IKHarvester – Informal eLearning with Semantic Web Harvesting

Jacek Jankowski, Adam Westerski, Sebastian Ryszard Kruk, Tadhg Nagle, Jarosław Dobrzaniński
Digital Enterprise Research Institute
NUI Galway, Ireland
<firstname.lastname>@deri.org

Abstract
Only recently, researchers and practitioners alike have begun to fully understand the potential of eLearning and have concentrated on new tools and technologies for creating, capturing and distributing knowledge. In order to support and extend those solutions we propose the idea of incorporating the informal knowledge into Learning Management Systems. Contributing to the body of research, problems of existing eLearning technologies are documented highlighting areas of definite improvement. Finally, semantic web harvesting technology as a solution is explored in the form of the knowledge acquisition tool called IKHarvester.

1 Introduction
Obtaining sustainable competitive advantage weighs heavily on the learning capability within organizations [1]. Just like many other industries, the learning and education industry has not been immune to eCommerce and Internet-driven change [2]. Even though there has been extensive research on knowledge management related to information technology [1], relatively little attention has been given to the area of eLearning [7].
eLearning has been identified as a growing market, a direct result of the increased demand for training [17]. It has been forecasted that world wide eLearning license revenue will grow at a compound rate of 15.6% each year creating a market worth over $685 million in 2009 [6]. Organizations have been investing more and more on training to respond to a growing need for new information and knowledge to facilitate organizational changes such as mergers, acquisitions, new business models, re-engineered and reinvented organizational forms [17]. Satisfying this demand, eLearning is seen as a revolutionary way to empower a workforce with the skills and knowledge it needs to turn change into an advantage [21]. Although considerable progress has been made, educators have just begun to exploit the transformational power of the Internet [7].
In addition, eLearning research is focused mainly on the formalized learning provided in highly structured Learning Management Systems (LMS) with minimal interaction or collaboration. Even though large sums of corporate revenue are put into formally educating employees each year, a lot of knowledge is gained through informal learning. Examples of informal learning include: conversations at the coffee machine or printer, assistance by more experienced employees to newcomers, and collaborative services, such as wikis, blogs, fora, and instant messengers. These are just some of the existing ways in which employees can quickly share their experience within the organization. Unfortunately, organizations are unable to harness the potential benefit of informal learning as many of the tools employed are primarily focused on formal learning. What is needed is a service, where knowledge, coming from different sources can be shared through the LMS.

1.1 Outline of the paper
This article is structured as follows. The next section describes problems of current eLearning systems concerning incorporating informal sources of knowledge. We present our solution in Section 3. In Section 4, we introduce Didaskon - the Semantic eLearning Framework; we focus on IKHarvester component for utilizing metadata from various online resources for further usage. We describe a strategy for evaluation of Didaskon and IKHarvester in Section 5. Finally, Section 6 recaps the results of the research in the field of informal learning and describes the future research.

2 Problem Statement
In this section we specify problems that this article aims to solve. We describe areas of possible improvement and difficulties of current eLearning systems concerning knowledge acquisition and incorporating informal learning.
2.1 eLearning systems

Learning Management Systems (LMS) are one of the core of e-Learning technologies. They integrate all the tools needed for storing knowledge, management, and utilization. Despite the rapid growth and numerous new solutions, contemporary LMSes still face a lot of problems. The most important and commonly addressed issues relate to human-machine interaction. This involves developing proper techniques that motivate and draw learner to interact with the learning material. We do not underestimate those issues. However, in our opinion, the contemporary LMS providers successfully take advantage of newest multimedia technologies to at least partially solve the interaction problems.

In this paper we approach the Learning Management Systems problems from a different perspective. We put more attention on the completeness and accuracy of learning material. Contemporary e-Learning systems provide content that is tailored to the needs of a generic user. The courses created by domain experts from a given e-Learning company reflect their personal perception of the topic. Since in e-Learning there is a lot less interaction between teachers and students, the feedback regarding courses is not as intense as in regular classes. This slows down or even prevents accurate updates of e-Learning material. The adjustments made might not reflect the problems learners experience. Furthermore the course content or course blocks are always based on some internal company sources. Teachers responsible for course composition tailor courses from available blocks or just use entire monolithic courses. The changes made in the actual learning material really reflect reaction or response on the proposed learning track from the users.

In traditional classes, in public or private schools, teachers follow a curriculum based on the guide lines from the supervising units. Nevertheless they adjust courses over the years based on their experience with pupils and the feedback they get during the classes. Additionally pupils broaden their knowledge through verbal collaboration during classes or outside of them. In the modern information society people often take advantage of the Internet to collaborate. They share their knowledge in an informal way on fora, blogs, wikis and other. Although the modern Learning Management Systems tend to miss this fact, we perceive all those informal knowledge sources as an invaluable data repository. In our opinion the data derived from such spaces could not only act as supplement for LMS courses but could also provide teachers with feedback about information their courses lack.

2.2 Knowledge acquisition

Although the potential of informal knowledge sources is big, the acquisition of this data and its reuse in the LMS is not an easy task. The large diversity of the Internet and information stored can cause a lot of problems.

Before any knowledge can be incorporated into LMS it has to be located. This task involves a problem of selecting a user or system that will locate the resource. Different LMS users might have different levels of credibility. If we choose to designate a LMS user who participates in a course then the input should be treated in a different way then one from a course supervisor. Furthermore, as mentioned before, human-machine interaction is an important issue in e-Learning. The tool that enables to indicate and import new knowledge into the system should be accessible in a way that will minimize the amount of additional effort.

Once a portion of knowledge is located it has to be somehow extracted and incorporated into the LMS. The informal sources have some structure but it is very loose. The blogs or wikis vary upon the engine used and even within the same engine the information structure can be customized. Additionally the content itself usually can be freely structured and contain the same elements under different names. Sometimes information sources are annotated with metadata but the standards can also differ depending on the solution. Furthermore even if the metadata format can be recognized as a particular ontology, or other data structure, those will most probably not match native LMS meta data standards and have to be properly mapped.

2.3 Existing solutions

There are few Semantic Web applications that can be considered to solve afore mentioned problems.

- Del.icio.us\(^1\) is the social bookmarking system that can be a substitute for the browser’s bookmarks.
- PingTheSemanticWeb.com\(^2\) is a service for sharing RDF documents.
- SIMILE Project\(^3\) (Semantic Interoperability of Metadata and Information in unLike Environments) empowers users to access, manage, visualize, and reuse digital assets.
- Zotero\(^4\) is an add-on for Firefox web browser. It helps with collecting, managing, and citing research material, mainly bibliographic resources.

---

1 Del.icio.us: http://delicio.us/
2 PingTheSemanticWeb.com: http://pingthesemanticweb.com/
3 SIMILE Project: http://simile.mit.edu/
4 Zotero: http://www.zotero.org/
All these tools are successful and they are good in acquiring metadata. However, they work differently and they all have some limitations. Some of those tools do not allow users to browse stored data besides viewing raw RDF documents. Some can only read embedded RDF. Some do not work with non-semantic sources, like Wikipedia. Therefore from our perspective it is difficult to exploit their advantages. These tools were just not designed for the eLearning purposes.

3 Solution

The afore mentioned knowledge acquisition problems are addressed in this section. However it has to be emphasized that knowledge location, extraction, and incorporation form a process which at some point is going to require user input. The work described in the section does not only point out the possible technical solution but it attempts to find balance between automatic processing, required user interaction, and accuracy of the entire process.

3.1 Idea description

In order to address the described problems of the contemporary LMS we propose a solution for harvesting data from informal knowledge repositories. This involves the analysis and definition of the entire process of knowledge acquisition. Our goal is to extract the useful knowledge published in an informal way throughout the internet and put it to use within the LMS in a structured way.

The process begins with the informal information discovery and search (see Fig. 1). We have decided to place this responsibility upon the user. We propose a tool integrated with the user interface. The applications captures the currently viewed resource. Human factor on the data selection stage enables an increase in the accuracy of imported knowledge. Additionally when a user decides to import a selected resource to LMS he can specify additional tags (such as subject, domain or a suggested course).

Although we focus on Social Semantic Information Sources (SSIS) as potentially the best source of information, we do not ignore unannotated web sources. On the data level SSIS are very valuable since the meta data is already structured and maintained on the source side. In this case the entire knowledge acquisition process gets down to ontology mediation. Nevertheless the large majority of informal knowledge (blogs, wikis, boards) published on the web is not annotated with Semantic Web standards or any other that would be shared for public. Until Semantic Web becomes popular it is required to take this data into account as well. In this case the only possible solution is to parse the text of web pages.

Once the data is extracted from the social web space it can be imported into the informal knowledge repository for further use. We propose this repository as a bridge between SSIS and LMS learning object repository. Such storage is needed to temporary hold imported resources until an arbiter (for instance LMS administrator) picks the relevant material that can be transferred into the learning system. Additionally an informal knowledge repository enables to mediation between the local ontology and the particular LMS ontology.

3.2 Architecture

3.2.1 User-side application - data lookup

While browsing the informal information sources any LMS user should be able to select a resource and request its addition to LMS. Apart of the resource location, the information passed to the data processor includes user information and additional (user provided) annotations. This data adds an extra meaning to resources at the data management stage. If the system is supplied with user awareness it can recognize the value of added content based on user status (for instance different credibility of an ordinary pupil and course teacher). In our solution we propose a web browser plugin to fulfill this role (see Section 4).

3.2.2 Data processor - data extraction and mediation

The data processor component function is to acquire informal information from the indicated resource location, recognize the data format and transform resource data into an internal format (in our solution a common level ontology). Depending on the resource data format the transformation process is different. If the resource has embedded RDF description and the processor can recognize the ontology, then the data can be directly mapped according to the predefined mapping. In case of non-semantic information sources the data has to be captured directly from the HTML code. In order to increase the accuracy of this process we propose to supply the processor with parsing templates; these can be delivered for different software engines of SSIS (e.g. different blog engines) and for different instances of such (for example different template for wikipedia and different for some other wiki).

3.2.3 Informal knowledge repository - temporary data storage

The data lifted to a common conceptual level by the data processor is stored in the informal knowledge repository for further management. This enables to alter and filter it with the help of human factor. Additionally data stored in one
Figure 1. Process of harvesting data from informal knowledge repositories.

common ontology can be easily exported to an LMS in the required format.

3.2.4 Administration frontend - data management

After being processed and converted as described above, the raw data selected by various users should not be imported directly into the LMS. Depending on the user base of a particular LMS the amount of data could flood the system. Additionally different types of users can provide data of different value for the LMS. For instance the course attendants are likely to provide large amounts of data of medium value; the course teacher would import data with more care and more useful for the course. Thanks to the user information and annotations possessed during the process of data lookup the domain expert or LMS administrator can filter information before transferring it to LMS.

3.2.5 Administration frontend - data export

The outcome of data management process is a set of resources that hold value to the LMS and should be imported into the system. All resources are stored in the informal knowledge repository and are encoded based on the same common level ontology. Therefore access to each of them is unified. According to the standards of eLearning the learning objects metadata in every LMS should be stored in LOM format. However, in practice, LOM is treated rather as guidelines and different LMS stores implement it in a different way. In our solution the data stored in a common ontology can be exported to different LMS thanks to the ontology mediation.

4 IKHarvester - Reference Implementation

We have described the informal learning. We shown that considerable amount of relevant information can be collected from them. We have briefly presented existing solutions for capturing metadata and pointed out their limitations. Then we introduced our approach for managing informal learning available on the Internet. In this section, we present IKHarvester (Informal Knowledge Harvester) which is the prototype implementation of our idea.

IKHarvester\(^5\) is a web service that provides two core features: harvesting Social Semantic Information Sources, and providing it for the eLearning frameworks. It benefits from the Semantic Web principles that demand rich descriptions of resource to be available online. Thus, the content of web pages is understandable not only with machines but also by machines.

4.1 Didaskon

IKHarvester has been designed as an informal knowledge repository for Didaskon\(^6\). Didaskon is a framework for automated, on-demand composition of a learning path for a student. The selection and the workflow scheduling

\(^5\)IKHarvester deployed on notitio.us service: http://notitio.us/ikh/
\(^6\)Didaskon: http://didaskon.corrib.org/DidaskonNET/
of learning objects is based on their description, semantically annotated user profile, anticipated knowledge after completing the course, and technical details of a client’s platform [9, 20].

In the other words Didaskon can create a learning path which best fits a specific learner. To achieve this, the system employs following information (preconditions):

- User’s profile that can store student’s needs, skills, learning history, technical details of his/her platform, etc. To manage such information we decided to use FOAFRealm⁷ - a distributed identity management system [13].
- Descriptions of learning objects stored in repositories registered in Didaskon. At the moment Didaskon uses LOSTRepository (Learning Object Repository)⁸ for storing formal LOs and IKHarvester deployed on notitio.us (see subsection 4.4) as a informal knowledge storage.
- Curricula prepared by experts, sets of learning objects that must/can take part in course’s building process.

**Figure 2. Didaskon context.**

Didaskon is built according to SOA guidelines; hence different components, e.g., IKHarvester, LOSTRepository, FOAFRealm can be easily integrated (see Fig. 2).

### 4.2 Data harvesting

The current version of IKHarvester captures RDF data from Social Semantic Information Sources, such as: semantic blogs, semantic wikis, and JeromeDL (the Social Semantic Digital Library) [12].

IKHarvester first attempts to discover RDF documents related to the given resource, which is indicated by a special HTML entry. An example referring to SIOC metadata:

```html
<link rel="meta" type="application/rdf+xml" title="SIOC" href="http://dobrzanski.net/index.php?sioc_type=post&sioc_id=20"/>
```

This notation informs a web browser that there is an RDF document related to currently viewed page, and it is available at the location defined with href attribute (here, [http://dobrzanski.net/index.php?sioc_type=post&sioc_id=20](http://dobrzanski.net/index.php?sioc_type=post&sioc_id=20)).

In addition to handling RDF data, IKHarvester uses Microformats which allow embedding RDF into HTML documents. Moreover, IKHarvester is capable of creating RDF descriptions from non-semantic information sources, such as a Blogger or Wikipedia⁹. For that reason, it scrapes the HTML code of an article in order to collect some data (for instance, a title, external links, see also links, references) from it.

In general, data captured from online communities, like blogs, wikis, and bulletin boards, can be described with SIOC ontology, whereas JeromeDL and MarcOnt ontologies are employed for describing bibliographic resources. The RDF document which represent an online resource is saved to the informal knowledge repository.

### 4.3 Providing data

Once the informal knowledge repository is filled with the data, it can be used by possible clients, including Learning Management System such as Didaskon. Therefore, IKHarvester exposes informal learning material in the form of learning objects [19]. Learning objects, in general, are something you can acquire, manage and use; they are reusable, modular, flexible, portable and compatible. We have followed SCORM CAM (Content Aggregation Model) instructions in defining the way of creating and managing learning objects. This standard endorses using Learning Object Metadata (LOM) standard for describing learning material.

There are nine categories of this information, each emphasizes different aspects [8]:

- **General** – general information about the LO as a whole
- **Lifecycle** – features related to the history and current state of the LO and those who have affected it during its evolution
- **Meta Metadata** – information about the metadata instance itself

⁷FOAFRealm: [http://foafrealm.org](http://foafrealm.org)

⁸LOSTRepository: [http://copernicus.deri.ie/Lost](http://copernicus.deri.ie/Lost)

• Technical – technical requirements and technical characteristics of the LO

• Educational – educational and pedagogical characteristics of the LO

• Rights – intellectual property rights and condition of use of the LO

• Relation – properties defining the relationship between the LO and other related LOs

• Annotation – comments on the educational use of the LO and information on the author of the comment and time when it was written

• Classification – describes the LO using a particular classification system

LOM standard describes learning objects very thoroughly with plenty of attributes. However, only part of them can be assigned with values taken directly from the description of the resource. Some attributes that pass educational, pedagogical and technical information for LMSs are bounded with default values, specific for each type of resources. We set those attributes to established default values based on the analysis of the type of the resource.

In Table 1, we exemplify how attributes of a post (first column) are mapped to SIOC ontology predicates (second column), and then to LOM attributes (third column). We do similar mapping of attributes for other types of resources.

4.4 notitio.us

notitio.us is a service for collaborative knowledge aggregation and sharing. It employs IKHarvester for retrieving RDF information about Web resources bookmarked by the users. Therefore, it is capable of indexing rich metadata, coming from various types of resources. In contrary to bookmarking services, such as del.icio.us, notitio.us keeps rich, semantically interconnected metadata shared by the users using Social Semantic Collaborative Filtering [11]. The resources can be shared with a bookmarking interface (SSCF). Rich metadata in notitio.us can be searched and browsed using TagsTreeMaps [15], a tags browser based on treemaps rendering algorithm, and MultiBeeBrowse [14], a collaborative browsing components. These components improve user browsing experience, utilizing metadata delivered by IKHarvester. Additionally, one of modules delivered by IKHarvester allows to expose aggregated metadata in LOM [3] standard, which turns notitio.us into a valuable source of learning objects based on informal knowledge, delivered by IKHarvester.

5 Evaluation Strategy

In the previous section we described our implementation of the idea of incorporating informal knowledge to eLearning. The question arises: Is this approach better then current eLearning solutions? To answer this question we need to look at Didaskon from different points of view and evaluate different aspects of our system, such as:

• Deriving from sources of informal knowledge - what improvement in a learning process can be achieved?

• Automatic process of creating learning objects - what impact this functionality has on costs of eLearning courses’ production? What about the quality of the created learning content?

To evaluate learning objects created from informal sources of knowledge that can be harvested by our IKHarvester, we decided to prepare an experiment. We will prepare two courses by enriching an existing one (e.g., KESP - The Knowledge Economy Skills Passport[11]) with respectively formal and informal learning objects. Then we will compare them in two ways described in the next subsections.

5.1 Evaluating Learning Effectiveness

To measure how effectively the training program accomplished its stated goals we decided to use Kirkpatrick’s four-level model for evaluating training programs [10] that has become a classic in the industry. However, even having this model, evaluating the learning process in terms of time and cost required is very expensive:

• It is hard to find people that are willing to learn material that is often needless for them.

• It is necessary to prepare good quality content that would be new for evaluation’s participants.

• According to Kirkpatrick’s model, second measurement is conducted three to six months after the training program.

Because of the aforementioned issues we are still working on evaluating learning effectiveness. Currently we are at stage of acquiring/developing learning objects and finding proper sources of informal knowledge that can be bound to the one eLearning course.

10notitio.us: http://notitio.us/

11KESP: http://www.kesp.ie/
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Predicate</th>
<th>LOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>sioc:Post</td>
<td>Educational.LearningResourceType=&quot;BlogPost&quot;</td>
<td></td>
</tr>
<tr>
<td>URI</td>
<td>Technical.Location &amp; General.Identifier.Catalog=&quot;URI&quot; &amp; General.Identifier.Entry</td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>dc:title</td>
<td>General.Identifier.Title</td>
</tr>
<tr>
<td>creation date</td>
<td>dcterms:link</td>
<td>Lifecycle.version=&quot;Date&quot;</td>
</tr>
<tr>
<td>description</td>
<td>sioc:content</td>
<td>General.Description &amp; Educational.Description &amp; Classification.Description</td>
</tr>
<tr>
<td>topic*</td>
<td>sioc:topic</td>
<td>General.Keyword &amp; Classification.Keyword</td>
</tr>
<tr>
<td>reply*</td>
<td>sioc:has_reply</td>
<td>Annotation.Entity=&quot;About author&quot; &amp; Annotation.Date=&quot;Date&quot; &amp; Annotation.Description=&quot;Content&quot;</td>
</tr>
</tbody>
</table>

### 5.2 Evaluating Learning Objects

Creating structured, engaging and interactive content that enhances learning is complex and expensive. However, as we state in this article, the Internet provides a variety of resources, including the ones that can be used in training. Automatic processes of creating learning objects from such resources will definitely lower the effort required to produce the eLearning content. But what about the quality of the created educational material?

In order to answer this question we decided to measure users’ satisfaction. We will use a Questionnaire for User Interaction Satisfaction (QUIS) developed by Shneiderman [18] and refined by Chin, Diehl and Norman [4].

### 6 Conclusions and Future Work

The water cooler effect can play a substantial part in the education of employees, students and individuals. Capturing this informal learning is a major challenge.

Within the eLearning domain, the water cooler is replaced with reference tools such as a Wikipedia, digital libraries and social tools such as blogs and bulletin boards. Current Web 2.0 and semantic tools go some way toward capturing this knowledge. However, no current tools are targeted for the eLearning domain. IKHarvester is a tool specifically designed to capture and track informal eLearning. Working in conjunction with a Learning Management System such as Didaskon, IKHarvester allows the user to manage their informal learning activity by capturing Social Semantic Information Sources and creating RDF description for non-semantic information.

IKHarvester has been implemented in two systems, the Didaskon LMS, and the notitio.us collaborative knowledge tool. Initial trials within these systems have provided positive results for capturing informal eLearning and a full usability survey is planned in near future.

Future plans include extending IKHarvester to operate on more types of online resources. We also plan to increase support for more wiki engines, such as MoinMoinWiki\(^{12}\), JSPWiki\(^{13}\) and IkeWiki\(^{14}\). Finally, we intend to support further semantic digital libraries, for instance BRICKS [5] and Fedora [16].

### 7 Acknowledgments

This material is based upon works supported by Enterprise Ireland under Grant No. ILP/05/203 and by Science Foundation Ireland Grant No. SFI/02/CE1/I131. The author would like to acknowledge Bill McDaniel, Stefan Decker, the DERI eLearning Cluster and the Corrib.org working group for fruitful discussions.

### References


\(^{12}\)MoinMoinWiki: http://moinmoin.wikiwikiweb.de/

\(^{13}\)JSPWiki: http://www.jspwiki.org/

\(^{14}\)IkeWiki: http://ikewiki.salzburgresearch.at/


