Abstract

The Internet Search Environment Number (ISEN) is a proposed new Internet standard analogous to the International Standard Book Number (ISBN) and the International Standard Serial Number (ISSN) that would allow Internet searchers to locate more authoritative and reliable information sources than can be identified using conventional search engines. ISEN is also a metadata creation environment and a search engine. For end users, ISEN offers metadata-based basic and advanced search options, as well as features and functionalities that support database providers and ISEN editors and catalogers. In addition, ISEN is an environment in which external associations, businesses, and organizations can develop value-added services that build on the rich metadata provided by the ISEN service. In general, databases are dynamically driven and routinely updated, producing many versions of the initial database content on a regular basis. ISEN is intended to facilitate the identification of a variety of ‘search environments’. A ‘search environment’ can be defined as a body of structured information with its own search algorithm. These search environments include not only databases, but digital libraries, institutional repositories, knowledge bases, online public-access catalogs (OPACs) and peer-to-peer (P2P) systems. At the core of the ISEN system is the Internet Search Environment Number (ISEN), which is a unique number based on the underlying technologies of Digital Object Identifiers (DOI), handles from the Handle System, and Universal Unique Identifiers (UUIDs). ISEN will be supported by metadata professionals who will provide the metadata for identified search environments.

Keywords: Internet; search environments; indexing; databases; meta searching; data mining

1.0 Introduction

The Internet Search Environment Number (ISEN) is a proposed new library standard that is a logical, though not directly related, development from the International Standard Book Number (ISBN), International Standard Serial Number (ISSN), and Domain Name Service (DNS). Whereas a DNS entry provides a name to a number, and ISBNs and ISSNs are associated with ‘names’, e.g., book and journal titles, ISEN corresponds to a Uniform Resource Locator (URL) and a database IP subnet address, i.e., a web server, or a database server, with a comprehensive set of metadata associated with an Internet-based resource.

To improve access to Internet-based databases, it is essential that their locations are known. The ISEN is for search environments.
Databases are dynamic and constantly updated. But they are still the same resource, a single instance of a changing, growing set of information. ISEN is intended to facilitate identification of a variety of search environments. A ‘search environment’ may be defined as ‘a body of structured information with its own search algorithm’. A ‘search algorithm’ can be defined as a matching operator used by an information retrieval system. An information retrieval system can be defined as one that “... does not inform (i.e., change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request.” These search environments include not only databases, but digital libraries, institutional repositories, or “… digital collections that capture and preserve the intellectual output of universities. ..”, knowledge bases, as well as online public access catalogs (OPACs), and peer-to-peer (P2P) systems.

Offline and online, “the” search environment is the concert of users connecting with the body of knowledge, the physical environment and people.

2.0 Description of the ISEN System

The ISEN system supports several activities for several classes of administrative users, as well as general searchers on the Internet. A database provider will have a web interface through which to create a draft metadata record for the database’s inclusion in the ISEN system. An ISEN editor, or classifier, will have an ISEN interface that supports review and modification of the ISEN metadata and classification of the database using the Library of Congress classification scheme, as well as a popular scheme developed by ISEN.

The ISEN system will allow searchers on the Internet to find more reliable, authoritative information than can be found with conventional search engines. ISEN will support quick and advanced search through an uncluttered search interface. A searcher will enter a quick search into a text box. Alternatively, the searcher will click on a link to an advanced search option, where the searcher can fill in a template for specifying a search more precisely.

In addition to these primary uses, ISEN also will provide an environment in which other companies will be able to provide value-added services based on the rich foundation of metadata provided by ISEN. These secondary uses include metasearch and data mining.

3.0 Functionality of the ISEN System

Section 3.1 describes how an ISEN number may be generated based on the underlying technologies of Digital Object Identifiers (DOI), handles from the Handle System, and Universal Unique Identifiers (UUIDs). The ISEN system can be seen as analogous to the domain name registry system (DNS), except that an ISEN name is resolved to a metadata record, rather than to an IP address. Section 3.3 describes these metadata records. The Dublin Core metadata fields are extended to include additional fields used by ISEN. Really Simple Syndication (RSS) is used to maintain accurate, identical copies of tri-located metadata records. The three copies are stored: 1) in the ISEN central database, 2) in ISEN database servers, and 3) on the web pages of the databases being indexed in ISEN. Section 4 describes ISEN’s classification environment, while Section 5 describes the search environment. Section 6 describes ISEN’s secondary role as an environment that provides rich metadata to support other services, e.g., metasearch, data mining.
3.1 Generating an ISEN

An ISEN will be a type of Digital Object Identifier, or DOI. A DOI is used to provide a persistent, unique identification number for digital content. The DOIs used by the ISEN system will consist of two components:

1. A publisher prefix that uniquely identifies a publisher, as is done with ISBNs.
2. A persistent, unique identifier generated by a process such as a UUID generator.

![Figure 1: ISEN number components](image)

ISEN plans to register databases as DOIs through an arrangement as, or via, a member of the International DOI Foundation (now under discussion and subject to mutual agreement).

ISEN is also a network monitoring tool that is used to ensure authentication and assurance. Building on existing Handle System tools, ISEN will develop a capability to confirm the availability of, URLs of an indexed database in order to ensure uptime.

3.2 Dynamic Tri-location of Metadata Records

The ISEN central database is a relational database. As ISEN registration increases, a distributed network of ISEN servers will operate in much the same fashion as the Domain Name Server system operates today, except that ISEN numbers, rather than resolving to IP addresses, will resolve to ISEN metadata records.

ISEN metadata records will be based on the fields provided by the Dublin Core, but extended with additional fields required for efficient, effective ISEN processing. Dublin Core is widely used throughout the information industry to represent metadata. The additional ISEN fields include:

- The provision for multiple classifications.
- A field showing the size of the database, both in terms of the number of records that it includes and in terms of bits.
- A field indicating which software is used in the database, e.g., Oracle.
- A field indicating the public, private, or non-profit status of the organization providing the database.
- A field giving information about the underlying geospatial or geographic information system supporting the database, when appropriate.

In addition to these Extended Dublin Core fields, ISEN will make use of the DOI data model, using its contextual ontology data dictionary and its capabilities for interoperable mappings.

One of the most important features of ISEN is tri-location of ISEN metadata records using RSS. The master copy of the ISEN record resides in the ISEN central database. Changes to this master copy, made by an ISEN editor, automatically propagate to the two other copies of the record residing on the ISEN database server(s) and on the web page of the database. If a change is made to either of the copies other than the master copy, the change will not
propagate. Instead, the ISEN editor will be automatically sent a message about the changed copy and will take any needed action.

4.0 ISEN from the Cataloger’s Perspective

Government, higher education, commercial and non-profit organizations will be solicited to submit their metadata and to obtain an ISEN license. The one-year license will include ongoing maintenance or changes of information in the record (for example, publisher changes, subjects added). Databases represented in the ISEN system will be classified according to both the Library of Congress (LC) classification scheme and a classification scheme developed specifically for ISEN. The ISEN-specific classification scheme will be designed by a professional taxonomist, but, unlike the Library of Congress scheme, will be targeted toward general users of the Internet. Vocabularies or term-tagging submitted by search environment owners will be generally considered valid unless they misrepresent the data in the form of metadata spam. Terms that search-environment owners ascribe to their data will serve as guides to representing the metadata in LC. In other words, the vocabularies of the submitters will be translated by ISEN catalogers into the official LC terminology. The original vocabularies submitted will be maintained in order to develop a folksonomy to accompany LC term navigation. So there are three sets of metadata: the LC subjects, the popular metadata and the “spam” metadata. Information and reporting can be derived from mining these databases.

Both sets of metadata, LC and popular, that will be revealed to the user will be presented to the searcher so that both popular and official taxonomy navigation methods will be offered as choices in directory navigation. The popular and official taxonomies will be mapped to each other and displayed to users to inform them about the difference between popular terminology (folksonomies) and official classification terminology. Showing users a shortcut to understanding how they describe what they are looking for in terminology that is structured will provide a translated link between folksonomies and taxonomies.

A database provider wishing to submit a database for inclusion in the ISEN system will type in the URL of the database at the ISEN database submission interface. A DC-dot like process will automatically generate a metadata record from the web page specified by the URL. The database provider will then review the metadata record and make corrections as needed. The record will then be submitted for further processing by ISEN catalogers.

ISEN will support both automated metadata generation and manual cataloging. Once a metadata record has been entered initially into ISEN by a database provider, an ISEN cataloger will review the metadata record and make any needed corrections. ISEN catalogers will also classify the database according to the Library of Congress classification scheme as well as according to the ISEN classification scheme. The ISEN cataloger will contact the database provider to obtain the database provider’s approval of the final version of the ISEN metadata record.

5.0 Searching with ISEN

The processes described in Section 4 lead to the creation of ISEN metadata records, which in turn support search by an end user, and new-resource alerting. The ISEN system supports three modes of search: standard, agent, and federated. An important feature of the ISEN system that adds to the convenience of all users of the system, both administrative and searchers, is a feature we call “blackdown”. If databases are cataloged using ISEN, we can tell which resources in the search environment are up or down. Then we can show the searcher that each resource is current and available. If for some reason a particular search environment is unavailable, this will be indicated in the Search Engine Results Page (SERP) by a black hyperlinked text to the URL instead of blue. The “blackdown” concept is based on the DOI because the DOI is essentially an unbreakable URL. It eliminates “404 Page not found” errors.
5.1 Standard Search

Standard search includes both quick and advanced search. With the quick search interface, a user can enter one or more words, or search terms, into a text box on the interface screen. If one or more of these terms matches terms in certain fields of the ISEN metadata record, the metadata record will be retrieved from the ISEN central database and displayed to the user. The fields from the metadata record that are searched during a quick search are: title, description, and keyword.

Alternatively, when the user views the search interface screen, the user may choose to perform an advanced search by clicking on the words “Advanced Search”. If the user does this, the advanced-search screen will be displayed. The advanced-search screen is a form in which the user can enter terms corresponding to several ISEN metadata fields: keywords, DB title, description, language, URL, person (admin, creator), and ontology term.

5.2 Agent Search

Agent-based searching is provided by software that takes a profile generated from an end user’s expression of interest and searches on behalf of the user. For example, consider a user who is interested in Indiana history. ISEN may include general-history databases, but nothing specifically related to Indiana history. The user’s profile can consist not only of keywords for general history, already represented in ISEN metadata records, but also keywords, called “wishwords”, that are not yet in ISEN metadata, that refer specifically to Indiana history. Then, if a new ISEN metadata record contains one or more of these Indiana history-specific keywords, the user will be alerted.

5.3 Federated Search

Federated search is similar to agent search in that a user provides a profile that is turned over to a software agent to run on a periodic basis. Federated search, though, provides that the query be run against a pre-determined federation of databases. Some of these may be for-fee databases. The mechanics of accessing these databases would be handled by ISEN behind the scenes, so that the user need not be concerned with these details once the federated search has been initially established.

5.4 Alerting/New-Resource Notification System

The ISEN system will provide an alerting, or new-resource, notification system for end users. This service will be based on the agent and federated-search functionality. User profiles can be generated by capturing the expression of interest of the queries as well as monitoring viewed results that the user chooses. The users may choose to have their queries and results monitored. This data can be used to create an avatar or agent that will scour the ISEN environment for items of potential interest to the user and be delivered to the user in the form of reports. This avatar learns the user’s preferences by observing the user’s online behavior and will assist in decreasing the labor of the user in having to explicitly call forth the user’s knowledge of fields of interest. Users may develop multiple avatars for multiple purposes. Each avatar would be trained like an individualized expert system to perform these subject-specific research tasks.
6.0 ISEN as an Environment for Metasearch and Data Mining

Many web pages currently include metadata fields. The metadata included within these fields, however, cannot be considered to be reliable.

ISEN, with its trained librarians, will be a trusted third party whose metadata can be trusted not only by searchers, but by all parties, including metasearch engines, data-mining products, and even by legitimate search-engine optimizers. Thus ISEN will provide rich, accurate metadata that will lead to improved search and also will enable enhanced related services.

7.0 Discussion

As long as there are search environments, there should be an ISEN. Search environments are reservoirs of structured information. Data included in a database are more refined, as opposed to data that exists in a semi-structured static web page. Irrelevant and unorganized online data increase over time. Too much of the information that is returned through standard searching is outdated, redundant, inaccessible, or irrelevant. Database information is structured for greater search effectiveness and exists in greater amounts on the Internet. The value is in knowing where refined resources exist.

A website with a database can stand above many other sites because the creators had the motivation to invest time, money and knowledge in setting it up. Costs attributed to having a database online also denote value: software cost, database administrator salary and data-entry management costs. ISEN would provide a registration service for a nominal cost that would include the services of a cataloger. It would provide an enhanced metadata-supported search environment on which other services could build, but ISEN itself would not market keywords.

8.0 Conclusion

The Internet Search Environment Number (ISEN) is for researchers who have difficulty finding the time to locate and search relevant, evaluated online databases. ISEN is the foundation of a portal that comprehensively catalogs the Internet's databases. ISEN focuses on the niche of cataloging databases. ISEN facilitates access and adds value by creating more effective and efficient Internet search experiences.

Additionally, ISEN is the foundation of a marketplace for information owners and vendors. ISEN provides a means to expose enough information about a given dataset that researchers and other users can: a) find the dataset in the first place, and b) make a well-informed decision about whether to purchase access to the dataset itself. ISEN offers a systematic approach to online data discovery that currently does not exist.

In short, ISEN provides an auditing tool for organizations and individuals to find refined and structured resources that have been obscured by the rapid and overwhelming growth of web pages. ISEN provides a new library standard in organizing databases that is historically overdue.

---

1 http://www.isbn.org/

iii http://issn.org/

iv http://en.wikipedia.org/wiki/ISSN

v http://en.wikipedia.org/wiki/Domain_Name_System


ix http://www.jcdl.org/

x http://www.arl.org/sparc/IR/ir.html

xi http://en.wikipedia.org/wiki/Knowledge_base

xii http://en.wikipedia.org/wiki/Peer-to-peer

xiii The Digital Object Identifier System®, http://www.doi.org/. DOI and DOI. ORG are registered trademarks and the DOI logo is a trademark of the International DOI Foundation.


xv OSSP uuid Universally Unique Identifier (UUID), http://www.ossp.org/pkg/lib/uuid/.


xix Dublin Core metadata editor, http://www.ukoln.ac.uk/metadata/dcdo/.