stocked by librarians, hence rarely cited by authors, and hence rarely indexed...

Information from the South can be divided into a formal sector (local journals and books) and an informal sector (indigenous knowledge). Both are important, and yet the formal sector is paradoxically currently less favoured than the informal one – note that the World Bank has jumped into indigenous knowledge not so long ago, deploying a dazzling array of experts and money, but the local formal sector remains outside the pale.

One project in this area is *ExtraMED*, which collects the full text of over 300 Third World biomedical journals, at present on a monthly CD-ROM (26 have been issued to date, with some 30,000 articles or 250,000 pages). There are a number of other full-text journals projects (notably African Journals Online, Bioline and WHO/BIREME’s SciELO) that could be viable sources of information from the formal sector.

In general, what is important in information projects in developing countries, is that locally relevant information needs to be distributed. "Of key importance for the poor in developing societies is not merely the provision of connectivity or access to knowledge, but whether relevant knowledge is disseminated".

Who should be doing the dissemination? All our projects are seeking to strengthen local skills in this area. The concept of distributing knowledge through "key informants", "infomediaries" or "local ‘doers of development’" has strong support in the literature. As the World Bank's former Chief Economist Joseph Stieglitz put it, "It is by the local selection, assimilation and adaptation of knowledge that local doers "make it their own"[13]. This is where the idea of "glocalization" – the globalization of local knowledge – is important, and where local libraries can have a global impact in the framework of digital libraries.

2.4 Essential information

"Essential information", is a concept borrowed from "essential drugs":

Essential drugs*

"Essential drugs are those that satisfy the health care needs of the majority of the population; they should therefore be available at all times in adequate amounts and in the appropriate dosage forms, and at a price that individuals and the community can afford."

Essential information

Essential health information is that which satisfies the health information needs of the majority of the population; it should therefore be available at all times in adequate amounts and in the appropriate formats, and at a price that individuals and the community can afford.

*From EB108/INF.DOC./2, discussed by WHO's Executive Board in May 2001.

It should be noted that, as with essential drugs, essential information is seen as a human right. It is information that is essential to human survival. Essential information comprises the things we need to know to survive, to be healthy, to plant the right seeds, to feed our families correctly. It includes information related to the basic minimum needs of humanity, information tools for trade and economic development, information essential to the development of backbone industries, basic science and survival services in health, education, welfare, agriculture and labour.

2.5 Internet Indicators

As digital libraries depend on the Internet, it is worth summarising the key Internet indicators very briefly. Although the following data is very US-centric, the table should be read as an indication of trends which are likely to be mirrored globally.

The Internet, 1999-2000: Negative and Positive Trends

Item	**1999**	**2000**	**Percent Change**
Total amount raised by Net IPOs[1]	1Q99: $50.3 billion	3Q00: $5.9 billion	-88%
Value of Net mergers and acquisitions[2]	1Q00: $51.7 billion[*]	3Q00: $9.3 billion	-87%
Median prevaluation of e-commerce startups[3]	4Q99: $80 million	2Q00: $20 million	-75%
The Standard 100[1]	Dec 31, 2000: 2595	Dec 6, 2000: 695	-73%
Average first-day gain of Internet IPOs[1]	1Q99: 113 percent	3Q:00: 51 percent	-55%
Dot-coms marketing-expenditure/revenue ratio[4]	3Q99: 79 percent	3Q00: 47 percent	-33%
Nasdaq[1]	Dec 31, 1999: 4069	Dec 6, 2000: 2876	-31%
Net venture capital funding[5]	1Q00: $16 billion	3Q00: $13 billion	-19%
Net venture capital rounds raised[5]	4Q99: 798	3Q00: 664	-17%

Internet-related IPOs[1]	1Q99-3Q99: 168	1Q00-3Q00: 161	-4%
U.S. interactive TVs[6]	1999: 1,143,400	2000: 1,118,100	-2%
Item	**1999**	**2000**	**Percent Change**
U.S. retail e-commerce[15]	$15 billion	$45 billion	+200%
Worldwide wireless Internet users[14]	6.3 million	18.1 million	+187%
U.S. business-to-business e-commerce[13]	$97 billion	$213 billion	+119%
U.S. DSL or cable-modem connections[6]	2.5 million	4.7 million	+88%
Number of web pages[12]	1.5 billion	2.7 billion	+80%
U.S. online advertising spending[11]	$4.6 billion	$8 billion	+74%
E-mail messages sent daily worldwide[10]	5.9 billion	9.7 billion	+64%
Worldwide online population[9]	245 million	338 million	+38%
U.S. average monthly time spent online at home[8]	8.2 hours	10.0 hours	+23%
U.S. ISP subscribers[6]	47 million	57 million	+21%
U.S. online population[7]	104 million	122 million	+17%

Source: Table from The Industry Standard's Metrics Report, 21/12/2000

[1] Bloomberg; [2] Omits AOL-Time Warner merger ($157 billion), Webmergers.com, October 2000; [3] VentureOne, August 2000; [4] The Standard based on an analysis of sales and marketing expenses expressed as a percent of revenues at public 20 Internet companies; [5] VentureOne, November 2000; [6] Telecommunications Reports International, November 2000; [7] Jupiter Research, July 2000; [8] Nielsen NetRatings, November 2000; [9] IDC, November 2000; [10] IDC, September 2000; [11] Internet Advertising Bureau, October 2000; [12] NEC Research, October 2000; [13] IDC, June 2000; [14] Ovum, September 2000; [15] Jupiter Research, October 2000.

In considering this table, it should be noted that the negative indicators arising in the transit from 1999 to 2000 are almost all to do with IPOs crashing and the consequent depression of the overall Internet stock picture. Net businesses were forced to lay off a total of 28,245 employees, and 61 dot-coms so far have closed their doors this year (IDC, November 2000). Investors are learning that dot-coms without a real product are ultimately without real value. To a large extent, the effect has been to get rid of the useless and to focus the economy on companies that have genuine products of real quality.

On the positive side, e-commerce, technological growth and the growth of information on the Internet continue to exhibit strong growth. This year the online population grew to encompass nearly half of Americans, and the Net became a more integral part of their lives. People at home spent an average of 10 hours online per month, an increase of 23 percent over last year. Use of the Web globally extended to 338 million people, growing by more than a third this year. The strong international appeal of wireless tripled the number of people using mobile Net access from 6 million to 18 million. U.S. broadband usage also started to come of age, registering an 88 percent increase to almost 5 million subscribers. The content accessed online swelled to 2.7 billion Web pages, while the number of e-mails sent per day worldwide increased by 64 percent.

2.6 The Semantic Web

> "The Semantic Web is about that move to information that is presented in a way that can be repurposed using a machine. That’s a whole new ball game, a whole new set of possibilities, a whole new explosion and a whole new impact on e-commerce"
>
> Tim Berners-Lee, *Business 2.0*, Christmas 2000

Tim Berners-Lee (the inventor of the World Wide Web, html and the http protocols) coined the term "The Semantic Web" in 1998 to denote the next evolutionary step of the Web. This has been described as "Associating meaning with content or establishing a layer of machine understandable data to allow for automated agents, sophisticated search engines and interoperable services… and enable higher degrees of automation and more intelligent applications." The ultimate goal of the Semantic Web is to allow machines the sharing and exploitation of knowledge in the Web way, i.e. without central authority, with few basic rules, in a scalable, adaptable, extensible manner. As such, it covers developments ranging from toasters that "talk" to refrigerators (and the online supermarket) to the automated customisation of the information seeking and distribution process.

This model suggests standard, interoperable systems allowing for information *mediation*, rather than information *aggregation*, which is a key concept in the digital library.

In other words, in the Semantic Web, the information remains decentralised with its producers, publishers and distributors, whether in partial aggregations or completely disaggregated to the producing individual. The move is away from global portals, vortals and gateways which seek to build massive resource bases and towards gateways that use standards-based interoperable systems to link and provide "information mediation" – ways of identifying and reaching information.

A multitude of tools, methods and systems have just appeared on the horizon which have the effect of empowering the Semantic Web in the area of information mediation. These will require the standardisation of input, notably in regard to the metadata that describes the content – such as the title, originator, rights holder and the like. However, a well-grounded standard exists for this in the <in*d*ecs> scheme, which has been designed and demonstrated to be fully compatible with such metadata schemes and standards as the Dublin Core and MARC. A locator methodology also exists in the Digital Object Identifier (DOI), which is also a full ISO standard (Z39.84). This is described more fully below.

So far, what has been missing up to now is really powerful software to automate the identification and description of information, to match it to user requirements, and to link it to standardised metadata and locator technology. The emergence of powerful contextual searching technologies provide this missing link. Users will be able to review related literature and content in all media, negotiate rights, comparison-shop, identify quality by brands or rankings, assess market data and statistical information based on actual use, build virtual communities based on common information interests and needs, and many other new capabilities which we can only dream of at present.

3. Open Access

The world of Open Access – where business models are turned upside down to allow end-users free access – has been developing rapidly and in complex ways. These are clearly beneficial for the idea of the digital library, particularly in respect of the developing world.. In this section, we will outline the main developments leading to this situation.

With the rise of the Internet as a medium of academic interchange, the original perception that everything was available somewhere for free turned from a work ethic into something approaching an ethical principle. "Information wants to be free", argued some, while others insisted that they had spoken to information and had been assured that it wanted to be copyright protected, encrypted, and sold to the highest bidder.

The open access concept has evolved from a number of sources. Among the precursors are certainly pre-print archives such as the Los Alamos physics e-Print Archive. Started in 1991 based on a server located in his office, Paul Ginsparg was processing some 24,000 submissions by 1998, according to Andrew Odlyzko, "about half of the volume of all mathematics papers published that year…but small compared to the perhaps 2 million papers in all STM (science, technology and medicine) areas". Odlyzko goes on to note that the success of the archive demonstrates that "scholars can embrace new technology in a short period and derive enough benefit that giving it up becomes unthinkable," but acknowledges that the substantial critical mass of papers required may not be present in most of the STM fields.

Nevertheless, experimentation with pre-print archives continues, to a mixed response from leading STM publishers, who are understandably not enamoured of such concepts as Stevan Harnad's "Subversive Proposal" – subversive, since Harnad originally suggested that researchers place their preprints in an archive, and then later replace them with the peer-reviewed, edited and polished versions produced by their eventual publishers. He has since retreated from this "subversive" concept, while still energetically supporting the value of authors "self-archiving" their work. If such local archives follow the emerging conventions of the Open Archives Initiative (OAI) (http://www.openarchives.org), then the resulting interoperability would enable users to easily locate a specific paper in whatever archive it is stored.

According to Walter Warnick, there are now some 7,000 scientific and technical preprint sites "although most of these lack formal data structures, such a metadata, to aid in the identification, description and retrieval of preprints. The PrePRINT Network search engine can be used to cross-search 4,000 preprint sites…Altogether, around 375,000 scientific and technical preprints can be searched via the PrePRINT Network and users can freely access and search the full text of this type of primary literature." As yet, however, the PrePRINT Network does not use OAI architecture.

Another key initiative on the road to Open Access was E-BioMed (now renamed PubMed Central), proposed in 1999 by Harold Varmus, then Director of the US National Institutes of Health. This was proposed to be a central free repository for medical science papers. Controversy has stalled this project, but it has not stopped it from encouraging the birth of successors – BioMed Central (www.biomedcentral.com), which offers to publish and distribute original research reports using a full peer-reviewed system and a non peer-reviewed depository. All of this research is placed in full and without delay on PubMed Central, and is thus made available free to all individuals though the web. Authors retain their copyright, and no charges or barriers to access are imposed.

A second "offspring" of PubMed Central is E-Biosci which follows the same basic model, and which was established by the *EMBO* to be a "European-based, digital information resource network with a global role"

Two other recent events should be mentioned. The first is the initiative of the Public Library of Science (http://Publiclibraryofscience.com), which issued a petition urging a boycott of scientific and scholarly journals that refuse to make their articles accessible online for free six months after a journal issue has appeared in print. The petition, circulated electronically, gathered signatures aimed at a target date after which the boycott would begin. Although some 24,000 authors from over 140 countries had signed by September 2001, including several Nobel Laureates, this is still a relatively small percentage of the total. Although it may not have much impact in practice, it is in any case a clear indication of the public mood.

The second is the victory of freelance journalists, after appeals that reached the US

Supreme Court, in securing their right to authorise and be paid for electronic uses of their works. Although the authors expected this decision to lead to enhanced royalties, it has had the effect of punching holes in databases, as publishers scramble to remove anything that may belong to authors who do not grant a flat permission to use all their works in electronic databases. "Rather than pay up or face billions in liabilities, publishers are deleting tens of thousands of freelance articles spanning decades. So who will bear the brunt of the extra work? The librarians, of course…"" Although this is an issue that impacts the news world, STM publishers are watching this development anxiously as well.

Finally, we should mention the Scholarly Publishing & Academic Resources Coalition (SPARC), which is an international alliance of over 200 college and research libraries aiming to build "a more competitive scholarly communication market place to address the high cost of information." Essentially, SPARC is a case of author power – where writers and editors establish their own journals in direct competition with publishers, seeking to undercut them in terms of price while maintaining quality.

Up to this point, the open access story has been told from the perspective largely of the writers. What have the publishers done to stem the tide?

One solution – the simplest – has been to make journal articles freely available through the publisher’s own web site. There are also a number of multiple-journal sites, such as the non-profit HighWire Press, which archives over 230 journals – more than 200,000 articles are freely available on the site.

An early initiative was the International Digital Electronic Access Library (IDEAL), which now includes over 300 journals published by Academic Press and Harcourt Health Sciences, totalling over 200,000 journal articles. The IDEAL Charter for Developing Countries was announced in February 2001, setting out the terms for reduced-rate access to low-income countries (following the World Bank definition as countries with a GDP per capita of $760 or less).

A second recent development has been the announcement that six of the world’s leading medical publishers (Blackwell Science, Elsevier Science, Harcourt International, John Wiley, Springer Verlag, and Wolters Kluwer) had joined forces with WHO in a venture to enable more than 100 of the poorest countries in the world to access scientific information free of charge through the Internet. The arrangement will allow almost 1,000 of the world's leading medical and scientific journals to become available through the Internet to medical schools and research institutions in developing countries for free or at deeply-reduced rates.

According to Dr Gro Harlem Brundtland, director general of the World Health Organization (WHO), "As a direct result of this arrangement, many thousands of doctors, researchers, and health policymakers, among others, will be able to use the best available scientific evidence to an unprecedented degree to help them improve the health of their populations. It is perhaps the biggest step ever taken towards reducing the health information gap between rich and poor countries."

With all this intense activity in last year, the stage is truly set to build digital libraries that, while being virtual, can also be virtually full of useful content, available in developing countries for virtually nothing.

4. Networking Activities

Digital libraries need networks to receive information and also to transmit information to their target user groups. Although there are many national, sub-regional and regional networks dealing with health information, and a large number of individual projects, two recent ideas are worth mentioning briefly which have tried to develop a global vision. These are the Information Waystations and Staging Posts (IWSP) project and the Health InterNetwork (HIN).

4.1 Information Waystations and Staging Posts

The first stage of the Information Waystations and Staging Posts activity was the **Health Information for Development** (HID) project, funded by the Bill and Melinda Gates Children's Vaccine Program at PATH, which compiled a ***Global Directory of Health Information Resource Centres*** (available at http://www.iwsp.org) between January 2000 and May 2001. HID was the fruit of 18 months of intensive work among non-governmental and international bodies, and in creating its Directory has carried out the first needs assessment regarding technology needs in this area. It was steered by 35 leading specialists on a Project Advisory Board and a Policy Advisory Council comprising AMREF, Gates/ PATH, the British Council, British Telecom, OSI (Soros), the Third World Academy of Science, UNDP, UNFPA, UNECA, and WHO.

Characterising Health Information Resource Centres

Average number of staff in centre: 8

Most are either self-supporting or Government/Ministry supported

65% of the respondents said that the collection and provision of information was their primary activity

84% of respondents had at least one computer. Most of these respondents (87%) wanted more computers and training.

Users of information

Specific target groups: 40%

Center’s own staff: 74%

General public as a whole: 72%

Primary/community health workers; 71%

Hospital staff, doctors, nurses: 68%

Researchers/teachers/students: 81%

Ministry personnel and institutions: 50%

Other centers/NGOs in the country: 81%

Other centers/NGOs abroad: 34%

Sources of information

Write/draw/produce own information: 78%

Use information from other centers:

—within the country: 79%

—from outside the country: 76%

Use local non-formal knowledge: 59%

Use local formal knowledge: 79%

Get information via:

—e-mail/Internet: 59%

—CD-ROM/diskette: 44%

Use of information received

Rarely change information: 69%

Adapt and modify information: 54%

Translate materials from foreign languages into local language: 40%

Provide information to users on demand by non-electronic means: 74%

Provide information electronically: 47%

Users come to the center to consult information: 79%

The **Information Waystations and Staging Posts Network** was created in July 2001 by linking the 550 centres in the *Directory* which have e-mail addresses. The IWSP Network aims to: 1) promote the IWSP methodology, 2) express the health information and ICT needs of its members, 3) act as a conduit for support to members from donors and development organizations, 4) act as a medium for research into health information issues, 5) act as a memory bank and evidence warehouse for best practice in health information, and 6) maintain the Directory and otherwise provide information mediation services.

Both of these preparatory stages serve as an initial needs assessment phase of the full **Information Waystations and Staging Posts** project, which aims to build capacity in technology and content management in the IWSP Network, and to provide locally appropriate content on health issues, particularly at the community level.

In the capacity building phases, thus, we are seeking funding to establish a range of regional seed projects. Selected resource centres will be upgraded into ***Information Waystations***, which ***are local points of access to health information received electronically****.* Each Information Waystation will have appropriate technology (PC, CD-ROM & databases, printer, modem, reliable satellite or land telephone); prepaid broadband Internet access; links to the network of other Information Waystations(sharing information with other centres in a two-way flow); and personnel who are trained in technical maintenance and database use.

Some Information Waystations will be selected for upgrading into **Staging Posts**, which ***act as "relay stations", translating and adapting information materials in order to make them locally appropriate.*** They will distribute information rapidly and widely, linked to health and education initiatives, and making use of appropriate local and external sources of information, particularly prototype publications provided electronically. They will share local information, both formal and non-formal/indigenous, in a two-way flow. Training will be provided in information handling and adaptation techniques.

Working Definition of a Health Information Resource Centre

"Health" is considered in its widest sense, covering population, family planning, nutrition, gender and water/sanitation/hygiene education, including centres dealing with all health and health-related topics (e.g., drug and/or poisons information, gender issues, poverty, environment, indigenous knowledge and local research). Centres may have support from government/Ministry of Health, private foundations, international organizations, NGOs, multi- and bi-lateral organizations, or be self-supporting. They may be at community or primary level, or at district and regional level, including libraries in research institutions, universities, teaching hospitals or governmental institutions. They may use traditional information resources and technologies (speech, performance), printed and graphic media, or computer and other information and communication technologies.

4.2 Health InterNetwork

The Health InterNetwork (HIN) Project is an initiative of the UN Secretary General to help bridge the digital divide as part of the UN Millennium Action Plan. It is envisaged that the Health InterNetwork will improve global public health by facilitating the flow of health information worldwide using Internet technologies and enable more effective health service delivery through access to high quality, relevant and timely information and better communication within the public health community.

The cornerstone of the project is the creation of an electronic/ Internet-based Health InterNetwork portal that will provide access to this information and to communication networks of policy makers, researchers and health service providers. Other key HIN project deliverables include:

- Content: Reliable and relevant local and international public health content (including electronic access to key publications and databases)
- Connectivity: 10,000 to 14,000 new public health information access points established across developing countries
- Training: skill development for information access and management.

The Health InterNetwork aims to bridge the "digital divide" by enabling health care workers, researchers and policy makers in developing countries to gain access to state-of-the-art health information using modern technology, including the Internet. Some 10,000 – 13,000 new health information access sites are expected to be made available in developing countries by the end of 2003. They will be located in urban and rural clinics, public health care facilities, hospitals and medical schools, where greater knowledge can help prevent disease and improve overall health.

It is envisaged that the HIN project will be implemented over a period of seven years including the pilot and planning phase, an operations and evaluation phase, and a transition phase to a broader scale resulting ultimately in a dynamic public health information network.

5. Technology

In providing a very brief overview of relevant technology, we will review keyword-based and contextual searching technologies, metadata and locator technology. All of these are of great relevance to digital libraries, and there is much developmental work underway in all of these areas.

5.1 Searching technology

These can be divided into keyword-based technologies and concept-based technologies.

Keyword-based Technologies

These technologies – exemplified by online search engines such as Yahoo, Lycos, Altavista and Google – involve the user typing a single word or several words into a search field in order to try and locate relevant information. As described earlier, the results are typically a mass of "hits" – references containing the word(s) in the search field. These can be presented either wholly unsorted, or with some effort to place those references that contain more instances of the word(s) at the top of the hit list.

Boolean searches rely on the searching skills of the individual. This means that some individuals can find what they are looking for with greater ease than others. Such skills are not necessarily anything to do with the training and intelligence of the searcher – for example, a student will not necessarily know that entering "falciparum" will lead him to articles on malaria. In general, blind luck plays a major role in most efforts at information resource discovery – particularly if one does not have a previous reference to start from.

Where there is a previous reference, one can use such techniques as following citation trails – branching out into resources identified by the references contained in a given paper. While effective within the confines of what it is capable of achieving, citation searches are limited by the narrow range of journals that are included – typically less than 3,000 journals are considered as exhausting all of the "impact" in science globally (Science Citation Index) – and necessarily leave out the most recent material that has yet to be cited. This is not the place to give a full critique of citation analysis, but it is sufficient to state that the basic principle is one of selection rather than inclusion. While it is argued that the journals selected are the "best journals" having the "highest impact factors", the fact is that researchers will read what is cited, cite what they have read, and librarians will in turn stock what is cited. Many people are convinced that valuable content is left out of these vicious circles, but to date there has been no systematic way to review such information. Certainly citation-based systems are not designed to achieve this.

Analogous to following citation reference trails is the link-vetting procedure applied by Google.com. Using a technique they call PageRank, Google uses the link structure of the Web as an indicator of an individual page’s value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. Google looks at more than the sheer volume of votes, or links a page receives; it also analyses the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important". According to Google, this is a way to find web pages that are both "important and relevant" to a search. Recognising the weakness of key word searching, "Google goes far beyond the number of times a term appears on a page and examines all aspects of the page’s content (and the content of the pages linking to it)" to determine if it’s a good match for a query. Recently Google has begun indexing content in .pdf files, "allowing searchers a significant peek into the "invisible Web," the large area of online content not covered by most search engines." This makes Google probably the most successful of the keyword-based search engines, although the arguments against both citation analysis and keyword-based searching still apply to Google.

It is worth mentioning another keyword-based attempt at tackling the deep web – that of Northern Light. Apart from providing a powerful variant of keyword-based searching of websites, Northern Light has indexed a large collection of documents which it can provide as a supplement (at a supplementary cost) to a series of web links. Again, using keywords means that the indexing of the non-web content is only as good, and the results as relevant, as keyword searching allows.

Attempts at adding relevance to Boolean searches rely on such techniques as counting the frequency with which keywords exist in documents – the more times a keyword is found in a document, the more relevant it is. This is clearly a very weak method at best, and completely fallacious at worst. Keywords can be repeated many times, even though the document may be about something completely different, or can occur rarely in documents that break new ground in the field.

Some search engines offer a "more like this" capability – but this is invariably based on further keyword counts, with all of the deficiencies mentioned above.

Essentially, keyword searches are about words, but not about the relationships between words. They are not grounded in meaning – and thus they have nothing do with semantics.

One approach to intelligent searching that has been attempted is what is called lexical analysis – picking apart the morphology of sentences, parsing them for grammar and building up rule-based meaning tables. Despite decades of effort, this has proved to be an elusive goal, owing to the intrinsic ambiguity of language, and the use of slang and poor grammar. Lexical analysis cannot deal with abstract thought (e.g., "it is interesting to note that" – what is the grammatical role of the word "it"?) and, since it cannot determine which noun is the most important one in a given sentence, cannot determine values (e.g., "The Rolling Stones stood beneath Ayers Rock and played to the audience" – this is about music, not mineralogy, but lexical analysis would not be able to determine this).

And once a lexical system makes a mistake, all that follows tends to go wrong as well, since such systems are based on yes/no decision trees. If you make a left turn when you should have taken a right, you will never get to your destination, unless you are very persistent or very lucky.

Finally, an attempt at bringing semantics into searching relies on using human intervention – getting people to rate or describe content items, so that when someone searches for "more like this", the search is based on what others have said or done. This is clearly as fallible as the people involved, and it relies on very simplistic notions of human behaviour (because I like blues, does it mean that I don't like classical music? Is a taste for Robben Ford in any way indicative of a taste for Prokofiev? Is the fact that I have purchased books on crystallography and Mongolia of any value to other keen crystallographers?). Most significantly, such systems are static and fail completely when new information is added, as none of the previous human interventions can have taken note of an item that didn't exist at the time.

Boolean and keyword-based systems of all kinds therefore represent the foothills of the semantic range and, even though they will continue to be used for certain purposes, they will soon be seen as outdated compared with concept-based technologies.

Apelon (www.apelon.com), the fusion of Lexical Technology and Ontyx, could also be mentioned. This is a healthcare-industry product that uses the UMLS thesaurus for biomedical searching. It seems to be based on the simple expansion of keywords for indexing, rather than powerful vector or Bayesian processing methodology. Although Apelon does not give details of its indexing software in its published information, its constituent companies have has focused on the UMLS thesaurus to a high degree, indeed obsessively, and it can be assumed that the technology is very closely welded to the structure of the thesaurus. They do not single out for mention the indexing technology, and again this suggests that they are using straightforward keywords, matching them with the UMLS and expanding the keywords into concepts. While this is better than the simple application of keywords, it does not do much for the analysis of relevance. Users are likely to have a broader range of hits, but without any of the sophisticated relevance determination provided by vector or probabilistic software.

Concept-based Technologies

At the time of writing, there are only two categories of genuine concept-based search technologies – the vector- and thesaurus-based software of Collexis (www.collexis.com) and the Bayesian-statistical methodology employed by Autonomy (www.autonomy.com), WebTop (www.webtop.com), Semio (www.semio.com) and Microsoft (www.microsoft.com in its Tahoe, or now SharePoint Portal Server 2001). While the Bayesian software shares some commonality of origins in the research institutes of Cambridge University, Collexis software was created independently. Both types of technologies inevitably have points in common, by virtue of both being concept-based.

According to company information, Autonomy's software is "informed by" Bayesian analysis, Claude Shannon's Information Theory and by work carried out on neural networks. Autonomy has a Dynamic Reasoning Engine (DRE). The DRE "exploits high-performance neural network technologies" and "advanced pattern-matching technologies" to extract concepts from text and create agents (similar to the Collexis® "conceptual fingerprints"), which are used for matching with other agents and thus identify related texts.

Collexis software (deployed in such public sector health research sites as www.shared.de) originated in a public sector activity and continues to remit a portion of its revenues to tropical diseases research. The mediation architecture consists of two engines (Collexis core components): an abstraction engine and a matching engine. The Collexis Abstraction Engine creates a conceptual fingerprint of a piece of text, which describes the text in a weighted listing of concepts. These concepts are related by means of the most powerful and comprehensive thesaurus for the given subject field in existence. The result is a tiny (400 bytes) fingerprint file which stores the concepts relating to the text in a detailed and semantically rich way, providing a true fingerprint of the text. All operations relating to searching and comparing conceptual fingerprints is carried out by the Collexis Matching Engine. This allows users to set the level of precision ("how close do you want the related conceptual fingerprints to be?"), adjust the concept terms, and build preferred search strategies.

As mentioned, the quality of results of contextual search technology is the "next generation" after keywords and Boolean searching.

5.2 Personalization

A key development activity that has been sweeping the electronic library landscape in recent times is the concept of user-centered, customisable interfaces to collections of library resources. These "MyLibrary" interfaces are portal-like systems with a declared primary purpose to reduce information overload. In a recent issue of the American Library Association's ITAL journal devoted to the topic, the Guest Editor, Eric Lease Morgan, Director, NCSU Libraries, Raleigh, North Carolina, explained:

> "It does this by first profiling new users of the system, prescribing selected resources for them, and then allowing them to add or delete items from the prescriptions. The entire system, both the user interface as well as the administrative interface for maintaining the resources, is driven by sets of CGI scripts running against a relational database. To use any part of the system, the user or the librarian needs only a Web browser."

Customisable interfaces to library resources are essentially database applications with Web front-ends. These interfaces not only provide the opportunity to improve the patron's library experience, but these interfaces also provide librarians with tools to practice librarianship better, namely public service and collection analysis.

The key to a successful MyLibrary portal is software that is able to search vast amounts of information, and match them to particular users needs in as customized a fashion as possible. The systems must be simple to use, and be as quick as possible in their identification of relevant data.

Consider the typical library need: libraries have their own collections and their own subscriptions to electronic information sources and resources. Equally, there is inevitably a wealth of information outside their library sphere. The ideal system should distinguish between these two situations. How to distinguish locally-held resources from those available outside was an important consideration in research on the DOI (see below). The DOI has to a large extent solved this issue by enabling multiple resolution, with the handles system resolving to the URLs of in-house content, as well as to the URLs of external sources.

5.3 Content Location Strategies

By itself, search technology, whether it is keyword- or concept-based, can only ensure that relevant materials are being identified. The next step is to take users to the content – whether to the actual full text of the content on the Internet, or to an abstract describing it and a further reference to where the full text may be purchased or otherwise downloaded, or to an order form that might lead to delivery by "traditional" postal methods, or to a database or directory listing.

For this, we need "metadata" – data about data, which describe the content – examples of metadata related to a journal article would be the article title, authors' name(s), affiliation, journal title and so on. The conceptual fingerprints need to be related to this metadata as a way of saying what the fingerprint identifies. This may seem obvious, but when you are planning to deal with millions, billions and even trillions of items, the effort to add this metadata must be scrupulous, done once and done correctly.

With such metadata added to the conceptual fingerprint, we need to be able to get to the source of the content. Where content is available locally in electronic form, whether as full text or as an abstract (or perhaps only as an online catalogue, OPAC or ETOC reference, it may be sufficient simply to link to the local instance of the work. Clearly this will to some extent depend on any licensing permissions and rights associated with the local instance. More broadly, the "source" may be in a member of a consortium – inter-library loan arrangements are examples of this. Finally, there is the case of linking to the rightsholder, whether it is a publisher or an individual. The best way to deal with the latter is by using the digital object identifier (DOI).

With these steps, all the possibilities of access and e-commerce are unlocked for users, libraries and publishers, authors and researchers, and a whole slew of third-party beneficiaries.

Below are brief accounts of the <in*d*ecs> scheme that has analysed the needs of the content industries, reviewed all other metadata projects and emerged with a reasoned and well-grounded set of principles for metadata, and of the Digital Object Identifier.

Metadata

> "We are on the verge of a metadata revolution. Get your data models

clean and prepare for an interesting ride".

Tim Berners-Lee, May 1999

Electronic trading depends to a far greater extent than traditional commerce on the way in which things are identified and the terms in which they are described. Identifiers themselves are simply names – names that follow a strict convention and are unique if properly applied, but names just the same. Such unique identifiers are particularly valuable in machine-mediated commercial environments, where unambiguous identification is crucial.

Some identifiers tell you something about the thing that they identify - the identifier "ISBN 1-900512-44-0" can be assumed to identify a book (always assuming that the ISBN rules have been correctly followed).

However, to find out which book it identifies, it is necessary to consult metadata – the identifier links the metadata with the entity it identifies and with other metadata about the same entity. Metadata is an integral part of making the identifier useful. Some of this metadata may be held in private systems (the publisher's warehouse system, for example) but some of it is more widely available (e.g. Books in Print).

The future of a significant proportion of intellectual property dissemination lies in the network environment, and it is inconceivable that it will not be mediated by machines. Metadata permits both recognition of the entity that is identified and its unambiguous specification; it also allows for the interaction between the entity and other entities in the network (and with metadata about those entities). This implies that the metadata that supports the management of that intellectual property must be machine interpretable.

<indecs>

Efforts at standardising metadata have been underway in different sectors and media for many years, and they have all been given added impetus and urgency by the development of online capabilities, most significantly the Internet. The MARC record is an example of a strong metadata system used in the library environment; recent efforts to generalise the library metadata discussion have evolved into formal systems, of which the Dublin Core is a strong example. Equally, extensive work has been carried out in other media – music, audiovisual, image – generally pushed along by the respective rightsholders,

The <in*d*ecs> (***In***teroperability of ***D***ata for ***E***lectronic ***C***ommerce ***S***ystems – http://www.indecs.org) project was sponsored by European Commission DGXIII, bringing together as partners and affiliates a global grouping of organisations with an interest in the management of content of all kinds in the digital environment. One of the Directors of DocDemon (Zielinski) participated closely in developing the <in*d*ecs> analysis when he was Chief Executive of one of the partners.

The <in*d*ecs> project took stock of all existing metadata efforts in the publishing and library worlds, as well as emerging metadata standards in other media, and particularly metadata relating to rights management. The framework elaborated is thus geared to be inclusive. It provides linkages between, and has been tested on, most of these other efforts.

The core of the project is the definition of a logically rigorous framework for "well-formed" metadata, which allows metadata developed in adherence to different schemes to interact or "interoperate" unambiguously. Without that interaction, different metadata schemes will operate as "trade barriers".

There are only two types of metadata that can be regarded as well-formed. The first of these are labels: the names by which things are called (of which "titles" are a subset). These are entirely uncontrolled and uncontrollable. Identifiers are a specialised type of label, created according to rules, but names nevertheless. The fact that they are created in accordance with a prescribed syntax makes them less prone to ambiguity than other types of label and therefore more readily machine-interpretable than completely free-form labels.

All other metadata (if it is well formed) needs to be drawn from a controlled vocabulary of values, which are supported by a data dictionary in which those values are concisely defined. This means that the values in one metadata scheme (or in one "namespace") can be mapped to those in another scheme; this mapping may not be exact – where two definitions in one scheme both overlap with (but are not wholly contained within) a single definition in another, for example. However, the use of a data dictionary avoids the sort of ambiguity that is inherent in natural language, where the same word may have very different meanings dependent on its context. Where precision of meaning is essential, human beings can clarify definition through a process of dialogue. This is not generally the case with computers.

In addition to the concept of "well-formed" metadata, <in*d*ecs> provides a set of "kernel metadata" that should be present in all descriptions of content of any kind. The set of kernel metadata implemented in the digital object identifier is given in the table below.

Element	**Definition**	**Status**	**Number**	**Allowed values**	**Possible genre qualifications**
DOI	A DOI	Mandatory	1 only	DOI	
DOI Genre	A class of entities with common characteristics defined by the IDF community	Mandatory	1 minimum	From DOI genre tables	
Identifier	A unique identifier (e.g. from a legacy scheme) applied to the entity	Qualified by Genre	1 minimum	Any alphanumeric string but must include an identifier type, e.g. ISBN.	Normative: mandatory unless the DOI Genre extension rules specifically state otherwise.
Title	A name by which the entity is known	Qualified by Genre	1 minimum	Any alphanumeric string	Normative: mandatory unless the DOI Genre extension rules specifically state otherwise.
Type	The primary structural type of the entity	Mandatory	1 only	From: *Work, Physical Manifestation, Digital Manifestation, Performance*	
Origination	The process by which the entity was made	Mandatory	1 minimum	From: *Original, Modification, Excerpt, Compilation, Replica*	
Primary agent	The name or identifier of the primary agents(s) (normally but not necessarily the creator).	Mandatory	All primary agents. (1 minimum, but all entities fulfilling the same agent role must be included.)	Identifier or Name from an agreed Genre namespace	The specification of the Primary Agent for any Genre is determined by the DOI Genre rules.
Agent role	The role(s) played by the primary agent(s)	Mandatory	1 minimum	Role code from an agreed Genre Namespace	

Characteristics of the <indecs> Framework

The framework recognises:

- metadata relating to any types of creation;
- the integration of descriptive metadata with commercial transactions and rights;
- that metadata should be created once, used many times for different purposes; and proposes:
- a generic attribute structure for all entities;
- events as the key to complex metadata relationships;
- a metadata dictionary for multimedia intellectual property commerce;
- unique identifiers (iids) to be assigned to all metadata elements;
- the need for transformation processes to express the same metadata at different levels of complexity for different requirements.

The <indecs> framework is designed to help bridge the gap between the powerful but highly abstracted technical models such as that expressed in the Resource Description Framework (RDF) and the more specific data models that are explicit or implicit in sector- or identifier-based metadata schemes.

For DocDemon and Collexis® Findware™, its significance is in providing a standardised "shorthand" for linking the resource discovery process based on conceptual fingerprints and the Collexis® Matching Engine and technology that locates the text. Apart from local systems of linking a reference to a text of text summary resident on a local server, generally by a simple URL, rightsholders have developed the digital object identifier (DOI) to provide a universal solution to content location.

The Digital Object Identifier

One of the key challenges in the move from physical to electronic distribution of content is the rapid evolution of a set of common technologies and procedures to identify, or name, pieces of digital content ("digital objects") . A widely implemented and well understood approach to naming digital objects is essential if we are to see the development of services that will enable content providers to grow and prosper in an era of increasingly sophisticated computer networking.

It became increasingly apparent during the 1990s that existing approaches to identification would prove inadequate to meet the need. Publishers, for example, could see the deficiencies of the ISBN as an identifier for electronic publishing, because of its limitation to identifying physical objects, and the difficulties with applying it to items smaller than "a book". At the same time, the only content identifier commonly in use on the Internet – the Uniform Resource Locator (URL) used to find particular pages on the World Wide Web – was clearly deficient, not least because it was not used to identify content but rather location. The location is transient, whereas what was necessary was a means of identifying content itself, persistently and without ambiguity.

From a research project begun by the American Association of Publishers in 1996 to consider ways of resolving this problem, what has developed is the complete DOI System. To date, over three million DOIs have been registered, and over 100 organisations are now DOI Registrants. Most of the DOIs have been issued as a by-product of the CrossRef project (www.crossref.org), which is the first (and to date only) DOI Registration Agency. One of the Executive Directors (Velterop) was a member of the CrossRef founding group.

Since 1998, the DOI has been managed by the International DOI Foundation (IDF), which was established to assume a leadership role in the development of a framework of infrastructure, policies and procedures to support the identification needs of content providers in the multinational, multi-community environment of the network. Both DocDemon Directors have participated in the development of the DOI, and one of them (Zielinski) was elected by DOI members to serve on the first Board of the IDF until April 1999.

Describing the DOI

The DOI can be described as "***persistent identifier*** of intellectual property entities". An "entity" is simply something that is identified. The equivalent term often used by the World Wide Web community is "resource". The DOI can thus be used to identify any of the various physical objects that are "manifestations" of intellectual property: for example, printed books and journal articles (but also CD recordings and videotapes). Critically, the DOI is a persistent identifier: even if ownership of the entity or the rights in the entity change, the identification of that entity should not (and does not) change. The responsibility for managing the DOI changes, but not the DOI itself.

The purpose of the DOI System is to make the DOI an ***actionable identifier***: a user can use a DOI to do something. The simplest action that a user can perform using a DOI is to locate the entity that it identifies. In this respect, a DOI may look superficially like a URL. However, the technology that underlies the DOI facilitates much more complex applications than simple location; and the DOI identifies the intellectual property entity itself rather than its location.

The DOI System has been designed to be able to ***interoperate*** with past, present and future technologies, including:

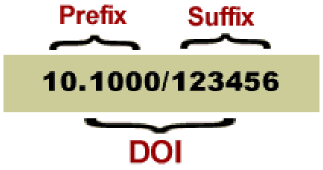
- So-called "legacy" identifiers, those we have used in the past and continue to use today – like the ISBN – can form an integral part of the DOI naming system. Businesses can use familiar naming or numbering systems in this new environment.
- The DOI even in its simplest implementation provides an actionable identifier on the Internet which is fully compatible with URLs and the World Wide Web, providing users with a persistent identifier that can overcome the problem of entities that are relocated from one place to another on the Web (because of change of ownership or simply for administrative reasons).

The more sophisticated and complex applications that are being developed as parts of the DOI System will be fully compatible with the standard environment of the Internet as it develops. The core Handle System technology that the DOI uses will always be "open standards" based.

Characteristics of the DOI System

The four primary components of the DOI System are:

- ***Enumeration:*** assigning an alphanumeric string to an intellectual property entity that the DOI identifies
- ***Description:*** creating a description (in metadata) o the entity which has been identified with a DOI.
- ***Resolution:*** making the identifier "actionable" by providing information linked to the DOI, and the technology to deliver the services that this can provide to users.
- ***Policies:*** The rules that govern the system.



Syntactically, a DOI consists of a prefix and a suffix. All DOIs start with the number 10, which indicates the string is a DOI. This is followed in the prefix by a number identifying the issuing agency. The suffix identifies the entity (defined below), and can be any alphanumeric string that the Registrant chooses. This can simply be a sequential number, or it can make use of an existing (legacy) identifier like the ISBN.

The DOI is covered by American National Standard ANSI/NISO Z39.84-2000: *Syntax for the Digital Object Identifier* (approved May 2000).

The DOI System is based on an underlying resolution technology called the Handle System. The Handle System is a general purpose distributed information system designed to provide an efficient, extensible, and secured global name service for use on networks such as the Internet. The Handle System includes an open set of protocols, a namespace, and a reference implementation of the protocols. The protocols enable a distributed computer system to store names, or handles, of digital resources and resolve those handles into the information necessary to locate, access, and otherwise make use of the resources. These associated values can be changed as needed to reflect the current state of the identified resource without changing the handle, thus allowing the name of the item to persist over changes of location and other current state information. Each handle may have its own administrator(s) and administration can be done in a distributed environment. The name-to-value bindings may also be secured, allowing handles to be used in trust management applications.

The Handle System was originally conceived and developed at the US Center for National Research Initiatives as part of the Computer Science Technical Reports (CSTR) project, funded by the Defense Advanced Projects Agency (DARPA) under Grant No. MDA-972-92-J-1029.