Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 207-211
© Firenze University Press

Enhancing end user searching of HealthInsite

Prue Deacon
HealthInsite Editorial Team
Commonwealth Department of Health and Ageing, Australia
prue.deacon@health.gov.au

Abstract

HealthInsite is the Australian federal government's Internet gateway to quality health information. The site was an early adopter of Dublin Core and makes extensive use of metadata in its navigation structure. HealthInsite has two search options utilising the Verity search engine: a simple text search and a metadata search. A third search option is the thesaurus search which is most likely to be used by information specialists. Additional functionality is being considered to improve subject searching for end users. This paper defines the research needed as background to developing the system specifications. The need to consider the whole information retrieval process is emphasised, and a clear role for metadata specialists identified.

Keywords: *Dublin Core, end user searching, HealthInsite, metadata, search engines, subjects, subject element, thesauri*

Dublin Core metadata and search engines

The main purpose of Dublin Core metadata is to promote relevant resource discovery by enabling more precise searching. However metadata cannot be implemented in isolation; it must be considered as part of an information retrieval system. There is no value in creating metadata if there is no system with the search functionality to utilise it.

In the early days of Dublin Core, there was some expectation that the public search engines would take it up. This might have happened if all implementations had used simple DC with no qualifiers and schemes. In practice, most implementations needed some complexity to give real value. Furthermore, different implementations needed complexity in different areas. Theoretically it is possible to dumb down any DC metadata to simple DC, but this is of limited value and certainly the public search engines have not rushed in to do so.

Thus DC implementations tend to be in relatively closed systems with limited interoperability. Such closed systems are small compared with the whole web and the value of metadata may well be less obvious. A good search engine performing text searching with appropriate ranking will achieve satisfactory results for many searches. A user who moves on to a metadata search may find that it appears to be no better than a text search. The user may even be confused by all the search options offered.

I believe that metadata can add considerable value in a closed/small system but that, to exploit it, you need to go beyond the standard search engine functionality. Metadata developers need to work closely with search analysts and system developers to get the most out of metadata.

In our gateway site, HealthInsite <http://www.healthinsite.gov.au>, we feel that improvements are needed in the user search functionality, particularly for subject searching. This is likely to require some new applications work which could be costly. Before starting we need to do some research into user experiences on HealthInsite, end user behaviour in general and the benefits that could come from different search applications.

HealthInsite background

HealthInsite is the Australian federal government's Internet gateway to quality health information. The site is managed by the Commonwealth Department of Health and Ageing. HealthInsite works through information partnerships with authoritative website owners ranging from government agencies to private non-profit organisations and support groups. Partners undergo a quality assessment process and then HealthInsite links to individual resources on their sites. Currently there are 54 partners and nearly 9000 resources; 50% of the resources are consumer health information, 30% are written for a health professional/provider audience and 20% are intermediate. HealthInsite also links to international health gateways with similar aims and quality assurance. HealthInsite was launched in April 2000 with a limit-

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952107163

ed coverage of health topics. This has now been considerably expanded. The next phase is to examine portal functionality, including the provision of access to services.

Our department was an early adopter of Dublin Core, first for the departmental website and then for HealthInsite. Our decision to use DC was in accord with thinking at whole-of-government level in Australia. We have been closely involved with AGLS development (Australian Government Locator Service <http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html>). For us, the advantages of Dublin Core are: simplicity; the delineation of key elements for resource discovery and display; and international and national recognition.

Simplicity is a crucial feature. We have tried to keep as close to simple DC as possible on the grounds that "the simpler the indexing structure, the easier it is to design search applications".

Figure 1 summarises our metadata specification.

Our modus operandi is that information partners maintain metadata on their own sites, usually embedded in the HTML coding of a resource but sometimes located in a separate directory. The metadata is harvested into the HealthInsite Oracle database and harvested again at regular intervals to pick

Element .qualifier	Scheme	Data format of content	Usefulness*
DC.Creator		text	metadata group 2
DC.Publisher		text	metadata group 2; display
DC.Rights		text	partner site administration
DC.Title		text	metadata group 1; display
DC.Title.Alternative		text	metadata group 1
DC.Subject	Health Thesaurus	text terms (controlled vocabulary); semi-colon delimiter between terms	metadata group 1
DC.Description		text	metadata group 1; display
DC.Language	RFC1766 / 3066	2-3 character codes; semi-colon delimiter between codes	limit
DC.Date.Created	ISO8601	formatted date	partner site administration
DC.Date.Modified	ISO8601	formatted date	limit; display; personalisation
DC.Date.Issued	ISO8601	formatted date	partner site administration
HI.Date.Review	ISO8601	formatted date	partner site administration
HI.Date.Reviewed	ISO8601	formatted date	partner site administration
HI.Date.Healthinsite	ISO8601	formatted date	personalisation
DC.Type	HI type	text term from menu	limit
DC.Type	HI category	text terms from menu; semi-colon delimiter between terms	limit
DC.Format	IMT	text term from menu	limit
DC.Format.Extent		numeric (size in Kb)	recorded, but not yet used
DC.Identifier	URI	URL	link
AGLS.Availability		text	recorded, but not yet used
AGLS.Audience	HI age	text term from menu	limit
HI.Complexity		text term from menu	limit
HI.Status		text term from menu	HealthInsite administration

*Usefulness code:

Metadata group 1: Title, subject, description grouped together in the Healthinsite metadata (power) search.
Metadata group 2: Creator, publisher grouped together in the Healthinsite metadata (power) search.
Limit: can be used to limit a search or for ranking/sorting the search results.
Display: title, description, publisher, date.modified are the elements displayed in search results sets.
Partner site administration: Elements for partners to use, if they wish, in managing their websites.
Link: used to link from the results set to the resource on the partner's site.
Personalisation: used in managing the personalisation features of HealthInsite.

Figure 1. Summary of HealthInsite metadata specification

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952107163

up any changes. In practice, things are a little more complicated. We assist many of our partners with creating the initial metadata records and we create the subject element for most records.

Figure 2 shows the metadata record for one of our partner sites.

The HealthInsite technical platform and Verity search engine

The technical platform for HealthInsite comprises: an Oracle database for metadata; a modular Cold Fusion application for presentation (soon to be replaced by the Spectra content management system); and a Verity search engine. When we implemented Verity, we were advised that some of the ideas we had for search functionality were beyond the scope of a search engine and had to be deferred for separate development.

As a search engine, our implementation of Verity can index text and index metadata. It enables searches based on full text (simple search) or restricted to text in groups of metadata elements (metadata or power search). It allows Boolean logic, truncation, limiting by various metadata elements and ranking/sorting by various metadata elements. The current implementation does not cater for spelling mistakes and synonyms.

The subject element in HealthInsite

HealthInsite is a subject gateway and it is known that most searches will be for subjects – the subject element is the focus for the rest of this paper.

Subject indexing in HealthInsite is very tightly controlled. We use the Health Thesaurus <http://www.health.gov.au/thesaurus.htm> which is a hierarchical thesaurus based on MeSH (Medical Subject

```
<META NAME=”DC.Creator” CONTENT=”Department of Human Services (Victoria)”>
<META NAME=”DC.Creator” CONTENT=”Centre for Eye Research Australia (CERA)”>
<META NAME=”DC.Publisher” CONTENT=”Better Health Channel”>
<META NAME=”DC.Rights” CONTENT=””>
<META NAME=”DC.Title” CONTENT=”Diabetic retinopathy”>
<META NAME=”DC.Subject” SCHEME=”Health Thesaurus” CONTENT=”causes; complications; diabetes
  mellitus; diagnosis; lasers; retinal diseases; risk factors; surgery; symptoms”>
<META NAME=”DC.Description” CONTENT=”Diabetic retinopathy is an eye disease caused by complications
  of diabetes. Everyone with diabetes will develop diabetic retinopathy. Regular eye exams when first diag
  nosed with diabetes, and then at least every two years, will reduce the risk of vision loss and blindness.”>
<META NAME=”DC.Language” SCHEME=”RFC1766” CONTENT=”en”>
<META NAME=”DC.Date.Created” SCHEME=”ISO8601” CONTENT=”2000-03-08”>
<META NAME=”DC.Date.Issued” SCHEME=”ISO8601” CONTENT=”2000-03-20”>
<META NAME=”DC.Date.Modified” SCHEME=”ISO8601” CONTENT=”2001-04-12”>
<META NAME=”DC.Date.ValidTo” SCHEME=”ISO8601” CONTENT=””>
<META NAME=”DC.Date.Review” SCHEME=”ISO8601” CONTENT=”2002-04-12”>
<META NAME=”DC.Date.Reviewed” SCHEME=”ISO8601” CONTENT=”2001-04-12”>
<META NAME=”DC.Type” SCHEME=”HI type” CONTENT=”document”>
<META NAME=”DC.Type” SCHEME=”HI category” CONTENT=”resource”>
<META NAME=”DC.Format” SCHEME=”IMT” CONTENT=”text/html”>
<META NAME=”DC.Identifier” SCHEME=”URI” CONTENT=”http://www.betterhealth.vic.gov.au/bhcv2/bhcar
  ticles.nsf/pages/Diabetic_retinopathy”>
<META NAME=”AGLS.Availability” CONTENT=””>
<META NAME=”AGLS.Audience” SCHEME=”HI age” CONTENT=”adult”>
<META NAME=”HI.Complexity” CONTENT=”easy”>
<META NAME=”HI.Status” CONTENT=”registered”>
```

Note that on the Better Health Channel site, this resource has an additional subject keyword string: bleeding eye, blindness, Centre for Eye Research Australia, CERA, diabetes, diabetic eye disease, diabetic retinopathy, Diabetes mellitus, Diseases and Disorders, Endocrine Diseases, endocrine, laser treatment, loss of vision, macula, macula vision, maculopathy, proliferative retinopathy, retina, Retinal diseases, Eye Diseases, sightless, vision, vision loss.

Figure 2. Metadata from HealthInsite for a resource on the Better Health Channel, a HealthInsite information partner - HTML syntax

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952107163

Headings) <http://www.nlm.nih.gov/mesh/mesh home.html>. Indexing is as specific as possible using preferred terms from this thesaurus. In the metadata record, the subject element looks quite simple. For example, from Figure 2:

<META NAME="DC.Subject" SCHEME="Health Thesaurus" CONTENT="causes; complications; diabetes mellitus; diagnosis; lasers; retinal diseases; risk factors; surgery; symptoms">

This subject line provides some useful words for resource discovery in the metadata search option. However, there are more sophisticated search possibilities. When the subject string is pulled into HealthInsite, the subject terms are associated with their hierarchy numbers. For example, diabetes mellitus has the numbers C.018.452.297 and C.019.246. This relates it to the broader terms "metabolic diseases", "nutritional and metabolic diseases" and "endocrine diseases" in the disease schedules of the hierarchy. It also relates it to the narrower terms "insulin-dependent diabetes mellitus" and "non-insulin-dependent diabetes mellitus"

Expert searchers, with knowledge of the thesaurus hierarchy and Verity, can use the full power of the thesaurus when searching. They can use the hierarchy as well as the related term structure to perform complete, but precise, searches. For example, in the HealthInsite navigation/browse facility, which is a topic-based structure, each topic contains an expert search which is performed dynamically on the latest version of the database.

For example, the topic "Drug treatments for heart disease" has the search

(c.014.280* <IN> THESAURUS_TREE_CODE or cardiology <IN> THESAURUS_TERM_NAME) and e.002.319* <IN> THESAURUS_TREE_CODE

In this search c.014.280* picks up "heart diseases" and all its narrower terms; e.002.319* picks up "drug therapy" and all its narrower terms.

This topic query technique is a major feature of HealthInsite, enabling considerable flexibility in adjusting topics without having to adjust metadata. It was evaluated in an earlier collaborative study (Deacon, Buckley Smith & Tow, 2001). These complex searches are clearly not an option for end users.

Currently HealthInsite has a thesaurus search option which allows users to navigate up and down the hierarchy (one step at a time) or to search on preferred terms. The interface is relatively limited, not self explanatory and may confuse the user. With some terms, the user would be much better to do a simple text search.

For example, a text search on nappy rash leads to 19 documents, of which the first 5 are highly relevant and the rest might have some useful information. It would take users 3 steps to get to the thesaurus search page. There they would find that nappy rash is not a valid thesaurus term. They would then have to work out what to do next.

In contrast to HealthInsite, one of its information partners, the Better Health Channel <http://www.betterhealth.vic.gov.au>, uses a controlled keyword scheme for subjects. In the metadata record in Figure 2, the keyword string is:

"bleeding eye, blindness, Centre for Eye Research Australia, CERA, diabetes, diabetic eye disease, diabetic retinopathy, Diabetes mellitus, Diseases and Disorders, Endocrine Diseases, endocrine, laser treatment, loss of vision, macula, macula vision, maculopathy, proliferative retinopathy, retina, Retinal diseases, Eye Diseases, sightless, vision, vision loss"

This has far more handles for resource discovery by an end user than the subject element in HealthInsite. The end user probably would not notice that some types of searches in the Better Health Channel site would lack precision.

In summary, while the metadata subject framework is essential for the topic-based navigation facility on HealthInsite, it is of limited use to end users doing their own searches.

What improvements could we make?

We feel that that the current situation is unsatisfactory for end users and that subject search functionality should be improved. These are some of the options:

- Bring the full librarians' functionality into an end user framework – like the subscription versions of Medline <http://www.nlm.nih.gov/databases/databases_medline.html>, or the public version (PubMed) <http://pubmed.gov>.
- Provide automatic synonym searching and spell checking.
- Make a link to the thesaurus application and add an application to help users construct searches. (The thesaurus is a database and there is an in-house application which enables full searching, with links between the preferred terms and hierarchy).
- Create standard limits to help users with text searches that retrieve very large results sets – for example, if a user searches on diabetes they could get the option to limit their search to prevention of diabetes.
- Create standard hedges to help people broaden a search. For example, a hedge for "heart" would contain the heart anatomy terms, all the heart diseases and cardiac surgery.
- Enhance the link between user searches and the relevant HealthInsite topics.
- Offer a librarian search service.
- Do nothing – it may be that end users do not really have a problem. If users get some information that

OPEN ACCESS This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952107163

they need, then it may not really matter to them if they have not found all the relevant items or if the results set is not very precise.

Most options require applications development or purchase, some at high cost. Because of the cost, we need to be very clear what we are trying to achieve and that it has real value before writing specifications.

Research required

The research plan is to study end user subject searching behaviour (both in general and on HealthInsite), to identify where users may require assistance on HealthInsite and to describe what sort of search functionality could provide this assistance.

It is well known that most users will try a simple text search first and many will not try anything more complicated. A literature search is needed, particularly to find evidence on user reactions to advanced or metadata searching.

From HealthInsite, we have three sources of information on end user searching:

- The data files of actual searches performed. These show the type of search (simple, metadata or thesaurus), the number of times the search was performed within a particular period and whether the search was successful or not. With around 2000 visitors a day to HealthInsite, these files are very large.
- User feedback on the site – users may advise us if they have had trouble trying to find information on a particular subject.
- Feedback from focus groups on the sort of facilities users want on HealthInsite. Consumer consultation is an important mechanism within the broader HealthInsite strategic planning process. Specific queries relating to end user searching and usability testing could be incorporated in the next rounds of consultation.

The main research task is to sample user searches from the data files, try the search on all three options (simple, metadata and thesaurus) and then evaluate the success of the search (recall/precision analysis) against the difficulty of performing it.

This will lead to reviewing the unsuccessful searches (including those identified in user feedback) to see what sort of assistance could be given and at what point.

Next, close liaison is required between content managers (metadata and search specialists) and IT staff to identify possible search functionality and its usability. This would involve looking at the options suggested in the previous section above. It will be necessary to assess whether the new functionality is convenient enough for the user to be persuaded to take the extra step beyond a simple text search. Furthermore, if a simple search is satisfactory, would the user be worse off by trying the new functionality?

It will be useful to review search options on other sites, although it is not always easy to ascertain the algorithms used.

There may be implications for the metadata specification or indexing rules – it is possible that a minor change to the metadata could have considerable benefits. Also, there may be new ways to use some of the other metadata elements to enhance subject searches.

Conclusion

This paper describes the metadata used in HealthInsite and shows that the subject element currently has more value for experts than for end users. The research planning phase of a project to improve subject searching for end users is outlined. When this research is complete, we will be able to decide what is feasible within our technical budget and then prepare the specifications for new search functionality. It is clear that this sort of system enhancement needs to be cognizant of the whole information retrieval process. All players should be involved – metadata, search & navigation and IT specialists, through to end users. The metadata experts in particular have a clear role to ensure the best use of metadata as well as to be flexible in considering adaptations to metadata standards.

References

Deacon, P., Buckley Smith, J. and Tow, S., 2001. Using metadata to create navigation paths in the HealthInsite Internet gateway. *Health Information and Libraries Journal*, 18, 20-29.

Web sites and resources

AGLS (Australian Government Locator Service). http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html

Better Health Channel.
http://www.betterhealth.vic.gov.au

The Health and aged care thesaurus. 5th edition. Commonwealth Department of Health and Aged Care 2001. (short title: The health thesaurus)
http://www.health.gov.au/thesaurus.htm

HealthInsite. http://www.healthinsite.gov.au

MeSH (Medical subject headings)
http://www.nlm.nih.gov/mesh/meshhome.html

Medline http://www.nlm.nih.gov/databases/databases_medline.html

PubMed. http://pubmed.gov

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952107163