

Optimising content search in DSpace ISS, the open access repository of the Istituto Superiore di Sanità.

Poltronieri Elisabetta,* Della Seta Maurella,** Di Benedetto Corrado***

* Istituto Superiore di Sanità, Rome (Italy). Publishing Unit

** Istituto Superiore di Sanità, Rome (Italy). Documentation Service

*** Istituto Superiore di Sanità, Rome (Italy). Data Management Service

Introduction

The development of institutional collections of published outputs openly available in Internet allows scientific literature to be globally disseminated and freely accessible over the web.

These digital archives, commonly named as institutional repositories (IR) represent the so called green road to Open Access, an innovative communication model of science aiming at removing any legal and technical barrier to scientific results.

Institutional repositories are distributed under open-source software which provide interoperability among systems at no cost, and ensure the harvesting of metadata recorded to fully spread scientific knowledge.

International authoritative directories as the *Registry of Open Access Repositories* (ROAR, <http://roar.eprints.org>), and the *Directory of Open Access Repositories* (OpenDOAR, <http://www.opendoar.org>) list an increasing number of IRs worldwide and provide tools and support to improve the quality of repository infrastructure. Their search interfaces also allow to perform queries and display results on repository contents, countries, software and language.

Moreover, web indicators are used to measure the global visibility and impact of the scientific repositories. This is the methodology applied by *Ranking web of world repositories* (<http://repositories.webometrics.info/index.html>) a valuable initiative of the Cybermetrics Lab, a research group belonging to the Consejo Superior de Investigaciones Científicas (CSIC), the largest public research body in Spain.

DSpace ISS: a cross-institutional repository

The Istituto Superiore di Sanità (National Institute of Health in Italy, ISS), the leading research institute in Italy in the field of public health, runs an institutional repository, DSpace ISS (<http://dspace.iss.it/dspace/>), which mainly collects the scientific works published by the internal researchers.

The repository, managed in close collaboration by the Publishing and Data Management units of ISS, also gathers collections of ISS partner institutions and actively promotes partnership to aggregate scientific literature produced by Italian research bodies in the public health field. To this end new communities and collections are gradually being created in the repository.

Building such a digital archive, compliant with the aims of the Open Archives Initiative (<http://www.openarchives.org>), was defined as a strategic objective in 2004, within the ISS Project to create a privileged reference point for online free access biomedical information produced by Italian research bodies.

After a review of the available systems, DSpace has chosen to implement the repository that was publicly released in December 2006. DSpace is an Open Source software resulted from the collaboration between MIT (Massachusetts Institute of Technology) and HP (Hewlett-Packard) and

used for preservation and distribution of digital material (text, audio, and video). It consists of a web application written in Java and follows the 3-tier architecture platform.

The ISS manages more than 29,000 publications using an in-house-developed software, based on Microsoft technology. Periodically, a dedicated software loads data from a not exposed database to the exposed DSpace database, from Microsoft SQL server to PostgreSQL. The DSpace version currently in use is 1.5.0. New features planned for the next version (2.0 to be released soon) include, among others, support for multiple metadata schemas.

DSpace ISS is designed to provide both data and services regarding primarily research articles published by ISS researchers. For this purpose efforts are being made to maximise the potential of the ISS internal bibliographic database, which maintains a regular record of the research output of ISS researchers. All records stored in this database are periodically processed for gradual migration to the DSpace ISS platform which includes over 25.000 items.

The bulk of the collection (about 17,000 items) are represented by metadata referring to papers published by ISS researchers on commercial journals, while about 4,500 refer to scientific works published in the series issued by the Institute and are available in full-text dating from 2001 onwards. The other stored material consists of metadata relating to papers produced by partner institutions (mostly belonging to Bibliosan, the Italian network grouping the libraries of the biomedical research centres). Other communities being considered for inclusion in the repository are disciplinary oriented. The existing collections in DSpace ISS regard the subject areas of “Endocrine disrupting chemicals diet interaction” and “Rare diseases and orphan drugs”.

In order to gain full access to valuable scientific outputs, full-text articles published on open access (OA) journals are gradually being posted to DSpace ISS, according to the ISS institutional policy for open access to scientific publications

(http://dspace.iss.it/dspace/bitstream/2198/353/1/Policy_ISS_EN.pdf), and to the editorial policies established by publishers.

Search options “MeSH oriented”

The Italian Medical Subject Headings (MeSH) translation was conceived in order to join NLM's Unified Medical Language System® (UMLS®), a project for the development of computer systems able to understand the meaning of the biomedical and health-related language.

The Metathesaurus, a repository of inter-related biomedical concepts, is the core component of the UMLS. It is a very large, multilingual vocabulary, which collects over 2 million terms for almost 900,000 concepts and 12 million relations among them. MeSH translations available from International MEDLARS Centres, including the Italian one, were integrated for the first time in the 2000 Metathesaurus, and updated in the following editions.

The first idea of translating MeSH arose in 1997, when data were downloaded from NLM for a pilot trial. Since no ad-hoc financial support for this project was granted, translators were recruited among ISS staff on the basis of personal curricula. In 2004 The NLM implemented, thanks to a two-year project, the MeSH Translation Maintenance System (Fig 1), which has made possible a direct access to the database, after migration of all vocabulary data from the old Model 204 maintenance environment to an Oracle-based client-server system.

MeSH Translation Maintenance System

User ID: MDS Translation Year: 2008 - 2009 Language: Italian

Work On Record

Descriptor : Influenza aviaria [Influenza in Birds] (D005585)

Concept	Term	DEL
● Concept: Influenza aviaria [Influenza in Birds] (M0008793)		
Influenza in Birds (T658126)		X
Avian Flu (T620741)		
Avian Influenza (T572291)		
Fowl Plague (T016932)		
Influenza, Avian (T572286)		
Influenza aviaria (ita0013504)		X
Influenza dei polli (ita0023616)		
Peste aviaria (ita0023617)		

View this record in:

- French
- Arabic
- Chinese
- Croatian
- Czech
- Dutch
- Estonian
- Finnish
- French
- German
- Greek
- Japanese
- Korean
- Latvian
- Lithuanian
- Norwegian
- Polish
- Portuguese
- Russian
- Slovakian
- Slovenian
- Spanish
- Swahili
- Swedish
- Thai
- Turkish
- Vietnamese

Fig. 1 Visualization of the MeSH terms in Italian with a scroll box for selecting other languages

MeSH structure has been refined with the introduction of concepts, representing classes of synonymous terms within a descriptor class. Translators can now work on-line and compare translations in other languages (Fig 2).

Descriptor : Cellule dendritiche [Dendritic Cells] (D003713)

Concept	Term	DEL
● Concept: Cellule dendritiche [Dendritic Cells] (M0005828)		
Dendritic Cells (T010879)		X
Cellule dendritiche (ita0012780)		X
○ Concept: Interdigitating Cells (M0459387)		
Interdigitating Cells (T010880)		X
Dendritic Cells, Interdigitating (T569486)		
Interdigitating Dendritic Cells (T690668)		
○ Concept: Interstitial Dendritic Cells (M0506227)		
Interstitial Dendritic Cells (T690680)		X
Dendritic Cells, Interstitial (T690681)		
○ Concept: Plasmacytoid Dendritic Cells (M0506155)		
Plasmacytoid Dendritic Cells (T690532)		X
Dendritic Cells, Plasmacytoid (T690533)		
○ Concept: Veiled Cells (M0459386)		
Veiled Cells (T010881)		X

Fig. 2 Record showing a group of related concepts within a certain descriptor class

Content retrieval through vocabulary control has been a primary goal to be achieved by the DSpace ISS team. The last implemented feature of the “Advanced search” function is a MeSH option included among the default search options. That will allow users to perform queries by inserting MeSH terms in

Italian or English. The “Advanced search” page in DSpace allows users to specify the fields they wish to search, and to combine their searches with the Boolean operators "and", "or" or "not". In case of a search query with the MeSH term “influenza in birds” the system will point to all records including this term among the keywords indexing the document. The term is marked by the tag *dc.subject.mesh* (dc stands for Dublin core metadata set) in the full item record (Fig 3).

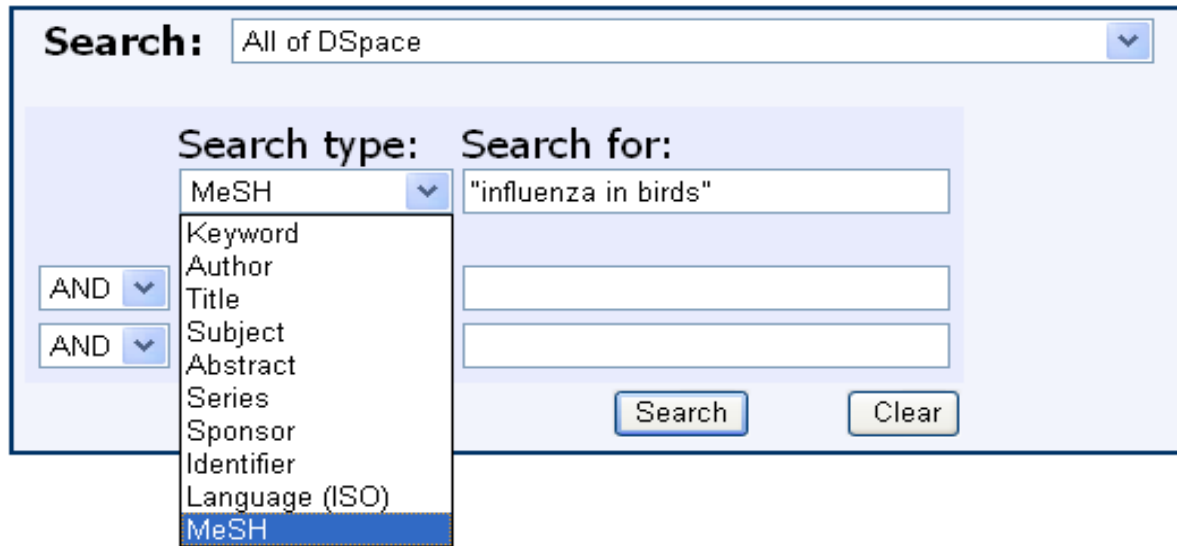


Fig. 3 Advanced search mask in DSpace ISS repository

Conclusions

Retrieving specific content through institutional repositories is being considered a desirable achievement to identify relevant information to the benefit of users. In this regard, controlled vocabulary has always proved to be a recommended tool to provide refined results in online searches. The adoption of MeSH terminology in repository systems based on open-source software, like DSpace ISS, represents a strategic choice to fully spread the use of terminological standard tools in the biomedical field. In the experience gained so far by the Istituto Superiore di Sanità considerable time and effort currently go into the technical development and customisation of DSpace ISS platform in order to enhance the search options of the system.

References

Bozec G. Community efforts help repository development. *Research Information*. 2006; Available at: <http://www.researchinformation.info/riaugsep06conference.html>. Last visited: 13th February 2009.

Ceccarini A, Della Seta M. The Italian translation of NLM MeSH: a collaboration between NLM and ISS. *Newsletter to European health librarians*. 2004;68(Aug):40-44.

De Castro P, Di Benedetto C, Poltronieri E, Roazzi P. The open access policy of the Italian National Institute of Health: steps forward to innovative publishing habits. *Journal of the European Association for Health Information and Libraries*. 2008;4(4):11-14.

Della Seta M. Un thesaurus bilingue per la biomedicina. La traduzione italiana dei Medical Subject Headings (MeSH). *Biblioteche oggi*. 2006;24(9):37-42.

Isa C, Mingolla A, Berretta C, Fruttini L, Ciappelloni R, Mari C, De Castro P, Poltronieri E, Roazzi P, Di Benedetto C. Integrazione dei sistemi informativi per l'accesso alle pubblicazioni scientifiche: il caso DSpace ISS e l'Istituto zooprofilattico sperimentale dell'Umbria e delle Marche. *ISTISAN Congressi*. 2008. 08(C12):109-110.