ELECTRONIC NEWS SERVICE FOR THE EUROPEAN CONSTRUCTION INDUSTRY

R. Amor, R. Marsh, and A. Hutchison
Building Research Establishment, Watford, UK

ABSTRACT: The Electronic News Service (ENS at http://www.connet.org/NS/Intl/) is an AltaVista-like service which is specialised for the construction industry. It has been developed as one of the services in the EC project CONNET (Construction Information Service Network, at http://www.connet.org/, Turk and Amor 2000). The ENS provides a searchable index of the contents of Internet sites relevant to the built environment. The database containing the source set of Internet sites to index has information on over 19,600 Internet sites across the world, categorised and classified by several criteria. This base set of Internet sites is drawn from all major lists offered to the construction industry (e.g., Yahoo, EEVL, UK-BRP, etc) and from published sources (e.g., Architect's Journal, Building magazine, etc). Over 35 major lists of site sources are utilised to build, maintain, and develop this set of resources for the construction industry.

This paper describes the ENS, the methods it uses to gather and index construction information across the world, and the services it offers to the construction industry. It also provides an analysis of the references gathered from the 35 major lists of resources which are established across the world. This analysis looked at the overlap that exists between the Internet sites referred to by each of these lists (which is remarkably small), the particular biases which appear in the lists (mainly towards English language and USA-based information), the currency of the sites in the lists (quite poor), and the predicted coverage of total construction-based Internet resources found in all of these lists.

KEYWORDS: Search engine, Internet news, Internet statistics, news-feed

1. INTRODUCTION

The growth in Internet usage over the last decade has been phenomenal. Recent surveys in the UK show that some 39% of construction companies have an organisational web site, and 69% of companies have access to the Internet (Business and IT Survey 1999). Assuming that similar percentages are accurate across Europe then of the approximately two million construction companies in Europe there are about 800,000 who have an organisational web site, even if this is just a page hosted on a domain specific portal.

Finding this information is a great problem. There are many approaches to indexing the web. Systems such as AltaVista (http://www.altavista.com/) have spiders which crawl the web trying to identify every possible link to information. Yahoo (http://www.yahoo.com/) like systems have human operators who check submitted sites to ensure a consistent classification into subject areas. The cross-system approaches, like AskJeeves (http://www.askjeeves.com/), provide a wider search by linking with many independent search engines. However, the growth in the Internet has far outstripped the ability of these generic search tools to handle the information flow. Recent analysis shows that the best search engine only covers 16% of the Internet's estimated 800 million publicly indexable web pages (Lawrence and Giles 1999). The top 11 search engines together cover approximately 42% of the total, so they are in no way comprehensive.

The approach followed in many domains to address this problem has been to establish portals which provide a link to quality information in their areas. In the built environment domains there are many such portals providing links to profession-specific sites (e.g., architecture, civil engineering), or to particular application areas (e.g., energy conservation), and even very comprehensive personal hotlists (e.g., Matti Hannus' site at VTT in Finland – http://cic.vtt.fi/links/). These sites tend to provide a restricted set of links to sites that have been shown to contain quality and useful information.

This project builds on top of these smaller initiatives in attempting to develop an index to the built environment that can be both searched and utilised to provide an active feed of updated and new information (news) for users. The ENS was developed as part of the EC funded CONNET project described below.

1.1 EC-CONNET project

The EC funded project CONNET (CONstruction information service NETwork, Turk and Amor 2000, CONNET 1999) was developed as part of the European technology transfer initiative to establish European Technology Transfer Networks (ETTN 1999). BRE has been working with partners in Finland (VTT and BII) and Slovenia (IKPIR) to develop the initial framework of this service for Europe. The project provides the construction industry with a platform to demonstrate the applicability of the EC's Electronic Technology Transfer Initiative to this diverse sector. The CONNET project provides the construction industry with an essential source of information, by creating a "virtual technology park", accessible to the whole industry, regardless of national boundaries.

A suite of five Internet-based information services has initially been developed, comprising a technical information centre; a waste exchange centre; a manufactured product service; a calculation and software centre; and an electronic news service. The scope of the other four services is as follows:

- The Technical Information Centre is working to draw together all of the publishers for the built environment within one nation, in this case the UK. This service currently indexes 20 of the over 220 publishers identified for the UK, providing access to the details of over 27,000 publications. Users can identify relevant publications from any publisher and then purchase through the site.
- The Waste Exchange Centre extends the current UK-based system to better enable the disposal and re-use of site waste across organisations both nationally and in Europe. Availability of, and requests for, waste materials are automatically matched in order to broker greater re-use of materials.
- The Manufactured Product Service enables Finnish and export market users to identify manufactured products which match their design specification, by incorporating product attributes into the selection system. Users are able to identify certified products and drag-and-drop CAD information into their designs.
- The Calculation and Software Centre provides the European entry point for information on all software products available for the architecture, engineering and construction domains (over 4,000 collated to date). Online demonstrations, online purchase, and even pay-per-use software is available.

2. THE ELECTRONIC NEWS SERVICE

The ENS service provides the industry with a free method of identifying sources of information based on the content of a web page or service-based classifications. Users are able to define profiles for news they have an interest in and to be periodically, and automatically, notified of new or modified web pages and sites which meet their criteria. Running the ENS within the CONNET network provides mechanisms to link together all

news services which are available, to provide answers to user requests across complementary systems, or even to take requests established for news and use them to identify other information sources of relevance (e.g., publications, software, products). This section describes the components of the ENS and its method of operation.
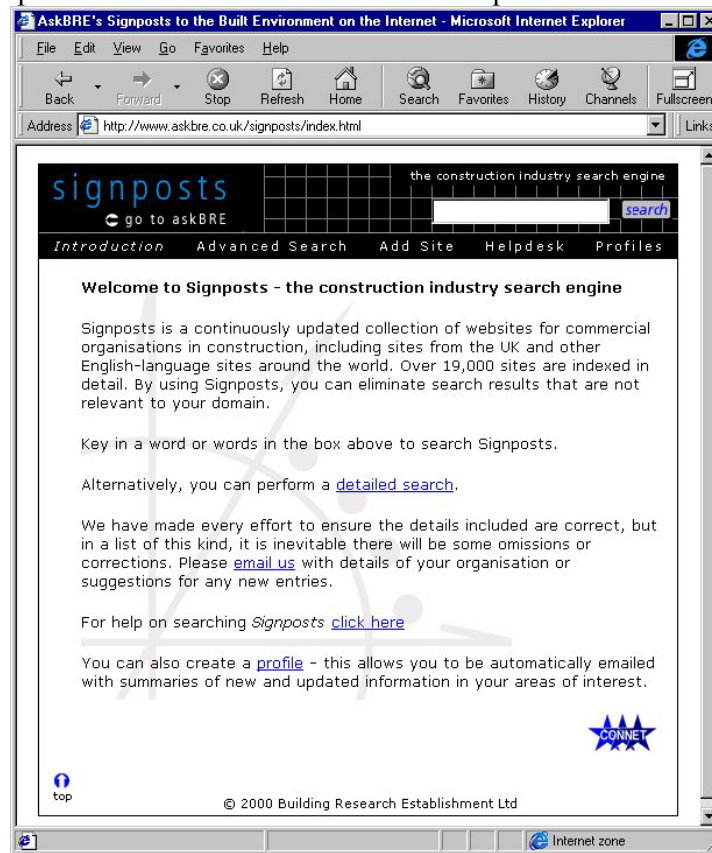

Figure 1. News Service interface and simple search

2.1 Data representation

The data representation required for Internet web pages is fairly standardised inside the range of available search systems. The SOIF (Summary Object Interchange Format) is the record format used by the Harvest software (Harvest 1996), which is a freely available system utilised by many universities, as well as in this project (see Section 2.2). This format has gained significant backing through the use of Harvest and the SOIF by Netscape as the basis for their Catalog Server product.

The ENS has a two tier representation of information about sites (see Section 2.2) which utilises both a relational database and the Harvest full-text web index. The relational database replicates a portion of the SOIF structure to provide access to top level information on a site including title, keywords, description, date last updated, whether it is operational or not, and whether it is currently being redirected.

2.2 Service structure

The Electronic News Service is based around three databases that contain details of web resources at varying levels of abstraction (see Figure 2).

- The lists database contains information on all the major lists of resources on the Internet. These include the various virtual libraries and personal hotlists. Currently over 35 lists of resources have been identified as indexable. For each of these lists the ENS records how

to extract sites from the list and the domain they serve. The ENS system visits each of these sites periodically and checks for any new additions to their lists. All sites gathered from these lists are fed into the main ENS database and indexed as to which list they were identified in.

- The main ENS database contains an entry for every known site. Currently this comprises over 19,600 sites across the world. The ENS periodically visits each site to ensure that it is still working and to gather top level information about the site. This information includes the site title, keywords, description, and all body text. Frame sets are recognised in a site and all referenced pages gathered for the top-level frame set. Counters are maintained for any problems with a site and after a fixed number of problems (currently five sequential visits which cause problems) the site is marked as defunct. Periodically all defunct sites are rechecked to see if they have come online again. The ENS also tries to identify which sites are redirected to a new URL and marks the site accordingly. This database uses Microsoft SQL Server version 7, which provides a full-text index of the gathered information as well as the normal relational database abilities.

- The Harvest database is built up from Harvest gathering all web pages from the working set of web sites identified from the ENS database. The Harvest system is instructed to visit each site and gather all referenced web pages that are further down the site hierarchy from the original site (i.e., staying on the same machine and following links down directories on that machine). Currently the ENS database is growing faster than the Harvest system running on BRE's bandwidth can keep up with, so there is no up-to-date index of every web page of every ENS database site.
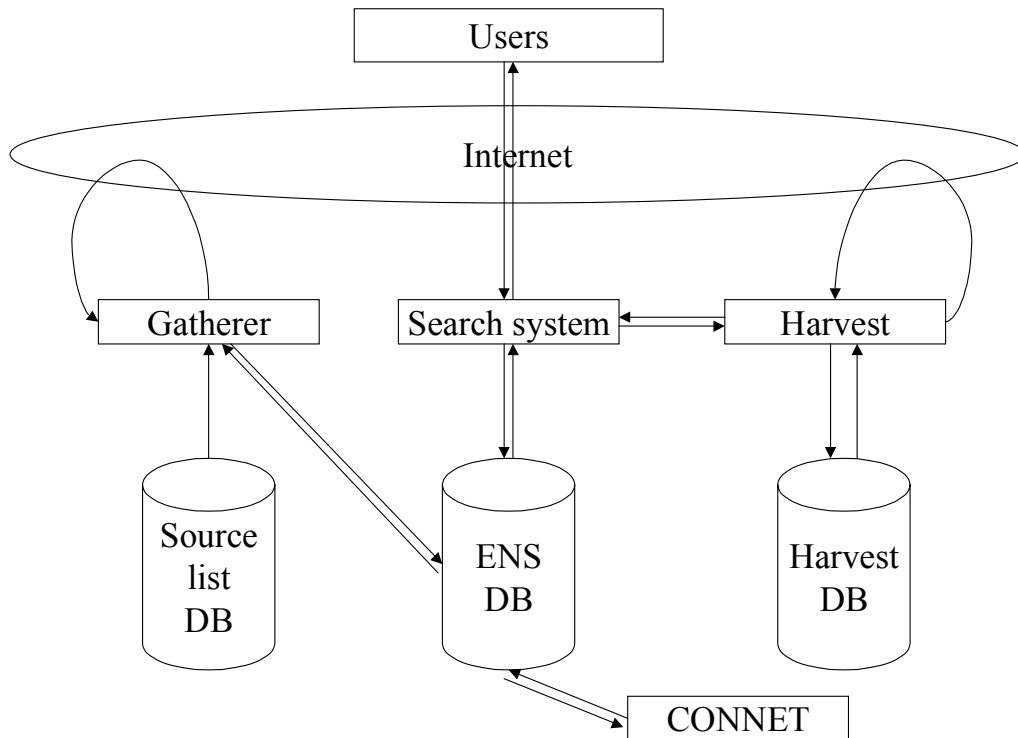

Figure 2. Structure of the ENS system

The ENS system also contains information about CONNET to allow the service to make use of the user profiling and tracking services offered by the central CONNET service (see Figure 2). The use of a full-text index allows matches for user queries to be given a score based on

weightings and the number of times search terms appear in the record. This score is used to rank the results for the user. The ranking for a matching record is weighted to give the highest ranking to items where the search term appears in the title with a lower weighting given when the search term only appears in other fields (e.g. in the web page's body).

Users interact with the system via a forms interface that allows them to enter search terms. From the search results page a user can: a) go directly to the named web site, for more information, b) they can pass their query to CONNET which can save it to the user's profile, or c) they can pass the query to another CONNET service, e.g. the technical information centre or the software centre.

2.3 Collecting information

The sites which are utilised by the ENS are drawn from two sources. These are existing lists of relevant sites and sites proposed by users of the system. The collation of sites by these means is described below.

2.3.1 Existing lists

Existing lists of resources for built environment industries are the major source of sites for the ENS. These sites include virtual libraries for a range of professions (e.g., architecture, civil engineering), organisations' lists (e.g., universities and professional bodies), construction portals' lists, and personal hotlists (e.g., Matti Hannus' list in Finland). The ENS periodically visits each of these lists and scans their contents for new sites. The ENS checks for the ROBOTS.TXT file on all sites it visits to ensure that it has permission to index the information on the sites. To date none of the sites visited restricts access to its lists in this way. The ENS parses all pages on a lists site and extracts all URLs it finds in the web pages. These are checked against the existing set in the database and new sites added, as well as ensuring that the name of the list site is recorded against each site found. This process does not work for sites which utilise a redirect mechanism on their list (i.e., where the list site records that a user has navigated to a particular site by trapping the click on a link) and this means that there are several major commercial list sites which can not be gathered by this means. In the future it may be possible to gather these lists through bespoke programs which follow every redirection to their ending web site.
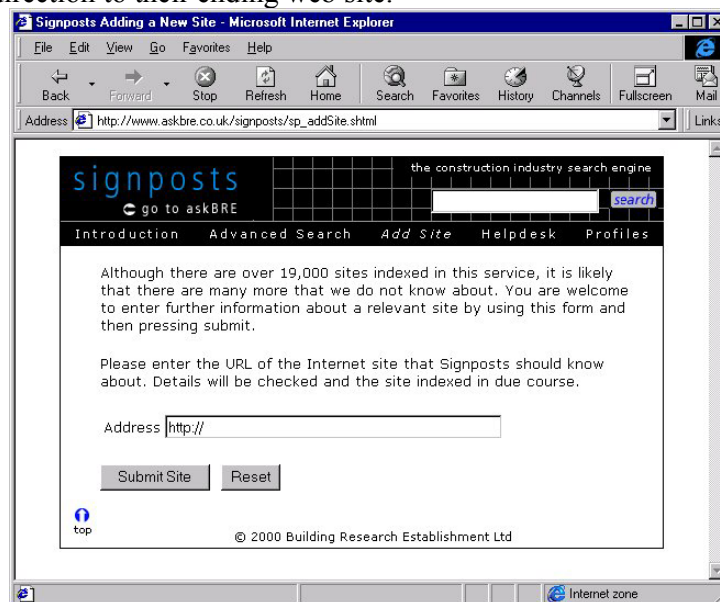

Figure 3. The user's interface for adding a site to the ENS database

2.3.2 Individual web sites

Users of the ENS system are able to propose that a site be added to the ENS database (see Figure 3). Any site proposed in this manner is put in a holding area to be checked by the maintainer of the ENS system before being added to the ENS database. This ensures that only built environment related sites are added and that miscellaneous sites do not end up in this search engine.

2.4 Searching the repository

The ENS provides two methods for accessing the index of sites. The simplest (see Figure 1) allows any set of words to be typed into a search box and submitted. A more advanced search (see Figure 4) allows the search to be restricted to particular domains (i.e., countries or organisation types) as well as restricting to the resources of a particular list (e.g., the Engineering Virtual Library) or groupings of resources, for example, conference lists. The advanced search also ties in a thesaurus tool, which expands the user's search query to cover related words for their search. This helps to ensure that all possible ways of describing a subject are used when searching (e.g., American and English terms for the same product sometimes differ).

Results from a search are ranked before being presented to the user. This tries to ensure that the most relevant results are shown first. The ranking algorithm takes advantage of the fact that the site's title, keywords, and body text are separately recorded. The ranking gives greatest prominence to matching words in the title, followed by the keywords list and then the main body text. Associated with this is a frequency ranking which ensures that the number of occurrences of a search term promotes the site in the ranking (e.g., 20 occurrences of 'timber' in a page is a greater sign of relevance than 10 occurrences).
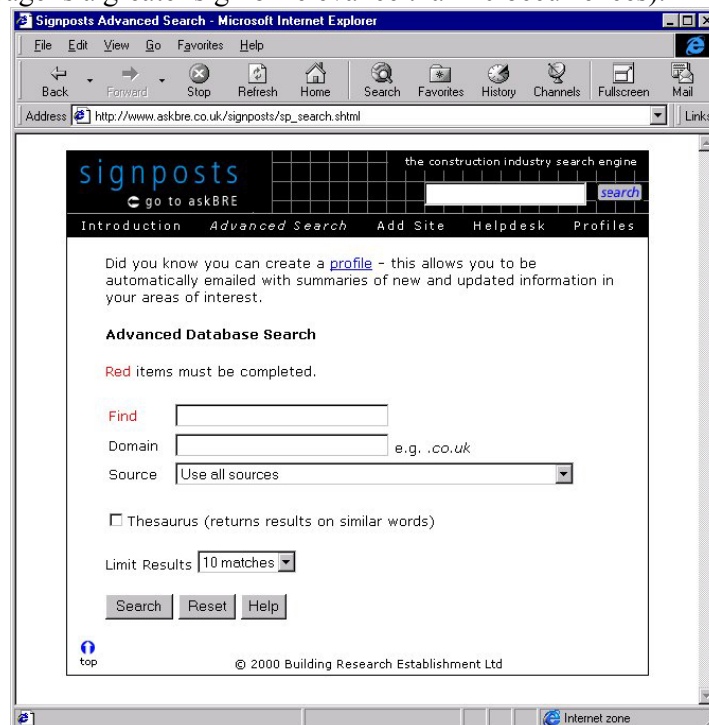

Figure 4. Advanced searching in the ENS

2.5 Notification to users

The CONNET infrastructure also offers support for active notification of users. In the case of the ENS a user can specify that they wish their search to be retained as a profile to be run periodically and the results of new or updated Internet sites to be emailed to them. CONNET maintains all profiles for a user across all services so the user can access and manage all their profiles across all information services from a single point.

2.6 Interoperability

One of the main aims of CONNET has been to ensure that information services developed are easily interoperable. To achieve this, each service has a defined API that can be accessed by other systems. This allows, for example, the CONNET system to search over all services to provide a collated set of information for any query (e.g., matching software, books, products, etc). It also means that any other CONNET service can offer the ability to identify related news. For example, from the publications service all related news can be identified after a publications search is completed. This API, in association with CONNET's information service infrastructure, also means that a search in one national system can be passed to all European systems of a similar type. For example, if a search for sauna sites returns a poor result in the UK system then a European wide search is likely to identify many relevant sites from Finland's news service (several of which are in English).

This also allows each individual service to become associated with a wider range of portals for the industry. For example, the ENS can run as a stand-alone service, but it can also be closely integrated with any other portal service.

3. ANALYSIS

In putting together the ENS system there were several problems encountered which led us to examine what is being provided to the industry from these lists (and therefore from the ENS). The biggest problem in setting up our indexing system was visiting sites which were defunct (i.e., would not respond or provided an error code). This happened frequently, and the ENS required a fair amount of coding to manage whether a site was defunct or not (e.g., revisiting periodically before deciding that a site is defunct). This section reports on the collated findings from this management work we had to perform on the ENS system.

3.1 Basic statistics

There are several basic statistics to be drawn from the ENS database, including:

Measure	Sites	Percentage
Total sites in ENS	19,636	2.5% of EC sites
Estimated sites in EC	800,000	
Defunct sites	5,000	25.5% of ENS
Working sites	14,636	74.5% of ENS
Titled sites	14,332	97.9% of working
Keyworded sites	6,503	44.4% of working

The analysis of these statistics shows that over a quarter of all sites listed in the 35 lists for built environment industries are currently defunct. This indicates that these list sites are not checking to see if their links are active with any great frequency. As a user of such a site this is likely to cause discontent as one in four links shown is not going to work. It is clear that almost all sites have titles, but surprising that such a small number of the top level entry points to sites have keywords. This lack of keywords will harm the chances of sites being found by the major search engines, where keywords are treated with greater importance than the body text on a site. It is also significant to note that only 2.5% of the estimated construction organisations with web sites are represented in this repository of sites. While it

is likely that the sites that are indexed are the more important sites for the industry, it shows the great effort that would be needed to extend any of the list sites to cover any great percentage of the industry's Internet information resource.

The following statistics are drawn from the Harvest index of all web pages on a site:

Measure	Value
Average number of pages per site	480
Average amount of text per site	704 kbyte
Average amount of text per page	1.46 kbyte

The analysis of these statistics provides some measure of the task of indexing construction resources on the Internet. The size of the index for the 14,636 sites which are currently working will be over 10 gigabytes and comprise some seven million web pages. However, if this is extended to the estimated 800,000 construction companies with web sites then the index required will be over 560 gigabytes and comprise some 384 million web pages. This extended analysis does not, however, match other estimates of the size of the Internet. Lawrence and Giles (1999) estimate 800 million indexable web pages, so 384 million belonging to construction is much greater than is possible. The likely explanation for this discrepancy is that the sites that are currently indexed are the biggest and most important repositories of information for the industry. The large number of sites which are not indexed are likely to be much smaller sites for smaller companies.

3.2 Spread of sites

As each web site in the ENS database is linked to the list from which it was found it is possible to identify how many list sites are associated with each URL. The analysis against the 35 list sites shows the following:

Linked by	Sites	Percentage
1	14,811	75.43%
2	3,574	18.20%
3	862	4.39%
4	228	1.16%
5	97	0.49%
6	33	0.17%
7	15	0.08%
8	9	0.05%
9	4	0.02%
10+	3	0.02%

This shows an incredibly low amount of overlap between sites. Almost every list site has a separate collection of sites, either because they are limited to the domain they are established for (e.g., architecture, civil engineering), or the country they are working in (e.g., USA, Canada, UK), or because of the small number of the total sites which are captured in any of these lists (i.e., if there are 800,000 construction-related organisations on the Internet).

3.3 Famous web sites

Turning around the analysis in Section 3.2 it is possible to identify the most popular sites from these sources. The analysis against the 35 list sites shows the following:

Links	URL	Title
16	http://www.bre.co.uk/	BRE - Constructing the future
11	http://www.ice.org.uk/	ICEnet : The Institution of Civil Engineers Homepage
10	http://www.fmb.org.uk/	BRIX - Builders Resource & Information Exchange
9	http://www.nrc.ca/cisti/journals/tocciv.html	CANADIAN JOURNAL OF CIVIL ENGINEERING
9	http://www.nrc.ca/irc/	Institute for Research in Construction
9	http://www.bfrl.nist.gov/	Building and Fire Research Laboratory

9	http://www.autodesk.com/	Autodesk - The Design Resource Leader
8	http://www.cibse.org/	Welcome to CIBSE On-Line
8	http://erg.ucd.ie/	Energy Research Group, UCD
8	http://www.ciob.org.uk/	The Chartered Institute of Building (CIOB) - International Headquarters
8	http://web.arch-mag.com/	Web Architecture Magazine . WAM HOMEPAGE . Internet into ARCHITECTURE
8	http://www.arch.buffalo.edu/pairc/	Cyburbia - The Planning and Architecture Internet Resource Center
8	http://www.aecinfo.com/	Bricsnet.com - The e-marketplace for the global building industry
8*	http://www.aecnet.com/	Welcome To Your WebSTAR Home Page
8*	http://www.aia.org/	www.aia.org - Address Change
8	http://www.asce.org/	ASCE - American Society of Civil Engineers

Whether any serious analysis can be drawn from this is unclear. Some of the highly linked sites in this list are surprising, and the figures here may indicate how active different organisations are in promoting themselves. Of interest are the two starred items, where both of these sites are now redirected. AECNet is now part of Cyburbia and the AIA now have a new web site. However, as noted in Section 3.1, many of the list sites are not checking and updating their lists of URLs (or being notified of changes by the organisations when they update their web sites).

3.4 Countries represented

Analysis of web site URLs to determine the spread of countries provides very few useful statistics. This is mainly due to the use of .com, .edu, and .org by organisations across the world and not just the USA. The statistics for these sites is as follows:

Code	Sites	Percentage
com	6012	30.6%
edu	1778	9.1%
org	1402	7.1%
net	731	3.7%
gov	534	2.7%
mil	60	0.3%
int	33	0.2%

This accounts for over 50% of the sites in the ENS database. For the remaining 9,086 sites the following statistics hold:

Code	Sites	Percentage
uk	4338	47.7%
ca	748	8.2%
au	524	5.8%
us	480	5.3%
se	401	4.4%
de	393	4.3%
nl	258	2.8%
fi	224	2.5%
fr	160	1.8%
jp	139	1.5%
it	121	1.3%
ch	107	1.2%
nz	96	1.1%

What these statistics show is the predominance of English language sites in the list sites. This will of course be skewed by the selection of list sites, and if there exist specific sites for particular nations the balance would be redressed by their inclusion.

3.5 List sites

The final analysis presented shows all of the list sites utilised in this analysis, and some basic statistics on what they provide:

Title	URL	Sites	Defunct
AEC Software Library	http://software.foraec.com/	989	13%
Architects' Journal		182	16%
Architectural Engineering Virtual Library	http://energy.arce.ukans.edu/wwwvl/wwwarce.htm	87	25%
Architecture Virtual Library	http://www.clr.toronto.edu:1080/VIRTUALLIB/arch.html	2130	44%
Australian AEC Network	http://www.arch.su.edu.au/kcdc/aec/index.html	224	33%
Bricsnet.com - The e-marketplace for the global building industry	http://www.aecinfo.com/	617	25%
Building magazine		103	14%
CALS related sites		20	25%
Civil Engineering Library	http://www.v-biblioteket.lth.se/civil.htm	867	14%
Concurrent Engineering Network	http://esoce.pl.ecp.fr/ce-net/	12	8%
Conference and presentation sites		303	29%
Construction Computing magazine		72	18%
Construction Industry Board	http://www.ciboard.org.uk/	115	13%
Construction Industry Trading Electronically	http://www.cite.org.uk/	29	14%
Construction Information Signpost	http://helios.bre.co.uk/cis/	1408	15%
Construction Sites - Napier University	http://www.bs.napier.ac.uk/www/construction.html	135	39%
Co-operative Network for Building Research	http://www.tce.rmit.edu.au/BCE/97LINKS/CNBR/cnbr.HTM	41	24%
CTI Centre For the Built Environment	http://ctiweb.cf.ac.uk/	788	23%
Cyburbia - Planning & Architecture Internet Resource Center	http://cyburbia.ap.buffalo.edu/	4284	27%
EDI related sites		46	33%
Edinburgh Engineering Virtual Library	http://www.eevl.ac.uk/	3760	14%
Energy Crossroads	http://eande.lbl.gov/CBS/eXroads/EnergyXroads.html	310	22%
ENS User		402	12%
European Construction Institute	http://www.eci-online.org/	630	24%
Galaxy Engineering	http://galaxy.einet.net/galaxy/Engineering-and-Technology.html	330	21%
Integrated CAD Research Information System	http://www.fagg.uni-lj.si/ICARIS/cr/	1077	44%
International Building Council	http://www.cibworld.nl/	123	14%
ISO-STEP and IAI-IFC related sites		119	37%
Journal and publication sites		1676	19%
Landscape Architecture Virtual Library	http://www.clr.toronto.edu:1080/VIRTUALLIB/larch.html	2126	44%
Paul Herrington's Useful Links	http://skuld.cage.curtin.edu.au/civil/links/	110	53%
ProC-E-Com Review	http://www.sbe.napier.ac.uk/ProConIT/ProCECom/reviews.htm	39	13%
Robert Amor's Hotlist	http://helios.bre.co.uk/ccit/people/ra_info/hotlist.htm	177	37%
Thomas Froese's Internet Bookmarks	http://www.civil.ubc.ca/~tfroese/Bookmarks.html	531	40%
University of Auckland Architecture Property and Planning	http://archpropplan.auckland.ac.nz/Planning/Planning.html	83	34%
VTT Construction IT Links	http://cic.vtt.fi/links/	1006	30%
Yahoo UK Lists	http://www.yahoo.co.uk/	647	6%

4. FUTURE DEVELOPMENTS

The ENS is being taken forward in a successor project to CONNET. This EC-funded project, called I-SEEC (Information Services to Enable European Construction Enterprises), is running in the financial year 2000/01. The I-SEEC project has partners in seven European countries (UK, Spain, Italy, The Netherlands, Finland, Iceland, and Slovenia) to ensure European take-up. I-SEEC aims to enhance the CONNET system by expanding the virtual technology park for construction, by Europeanising the services that already exist, and by adding a number of services which may already exist (though are not electronically accessible), or which are currently being developed by organisations located in EU member states or associate states. As one of the services to be taken into Europe, the ENS system will be adapted to allow a national focus to be presented for each member state. This will incorporate an interface in the national language and higher ranking to sites which are either in that nation or in the language(s) of that nation.

ACKNOWLEDGEMENTS

REFERENCES

AskBRE (2000). askBRE Information Services for the UK Built Environment, http://www.askbre.co.uk/.

Business and IT Survey (1999). Business and IT Survey 1998/9, The Chartered Institute of Building, UK.

CONNET (1999). Construction Information Service Network, http://www.connet.org/.

ETTN (1999). European Technology Transfer Network, EC, http://ettn.jrc.it/.

GENIAL (1999). Global Engineering Networks: Intelligent Access Libraries, http://wwwgen.uni-paderborn.de/GENIAL/.

Harvest (1996). Harvest software, http://www.tardis.ed.ac.uk/harvest/.

Lawrence, S. and Giles, C.L. (1999). Accessibility of Information on the Web, Nature, No. 6740, July 8, pp.107-109.

Turk, Z. and Amor, R. (2000). Architectural foundations of a construction information network, International Journal of Construction Information Technology, 7(2), pp. 85-97.

WordNET (1999). WordNET thesaurus system, http://www.cogsci.princeton.edu/~wn/.