HARVESTING AND AGGREGATION OF DIGITAL LIBRARIES USING THE OAI FRAMEWORK

Alon Kadury and Ariel J. Frank
Department of ComputerScience, Bar-Ilan University, Ramat Gan, Israel

Keywords: Digital library, Harvesting, Aggregation, SDL, FDL, HDL, OAI, OAI-PMH.

Abstract: Digital Libraries (DLs) are an important tool for quality information retrieval over the Internet. However, with the information explosion on the Internet and the increase in the number of DLs, users might need to search several DLs before finding the relevant information looked for. Federated DLs (FDLs) and Harvested DLs (HDLs) can solve this problem. To overcome the lack of uniformity and interoperability problems between DLs and to develop standards that can ease the distribution of information between them, an initiative called OAI established the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) protocol. While this protocol enables the construction of FDLs, HDLs are hardly found. This paper presents the RIDDLE model for construction and aggregation of HDLs using the OAI-PMH protocol with some modifications. The use of RIDDLE can provide for high quality information retrieval from digital libraries on the Internet.

1 INTRODUCTION

The rapid growth nowadays of available information on the Internet increases the risk of “Information Explosion” and the difficulties in retrieving relevant information rise. Ad hoc and unsound use of Web Search Engines (SEs) is akin to searching for a needle in a haystack, with similar outcomes.

On the other hand, in appropriate circumstances, wise use of relevant Digital Libraries (DLs) provides high quality information retrieval of authoritative results, since DLs are quality oriented. However, the availability, Awareness and discovery of DLs on the Internet are still lacking (Yom Tov & Frank, 2006).

Moreover, even if DLs are used, users might need to search several DLs before finding the relevant information looked for. Here we investigate some of the arisen questions, and propose a promising solution for increasing the availability and discovery of DLs using harvesting and aggregation techniques.

In section 2 we elaborate more on digital libraries, their types and use. Section 3 discusses the OAI-PMH protocol and presents the problems encountered in construction of HDLs using the OAI-PMH protocol. Section 4 introduces the proposed RIDDLE model for construction and aggregations of HDLs, and describes an initial prototype. Section 5 concludes with a discussion and future directions.

2 DLS ON THE INTERNET

Digital Libraries (Arms, 2000) are both a direct extension and complement of classical (analogical) libraries. Here we define a DL as having six major characteristics (Hanani & Frank, 2000; Sharon & Frank, 2000). DLs exist on the Internet for over a decade, and their number is growing. The advantages of DLs holding digital collections are that they provide quality, up-to-date materials, offered with rich library services.

2.1 Types of Digital Libraries

There are various classifications of digital library types in the literature. Here we classify DLs into the following three types (Sharon & Frank, 2000):

1. Stand-alone Digital Library (SDL) – a regular library implemented locally in a fully computerized fashion, with networked access.
2. Federated Digital Library (FDL) – a logical federation of entire autonomous
3. Libraries, based on common focus and topic, on the network.
4. Harvested Digital Library (HDL) – a virtual library providing mainly metadata based access to relevant items that are distributed over the network.

Kadury A. and J. Frank A. (2007).
HARVESTING AND AGGREGATION OF DIGITAL LIBRARIES USING THE OAI FRAMEWORK.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Web Interfaces and Applications*, pages 441-446
DOI: 10.5220/0001284704410446

As can be expected, SDLs are the most prevalent type of DLs. However, with the ever increasing number of SDLs, users looking for certain information may be compelled to look into several SDLs before they can fully locate relevant information on the topic they are looking for.

In order to alleviate this difficulty, several solutions have been proposed in order to generate a single entry point that would transparently provide coverage for relevant SDLs. Such solutions include:

1. Use of vertical search engines.
2. Access to a FDL library.
3. Search through a HDL library.

Use of vertical search engines can lead to finding relevant results but they also suffer from the aforementioned SE disadvantages (Xiaoming, 2001; Yom Tov & Frank, 2006).

An FDL federates several autonomous SDLs by logically constructing a flat composition of all their contents to form a unified library. Example FDLs are the Networked Computer Science Technical Reference Library NCSTRL (www.ncstrl.org) and the National Science Digital Library NSDL (nsdl.org).

The FDL provides a transparent, uniform interface to all the underlying SDLs' contents, while overcoming lack of uniformity and any interoperability problems between them. Since different SDLs are involved, FDLs tend to have relatively coarse granularities. Moreover, each additional SDL that is federated increases the granularity of the FDL even more.

An HDL, on the other hand, filters SDL resources that are relevant to its focused library topic by harvesting only their metadata into the HDL. A framework for generating HDLs is the Katsir system described in (Hanani & Frank, 2000). Katsir was based on the Harvest system that used the SOIF format for summarizing varied resources into metadata records that were kept in the constructed HDLs. Example HDLs are the National Nanotechnology Initiative (www.nano.gov/) and SourceBank (www.sourcebank.com/).

HDLs tend to have relatively fine topic granularities, which isn't increased as additional relevant resources on the same topic are harvested. Retrieving quality information on a focused topic is an advantage of HDLs over FDLs which tend to have coarser topic granularities. Moreover, as need arises, HDLs can be easily composed to form aggregated HDLs with any coarser topic granularity required. In principle, this aggregation can continue for as much as needed, forming a hierarchy of DLs represented and accessed as a relevant topics tree.

3 HDLS AND THE OAI-PMH PROTOCOL

When developers attempt to generate FDLs and HDLs they are usually confronted with lack of uniformity and interoperability problems between the different SDLs involved. Major problems are different metadata formats used by DLs, lack of consistency in the way requests are sent and responses received in the varied user interfaces and in the sharing of library resources (Suleman & Fox, 2002).

3.1 The OAI-PMH Protocol

To solve the lack of uniformity and interoperability problems and to develop standards that can ease the distribution of information between varied repositories, OAI (Open Archives Initiative) was initiated. OAI established the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) protocol (OAI-PMH, 2002). The protocol attempts to solve these problems by setting a standard for sharing information between many repositories by use of a single metadata format.

OAI-PMH standardized the way repositories, known as data providers, are queried (by defining 6 HTTP queries with different parameters) and the way the answers are received (based on pre-defined XML schemes) by the service providers. In addition, the protocol specified DC (Dublin Core) as the basic metadata standard that all repositories need to adhere to (DC, 1998).

OAI-PMH clearly supports the construction of FDLs (as service providers) based on gathering information from SDLs (as data providers). Moreover, the protocol was designed to provide selective harvesting by the use of Sets (OAI-PMH, 2002) which seems to also provide for the construction of HDLs (as service providers).

The idea of Sets is as follows. Instead of filtering metadata records one by one, it would be efficient to filter a group of related items at once. This is made possible in the protocol through the use of Sets. A set is an optional construct for grouping items for the purpose of selective harvesting. SDLs may organize items into sets. Set organization may be flat, i.e., a simple list, or hierarchical. The number of sets and their organization is up to the SDL developers. An item record need not be affiliated with a set, even if sets are defined, or it can be affiliated with one or more of them.

3.2 Problems Constructing HDLs using OAI-PMH

The use of the OAI-PMH protocol is accelerating and it is employed by many SDLs and FDLs that have implemented it (Van de Sompel & Lagoze, 2002). A few hundred OAI-based FDLs exist on the Internet; these can be found in the OAI service providers list (www.openarchives.org/service/listproviders.html). Considering the advantages of HDLs, and the availability and benefits of OAI-PMH, we would have expected an increasing number of HDLs on the Internet. However, it is interesting to note that there are nearly no HDLs that employ OAI-PMH. Even those found, turn out to make limited use of the protocol capabilities. The arisen question of course is why that is so.

The major difference between FDLs and HDLs is in the filtering and the selective harvesting done by HDLs. Consequently, this leads to investigation of the OAI-PMH protocol capabilities that can serve in these goals. The outcome of this investigation should lead to a potential solution for the harvesting and aggregation of HDLs in the OAI-PMH framework.

As part of our analysis, we found that OAI-PMH enables filtering of metadata at three levels: item, set and library, as described below.

Item-level Metadata

The item-level metadata is a summarized record of each resource of the repository. The protocol compels SDLs to expose their item-level metadata as DC format, at the least, to the service providers. The DC contains 15 elements (DC, 1998). All elements are optional, and all elements may be repeated. This DC metadata, comprised of fields with known meanings, is of course an excellent source for filtering the resources that get selected for the HDL.

Previous research (Dushay & Hillmann, 2003; Lagoze et al., 2006; Tennant, 2004) has already pointed out varied DC related problems that exist in OAI repositories DC.

Group-level Metadata

Instead of filtering metadata records one by one, it would be efficient to filter a group of related records at once. As aforesaid, this is made possible in the protocol through the use of Sets (Dushay & Hillmann, 2003; Tennant, 2004).

In an inspection regarding Sets (Kadury, 2006), a total of 164 sample OAI libraries were checked for their use of sets and the structure and name compatibility between them. However, nearly no SDLs that make use of sets were found. The investigation exposed that many SDLs do not use sets at all. Of those that used sets, the use was low and there was nearly no compatibility between the extant sets. Even libraries on similar topics defined sets in different ways using different names.

It is clear from this that attempting to use selective harvesting using OAI-PMH for constructing HDLs will currently be ineffective.

The difficulties and incompatibility in the use of sets entail mainly from the following reasons:

1. Sets are defined in the protocol in a very general manner.
2. No standard way to name sets and to describe them by use of the optional Description field.
3. No instructions on organizing sets and constructing a hierarchy between them.
4. It is not clear when items should be associated just with a single set or with several sets.

The above can explain, for example, why we have found several SDL libraries that have only one set that is just named after the library itself.

Library-level Metadata

Having metadata about the DL itself can also be useful when attempting to filter entire SDLs. Based on this library-level metadata, we could initially check if the SDL itself is at all relevant for us.

To achieve this, we could enhance the Identify query to receive also the library metadata. The returned information includes several mandatory fields about the protocol parameters used and on the identified library itself.

In addition, there is an optional Description field that can contain a textual description of the library and any additional details. Note that there is no fixed syntax for this field. However, there are a few suggestions regarding the information it should include based on tags that describe the repository (OAIGuide, 2002).

From the inspection done, there are libraries that make use of this field but usually the information contained in it is general – not very detailed and henceforth not very meaningful. This means that the Description field, as currently used, isn't useful in HDL construction.

4 THE RIDDLE MODEL

Based on this investigation, we can see that OAI-PMH has several features that could be used for construction of HDLs. However, as they are defined, they are too weak and need to be enhanced for effective filtering and harvesting of SDLs as part of the construction of HDLs.

To reach this goal, we propose here the RIDDLE (Resource Inquiry and Discovery in a DL Environment) model for the construction and aggregation of HDLs (see Figure 1).

It is noteworthy that the OAI-PMH designers envisioned an open protocol with a low entry level to describing resources by metadata records. They tried not to pre-enforce division of records into sets, for example. The idea was to have a light and flexible protocol that could be suited to a wide range of applications. Henceforth, the RIDDLE model preserves this spirit while suggesting enhancements that enable better construction and aggregation of HDLs, as described below.

4.1 Model for HDL Construction

The RIDDLE model (see Figure 1) supports the construction of HDLs from SDLs (at layer 3) and their further aggregation (at layer 4). The model is based on the following proposed OAI-PMH enhancements, introduced at the previously described 3 levels: item, set and library.

Item-level Metadata

The solution to the aforementioned item-level metadata problems is the mandated use of extended DC (DCMI, 2005) that can provide a better detailed description of the SDL resources. That is, extended DC needs to be defined, at the least, for each resource; DC itself isn't enough. Exemplary extended DC fields are abstract and audience. Thus extended DC metadata will enable better filtering at the item-level of SDLs.

Group-level Metadata

The solution to the aforementioned group-level metadata problems is to expand the protocol definition of Sets through use of a naming standard for uniform naming of the topics of the sets. The naming standard proposed to be used here is DDC (Dewey Decimal Classification) (DDC, 2003) that is used for classification of topics in classical libraries and within topic trees.

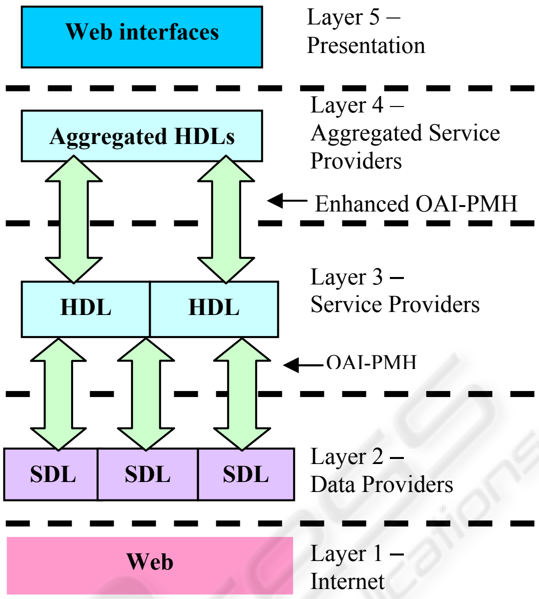

Figure 1: The RIDDLE model for construction and aggregation of HDLs on the Internet.

The DDC standard was chosen for this purpose since it is in popular use in the library system, is easy to use and provides for a detailed topics tree (DDCServices, 2006). An example DDC topics subtree is given in Figure 2. We propose to expand the Sets definition with an additional DDC field that indicates its DDC mapping. This field will enable efficient filtering at the group-level of SDL sets.

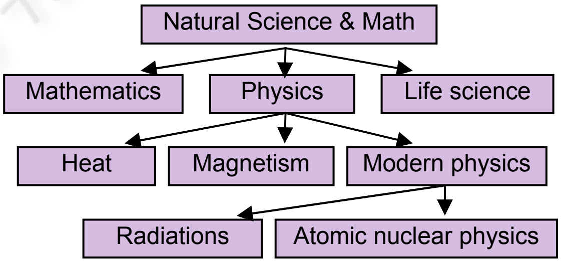

Figure 2: An example of a DDC topics subtree.

Library-level Metadata

The solution to the aforementioned library-level metadata problems is to mandate the use of extended DC also in the Description field. This will enable better filtering at the library-level itself.

It is noteworthy that with this proposed scheme, the library itself is described in the same way as the library resources are. Henceforth, we can use the item-level filtering tools also for the library-level itself. This enables the aggregation of HDLs as suggested below.

4.2 Model for HDL Aggregation

Once we have HDLs, it now becomes possible to aggregate HDLs into coarser granularity HDLs, as needed. The idea is to take relatively fine granularity HDLs and construct from them relatively coarser HDLs, in a hierarchical manner. The aggregated HDL need not construct any new libraries but just represent the logical composition of extant ones.

As envisioned, the aggregated HDL relies on a hierarchical structure that is composed based on extant hierarchies. An example of an aggregated HDL can be seen in Figure 3. The resultant structure enables hierarchical browse and search that is based on the extant topics tree of libraries.

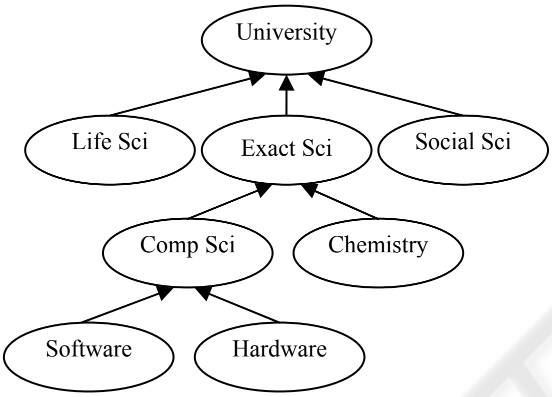

Figure 3: An example of an aggregated HDL.

To provide for this hierarchical structure we propose to use DDC, as this standard was already used for the Sets definition in HDL construction. This enables each library to designate its DDC topic and so denote its correct location in the DDC topics tree (see Figure 2).

If need be, the DDC topics tree structure can be expanded (Kadury, 2006), so this solves the problem of a library whose topic name doesn't fit an existing DDC name, or the inclusion of several libraries with the same DDC topics name.

Each HDL will have a library-level metadata record, like SDLs have. This record should be delivered then in response to an Identify query sent to an HDL. This will provide the needed library-level metadata on the libraries that need be aggregated beneath it.

If the HDLs would be enhanced with OAI-PMH interfaces (as data providers), they could also enable the harvesting of their data by the aggregated HDLs. While OAI-PMH enables offline harvesting, with few modifications (ibid) online harvesting could be made possible and beneficial in certain cases.

4.3 The RIDDLE Prototype

The initial prototype implementation realizes the functionalities of the RIDDLE model (Kadury, 2006). The prototype supports the construction of both FDLs and HDLs, and the further aggregation of HDLs. The aggregated HDLs can be searched by browsing a DDC topics tree or by free textual search, which searches the HDLs' library-level metadata.

The prototype supports several user interfaces used for displaying the search results: a common list oriented interface, and a Google-like interface that lists HDLs that answer the search criteria, instead of showing the list of sponsored links.

A suitable test collection of FDLs and HDLs was generated using the OAI-PMH protocol with needed manual cleansing of the metadata records and their augmentation with extended DC fields (ibid). Several tests where carried out on the RIDDLE prototype, in order to check the quality of information retrieval from HDLs versus using several FDLs, by checking precision, relative recall and efficiency (by F-measure which is the weighted harmonic average of precision and recall). Tests on the ease of discovering and using aggregated HDLs were also done.

Comparisons were made between tasks that were first carried out on FDLs and then on HDLs (ibid). The summarized results show that HDLs received better marks on all measures compared to the FDLs (see Figure 4).

In addition, several experiments were carried out with a group of users that tried to separately locate information using several FDLs and several HDLs. These experiments also exhibited that the users preferred HDLs over FDLs. The users preferred using a single HDL interface rather than several FDL interfaces and liked the ease of locating the needed information from the HDL results.

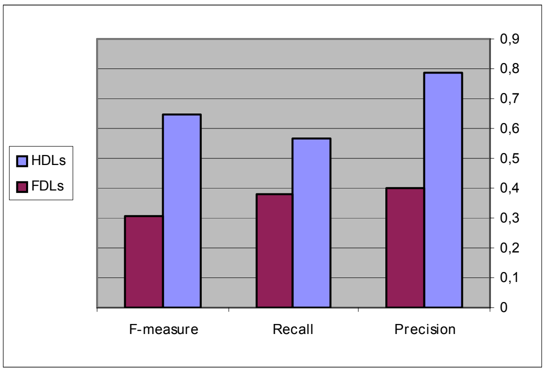

Figure 4: Comparison of precision, recall and F-measure.

5 DISCUSSION AND FUTURE DIRECTIONS

The major problems pointed out in this paper are the lack of uniformity in SDLs using OAI-PMH that prevents effective creation of HDLs, thus motivating the need for the RIDDLE framework. The RIDDLE model suggests the changes needed in the OAI-PMH protocol to enable the efficient construction of HDLs and their aggregation as needed. The major changes promote the repeat use of both the extended DC and DDC standards to provide for better metadata at all library levels that enables selective harvesting and aggregation.

Initial testing indicates that use of an HDL is more efficient when compared to the use of several separate FDLs. RIDDLE received high scores from the end users. In addition, the users found the process of locating aggregated HDLs in the DDC topics tree easy to use. In general, the users preferred RIDDLE due to its ease and effectiveness in searching and retrieving results. The initial experiments are promising but the current prototype needs to be enhanced for wider experiments and for public use.

In this paper we presented the changes required in the OAI-PMH protocol to enable the construction and aggregation of HDLs using RIDDLE. Introducing these required extensions into the OAI-PMH protocol can enable efficient construction and aggregation of HDLs, and consequently better information retrieval from digital libraries on the Internet.

ACKNOWLEDGEMENTS

We kindly thank Dr. Yaffa Aharoni and Taly Sharon for insightful contributions and support, and Nir Yom Tov and Elina Gershenovich for their active participation in the related digital libraries research.

REFERENCES

Arms, W. Y. (2000). *Digital Libraries*. MIT Press, Cambridge.

DC, Dublin Core Metadata for Resource Discovery. (1998, September). The Internet Engineering Task Force (IETF). RFC2413. Retrieved August 2006, from www.ietf.org/rfc/rfc2413.txt

DCMI, DCMI Metadata Terms. (2005, June). Dublin Core Metadata Initiative. Retrieved August 2006, from dublincore.org/documents/dcmi-terms/

DDC, Introduction to Dewey Decimal Classification. (2003). OCLC, Forest Press. Retrieved August 2006, www.oclc.org/dewey/versions/ddc22print/intro.pdf

DDCServices, Dewey services. (2006). OCLC, Cataloging and Metadata. Retrieved October 2006, from www.oclc.org/dewey/

Dushay, N., Hillmann, I. D. (2003, October). Analyzing Metadata for Effective Use and Re-use. Dublin Core Conference, Seattle.

Hanani, U., Frank, A. J. (2000, July). Katsir: A Framework for Harvesting Digital Libraries on the Web. ECIS 8th European Conference on Information Systems, Vol. 1, 306-312, Vienna, Austria.

Hanani, U., Frank, A. J. (2000, November). The Parallel Evolution of Search Engines and Digital Libraries: their Convergence to the Mega-Portal. ICDL'00, Kyoto International Conference on Digital Libraries: Research and Practice, Kyoto, Japan, 269-276.

Kadury, A. (2006). Harvesting and Aggregation of Digital Libraries in the OAI framework. Master Thesis, Department of Computer Science, Bar-Ilan University.

Lagoze, C., Krafft, D., Comwell, T., Dushay, N., Eckstrom, D., Sayloy, J. (2006, June). Metadata aggregation and "automated digital libraries": A retrospective on the NSDL experience. JCDL Joint Conference on Digital Libraries, North Carolina, USA, 230-239.

OAI-PMH, The Open Archives Initiative Protocol for Metadata Harvesting, Protocol Version 2.0. (2002). Retrieved August 2006, from www.openarchives.org/OAI/openarchivesprotocol.html

OAIGuide, Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting. (2002), Open Archives Initiative. Retrieved August 2006, from www.openarchives.org/OAI/2.0/guidelines.html

Sharon, T., Frank, A. J. (2000, August). Digital Libraries on the Internet. 66th IFLA Council and General Conference, Jerusalem, Israel, 13-18.

Suleman, H., Fox, E. (2002). The Open Archives Initiative: Realizing Simple and Effective Digital Library Interoperability. Journal of Library Administration, Vol. 35, Issue 1, 125-145.

Tennant, R. (2004, July). Metadata Bitter Harvest. Library Journal, Vol. 32. Retrieved August 2006, www.libraryjournal.com/article/ca43444322

Van de Sompel, H., Lagoze, C. (2002). Notes from the Interoperability Front: A Progress Report on the Open Archives Initiative. ECDL 6th European Conference on Digital Libraries, 144-157.

Yom Tov, N., Frank, A. J. (2006, October). Harnessing Search Engine Technologies to Increase Awareness and Discovery of Digital Libraries. 4th IEEE International Conference on Information Technology: Research and Education, Israel.

Xiaoming, L. (2001, December). DP9 Service Providers for Web Crawlers. D-Lib Magazine, Vol. 7, No. 12. Retrieved October 2005, from www.dlib.org/dlib/december01/12inbrief.html