The Learning Registry: Building a Foundation for Learning Resource Analytics

Marie Bienkowski
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
+1 650 859 5485
marie.bienkowski@sri.com

John Brecht
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
+1 650 859 2325
john.brecht@sri.com

Jim Klo
SRI International
4111 Broad Street
San Luis Obispo, CA 93401
+1 805-542-9330 x121
jim.klo@sri.com

ABSTRACT
We describe our experimentation with the current implementation of a distribution system used to share descriptive and social metadata about learning resources. The Learning Registry, developed and released in a beta version in October 2011, is intended to store and forward learning-resource metadata among a distributed, de-centralized network of nodes. The Learning Registry also accepts social/attention metadata—data about users of and activity around the learning resource. The Learning Registry open-source community has proposed a schema for sharing social metadata, and has experimented with a number of organizations representing their social metadata using that schema. This paper describes the results and challenges, and the learning-resource analytics applications that will use Learning Registry data as their foundation.

Categories and Subject Descriptors
J.1 [**Administrative Data Processing**] *Education*; K.3.1 [**Computer Uses in Education**] *Computer-managed instruction (CMI), Distance learning;* H.3.5 [**Online Information Services**] *Data sharing, Web-based services*

General Terms
Performance, Design, Experimentation, Standardization.

Keywords
Learning analytics, attention metadata, social metadata, learning-resource analytics.

1. INTRODUCTION
Learning resources are available from many agencies: federal and national governments, state/provincial agencies, local school districts, post-secondary institutions, and for-profit, and not-for-profit organizations. These resources are distributed across a variety of repositories, using a variety of metadata standards and access mechanisms. Open learning resources should be made publicly available to all—and in principle they are—but in practice users must visit each organization's repository individually and contend with their various interfaces. Harvesting these resources for use elsewhere can involve complicated metadata crosswalks, and metadata often is incomplete and out of date. A searcher should be able to search across repositories for all of the learning resources available on a particular topic or find ones that have been authored by a particular person or institution. Federated search or registries that collect metadata could provide a single access point to these repositories but require a centralized authority to maintain and prove difficult to upkeep. Updates to descriptive metadata could be obtained from information published on the web by scraping websites, use of microformats on learning resource pages [1], from search terms entered into repositories, or from analysis of online curriculum that uses embedded resources.

Yet even if federated search and automated collection of metadata were achievable, searching for learning resources based on descriptive metadata alone (e.g., keywords, author, publication date) may not yield adequate results. Newer methods of locating relevant items take into account characteristics of the searcher to provide recommendations. Online commerce and social networking sites using analytics have demonstrated the value and efficiency of recommendations, but these require that the search mechanism have access to both the properties of the object of search (metadata), and the properties and actions of the searchers and their community. In education, this richer set of properties and actions is increasingly coming from user interactions with portals and repositories, and includes data such as counts of use (in the classroom or online), contexts of use (e.g., what sorts of classrooms/students/teachers), and reflections by users (e.g., ratings, descriptions). However, this rich social metadata—teachers downloading, favoriting, rating, commenting upon—is, at present, locked inside portals and cannot be shared across a broad network of education stakeholders (including researchers). The Learning Registry provides a unique opportunity to share data across these information siloes

The Learning Registry is an infrastructure that supports learning resource discovery, sharing, and amplification. Consider the case of an open educational resource (OER)[1] about robotic planet exploration (i.e., rovers) developed by a design museum. This design OER could be hosted on the museum's website in their education corner and include alignment to standards. As the OER becomes better known, links to it may appear in teacher portals, resource repositories/aggregators, or national or state-level curriculum. Related resources, for example, about an actual rover

LAK'12: 29 April – 2 May 2012, Vancouver, BC, Canada.

[1] Note that the Learning Registry is not limited to only open/free resources.

sent to Mars, may be available. This rover-design OER may also be mentioned in blogs, discussion forums, and social media sites. Data about this OER (including related OERs) is scattered ever more widely as it grows in popularity.

Learning Analytics research and development efforts will build learning environments that capture different kinds of data about this rover resource: aligned to a standard; used in an online design course; commented upon by an educator. A widely crowd-sourced dataset is necessary to conduct the types of analytics needed for educational improvement. Where does all of this disparate data get aggregated? In the absence of a solution such as the Learning Registry, one option is to centralize all activity about a resource in one site so that attention/social metadata can be captured there. Another option is to create n-way connections among metadata and social metadata sources to share data. A third option might be to release a browser plugin that acts as a sensor to help track user activity and engagements with and around learning resources (e.g., with a learning management system, or LMS). Web crawlers could scrape websites for (static) information about resources; harvesters could extract metadata from repositories; or programs could harvest attention metadata via APIs to social media sites.

Yet even with all these approaches, the "big data" still needs to go somewhere. A for-profit company could collect this data and monetize it, but the Learning Registry is working to foster the alternative: that individuals and groups will contribute data to this distributed and vast timeline-oriented dataset for use by researchers and developers to improve teaching and learning with online resources. The Learning Registry makes possible answering empirical research questions such as: how effective is a resource and for whom and what? Learning Registry data could be used to answer questions from analytics: if Learner A uses resources U, V, W; Learner B uses U, V, X; and Learner B shows more competence, then resource X may be worth noting.

2. DESIGN OF THE LEARNING REGISTRY

The Learning Registry was envisioned as a store-and-forward network based loosely on the NNTP model (e.g., see [2], [3]): a network to which providers of learning resources, metadata and social metadata can distribute information for consumption and amplification by the community. To support these providers, the design must accommodate a large volume of data expressed in a variety of metadata standards. In this section, we provide a brief overview of the design of the Learning Registry. More detailed information can be found in [4] and at the starting points at www.learningregistry.org.

2.1 Storing and Distributing Learning Resource Information

The Learning Registry accepts, stores, and provides access to learning resource descriptions—metadata or social metadata—as documents expressed in JSON notation. The storage and access mechanism used internally is CouchDB (couchdb.apache.org), a lightweight, open-source document-based database. CouchDB provides data access in the form of views generated by MapReduce functions written in Javascript. Couch's replication mechanisms make it easy to stand up a network of Couch nodes, which serves the goal of decentralizing the Learning Registry and eliminating single points of failure. Replication can also be leveraged by high-volume users to create their own local repository containing all or some of the Learning Registry data.

On top of CouchDB, the Learning Registry provides a layer of services (written in Python) as APIs to publish data, to query documents, and to retrieve documents. These services are the principle public interfaces to the Learning Registry, though developers are welcome (and encouraged) to provide other services in addition to or on top of these services.

2.2 Submitting Descriptive Metadata

By design, the Learning Registry defines a loose format for the submission of metadata about resources, and does not specify what metadata schema should be used. The Resource Data Description (RDD) document can be used to submit metadata or social metadata (either linked to via a URI or embedded directly as the payload), and can contain a reference to a URI that gives the schema for the metadata. The RDD is a thin wrapper around the submitted metadata; in this way, the Learning Registry can be agnostic about metadata formats and, as the field develops, about formats for social metadata. Early users have used the National Science Digital Library (NSDL) Dublin Core schema [5] or IEEE LOM [6], but nothing precludes the use of newer schemas such as LRMI (lrmi.net) or custom ones (e.g., [7]). We expect that services built on top of the Learning Registry can provide extraction or crosswalk services across RDDs that use disparate standards, or can assemble metadata fields from different schemas into custom views.

Documents submitted to the Learning Registry are assigned a unique document ID. This document ID is not a unique identifier for the *resource* described by the document: multiple documents in the Learning Registry can describe the same resource. We have found it challenging to create a unique way to identify a resource so that these descriptions (metadata or social metadata) can be used for analytics. The document ID of a metadata description is specific to just that one submission of metadata, and, indeed, social metadata may be submitted for a learning resource before the resource itself has been registered with a metadata submission.

For the present, the Learning Registry uses the resource's URL as this global identifier (called *resource_locator)*. Using URL this way is already posing difficulties because in practice multiple URLs might refer to the same resource. (e.g., PBS Learning Media, www.pbslearningmedia.org, desired to redirect its users to local PBS affiliates, turning the URL into something like ca.pbslearningmedia.org). While a solution such as OpenURL could provide a standard URL to match resource metadata with social metadata, it requires a central service. Over time, the Learning Registry community may develop more consistent URL conventions, adopt OpenURL, provide translation services, or simply live with "islands" of data created by non-uniform naming. The Registry may also eventually support assertions that could allow community members to make formal statements to the effect: "I assert that this URL, *x*, and that URL, *y*, refer to the same Learning Resource." and thus allow combined analytics for *x* and y.

Uniquely identifying resources was also a difficulty. In our experimentation, we found that publishers assumed that their local URL was the canonical version. For example, a site such as learnalot.com would submit http://learnalot.com/resource-1, which is where the reference to the learning resource is hosted on their site. However the resource may have originated from http://federalagency.gov/resource-1) and, in order to aggregate all paradata for this resource, a canonical locator must be used. OpenURLs, which are used in many learning contexts such as Google Scholar, rely on a query mechanism to locate a context-

sensitive URL for a resource, relative to the requestor. There is currently no way to express an OpenURL in a canonical format such that all context-sensitive copies can have an association expressed within the Learning Registry

2.3 Submitting Social Metadata

The NSDL Com_para format, created for the STEM Exchange prototype [8], was an early starting point for the social metadata, or "paradata" format. After discussion with the "attention" metadata developers (e.g., [9]), the Com_para format was modified. Then, in an effort to move to a JSON-based format, the Activity Streams format (activitystrea.ms) was investigated and ultimately extended for use in the Learning Registry: the extension was to allow aggregations of data across actors or resources. Our intent is to collect any and all data we can at any grain size, without enforcing a specific vocabulary. Because we are building a data store for analytics, in principle we are not concerned with data volume or expiration.

The current plan is to solicit ratings data, such as *rated*, *favorited*, and *bookmarked*. We are also encouraging the submission of usage data, such as *downloaded*, *viewed*, or *aligned* to a standard. Other activity data could be enhancing descriptive metadata, such as *commenting* and *tagging*. The schema for social metadata is [actor], [verb], [object] with [modifiers] allowed for [verb] that are most commonly used for measures and dates.

A challenge in the submission of social metadata is that the characterization of the actor is left to the publisher of the data. This potentially creates a bias in the data and may limit the ability to do relationship analysis. For example, if a publisher asserts "a student watched video *Y*," later analysis cannot determine, from the generic "student," grade-level, cultural background, or municipality. Our strategy is to build early rudimentary analytics to show the community what can be done with them, and to encourage best practices in social metadata expression.

2.4 Retrieving Stored Data

The Learning Registry offers two core functions for data retrieval: obtain and harvest. *Obtain* is used to gather all RDDs at a node, or a subset based on the *resource_locator* present in the RDD. *Harvest* is based on OAI-PMH Harvest [10] to gather RDDs or payload data for specific date ranges. Additionally, the *Slice* service allows users to retrieve documents based on properties of RDDs. *Slice* is not a deep search of full metadata, paradata (another name for "social metadata"), or the resource itself, but a means of querying high-level properties included in the RDD "wrapper" around such payloads including (1) identity of the resource owner, metadata author, or submitter; (2) date of submission, and (3) "tags" (including keywords, schema format, and the resource data type, i.e., paradata or metadata). We face challenges in our application of CouchDB's views in storing data. Because CouchDB is not a relational database, it relies on extensive indexing to populate views. As such, users would submit data and not see it immediately via the *Slice* interface, due to the slowness of updating views. This issue can be addressed by managing user expectations or by developing alternatives to *Slice* that update their views more quickly.

2.5 Special Tools and Issues

The Learning Registry community has expressed interest in building out tools and services to connect various data collection platforms to the Learning Registry: e.g., Basic Learning Tool Interoperability (BLTI) interfaces for Learning Management Systems (LMS); usage data from LMSs; social metadata from portals and repositories; and descriptive metadata of interest to K-12 teachers (e.g., alignment to the Common Core). We anticipate these being built out over the coming year. As an example, the Learning Registry has created an OAI-PMH "data pump" to support a one-time extraction of Dublin Core and NSDL_DC metadata from repositories that support an OAI-PMH harvest end point. This script extracts, from the harvested metadata, various field values to create tags (called "keys" in the RDD) for the submitted metadata.

An issue for the Learning Registry is ensuring that valid information is submitted. The Learning Registry thus requires that documents submitted must be digitally signed using a key signature. Our approach to identity [11] relies, at present, on a PGP-based signature attached to the document. (Signing submissions turned out to be difficult for some users, and we also found the need to create a PGP key store for some users.)

3. INITIAL EXPERIMENTATION

Our first experiment with the Learning Registry replicated an existing linkup between NSDL (nsdl.org) and a teacher portal. NSDL provided an OAI-PMH protocol to transfer data to a teacher content management system (CTE Online, cteonline.org) that could harvest math resources and send usage data back. This basic exchange was replaced with Learning Registry functionality to demonstrate feasibility. This early proof of concept led us to create the OAI data pump, which became a useful tool for data extraction.

The Learning Registry beta version was opened to an early-adopter community in the September-October 2011 timeframe. This was done to ensure that the concept was viable, to see what problems people faced as they published and extracted their data, and to learn how expressive they found the paradata specification.

To demonstrate the utility of data submitted, we created an implementation of a search amplified by Learning Registry data. The AMPS Chrome browser extension[2] inspects the results of an Internet search performed on google.com and then injects related activity data and standards alignment data from the Learning Registry. AMPS utilizes a simple mapping from resource URLs to data stored/referenced in the Learning Registry (emphasizing the importance of canonical URLs).

AMPS provided a good-proof-of-concept for showing how the Learning Registry can be used to connect data from disparate sources. For example, ISKME submits Achieve rubric data on alignment of open resources to the Common Core Standards, and ratings, e.g., for quality of assessment and interactivity. These data can be correlated, via canonical URL, to ratings data from a distant teacher portal and surfaced in a search, using this extension.

Submitters wanted a visual way to see connections among disparate data submitted. The Learning Registry Visual Browser (demolearningregistry.sri.com/browse) provides a browser-based interface to explore data. The user can enter a search term and the browser queries the Learning Registry for documents containing that term as a tag or identity. A list of summaries of these documents is displayed, as an ordinary search engine might, but the browser also displays a cloud of related terms, which allows the user to easily explore the (semantically) nearby "space" of documents in the Learning Registry. (The visual browser for

[2] Available from https://github.com/jimklo/AMPS-Chrome

different early-adopters can be accessed from various community pages on the learningregistry.org website.)

4. LEARNING ANALYTICS ON LR DATA

The alternatives to a system such as the Learning Registry are for searchers to visit individual websites and repositories, for organizations to build federated search or to make n-way connections among resource providers, or for web crawlers to scrape for descriptive and social metadata. As we have described, these are all barriers to resource locating, sharing, and amplifying. The Learning Registry data store provides a unique value to learning-resource analytics that is now not possible. In this section, we describe applications for learning analytics, and for each, how Learning Registry data could be used.

- Relationship Mining: The Learning Registry could surface relationships between people based on their attention to resources. What institutions, portals, or groups of users have shared/curated the same resource?
- User knowledge modeling: Learning Registry data could be used to compute what a student might be *expected* to know. In contrast to the coarse characterization of ranges of grade levels present in most metadata, Learning Registry assertions could be specific about the grade or level at which resources were successfully used.
- User experience modeling (Are users satisfied?): Learning Registry *ratings* social metadata could be used to compute satisfaction and thus provide feedback to developers.
- User profiling (What groups do users cluster into?): Inverting the [actor] [verb] [object] syntax, we could compute the types of actors who use resources. Submitters of social metadata can also be clustered in categories based on how and when they submit.
- Domain modeling (How is content decomposed into components and sequenced?): Curriculum construction tools could gather data on sequencing of learning resources or alignment to standards.
- Trend analysis (What changes over time and how?): Trends in *attention to* different resources could be computed if a sufficiently fine-grain size of social metadata is submitted, because the social metadata specifies a date or date range.
- Recommendations (What next actions/resources can be suggested for the user?): The Learning Registry can support recommendations by clustering users or by building a social network graph and then recommending resources among a cluster or network.
- Feedback, Adaption, and Personalization (What actions should be suggested for the user? How should the user experience be changed for the next user?): Learning Registry data could provide feedback to developers about the utility of their resources, about who adapts them and how, and could eventually cause "widespread sharing" of learning resources to learners at the appropriate time.

The Learning Registry community is now creating analytics-based applications such as these using its unique timeline-organized dataset.

5. ACKNOWLEDGMENTS

SRI International's work on this project is supported by the US Department of Education (ED-04-CO-0040/0010). The open-source Learning Registry project was conceived by Steve Midgley, US Department of Education, and Dan Rehak, Advanced Distributed Learning Initiative (ADL), U.S Department of Defense. Susan Van Gundy, UCAR and NSDL, developed the first paradata specification and Aaron Silvers (ADL) amplified it based on the activity streams specification. Joe Hobson of Navigation North worked out many of the bugs in submitting metadata and paradata into the Learning Registry. Pat Lockley conceived of the Chrome search plugin. Technical support is provided by Lockheed Martin Global Training and Logistics under contract to the US Dept. of Defense Many other people and organizations contributed input and submitted data for the early proof-of-concept. A complete list of community members can be found at www.learningregistry.org.

The Learning Registry code is stored in an open GitHub repository, https://github.com/LearningRegistry/.

6. REFERENCES

[1] Soylu, A., Kuru, S., Wild, F., and Mödrichter F. 2008. e-Learning and Microformats: A Learning Object Harvesting Model and a Sample Application, In *Proceedings Mupple'08 Workshop*, 57-65.

[2] Chang, L. K., Liu, K-Y. Wu, C-A. and Chen, H-Y. 2005. Sharing Web-Based Multimedia Learning Objects Using NNTP News Architecture. Proceedings of the *Fifth IEEE International Conference on Advanced Learning Technologies* (ICALT'05).

[3] Smith, J., M. Klein, and M. Nelson. 2006. Repository Replication Using NNTP and SMTP. *Research and Advanced Technology for Digital Libraries*, J. Gonzalo, et al., Eds. Springer Berlin / Heidelberg, 51-62.

[4] Jesukiewicz, P. and Rehak, D. 2011. The Learning Registry: Sharing Federal Learning Resources. Presented at the *Interservice/Industry Training, Simulation, and Education Conference* (I/ITSEC).

[5] NSDL. n.d. NSDL_DC Metadata Guidelines, http://nsdl.org/contribute/metadata-guide

[6] LOM. 2002. IEEE Standard for Learning Object Metadata, IEEE Computer Society, September 2002.

[7] Fait, H., and Hsi, S. 2005. From Playful Exhibits to LOM: Lessons from Building an Exploratorium Digital Library. *JCDL'05*, (June 7–11, 2005) Denver, Colorado, USA.

[8] NSDL n.d. Paradata. http://nsdlnetwork.org/stemexchange/paradata

[9] Najjar, J., Meire, M. and Duval, E. Attention Metadata Management: Tracking the use of Learning Objects through Attention.XML. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*. (2005). 1157--1161.

[10] OAI-PMH, (2008). The Open Archives Initiative Protocol for Metadata Harvesting, V2.0, www.openarchives.org/OAI/openarchivesprotocol.html

[11] Bienkowski, M. and Klo, J. 2011. Identity in the Federal Learning Registry. Position Paper for *W3C Workshop on Identity in the Browser* (Mountain View, CA, USA, May 24 – 25, 2011). http://www.w3.org/2011/identity-ws/papers/idbrowser2011_submission_27.pdf.