The Learning Registry – Wider Potential

David Kay, Sero Consulting, October 2012

This Report

This report seeks to examine the wider potential affordances of The Learning Registry as an architecture or conceptual approach, looking beyond the core educational technology focus to the broader information environment and the associated JISC community. It represents personal but hopefully useful observations on the Learning Registry at a point in time (October 2012), which should be weighed against the more detailed inputs (covering both technology and practice) to the JISC Learning Registry Node (JLeRN) project.

Section 1 – Background

1.1 - The Problem Space

In 2010, the Learning Registry (LR) project originated in the US in response to particular needs identified in US education and training, particularly relating to surfacing, characterizing and sharing / reusing learning resources.

Whilst those challenges may have had particular manifestations in the minds of Department of Defense (DoD) and Education (DoE) stakeholders who supported the LR project, they were symptomatic of widely recognized and persistent difficulties in describing learning resources. These amount to a debilitating lack of consensus about how to characterize / homogenize / harmonize learning resource metadata, resulting in a chaotic **mess**, largely incapable of supporting content reuse and analysis in any scale beyond local.

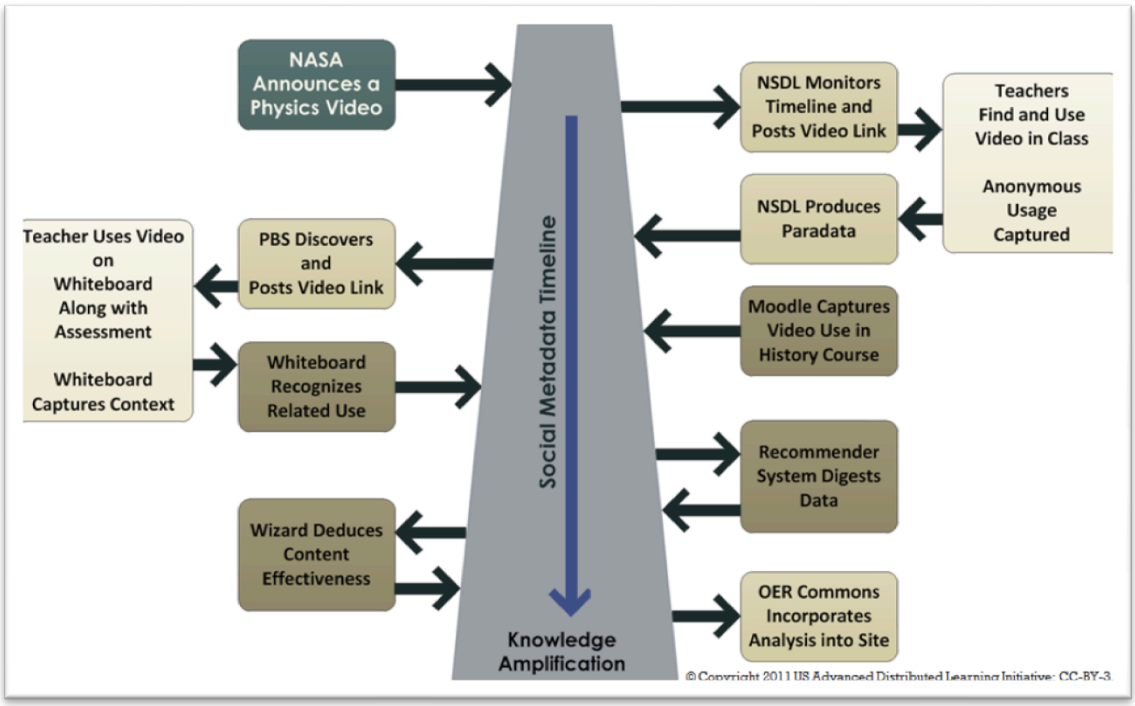
Paradata, resource usage data with **context**, may be a vital part of the jigsaw as it allows resources to become increasingly 'well-described' on the basis of their utilization (Who, Where, When, How, etc…). However, it is only a format that is subject to the quality of the data itself, particularly regarding the use of consistent and persistent identifiers to 'link' paradata.

In enabling analysis and reuse, paradata has the potential to be more powerful at **scale** (introducing statistical reliability and exposing the long tail of resources and of usage). It may therefore benefit from the ability to network datasets across the community (subject, national, international), as addressed in the LR concept of Nodes.

1.2 - The Vision

The LR project is therefore of particular interest because it has developed an 'approach' (model, architecture…) to address those key issues of mess, context and scale. In simplified terms, the LR proposes that metadata mess is addressed by a flexible approach to data attributes and schemas (whilst requiring

consistent resource identifiers), that context is evidenced by paradata, and that scale is enabled by orchestration between networked nodes.
Central to this vision is the potential of paradata within a distributed resource network connected, where usage episodes are connected by consistent use of resource identifiers (such as URLs). The resulting potential to track the availability, use and enhancement of a resource is illustrated this ADL diagram of the shared metadata timeline. It is predicted that this cycle will in turn enable more fruitful analysis, more accurate recommendation and wider amplification.



1.3 - The JLeRN Project

Having tracked the development of the Learning Registry since 2010, JISC funded the JLeRN (JISC Learning Registry Node) project in 2012 to assess the LR approach and the supporting software by going though the process of setting up a UK node and to examine its potential through community experimentation and dialogue.

The end of this process (October 2012) presents a good opportunity to question the potential of the vision and of the technical approach. Key questions might be framed as follows:

- In terms of operational **vision**
  - Is the LR vision dependent of shared practice and operational scale that will continue to elude the learning community
  - **Or** do the envisaged uses cases (e.g. as above) and the timing (in this era of social data) have a compelling resonance?

- In software **product** terms

    - Does the LR add enough to the underlying technology stack to establish a necessary and valued role?
    - **Or** is the LR simply an exemplification of what can be achieved using increasingly malleable lower level components?

- In terms of market **engagement**
    - Are the teaching and learning audience too narrow and the post-project governance too uncertain to elicit the ongoing commitment needed to deliver the power of the LR approach?
    - **Or** is there a wider value in the LR approach potentially involving other domains that would bring critical mass and a sustainable trajectory?

The JLeRN project has indicated some likelihood that **the vision** has currency with not only learning technologists but also with practitioners responsible for course development and learning delivery. This paper is particularly concerned with take-up and sustainability and therefore with the latter questions about product differentiation and the potential breadth of engagement.

Section 2 – Enough product?

This section considers potential in terms of the software **product**:

- Does the LR add enough to the underlying technology stack to establish a necessary and valued role?
- **Or** is the LR simply an exemplification of what can be achieved using increasingly malleable lower level components?

2.1 - Functional requirements

The bottom line is to understand the extent to which the LR functional capabilities add value to the underlying IT tools (notably the noSQL Couch database). Crudely expressed, if the LR only adds 'nice to have' features and / or compounds cost and risk in an unlikely to be completed / maintained layer, then the sustainability proposition is likely to be weak.

Here is a simple headline list of what we might expect from a toolset in the problem space being addressed by the LR:

Data	
Metadata	Description of objects, such as Title, Author
Paradata	Use of objects, such as Activity, Actor, Context, Date, Volume
Identifiers	Persistent and consistent IDs (DOIs, URIs) to link submissions
Files	The objects themselves or related assets
Vocabularies	Standardisation of terms used in metadata and paradata
Operations	
Ingest	Getting the data into the system
Storage	Storing the data

Replication	Replicating the data to other instances
Search	Selective retrieval of data
Index	Indexing based on full text or facets to optimize retrieval
Notification	Notifying other instances or users of changes
Annotation	User generated annotation of records, such as notes and ratings
Presentation	Presentation useful to humans, such as listings and visualizations
Authorisation	Control of access based on appropriate granularity
Setting	
Scale	The number of records may be very large
Distribution	The data may be curated across a network of nodes
Heterogeneity	The data may vary considerably within and between nodes
Services	The application presents itself as a range of services

2.2 - Learning Registry capability

How does the LR measure up to such a functional checklist? The following table is only indicative and is particularly open to comment from developers and implementers:

- It is suggested that much of the benefit is derived directly from the underlying CouchDB capability – see the LR Scope column;
- Meanwhile, the major value added is in the APIs developed by the project;
- However – as illustrated in the Downstream Challenges column – there is significant work to be done in terms of (a) core APIs especially in the retrieval area and (b) extensions that bring the LR to life such as recommender services.

Data	**LR Scope**	**Downstream challenges**
Metadata	Couch	
Paradata	Couch	
Identifiers	(Consistency assumed)	May be a showstopper as data accumulates
Files	Couch (Referenced or Attached)	
Vocabularies		Likely to be major depending on data structures and ingest implications
Operations		
Ingest	LR API	
Storage	Couch	
Replication	Couch plus	OAI-PMH potential
Search	LR API	
Index		Solr-like expectations
Notification	RSS	Recommender expectations
Annotation		Talked about
Presentation		Tools to be integrated (e.g. for visualization)
Authorisation		Could be major depending on granularity

		required
Setting		
Scale	Couch	
Distribution	Couch plus	
Heterogeneity	Couch	
Services	LR	But a work in progress, as above

In summary, there are notable caveats in the feedback from some participants in the concluding JLeRN workshop:

- That there is nothing experienced developers couldn't do in some way with CouchDB and / or other tools if the LR didn't exist;
- That the most adventurous and potentially rewarding of LR features, the value of paradata at scale and of the network of nodes, remain largely untested;
- That, whilst the LR concept adds significant value, such a layer adds operational overheads in terms of commitment and governance;
- That, whilst the costs of provisioning the LR network (such as providing, setting up and maintaining node servers) are low, they represent potential stumbling blocks for enthusiastic individuals or departments.

We might conclude that, whilst such reservations can be contested, some detailed work is required to confirm whether ongoing investment in the LR software model (as opposed to the vision) is truly worthwhile.

Section 3 – Enough traction?

This section considers potential in terms of the market **engagement**

- Are the teaching and learning audience too narrow and the post-project governance too uncertain to elicit the ongoing commitment needed to deliver the power of the LR approach?
- **Or** is there a wider value in the LR approach potentially involving other domains that would bring critical mass and a sustainable trajectory?

3.1 - Possible domains

Whether or not the teaching and learning audience is too narrow, it is instructive at this stage to consider whether the type of functionality described above has a good fit with similar problem spaces in other operational domains, especially in Further and Higher Education.

The following domain examples bear some similarities.

Domain	Requirement potentially aligned to the LR
Libraries	Paradata has an increasing profile in libraries for recommendations, collection management and student early warning systems. Whilst any above-campus / shared service value is less well established, the potential is being considered in current Mimas and JISC work. For more on library activity data, see http://activitydata.org. In addition the network model proposed by the LR may have relevance to rethinking how aggregation processes (such as those used by Copac or SUNCAT) might optimally be designed.
Other Curatorial Domains	Archive and museums would be expected to have a similar interests and motivations to libraries, though they are less likely to prioritise the resourcing of such developments. The aggregation of paradata is across institutions is however hardly relevant as each curates its own unique resource.
Repositories & Research Data	The research data and content landscape is currently a hot topic, with partial and sometimes integrated solutions offered by institutional repositories and CRIS applications. The elephant in the room is the research data itself. The overarching potential for research paradata and associated analytics is relatively unexplored, though there are clearly important working parts relating to bibliometrics / citations, the REF and social network activity – see 'Analytics for Understanding Research' in the CETIS Analytics Series (to published 22 November) The diverse nature of the data suggests good fit with the LR vision, though the above-campus dimension will be less important to institutions than to the researchers themselves.
Estates	Estates covers a multitude of responsibilities, including environmental controls, security and utilization / footfall. The data is potentially quite large and certainly diverse, ranging from turnstiles to sensors, and is unlikely to be standardized. Whilst there will be no interest in sharing beyond the institution, the LR approach to this challenge may resonate more that an enterprise BI solution.

More generally, there may be fit with any operational requirements that need to accumulate and analyse:

- Diverse analytics data of any type
- Heterogeneous / specialised metadata
- Distributed networks of documents / digital assets, as in management of software downloads across forges and other mirror sites
- Widely sourced activity data about demographics (people) and geographies (places)

3.2 - Parallel universes

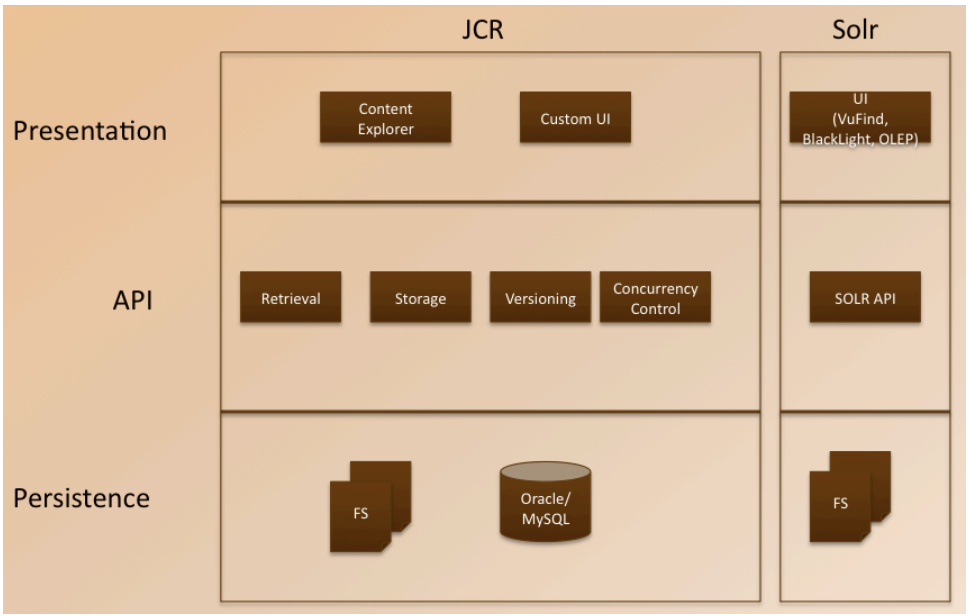
As above, it is clearly possible to indentify domains, not least in the HE world, where the LR approach has an interesting fit. However it is essential to recognise:

- The challenge - the extent of intersection / fit typically only relates to part of the LR model, with the possible positive exception of libraries;
- The challengers - There are other overlapping approaches to aspects of the problem space that may gain traction, though they may be less visionary; some, like the LR, are inspired by domain challenges, whilst others offer a generic approach, notably enterprise Data Warehouse / Business Intelligence solutions from such as IBM and Oracle.

Assuming that the price tag and complexity of generic enterprise solutions suggests opportunity for community sourced alternatives, we conclude by referencing some examples found in the HE library and repository domains, which like the LR have some potential for wider application.

Kuali Open Library Environment (OLE) DocStore
In presenting its recently developed DocStore application (based on Apache JackRabbit and Solr) at Kuali Days 2012, the Kuali OLE library partners presented familiar rationale for a generic metadata and content store:

- Our business is changing
- Metadata standards are in flux
- Metadata diverges by content type
- Libraries are being asked to expand their stewardship

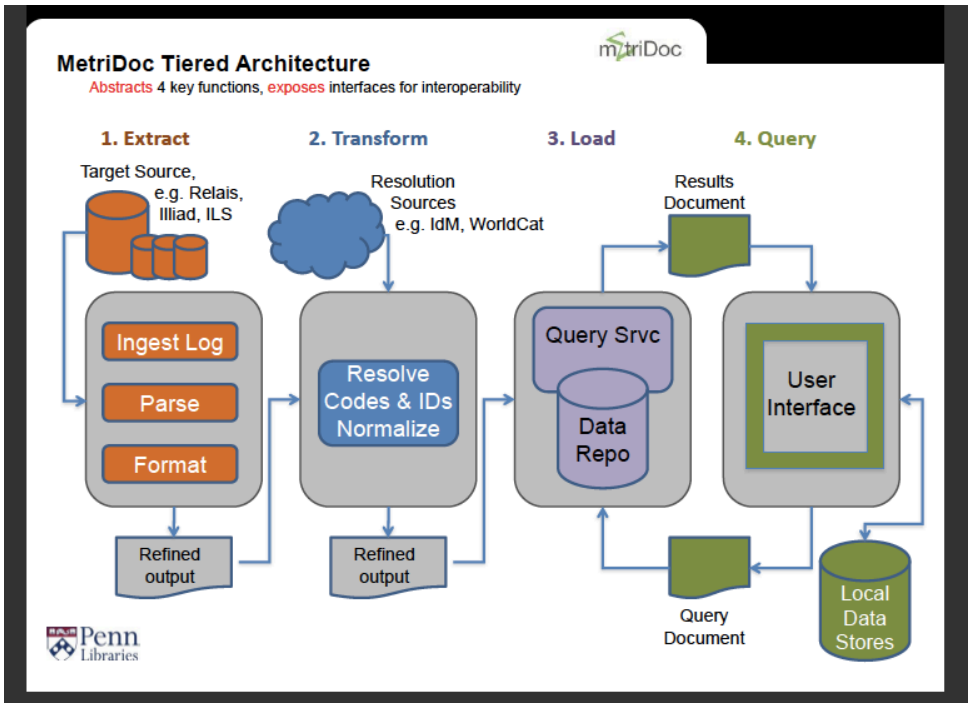


The Kuali OLE DocStore model

Metridoc from Penn Libraries

As in the learning resource world, these pressures (which can represent service opportunities) typically go hand in had with an imperative to develop analytics capabilities. For example, University of Pennsylvania Libraries (a Kuali OLE investor) has developed the MetriDoc open source model for collecting, transforming and analyzing activity data from multiple sources. In their own words:

> *MetriDoc is an extensible framework that supports library assessment and analytics, using a wide variety of activity data collected from heterogeneous sources. Data points are derived from fund accounting systems, discovery tool logs, publisher COUNTER reports, resource sharing systems, and authentication logs. But these sources are only the beginning of potential MetriDoc targets, which Penn is gradually expanding, guided by management and planning needs. In the near term, MetriDoc will consume book circulation data and research consultation/library instruction inputs. As Penn brings on new or replacement technology to support user services, the framework will expand to incorporate additional activity targets.*



The Penn Libraries MetriDoc model

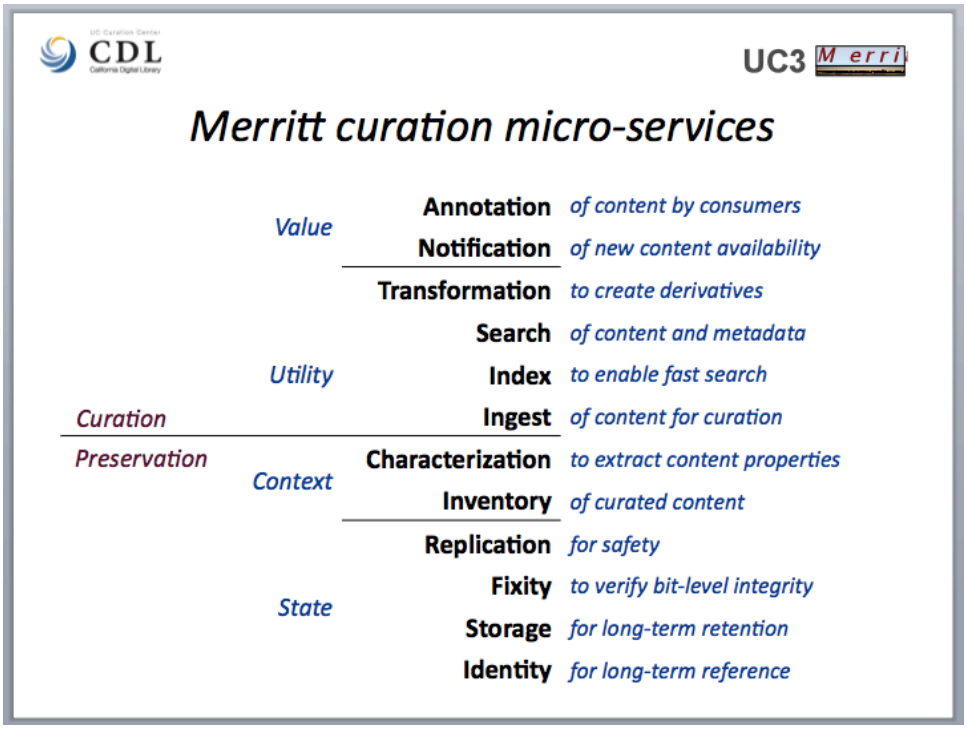
Just like the LR, neither the OLE DocStore nor the Penn Libraries MetriDoc hub address the full requirements set suggested in Section 2. However, both are based on Open Source Software, address data heterogeneity, can be applied in above-campus settings and are linked to potentially sustainable governance and institutional investment models.

California Digital Library (CDL) Micro-Services

The CDL micro-services approach represents a way of thinking about problem spaces that have a lot on common with the LR territory. CDL is particularly interesting on account of the way it breaks down the problem in to a set of

functions (micro-services) that can then addressed optimally, where possible using common and highly invested technology components (not least the filesystem). This approach has influenced such as the Oxford Data Flow project (http://www.dataflow.ox.ac.uk/index.php/about).

It would surely be of value to undertake an analysis of the space addressed by the LR based on the micro-services vision, and therefore to determine the necessary working parts and how they might be sourced, without the presumption of building a single end-to-end system.



Linked Data

Finally, it seems unreasonable to conclude without at least hinting at the possibilities for Linked Data – not as a solution in itself, but as adding value in description and connection of the emerging web of metadata and paradata.

The principles of linked data may be defined as:

- using URIs as names for things;
- using HTTP URIs to enable look up of those names;
- providing useful RDF information related to the URIs;
- including RDF statements that link to other URIs to enable discovery of other related concepts of the Web of Data

Linked Data may require far too much premeditation on the part of everyday actors in the learning and teaching world, but it is hard to deny that it confronts head on the issues of reliable connection that are likely to bite any deployment of the LR at scale.