The UK Technical Information Centre

R. Amor and A. Hutchison
BRE, Bucknalls Lane, Garston, Watford WD2 7JR, UK
hutchisona@bre.co.uk*, Phone +44-1923-664556, Fax +44-1923-664689*

ABSTRACT: The UK Technical Information Centre (TIC, URL-3 1999) is one of a set of Internet-based services developed in the EC funded project CONNET (URL-1 1999). The TIC aims to provide a single source of information on any purchasable publication of relevance to the built environment in the UK. Over 220 publishers have been identified and it is estimated that information on about 50,000 publications will be collected. This paper examines the issues raised in structuring and integrating the publication information from such a wide range of publishers. Particular attention is paid to data models and integration, the impact of classification systems, and coping with non-standard transfer formats. The structure of the TIC is detailed and strategies for searching across the published base data are examined along with interoperability issues and the benefits these bring to evolving portals.

1 INTRODUCTION

The UK is not unique in having a wide range of publishers who serve various segments of the built environment market. The publishers range from standards organisations (e.g., BSI and HMSO), to generic publishing houses (e.g., Butterworth-Heinemann and Pearson Education), to professional organisations (e.g., BSRIA, ICE, and IChemE), through to specialist organisations (e.g., Association for Specialist Fire Protection). Initial lists have identified over 220 UK publishers of relevance to the built environment with between 5 and 15,000 publications each of relevance to this domain.

Currently, it is not possible to access details of all publications from all of these publishers. While many have web sites, it is infeasible to visit the majority of these sites to identify potentially relevant publications, even if all the relevant publishers are known in advance. It is also clear that major online bookshops do not index many built environment publications. For example, BRE has over 1000 publications currently in print. However, Amazon (USA and UK) list less than 100 of these publications. Specialist bookshops in this area concentrate on particular domains (e.g., RIBA bookshop) and also only stock a small percentage of published titles which are relevant to their major market.

The driver for addressing this issue is that these publications are absolutely vital for all practitioners in the industry. The scope of the publications cover technical, regulatory and legislative information and advice. This includes standards, best practice guides, directives, technical notes, new construction techniques, new certified processes, etc. which in total form a major portion of the knowledge base of the professionals in the built environment industries.

Finding the right publication at the right time enables problems and solutions to be addressed accurately and with high quality.

The TIC addresses these issues by creating a single point of access through which it is possible to identify the wide range of high quality, purchasable, technical information that is of benefit to the whole construction industry. The TIC was developed as part of the EC funded CONNET project (see below) to help enable technology transfer in the built environment industries. The remainder of this paper addresses the representational and process problems encountered in establishing the TIC for the UK.

1.1 The EC CONNET initiative

The EC funded project CONNET (CONstruction information service NETwork, Turk and Amor 2000, URL-1 1999) was developed as part of the European technology transfer initiative to establish European Technology Transfer Networks (ETTN, URL-4 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 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 3,500 collated to date). Online demonstrations, online purchase, and even pay-per-use software is available.
- The Electronic News Service enables members of the construction industry to register an interest in specific topics and to be notified of any Internet published news that matches their interest. The news sources are drawn from the main information providers and professional institutes in the industry, both within Europe and internationally. Currently over 19,000 Internet sites have been identified and indexed for this service.

2 REPRESENTING BIBLIOGRAPHIC INFORMATION

One of the major issues in establishing a gateway to publications is the representation to be used and how this relates to the various representations chosen by publishers. This section describes the formats offered by publishers, the classifications they use, and the choice of a standard representation for Internet publication representation.

2.1 Publisher representations

The problem of managing information from publishers has two aspects. One is the representation they have chosen for their data, and the second is the transfer mechanism they employ to send the information.

2.1.1 Publisher data representations

A surprising large number of publishers were found to have utilised different representations for the metadata on their publications. For the larger publishers (e.g., BSI with Perinorm) this represents the structures that they have built up over many years to represent publications in their specific domains. For the majority of smaller publishers the structure represents the particular bibliographic information system they purchased within their organisation and any extensions they made to it for their management purposes. As can be seen from Table 1 there is a great variance in what is represented and the detail into which it is subdivided.

Table 1: A selection of the data representations

Publisher or Format Type	**Fields**
BSRIA	ID, Title, Author, Abstract, Availability, ISBN, Keyword, Source, Member price, Non-member price, Publication date
BSI (Perinorm format)	Origin code, Update flag, Document identifier, Publication date, Issuing body, Status, Effective date, Confirmation, Classification, Committee reference, Expiry date, Certification, Original language, Title, Available from, Translations, Pages, Price, Full text address, Abstract, Descriptors, Notes, Published in, Handbook, Replaced by, Replaces, Amended by, Amends, Draft superseded, International relationship, Cross references, Legislation, Sectional list, Withdrawal date
Butterworth-Heinemann	ISBN, Title, Author, Author biography, Pages, Publisher, Binding, List price, Publication date, Description, Publisher comments, Table of contents
RICS	ID, Title, Subtitle, Edition number, Edition text, Volume, Series title, Stock code, ISBN, Publication date, Binding, Extent, Member price, Non-member price, Imprint, Authors, Editors, Blurb, Subject code
The Stationery Office	ISBN, Title, Publication status, Publication date, Publisher price, Physical description, Series title, Series number, Corporate name, Note, Index note, Subject heading
Amazon format	ISBN, Title, Author, Photographer, Illustrator, Editor, Translator, Subject, Edition, Volume, Pages, Discount, Publisher, Binding, List price, Publication date, Distributor, Description, Excerpt, Table of contents, Review, Back cover, Inside flap, Author biography, Author comments, Publisher comments
Book Data format	ISBN, Title, Publisher, Distributor, Price, Availability, Format, BIC classification, Description, Full description, Table of contents
UKMARC format	Record control number, Information codes, Amendment message, ISBN, ISSN, Source, Library of Congress classification, Dewey, Personal name, Title, Second level title, Edition, Publication, Distribution, Physical description, Terms of availability, UK price, Non-UK prices, Product information, Series, Contents note, Detailed summary note, Summary note, Change of control number, Audience note, Numbers on the item, Original language, Awards note, Personal name subject heading, Other subject headings, Personal name added entry, Corporate name added entry, Conference name, Promotional information

2.1.2 Publisher transfer formats

Equally surprising was the discovery that transfer formats were also very varied. Instead of being able to rely upon what was believed to be the ubiquitous comma-delimited format for data transfer, the majority of publishers only offered bespoke data transfer formats as supported by their bibliographic information system. This included SGML, tagged fields, (i.e., where every field is prefixed with a unique tag code), fixed format files (i.e., the position on a line denotes the field being represented), XML, as well as comma-delimited files. As described in Section 4, this has required the development of bespoke parsers for a large proportion of the publishers.

2.2 Classification systems

A large number of the specialist publishers have developed bespoke classification systems to match their requirements in their specialist domains. When analysing the data received from the publishers this has highlighted a plethora of classification systems in use. It is quite clear that there is no single classification system which can be used as an intermediary for all of the publishers, e.g., some have hundreds of codes for steel elements, while others have only a few. A decision was taken early in the project that it would be impossible to try and map classification systems and codes. It was decided to utilise UNICLASS (NBS 1996) as a core classification system for the UK TIC system, as with its table-based approach conforming to ISO/TR 14177 (1994) and its UK-specific nature, it provided the greatest range of classification views of any currently published classification system. It is envisaged that other national TIC systems would incorporate their nation's favored classification system.

However, to map between classification codes of different publishers a very pragmatic solution was devised. This solution takes the descriptor for the classification code in one system and identifies the closest matching descriptor in the second system, based on words contained in both. This provides a set of possible codes with differing ranks, based on the closeness of match, which can be used to query for publications in the second system. In practice this has proved remarkably robust, though there are many obvious examples where this process does not work correctly. In the TIC we have also attempted to include classification descriptors as well as the codes in the metadata which is indexed. For advanced searches this provides a ranking mechanism which will promote publications where the search terms are in the classification for the item.

2.3 Standard representations

The selection of a standard representation for publications in this system is a trade-off between successful commercial models and emerging standards. Commercially, the major formats used are BookData, where the majority of ISBN publications are recorded, or Amazon, which has the most successful model for Internet book sales (Amazon also takes a major feed from BookData). However, Internet standards are quickly evolving for various domains and with the advent of XML there are many data definitions being promoted. The one which has been selected by W3C for electronic resources is the Dublin Core (URL-2 1999). Its described intent is:

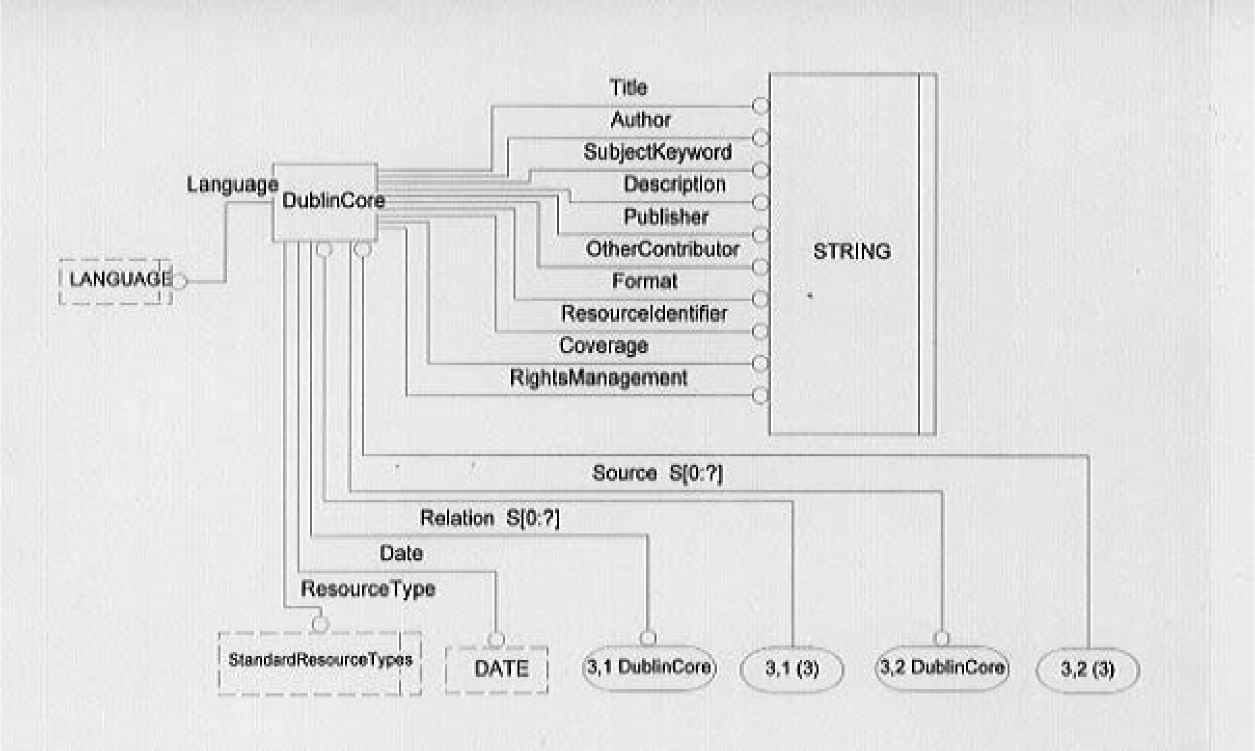
> *"The Dublin Core is a metadata element set intended to facilitate discovery of electronic resources. Originally conceived for author-generated description of Web resources, it has attracted the attention of formal resource description communities such as museums, libraries, government agencies, and commercial organizations. The Dublin Core Workshop Series has gathered experts from the library world, the*

networking and digital library research communities, and a variety of content specialties in a series of invitational workshops…." (URL-2 1999)

The major characteristics that were aimed at for this model (as taken from URL-2 1999) were:

- **Simplicity:** The Dublin Core is intended to be usable by non-catalogers as well as resource description specialists. Most of the elements have a commonly understood semantics of roughly the complexity of a library catalog card.
- **Semantic Interoperability:** In the Internet Commons, disparate description models interfere with the ability to search across discipline boundaries. Promoting a commonly understood set of descriptors that helps to unify other data content standards increases the possibility of semantic interoperability across disciplines.
- **International Consensus:** Recognition of the international scope of resource discovery on the Web is critical to the development of effective discovery infrastructure. The Dublin Core benefits from active participation and promotion in some 20 countries in North America, Europe, Australia, and Asia.
- **Extensibility:** The Dublin Core provides an economical alternative to more elaborate description models such as the full MARC cataloging of the library world. Additionally, it includes sufficient flexibility and extensibility to encode the structure and more elaborate semantics inherent in richer description standards
- **Metadata Modularity on the Web:** The diversity of metadata needs on the Web requires an infrastructure that supports the coexistence of complementary, independently maintained metadata packages. The World Wide Web Consortium (W3C) has begun implementing an architecture for metadata for the Web. The Resource Description Framework, or RDF, is designed to support the many different metadata needs of vendors and information providers. Representatives of the Dublin Core effort are actively involved in the development of this architecture, bringing the digital library perspective to bear on this important component of the Web infrastructure.

Figure 1. EXPRESS-G diagram of Dublin Core



These characteristics seemed to provide for the requirements of the TIC. The TIC requires a model which is not difficult for users to understand and for the smaller publishers to work with when supplying information. The model needed to be close to common representations used by publishers' bibliographic systems, or other commercial standards, and it needed to be a recognised standard to future-proof the developed system. The data model implemented for Dublin Core in the TIC is shown in Figure 1.

3 MAPPING AND MAINTENANCE OF INFORMATION

3.1 Field-based mapping

All of the data formats used by publishers (contacted to date) have a structure that is equivalent to that of the Dublin Core, or more complex. This simplified the major mapping process that needed to be supported to one of concatenating the detailed fields from publishers into a single field in Dublin Core. For example, title and subtitle from The Stationery Office are placed in the Dublin Core title field, or author title, surname, and forename from Pearson Education are placed in the Dublin Core author field. Moving from a more complex data structure to a simpler structure also means that the data being concatenated can be formatted in a form which is required by Dublin Core (see syntax mapping issues in Section 3.2) and is easy for the user to read when viewing the metadata.

3.2 Syntax mapping

Alongside the mapping between structures there is an issue regarding how to map between different syntactic forms for fields. For example, author fields are often structured quite differently ranging from surname and initials to having separate sections for title, first name, middle initials, and then surname. While it would be possible to write a parser for every varying field structure and then attempt the mapping to a common form this was determined to be too expensive for the benefit achieved. In the TIC the syntax is left unchanged between the publisher's format and the Dublin Core central representation. As all searches are for words within a field this seems to have very little impact on the operation of the system.

3.3 Publisher maintenance issues

To maintain the integrity of the system the data from publishers must be kept as up-to-date as possible. There are currently two methods employed to ensure that data is kept current. For publishers with large and frequently changing catalogues (e.g., BSI) a monthly importation of their catalogue is offered with arrangements to gain a tape or ftp'd file of their data. This data is then parsed and mapped into the Dublin Core format and replaces their current records in the TIC. For other publishers, or to effect immediate changes, a secure Internet interface is offered which allows publishers to manage their whole catalogue online. Through this interface the publisher can add or remove publications, change their price and availability status. The publisher can also upload cover images, etc to be associated with the publication information.

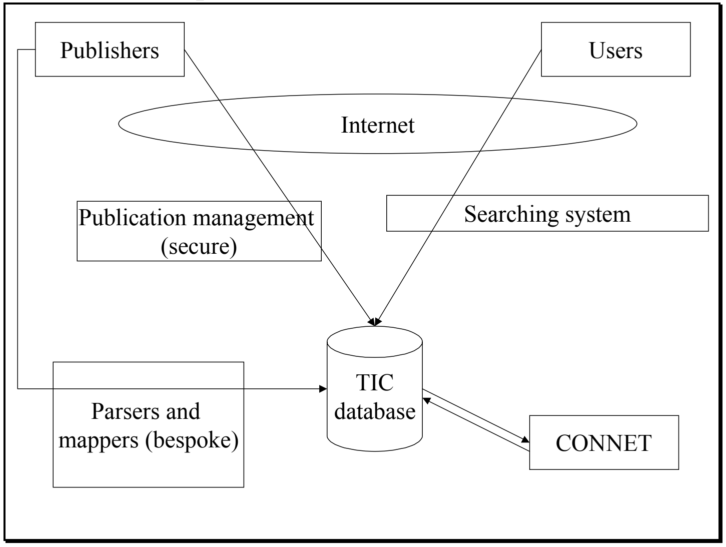
4 THE TECHNICAL INFORMATION CENTRE

4.1 Structure of the system

The Technical Information Centre is based around a central database that contains details of all the publications and publishers. The database 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). This database is implemented using Microsoft SQL Server version 7 which allows full-text indexing to be used on the fields used for user queries and can also give a score to each result based on weightings and the number of times search terms appear in the record, this score can be 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 abstract).

Users interact with the system via a forms interface that allows them to see which publishers have publications in the system and to search the database for publications. From the search results page a user can: a) go directly to the publisher's web site, for more information on the selected title, 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 Technical Information Centre or another CONNET service, e.g. the international news service or the software centre (see Sections 4.3 and 4.4).

Figure 2. Framework of the TIC implementation


Publishers
Users
Internet
Publication management (secure)
Searching system
TIC database
Parsers and mappers (bespoke)
CONNET

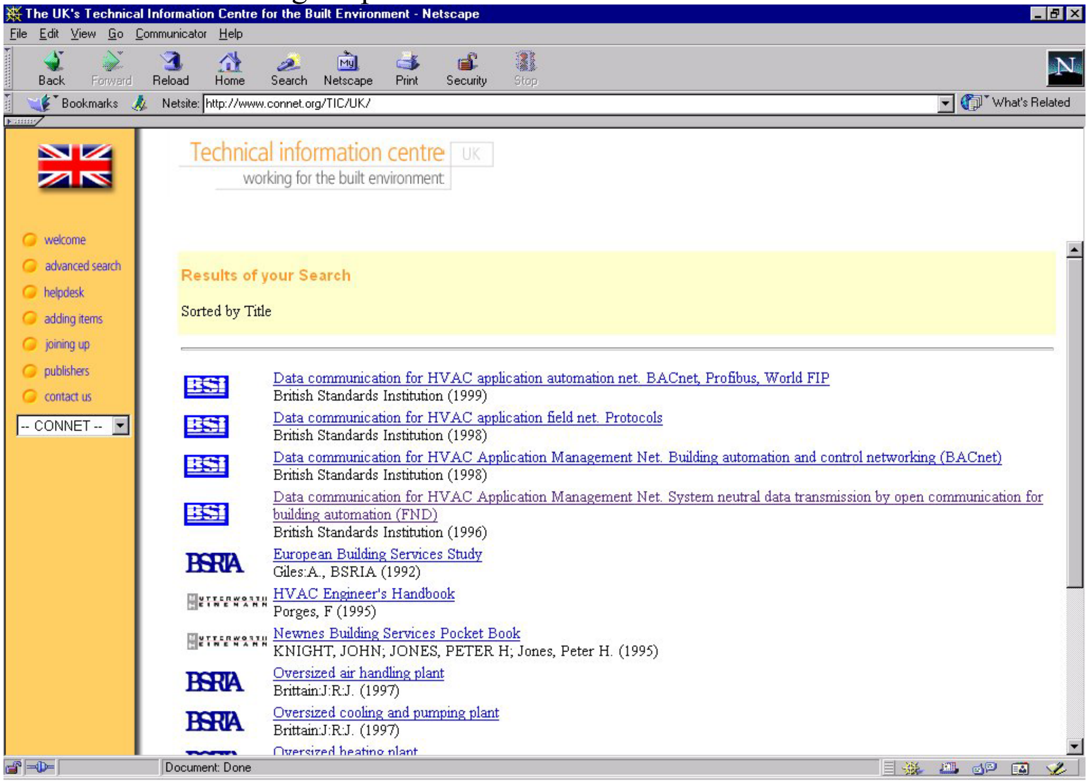
4.2 Searching for information

The TIC provides users with two options for searching – a quick search and an advanced search.

The quick search provides the user with a single text box to enter a query, this search term is then used to search all of the searchable fields in the database. For this quick search all of the options available in the advanced search are set to their default settings i.e. return a maximum of 10 results per page, the thesaurus is switched off and the results are ordered by rank score (see Figure 3 for the output of a search).

The advanced search form allows the user to search individual fields (title, author, publication date), or all fields (as in the quick search), and additionally, the publisher can be selected from a drop down list. There is also the option to supply two additional criteria against the minor Dublin Core fields comprising keywords, abstract, other contributor, format, resource identifier (usually ISBN), coverage, rights management and resource type. As well as specifying the fields to be searched the user can also choose to include synonyms for their search terms (see Figure 4). If they select this option their search terms are passed to WordNET (URL-6 1999) which returns synonyms which are included in the query. WordNET was chosen as one of the only thesaurus systems available electronically, however, it is not construction specific. The only online construction thesaurus identified to date is a Canadian system (URL-7 1999) whose linkage over the Internet was deemed too slow for the TIC. The advanced search form also allows users to specify the maximum number of results that they want returned to a page (10, 20, 50, 100 or all) and how they want the results sorted (i.e., ranking, title, author, date published or by publisher).

Figure 3. Results of searching for publications


The UK's Technical Information Centre for the Built Environment - Netscape
File Edit View Go Communicator Help
Back
Forward
Reload
Home
Search
Netscape
Print
Security
Stop
Bookmarks
Netsite: http://www.connet.org/TIC/UK/
What's Related
Technical information centre UK
working for the built environment
welcome
advanced search
helpdesk
adding items
joining up
publishers
contact us
-- CONNET --
Document: Done

Results of your Search

Sorted by Title

Data communication for HVAC application automation net. BACnet, Profibus, World FIP
British Standards Institution (1999)

Data communication for HVAC application field net. Protocols
British Standards Institution (1998)

Data communication for HVAC Application Management Net. Building automation and control networking (BACnet)
British Standards Institution (1998)

Data communication for HVAC Application Management Net. System neutral data transmission by open communication for building automation (FND)
British Standards Institution (1996)

European Building Services Study
Giles:A., BSRIA (1992)

HVAC Engineer's Handbook
Porges, F (1995)

Newnes Building Services Pocket Book
KNIGHT, JOHN; JONES, PETER H; Jones, Peter H. (1995)

Oversized air handling plant
Brittain:J:R:J. (1997)

Oversized cooling and pumping plant
Brittain:J:R:J. (1997)

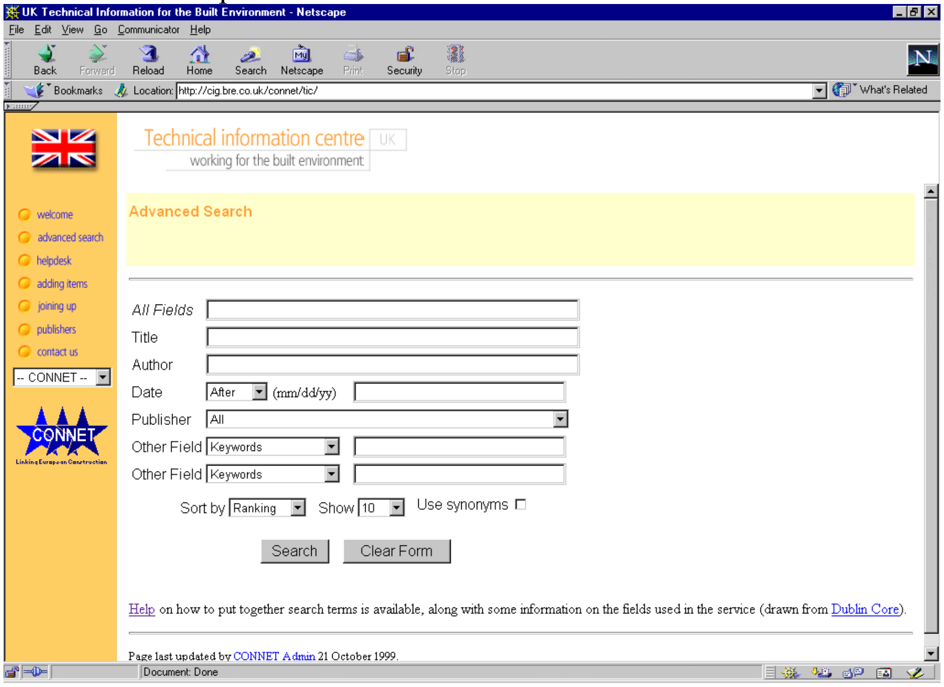
Oversized heating plant

4.3 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 which 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, news, products, etc). It also means that any other CONNET service can offer the ability to identify related publications. For example, from the news service all related publications can be identified after a news 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 publications returns a poor result in the UK system then a European wide search is likely to identify many relevant publications from Finland's technical information centre (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 TIC can run as a stand-alone service, but it can also be closely integrated with any other portal service. This is very similar to the model offered by Amazon where many sites allow a similar search to be enacted for books (e.g., from AltaVista searches through to Amazon books on the same subject).

Figure 4. Advanced search options in the TIC



UK Technical Information for the Built Environment - Netscape

Location: http://cig.bre.co.uk/connet/tic/

Technical information centre UK

working for the built environment

Advanced Search

All Fields

Title

Author

Date After (mm/dd/yy)

Publisher All

Other Field Keywords

Other Field Keywords

Sort by Ranking Show 10 Use synonyms □

Search Clear Form

Help on how to put together search terms is available, along with some information on the fields used in the service (drawn from Dublin Core).

Page last updated by CONNET Admin 21 October 1999.

Document: Done

4.4 Notification services for users

The CONNET infrastructure also offers support for active notification of users. In the case of the TIC 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 publications 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.

5 FUTURE DEVELOPMENTS

5.1 Commercial viability

The EC's requirement in the CONNET project was for the development of services which were viable after the end of the project. To this extent various business models for the TIC were examined and a decision made to further develop the current system by overlaying an E-Commerce component. At the end of March 2000 the updated TIC will be launched with full online purchase of all UK built environment publications available from the site. The E-Commerce extension is being developed under the askBRE initiative (URL-5 2000) and will see BRE taking online credit-card transactions for publications and passing the orders through to each individual publisher to fulfil. The percentage cut that is taken for each purchase will enable the further development and maintenance of this service for the construction industry, both in the UK and across Europe.

5.2 I-SEEC continuation project

The CONNET project has won further funding from the EC in 2000 in a project called I-SEEC. This will enable the technology transfer services developed in CONNET to be established across Europe (7 countries involved to date) and for a range of further quality information services to be added in the different European nations. It will also enable a more robust infrastructure to be established including further support for classification mappings and multi-language translation.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support of the EC, under the ETTN (European Technology Transfer Networks) initiative, in the development of this service.

REFERENCES

ISO/TR 14177 (1994) Classification of information in the construction industry.

NBS (1996) Uniclass, United Classification for the Construction Industry, NBS Services, UK.

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

URL-1 (1999) http://www.connet.org/ EC funded CONNET gateway initiative.

URL-2 (1999) http://purl.org/dc/ Dublin Core Metadata initiative.

URL-3 (1999) http://www.connet.org/TIC/UK/ UK Technical Information Centre

URL-4 (1999) http://ettn.jrc.it/ European Technology Transfer Network

URL-5 (2000) http://www.askbre.co.uk/ askBRE Information Services for the UK Built Environment

URL-6 (1999) http://www.cogsci.princeton.edu/~wn/ WordNET thesaurus system

URL-7 (1999) http://www.cisti.nrc.ca/irc/thesaurus/ Canadian Thesaurus of Construction Science and Technology