

## MIT Open Access Articles

*Harvesting latent and usage-based metadata  
in a course management system to enrich  
the underlying educational digital library*

The MIT Faculty has made this article openly available. **Please share**  
how this access benefits you. Your story matters.

**Citation:** Kortemeyer, Gerd, Stefan Dröschler, and David E. Pritchard. "Harvesting Latent and Usage-Based Metadata in a Course Management System to Enrich the Underlying Educational Digital Library: A Case Study." *International Journal on Digital Libraries* 14.1–2 (2014): 1–15.

**As Published:** <http://dx.doi.org/10.1007/s00799-013-0107-6>

**Publisher:** Springer Berlin Heidelberg

**Persistent URL:** <http://hdl.handle.net/1721.1/104920>

**Version:** Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

**Terms of use:** Creative Commons Attribution-Noncommercial-Share Alike



# Harvesting latent and usage-based metadata in a course management system to enrich the underlying educational digital library

## A case study

Gerd Kortemeyer · Stefan Dröschler ·  
David E. Pritchard

Received: 14 December 2012 / Revised: 18 October 2013 / Accepted: 8 November 2013 / Published online: 29 November 2013  
© Springer-Verlag Berlin Heidelberg 2013

**Abstract** In this case study, we demonstrate how in an integrated digital library and course management system, metadata can be generated using a bootstrapping mechanism. The integration encompasses sequencing of content by teachers and deployment of content to learners. We show that taxonomy term assignments and a recommender system can be based almost solely on usage data (especially correlations on what teachers have put in the same course or assignment). In particular, we show that with minimal human intervention, taxonomy terms, quality measures, and an association ruleset can be established for a large pool of fine-granular educational assets.

**Keywords** Course management system · Recommendation system · Educational digital library taxonomy · Content sequencing · Online assessment

## 1 Introduction

Comprehensive metadata and quality measures are the most important features that should distinguish assets in digital libraries from the vast number of assets available on the “open web”. In educational settings, metadata should guide users toward appropriate assets, provide data on the reliability of

the assets, and establish meaningful connections between the assets. The goal for metadata-based recommender systems has been pursued in the educational realm for over a decade (e.g., [1]), but unfortunately, still today many digital libraries fall short of this promise. Thus, these systems fail to have significant impact on the day-to-day operation of most schools, colleges, and universities, where insular course management solutions and traditional textbooks are prevalent [2]. More often than not, searches on the “open web” yield more useful results than those within the confines of the library. Some studies find that perceived usefulness and usability are driving factors in digital library technology adoption [3], while others find that behavioral intentions tend to be stronger factors (see Turner et al. [4] for review). In any case, studies regarding usability of digital libraries have increasingly gained importance (e.g. [5,6]).

Metadata in current digital libraries generally includes controlled keywords and established taxonomies, which are maintained by dedicated staff. In addition, rapid review mechanisms need to be in place through editors or peer-review. Unfortunately, all of these approaches have tradeoffs [7]. Particularly for an open-source free system, many of these resources are not available on a sustainable base. Such a system depends on the goodwill of its authors, who cannot realistically be required to write metadata (which does not help them personally) and who cannot be expected to keep their metadata up to date when they no longer use their assets. An alternative approach attempts to offload metadata construction and maintenance from the authors to the user community, making use of social networking mechanisms and user tagging, and leading to the so-called “folksonomies” [8,9]—however, even in this approach, user motivation is essential and not always a given [10].

There is thus a delicate balance in the competition between the “open web” and digital libraries: demanding high-quality

---

G. Kortemeyer (✉) · S. Dröschler · D. E. Pritchard  
Massachusetts Institute of Technology, Cambridge, USA  
e-mail: korte@lite.msu.edu

*Present address:*  
G. Kortemeyer  
Michigan State University, East Lansing, USA

*Present address:*  
S. Dröschler  
Ostfalia University of Applied Science, Wolfenbüttel, Germany

rigid metadata and implementing strict quality measures at the time of acquisition from volunteer authors may result in the digital library never reaching critical mass, while being lenient on metadata and quality measures deprives the digital library of many potential advantages. The subject of our case study, the open-source free LON-CAPA system, has reached critical mass, alas at the expense of metadata quality.

LON-CAPA currently holds approximately 446,000 educational assets that are used by 150,000 students per year world-wide. Most authors in the system are active instructors. A major shortcoming of the system is that the author-provided metadata associated with its assets is ill-defined, at times erroneous, and extremely sparse, which makes it hard for instructors to locate appropriate materials in the digital library layer of the system. We present techniques and experiences on how to improve the quality of these metadata by complementing and refining them with latent and usage information collected over 10 years. The goal of this effort was to provide instructors with guidance for selecting appropriate assets for their teaching venues. We argue that this latent and usage-based metadata is richer than the classic static metadata, and that the presented mechanisms are universally applicable for learning content management.

LON-CAPA is different from most digital library systems, as the tools to create, sequence, and deploy content assets are part of the architecture; the system has integrated sequencing tools and a complete course management system as frontend. In contrast, in most systems, the deployment of an asset is disconnected from the repository: assets get downloaded from a repository and uploaded into a course management system, where they are sequenced and deployed (see Klebl et al. [12, 13] for a review of different educational library systems and architectures). We argue that LON-CAPA's integrative approach is the key to collecting asset metadata that can be richer and more useful than classic static metadata, as it allows for automated feedback of metadata back into the digital library layer.

In this case study, we first describe versioning, cataloging, and quality mechanisms which were established, and we evaluate the resulting state of the metadata after a decade of operation. Our study does not follow the line of machine learning-based strategies using instructor data (e.g. [7, 11]), but takes advantage of learner usage and quality information gathered to a large degree from the integrated course management system. We present different efforts to improve metadata by heuristically constructing new information based on usage and other latent information: how can this dynamic and fluid data be turned into useful metadata for searches, browsing, recommendations, and quality control?

In the evaluation phase of our study, we attempt to answer the following questions:

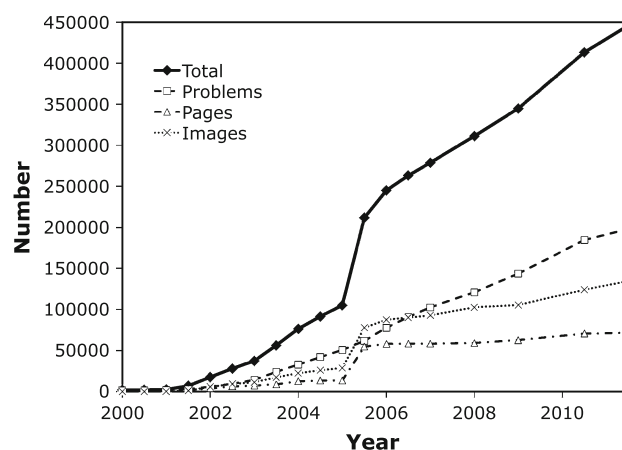
- To which degree can correct taxonomy term assignments be made based on usage and latent information from a distributed course management system?
- To which degree can useful ranked recommendations to instructors be based on usage information from a distributed course management system?

## 2 Model system: LON-CAPA

### 2.1 System overview

LON-CAPA is an open-source learning content management and assessment system that has been in production since 1999 [14], and its shared resource pool has been growing approximately linearly since (Fig. 1). In fact, LON-CAPA grows linearly in most any respect: institutions, learners, authors, etc. On the one hand, we are encouraged by the steadiness of this growth, which we attribute to the low entry barrier, while on the other hand, from a mature networked community, a different higher-order growth mode might be expected. We surmise that the fact that the system does not really “catch on” is partly due to the fact that the underlying digital library is not sufficiently exposed and thus is of limited usefulness.

The library is dominated by online homework and exam problems, as LON-CAPA has a particularly powerful assessment engine. The system is used at 160 institutions, both secondary and postsecondary, and is distributed: institutions need to provide their own instance of the system. At the same time, all institutions share access to the same distributed content library. Content assets are stored at a low granularity level (one page, one image, one homework problem), and the system allows instructors to assemble content into

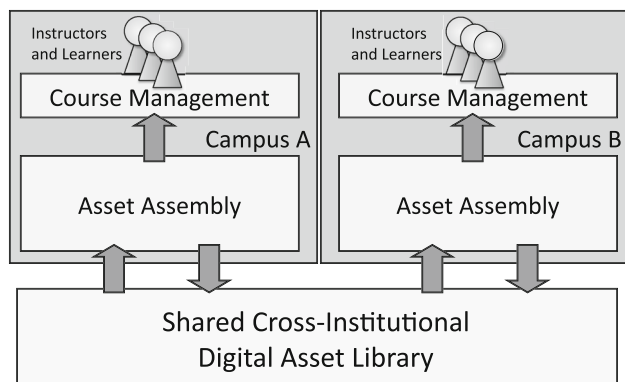


**Fig. 1** Growth of the LON-CAPA asset pool over the years. The steady (though mostly linear) growth may partly be due to the low entry barrier for contributing new materials

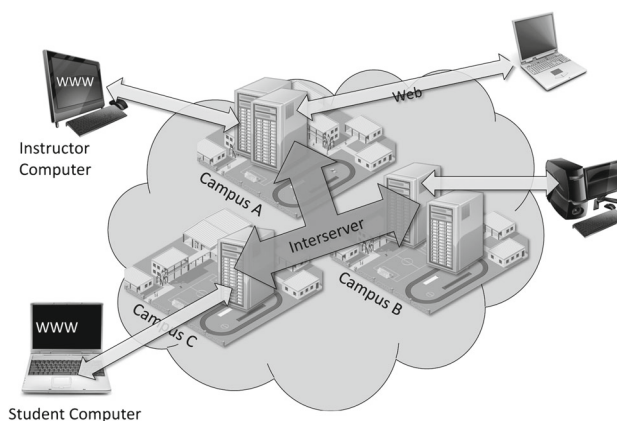
higher granularity objects (modules, chapters, etc.). Content gathered from the asset pool can immediately be deployed within the integrated course management system. Figure 2 shows an overview of the logical architecture of the system. This architecture is different from most other digital libraries: many libraries merely link collections and essentially manage metadata for assets stored elsewhere, while other systems may hold the actual asset itself, but are not the deployment platform—assets instead need to be downloaded for use, for example into a commercial course management system.

## 2.2 Network architecture

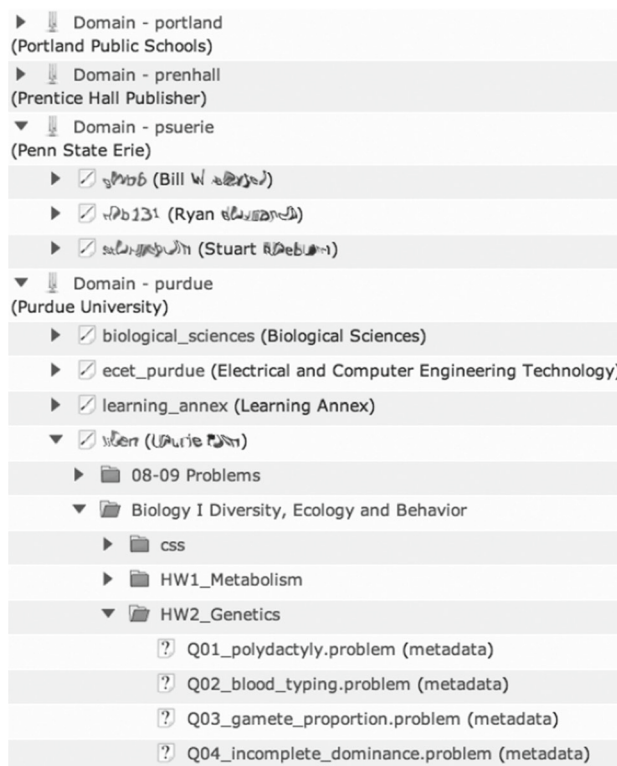
LON-CAPA is truly distributed: user data, as well as the original copy of contributed assets, are stored at the user's instance of the system. Its architecture satisfies data privacy issues and provides a high level of scalability. Content replication and subscription mechanisms avoid bottleneck situations when an asset generated at one institution is used at another one (Fig. 3). Assets are generated in a private staging area, which is server space into which authors can upload content and inside of which they can use built-in editors to edit content. Inside this space, authors can establish nested subdirectories to organize their content. For an asset to become available to the participants in the network, it needs to be published by the author under a system-wide persistent URL path. Paths are organized in a subdirectory-like fashion, based on the asset's position in the staging area, so a typical URL path may look like /msu/jdoe/physlib/force/newton.html (see Fig. 4 for a screenshot of browsing this file space). During publication, the author can also set usage permissions: who can use their asset, and who can make derivative works; the system then enforces these permissions. After publication, assets can be changed and re-published, at which point the system creates a new version, while keeping all previous versions. A published resource cannot be deleted (avoiding stale links), but it



**Fig. 2** Logical architecture of LON-CAPA. The course management (CMS) and digital library layer are closely coupled



**Fig. 3** LON-CAPA network architecture. The network architecture poses the additional challenge that no metadata catalog can be centrally managed and updated



**Fig. 4** Subdirectory structure of LON-CAPA's asset pool

can be marked as “obsolete”, from when on it does not show up anymore in searching and browsing.

Unfortunately in the framework of this project, this distributed architecture poses additional challenges: there is no central static metadata catalog, which could be managed and maintained by librarians or even designated faculty members. Thus, in this project, for the first time we needed to collect (“harvest”) static metadata and usage data from across the network. In future projects, we need to investigate how

to re-distribute the generated new metadata across the network.

### 2.3 Content publication and quality control

The majority of the assets in the system are authored by educators. Most of these educators, who are actively teaching, are unwilling or unable to submit comprehensive curricula, but are willing to add assets to an existing critical mass of applicable assets, and expect return on their investment by being able to immediately use their new assets in the context of the existing ones—this requires rapid publication mechanisms. Thus, early on, a decision was made that there is no formal review mechanism to not create bottleneck situations. In traditional library systems, a staff of librarians can address the associated quality control and metadata challenges, however, in an open-source freeware system like LON-CAPA, other automated mechanisms needed to be implemented:

- Only bonafide institutions, whose authenticity and integrity have been verified by a board, are admitted as members in the shared asset pool.
- Authoring access is generally limited to instructors at member institutions, which has the positive side effect of establishing some basic quality control.
- During publication, documents are parsed for some basic metadata—a mechanism we describe in Sect. 2.4.
- Usage information is collected on each asset—by collecting information on which instructors chose to use an asset in their classes, the system effectively has established peer-review.

In this case study, particularly the last point is crucial, as this implicit peer-review (or “peer-approval”) is one of the major keys to establishing the bridge between an educational digital library and actual teaching practice. In a library system that only holds links to external assets, there is no way to keep track of the actual asset usage; in a library that holds the assets just for download, one can keep track of the number of downloads, but not of the actual usage: was a particular asset used at all? was it used in a class of 20, or in a class of 200? for how many semesters was it used?

### 2.4 Existing metadata mechanism

Already more than a decade ago, efforts were put into place to establish standards for educational metadata [15, 16]. Such metadata schemes can be successful, but require a careful process to incorporate [17], which we were only able to sustain for a fraction of our assets. LON-CAPA supports a superset of the Dublin Core metadata scheme [18]. When publishing an asset, authors get to a screen where this metadata can be entered. Unfortunately, already early on it became appar-

ent that authors are perfectly willing to spend hours writing quality content, but do not seem to be willing to invest a few minutes into providing metadata. Particularly since many authors seem motivated primarily by generating content for their next lesson, homework assignment or exam, the reusability of content fostered by quality metadata does not appear to be an immediate priority. We thus implemented a number of features to suggest metadata, which would be filled in and remain in place if the author simply pressed “publish” without further review:

- The system gets author name, institution, timestamps, etc., from the system environment.
- The system picks up all meta-tags embedded in the document and fills in the fields.
- The system fills in the format field from the MIME type of the document.
- The system attempts to extract a possible title for the document from the XHTML tags.
- The system scans the document for possible keywords, which it identifies by discarding all formatting commands and all words that correspond to a table of non-keywords (e.g., “the”, “it”, “figure”, “section”; this list was initially generated by collecting all words from a number of assets and having a human mark all words that were not keywords). During publication, the author is presented with a checkbox list of all possible keywords, where possible keywords that were frequently used in the past are already pre-checked.
- Authors are frequently using decent subdirectory structures in their staging area, as that organizational concept is familiar from any personal computer (Fig. 4). The system provides the ability for authors to define metadata fields on a directory level, which are then inherited during publication in a cascading fashion, where lower directories override or add to the fields (depending on the field). Thus, we make it easier for authors to take advantage of their existing organization during metadata cataloging, and this mechanism makes it more efficient to publish large collections of assets.
- In addition to Dublin Core, during publication all assets that the asset depends on or links to are noted, as well as all response types of embedded assessments, etc.

We call this metadata “static”, since it is collected once at publication time, and only changes when new versions are published. Still, even with all of the above mechanisms in place, metadata remained sparse. Only 153,000 of the 446,000 assets have an assigned title, i.e., 66 % of the assets do not even have the most basic metadata field. Only 203,000 of the assets have an assigned subject; as a result of the non-controlled vocabulary, there were 15,000 unique subject fields, ranging from general subject assignments like



“Physics” and “Botany”, to specific topics like “Math Diagnostic Test” and “The Shapes of Molecules”, to subject assignments that resemble keyword lists, such as “momentum, impulse, elastic, inelastic, collisions”.

310,000 Assets have an average of 5.3 assigned keywords, while 136,000 assets have no keywords. Altogether, there are 1,636,000 assigned keywords, however, only 25,000 unique keywords. While this latter number is high in absolute terms and reflects the non-controlled vocabulary, the 1:65 ratio of unique to assigned keywords shows that the above-described mechanism of automatically assigning the previously most frequently assigned keywords in part served the function of controlling the vocabulary.

Unfortunately, in our case study, noise was introduced into the vocabulary through the multilingual nature of the system. As LON-CAPA is also used outside the English-language realm, several assets have metadata in languages other than English. During the first years of operation, the encoding of foreign languages was ISO, while in later years, the system completely switched to UNICODE. As a result, our foreign-language metadata still has mixed encodings. The same is true for special characters in metadata fields, for example Greek characters (e.g., “ $\gamma$ -radiation”), mathematical symbols (e.g., “ $\pm$ ”), and accented characters (e.g., “Mössbauer effect”), which in addition at times were coded in L<sup>A</sup>T<sub>E</sub>X-format. These inconsistencies led to further degradation of the static metadata.

Finally, one of the biggest challenges to overcome is that due to the automated nature in which keywords were harvested from assets, the most salient keywords may be absent from assessment content. Very often, the main concept required to solve homework and exam problems is missing from the text. For example, a problem may have the keywords “collision”, “mass”, and “velocity” extracted from its text. However, the problem very likely is actually about momentum conservation, but the word “momentum” is nowhere in the text: the student is supposed to figure out this solution strategy himself or herself.

Currently, when doing searches in the system, only this static metadata is considered, while browsing is based on the URL paths (Fig. 4). The simple usage-based information (number of accesses, difficulty, etc.) can be used for ranking during searches. Users found this mechanism unsatisfactory: for example, searching for the subject “Physics” would miss the majority of physics resources, and browsing for physics resources requires prior knowledge of where to look. Ironically, even faculty who are notoriously bad at providing metadata for their own assets have complained about how difficult it is to find assets. As asset sharing is one of the main features of LON-CAPA, the situation needs to be remedied: instructors need to be able to find materials, and need to have multiple routes to discover new materials including recommendations by the system.

## 2.5 Additional data collection mechanisms

Since LON-CAPA’s architecture is designed to cover the entire life cycle of assets (Fig. 2), we can keep track of their usage. We currently store in which courses an asset is used, which asset is used before it, and which asset after it. Also, we know how many accesses an asset had. For assessment assets, we store each submission and transaction, and which assignment the asset had been part of. When an instructor internally calculates statistics, we store degrees of difficulty and degrees of discrimination alongside the assets. Through these mechanisms, we are taking advantage of the feedback that can be gleaned from the assembly and deployment (course management) layer of the architecture to enhance the metadata of the assets in the digital library layer. We call this metadata “dynamic”, as it constantly accumulates with asset usage. This data is available for 168,000 assets based on their usage in 7,700 courses, and it is attached to the asset metadata. Unfortunately, this dynamic metadata is currently insufficiently exposed to the user: the current version of the software uses this information only for the ranking of search results, and only on demand displays it to the user for one asset at a time.

Over the last decade, 7,700 courses were run in the system, with a total of 965,000 student enrollments. For each of these, all asset accesses were logged anonymously. A total of 73,520,000 homework and exam problems were served to students in 73,600 assignments. This data, currently, is not used at all.

## 2.6 Overall status of existing metadata

In our case study of this author-driven digital library, it is helpful to use the simple analogy of a supermarket to describe the status quo. The aisles in this hypothetical supermarket are currently not organized by product classification (bread, cereal, dairy, etc.), but by manufacturer (Kraft<sup>TM</sup>, General Mills<sup>TM</sup>, Sony<sup>TM</sup>, etc.). Thus, when browsing for a particular kind of product, the customer would need to know who produces it. The store clerk in this supermarket is unable to answer general questions (“Where can I find dairy?”), but can only answer very specific questions at unpredictable levels of detail (“Where can I find yoghurt?”, “Where can I find something black?”). Finally, if the customer is interested in the quality of the products, this information is known but not globally available. In our supermarket analogy, information on how well a product fulfilled its purpose is printed on the individual packages, but currently the general question which one of a set of products is the best for a particular purpose cannot be answered by the system.

### 3 Constructing metadata from usage data

As the static metadata in our system is sparse and inconsistent (see Sect. 2.4), it is our goal to enhance it by evaluating our dynamic metadata (Sect. 2.5). In addition, in order for instructors to more easily discover and locate appropriate assets, we want to lay the foundation for a recommender system similar to that of online bookstores.

As opposed to our analogy of the physical supermarket, we have the luxury that we can overlay as many organizational structures for our products as seems beneficial. The first of these organizational structures is analogous to sorting the products into reasonable sections (food products, hardware, stationary, pharmacy, etc.) and below that into aisles (food:bread, food:dairy, ..., hardware:tools, etc.) and shelves (hardware:tools:hammers). As straightforward as this seems, the challenge is to first construct and then populate this taxonomy based on the very limited and uncontrolled data that we have on the products (the supermarket equivalent would be an uncontrolled set of keywords like “yoghurt”, “metal”, “heavy”, “black”, “grain”, “milk”, “nutritious”, etc.).

In our analogy, the next challenge is to educate our store clerk, so that customers can get recommendations based on what other customers bought in the same context, and based on the quality of the products.

#### 3.1 Taxonomy

Our first goal was to find meaningful subject fields, and in fact implement an up to three-level taxonomy that corresponds to the way materials are usually taught in high school and college, e.g., “Physics:mechanics:force”. The first taxonomy level would be the course subject (“Physics”, “Chemistry”, “Biology”, etc.), the second level usually the course topic (“Mechanics”, “Genetics”, “Organic chemistry”, etc.), and the third level the particular topic (“Force”, “Capacitance”, etc.). We failed to find established taxonomies for educational assets (for example, taxonomies provided by professional societies or journals are too research-oriented for our purpose, while the Library of Congress classifications are too general for our purpose). The metadata scheme we constructed was thus based on the table of contents of standard textbooks, as well as the table of contents of some of our courses.

Given our content, we decided on 234 classification terms at different levels; Table 1 shows an excerpt of our scheme. We expect to be modifying the taxonomy in the future, but that in most cases this would be an addition of new classifications or a renaming and regrouping of existing classifications.

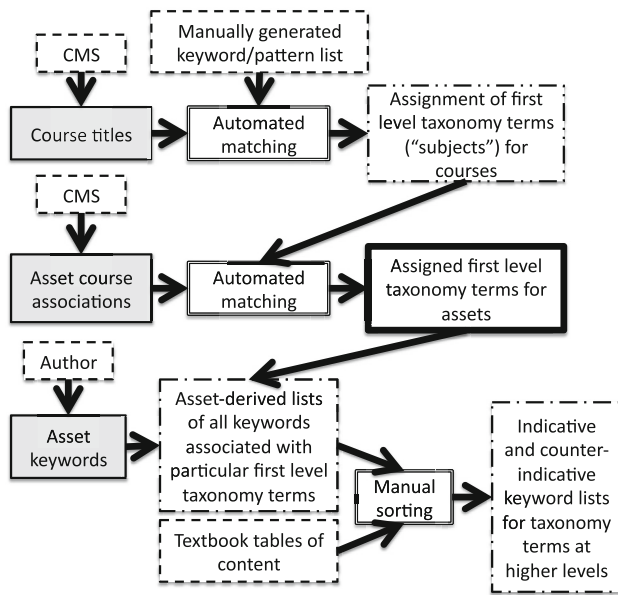
We decided that the language of this taxonomy would be English, independent of the language of the document; the localization of the platform should show the taxonomy terms in the language of the user, so translations of the taxonomy

**Table 1** Excerpt of the taxonomy we currently use

First level	Second level	Third level
Accounting	Payroll	
Accounting	Statements	
	...	
Biology	Anatomy	Circulatory
Biology	Anatomy	Digestive
	...	
Chemistry	Bonding	Covalent
Chemistry	Bonding	Ionic
Chemistry	Bonding	Polarity
	...	
Geology	Climate change	
Geology	Deformation mountains	
Geology	Earth history	
	...	
Mathematics	Calculus	Derivatives
Mathematics	Calculus	Extrema
	...	
Physics	Data	
Physics	Electromagnetism	Accircuits
Physics	Electromagnetism	Capacitance
Physics	Electromagnetism	DC circuits
Physics	Electromagnetism	Electrostatics
Physics	Electromagnetism	EM waves
Physics	Electromagnetism	Inductance
Physics	Electromagnetism	Magnetism
Physics	Electromagnetism	Potentials
	...	
Statistics	Bayesian	
Statistics	Descriptive	Distributions
Statistics	Descriptive	Means

terms would need to be provided as part of the localization package.

While we expected that the goal of a recommendation system could only be achieved by evaluating the assets in usage context, our initial hope was that the relatively simple goal of assigning taxonomies to the assets could be achieved based on the assets’ metadata alone—but even that turned out to be impossible. We thus needed to heuristically extract the taxonomy assignments from the assets’ usage contexts, and resorted to a bootstrapping approach where we built up the taxonomy assignment through several iterative steps. We know which assets instructors used within the same context (chapter, module, assignment), and thus get combinations of related keywords. While the keywords assigned to the individual assets may be too sparse to draw any conclusions, the superset of the keywords of contextually related groups of



**Fig. 5** Process that led to the keyword list in Table 3

assets can be rich enough to establish a taxonomy classification for these groups. Finally, we know how each manufacturer organized their own products, i.e., their subdirectory structure (URL path).

### 3.1.1 First-level taxonomy term assignments

While assets do not have to have assigned titles and subjects, courses in the system need titles for students to be able to identify them. The course titles often include American-style course codes, such as “FS 07 PHY 231-Introductory Physics I”, etc. Over the last decade, 7,715 courses were run in the system, and the vast majority of the assets in the system would have been used in at least one of them. In spite of the importance of titles, only 7,450 courses actually had titles—we assume that untitled courses were not actually used. Manually classifying all course titles seemed prohibitive and also not extensible into the future, so an automated process was required. In this process, course subjects were heuristically assigned based on a lookup table including keywords and patterns. Populating this lookup table with appropriate keywords and patterns (regular expressions) required some knowledge of course code and title assignments in North America and Germany, where the majority of LON-CAPA user institutions are situated. This process is illustrated in the top rows of Fig. 5.

Using this mechanism, we were able to identify possible subject areas for the majority of courses: while 2,021 courses remained unclassified, 5,390 courses had exactly one subject identified, 244 courses had two possible subjects, and 58 had three possible subjects, and two courses had four possible subjects. A large number of multi-classified

**Table 2** Course and preliminary asset subjects heuristically deduced from course titles

First level	Course frequency	Asset frequency
Accounting	39	412
Advertising	5	1,246
Astronomy	133	4,814
Biochemistry	42	0
Biology	589	27,921
Biophysics	8	0
Chemistry	1,567	32,858
Computer science	106	1,403
Ecology	7	182
Engineering	17	301
Finance	23	1,783
Geology	55	538
Geometry	129	538
History	5	48
Mathematics	526	9,646
Medicine	96	2,516
Nursing	3	8
Philosophy	2	10
Physics	2,515	70,557
Psychology	40	535
Statistics	143	2,960
Zoology	10	261

courses got classification assignments like {Mathematics, Statistics} or {Biology, Chemistry, Biochemistry}, which “made sense”. Some combinations like {Chemistry, Geology} made less sense until one looked at the associated title: “Waves and Electricity for Chemistry & Earth Sciences Students”—likely in reality this is a physics course, but simple heuristics would not be able to identify this. The middle column of Table 2 shows the frequencies of course subject assignments.

For every asset in the library, we know which courses it was used in. We decided to assign a preliminary first-level taxonomy term for each asset based on the subjects of the majority of courses it was used in; if an asset has an equal number of subject assignments across courses, we arbitrarily chose one of them. The rightmost column of Table 2 shows the frequency of subject assignments for the assets. We were thus able to assign preliminary first-level taxonomy terms to 158,537 assets or 36 % of the assets based on the titles of the courses in which they were used.

### 3.1.2 Keyword list

We then looked at the complete lists of keywords associated with assets in a certain first-level taxonomy to compile



**Table 3** Excerpt of the keyword list associated with certain taxonomy labels

Taxonomy label	Indicative keywords	Counter-indicative keywords
Physics:mechanics:linearkinematics	Motion, kinematics, speed, velocity, acceleration, distance, displacement, position	Angular, angle, force, forces, friction, work, atom, quantum, momentum, inertia
Physics:mechanics:rotational kinematics	Rotation, turn, turning, angular, speed, velocity, acceleration, angle, angles, degree, degrees, radians, displacement, balance	Work, force, torque, atom, quantum
Physics:mechanics:lineardynamics	Force, forces, free, diagram, acceleration, mass, newton, weight	Torque, angle, angular, charge, magnetic, atom, quantum, work
Physics:mechanics:rotational dynamics	Torque, angular, acceleration, inertia, moment, rolling, rotate, rotation, rotating, rotational, torques	Atom, quantum, momentum
Physics:mechanics:linear momentum	Momentum, velocity, mass, collision, collisions, elastic, inelastic, impulse	Angular, atom, quantum ,inertia, torque
Physics:mechanics:angular momentum	Momentum, velocity, angular, inertia, moment, torque	Atom, quantum
Physics:mechanics:energy	Energy, force, distance, work, potential, kinetic, gravitational, gravity	Entropy, charge, atom, quantum, electric
Physics:modern:quantum	Spin, quantum, level, energy, black, body, bohr, Heisenberg, atom, atoms, electron, electrons, wave, state, states, uncertainty, spectrum, line, photon, emission, absorption, emitted	Isotope, nuclear, neutron, neutrons, compound

lists of keywords that could be used to classify the assets on the next levels of the taxonomy. Besides establishing the aisles and constructing the patterns for course topic determination, this is the third step that required human input. In our simplistic analogy, the system can identify that {yoghurt, cheese, milk} go together, and even that it is likely an aisle in the food section (based on course association), but not that it is talking about dairy products. For example, in chemistry, we found 1,013 unique keywords associated with the 32,858 resources. While the most frequently used keywords [in the case of chemistry, “reaction” (assigned 2,516 times) and “solution” (assigned 2,032 times)] provided no basis for further distinction, less frequently used keywords allowed distinctions [e.g. “hybridization” (assigned 211 times)]. It soon became clear that in addition to lists of keywords indicating a certain subtopic, we would also need a list of counter-indicating (“veto”) words that indicate that a resource would not be in a certain subtopic. These counter-indicating keywords frequently were used to indicate that an asset actually belongs to a more advanced subtopic than the remainder of the keywords would suggest. Figure 5 illustrates the whole process up to this stage, and Table 3 shows an excerpt of the list thus compiled. A good example of a counter-indicative keyword is “quantum” in the entry for “physics : mechanics : angularmomentum”—quantum physics uses the same terminology of “angular momentum” as classical mechanics, yet is definitely a more advanced topic.

### 3.1.3 Second- and third-level taxonomy term assignment based on keyword heuristics

The second- and third-level taxonomy turned out to require multiple sources:

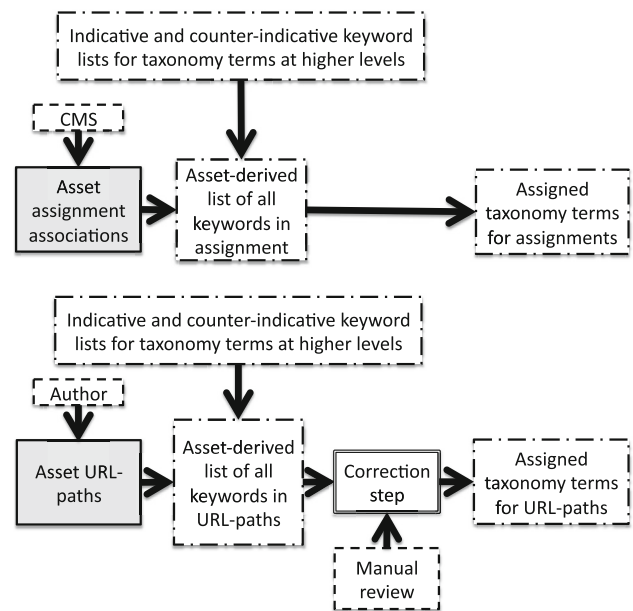
- The keyword list immediately associated with the asset.
- For problems: the assignment that the asset was part of. Since most homework assignments are on a particular topic, it can be surmised that all problems found within the same assignment should have the same taxonomy label.
- The URL path of the asset. This was an unexpected source, but as Fig. 4 indicates, authors frequently sort their assets by subtopic into the filesystem. For example, the URL path “Biology I Diversity, Ecology, Behavior / HW2\_Genetics / Q04\_incomplete\_dominance. problem” at the bottom of the figure contains valuable taxonomy data, and it can also be surmised that all assets found within the same subdirectory should have the same taxonomy label.
- Manual revision of URL path associations.

The keyword lists immediately associated with the assets were analyzed first, going back to the original static metadata. Using the scheme keyword list described in Sect. 3.1.2, we attempted to classify the assets directly. For each

taxonomy assignment, we calculated a simple “agreement index”, which was the difference between the number of found indicative keywords and the number of found counter-indicative keywords. We assigned the taxonomy term or terms with the highest “agreement index”, but demanded a minimum index of two (i.e., there had to be at least two more indicative than counter-indicative keywords). Not surprisingly, due to the poor quality of the individual keyword lists, this method yielded meager results: 433,032 assets were not classified at all, 8,037 had one taxonomy label, 3,077 had two labels, and 1,980 assets had more labels. The “agreement index” was stored alongside the assigned taxonomy terms.

We also compiled the keywords from all problems in a problem set (e.g., a homework assignment), and then assigned a taxonomy label to the problem set as a whole using the above algorithm. 17,178 of the 73,634 homework assignments did not receive a taxonomy label, but 38,786 assignments received exactly one taxonomy label, 9,502 received two taxonomy labels, and the remainder more than two labels. We then used the assignment taxonomies to “vote” on the taxonomies of its constituent problems: 369,151 assets did not receive a taxonomy label (this includes non-problems), 32,423 of the 198,020 problems received one taxonomy label, 16,280 problems received two labels, and 28,273 problems received more than two labels. The number of problem sets (homework assignments) on which the problem had a particular taxonomy label was stored alongside as a confidence measure. For example, a particular problem may have been on six problem sets that were classified as “physics:mechanics:lineardynamics” and on two problem sets that were classified “Physics:Mechanics:Linearkinematics”.

A third effort focused on the URL path of the assets, assuming that authors would sort their assets approximately according to a scheme approximating a taxonomy (see Fig. 4). The same keyword-file as in the previous efforts was run over the URL paths of the assets in an effort to determine a taxonomy at every level of the URL path. In the above example, “Biology I Diversity, Ecology, Behavior / HW2\_Genetics / Q04\_incomplete\_dominance.problem”, the labels “biology” and “ecology” would be attached to every asset within and underneath the subdirectory “Biology I Diversity, Ecology, Behavior”, while “biology:genetics” would be attached to every asset within and underneath the subdirectory “Biology I Diversity, Ecology, Behavior / HW2\_Genetics”. In addition, a limited number of URLs were manually reviewed, where a script queried likely taxonomies for different levels of the URL path from the human reviewer, thus establishing a data structure that would complement or override the results of the directory-based effort. As a result, 219,484 assets were classified according to their URL path. One might argue that the URL path is not truly latent, since after all it was assignment by the author. However, the author



**Fig. 6** Process that led to the taxonomy terms for problem sets (assignments) and URL paths

did not provide this information with the intent of providing metadata, it was merely established as a byproduct of the authoring process (Figs. 6, 7).

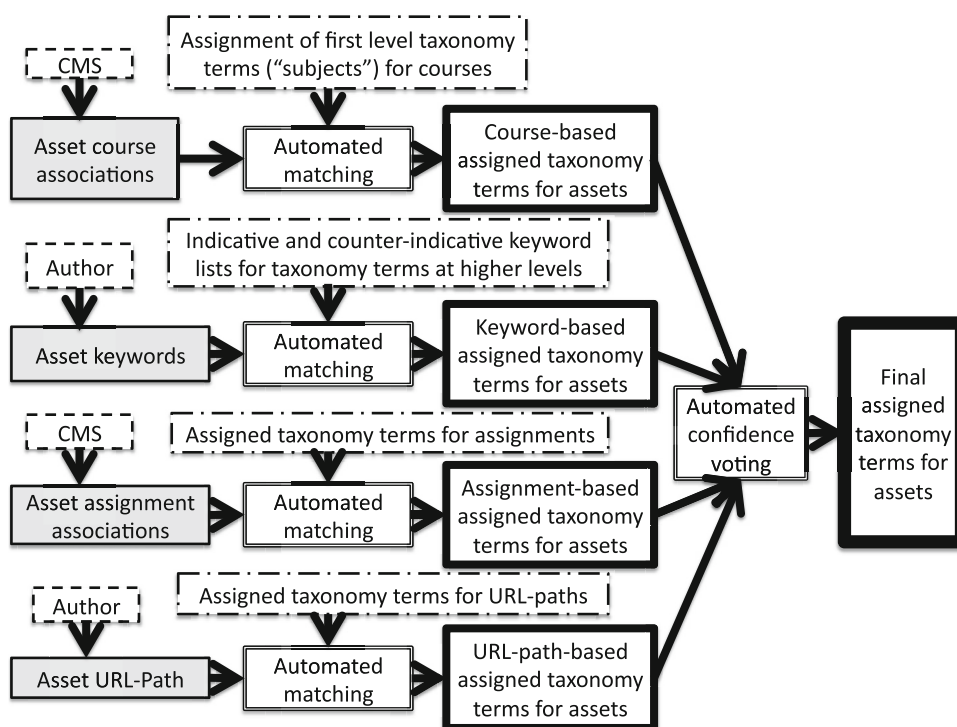
Finally, the results of all of the above mechanisms were combined using a confidence voting mechanism. In the end, 120,972 assets still remained unclassified, but 321,256 assets had one taxonomy, and 3899 assets had two taxonomies; no assets with three taxonomies were left. Of the assigned taxonomies, 233,800 were first-level, 35,325 were second-level, and 59,929 were third-level taxonomies.

### 3.1.4 Relative effectiveness of methods for automated taxonomy term assignments

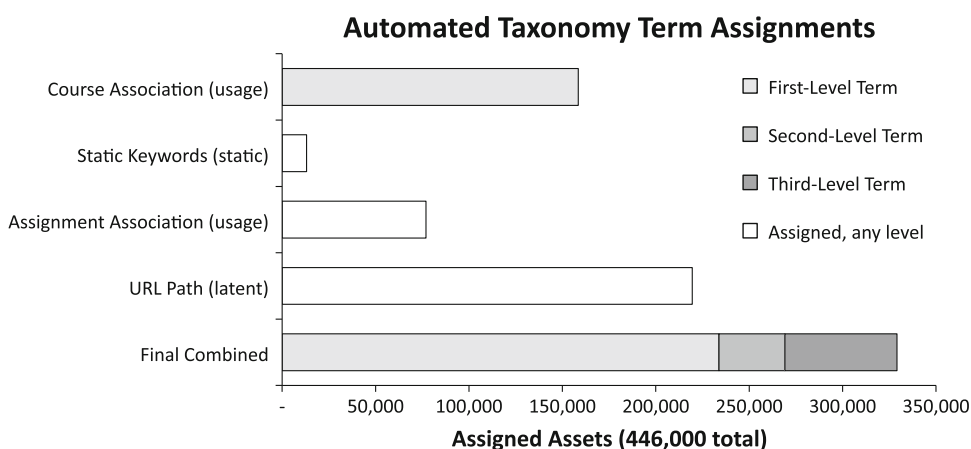
As pointed out, the construction of the taxonomy was essentially a bootstrapping process, where one step built upon another. It still makes sense to consider relative effectiveness, and Fig. 8 shows the yields from the different assignment mechanisms. The simple usage-based classification based on course association (Sect. 3.1.1) already went a long way toward establishing the first-level taxonomy. Of the methods to establish higher-level taxonomies based on keyword taxonomies (Sect. 3.1.3), not surprisingly for this case study, the method based purely on static keywords of individual assets yielded very little classifications, while the usage and latent data contributed much more strongly. The eventual classification, based on a confidence-weighted combination of these methods, finally yielded satisfactory results.

Since we know more about assessment problems than about any other content type, not surprisingly their classi-

**Fig. 7** Process that led to the final combined taxonomy terms assigned to the assets



**Fig. 8** Automated taxonomy assignments at different stages of the process illustrated in Fig. 7



fication was also more successful. While of the assets that could only be classified at the first-level, 26 % are problems, of the assets that received second-level and third-level taxonomies, 83 % are problems. Given the minimal reliance on author assigned keywords, the process underlined the importance of usage and latent data in the classification process.

### 3.2 Association data

Having established the aisles of our supermarket, we are now turning toward educating the clerk. This is the point where our simplistic supermarket analogy and the historical roots of the data mining techniques we are using overlap, as so-

called “market basket analysis” was arguably one of their first notable applications. At the point of sale, analyzing the basket contents of shoppers, this technique collects so-called frequent item-sets, e.g., that bread is frequently bought at the same time as peanut butter and jelly, all without “knowing” anything else about either of these items. In a physical store, this analysis allows adjusting the layout of the store to move sets of frequently linked items into physical proximity: customers who buy the one product see another product that they are likely to desire at that time. In an online store, items have no physical location, and the store can provide these kinds of recommendations much more effectively. In both scenarios, customers may discover desirable products, things they did not even know existed. In the LON-CAPA

asset pool, the goal is to help instructors discover assets that other instructors have used effectively in the same context.

Asset assembly and sequencing happen within the system itself. It is thus possible to capture usage data for each asset, including the information which modules, chapters or assignments it has been used in. These content units are the equivalent of a market basket. In reality, unfortunately, this data is only readily available for the homework assets, as homework transactions are logged more comprehensively. Frequent item-sets were mined from the de-identified transaction data of 138 million homework transactions and led to 2.8 million weighted association rules between assessment assets. Each of these rules is of the type “asset *A* has been used *n* times in the same context as asset *B*”.

At first glimpse, it seems disappointing that there are *only* 2.8 million rules, as there could in principle be approximately 20 billion associations between the 198,020 homework problems. However, it is very reasonable to assume that the vast majority of these relationships should legitimately be zero: if these rules are any good, these first-order rules should show zero relationship between an asset in cell biology and an asset in high energy physics. In principle, one could combine frequent item-sets into second- and third-order relationships of the type “asset *A* has been used *n* times in connection with *B*, which in turn has been used *m* times in connection with *C*”, i.e., construct frequent 2-item-sets or even frequent 3-item-sets, but we did not consider those yet, as the first-order relationships already yielded satisfactory results.

### 3.3 Item response theory and time-on-task measurements

LON-CAPA already routinely gathers difficulty and discrimination data whenever an instructor calculates descriptive course statistics and stores this data as dynamic metadata alongside the resources. However, transaction data is much richer than that. We have begun to calculate item response theory (IRT) [19] parameters based on the raw de-identified transaction data, and found that these are a rich source of analytics. These items, alongside confidence measures (“error bars”), can be used as yet another quality and selection criterion as faculty assemble course materials. Also, time-on-task information can be gathered from the transaction logs, with the usual caveats applicable due to “multi-tasking” and guessing behavior of the learners [20]. Finally, all of these statistics will be noisy due to copying and cheating [21], so confidence measures are essential.

## 4 A recommender system

Combining the extracted taxonomy with search and browse mechanisms, as well as association data, led to the construction of a prototype recommender system for LON-CAPA.

This system is targeted toward the instructor who desires to select and sequence assets for his or her students. Figure 9 shows a screenshot of this system, as it would be called if the faculty member clicks “Import” from inside a module, chapter or assignment of their course. The general functionality is similar to that of many online bookstores, users can browse (using either URL tree of taxonomy hierarchies) or search, apply filters, get ranked lists of assets to choose from, and get recommendations based on their cart, the asset they are viewing, or the current position in hierarchical trees.

The top row in Fig. 9 includes a standard search bar, as well as the extracted taxonomy and a content type filter (e.g., image, page, problem, etc.). By default, only the top-level taxonomies are shown, but as the user selects a branch, deeper levels are displayed. The top bar also contains the link to the “shopping cart”, which, combined with the assets that are already in the module that the Recommender was called from, forms the base for the subsequent rankings and recommendations.

The remainder of the Recommender screen then shows applicable resources, sorted by association. The listing here is contextually generated based on searches, taxonomies, and cart/module associations. Items that the user picks get added to the cart. From here, the user can also directly jump into a particular taxonomy level (third column), or into a particular author directory (fourth column).

Once the user is done with picking assets, the contents of the cart get added to the current module, chapter, or assignment.

Essential for the success of the Recommender is the combination of current context (the position in the course where the instructor presses “Import”) and past usage of the assets by other users, once again enabled by combining the digital library, sequencing, and course management functionality (Fig. 2).

We are evaluating two factors when making recommendations for potential new assets:

**Number of associations:** To calculate this number, we first form the superset of the assets in the current folder (into which the instructor wishes to import) in the current course and the current content of the “shopping cart”. We then count how often the potential new asset appears in the same folders as any of these assets in any of the courses network-wide. This factor evaluates the association of the potential new asset with the current context.

**Number of Uses:** Number of users (students or instructors) who interacted with the potential new asset, establishing reliability.

For ranking purposes, we then established a “quality function”



**Fig. 9** Prototype of the LON-CAPA recommender system. In this example, the instructor was looking for problems (see filter setting) in the context of kinematics

Search

accounting (640) advertising (2,000) astronomy (6,000) biochemistry (2,200) biology (50,000) botany (89,000) chemistry (36,000) computerscience (1,200) design (450) ecology (1,700) engineering (2,200) finance (1,600) geology (2,400) geometry (550) history (260) languages (1,3) mathematics (12,000) medicine (5,100) nursing (8) nutrition (590) philosophy (1,0) physics (110,000) psychology (530) statistics (3,100) zoology (220)

Filter Lists Problems

Checkout  
Your cart (8 item(s))

Recommendations

Next

Add	msu-prob16.problem	physics/mechanics/linear/kinematics	/ msu / physicslib / msuphysicslib / 05_1D_Motion / msu-prob16.problem
Add	msu-prob54.problem	physics/mechanics/linear/kinematics	/ msu / physicslib / msuphysicslib / 05_1D_Motion / msu-prob54.problem
Add	msu-prob55.problem	physics/mechanics/linear/kinematics	/ msu / physicslib / msuphysicslib / 05_1D_Motion / msu-prob55.problem
Add	msu-prob38.problem	physics/mechanics/linear/kinematics	/ msu / physicslib / msuphysicslib / 08_2D_Motion_and_Motion_in_a_Circle / msu-prob38.problem
Add	msu-prob40.problem	physics/mechanics/linear/kinematics	/ msu / physicslib / msuphysicslib / 05_1D_Motion / msu-prob40.problem
Add	msu-prob38.problem	physics/mechanics/linear/kinematics	/ msu / physicslib / msuphysicslib / 05_1D_Motion / msu-prob38.problem
Add	msu-prob53.problem	physics/mechanics/linear/kinematics	/ msu / physicslib / msuphysicslib / 05_1D_Motion / msu-prob53.problem
Add	msu-prob01.problem	physics/introduction/mathematics	/ msu / physicslib / msuphysicslib / 07_Vector_Calculus / msu-prob01.problem
Add	msu-prob10.problem	physics/mechanics	/ msu / physicslib / msuphysicslib / 06_Vectors_Scalars / msu-prob10.problem
Add	msu-prob09.problem	physics	/ msu / physicslib / msuphysicslib / 01_Math_1 / msu-prob09.problem
Add	msu-prob33.problem	physics/mechanics/linear/kinematics	/ msu / physicslib / msuphysicslib / 05_1D_Motion / msu-prob33.problem
Add	SpeedTimeHist.problem	physics/mechanics/linear/kinematics	/ msu / kashy / physicslib02 / 05_1D_Motion / SpeedTimeHist.problem
Add	msu-prob16.problem	physics/mechanics	/ msu / physicslib / msuphysicslib / 03_Units_Scaling / msu-prob16.problem
Add	msu-prob46.problem	physics/mechanics/linear/kinematics	/ msu / physicslib / msuphysicslib / 05_1D_Motion / msu-prob46.problem
Add	msu-prob11.problem	physics/mechanics/linear/momentum	/ msu / physicslib / msuphysicslib / 16_Momentum / msu-prob11.problem
Add	msu-prob03.problem	physics/mechanics/linear/dynamics	/ msu / physicslib / msuphysicslib / 09_Force_and_Motion / msu-prob03.problem
Add	msu-prob43.problem	physics/mechanics/linear/kinematics	/ msu / physicslib / msuphysicslib / 08_2D_Motion_and_Motion_in_a_Circle / msu-prob43.problem
Add	AccelerationHist.problem	physics/mechanics/linear/kinematics	/ msu / kashy / physicslib02 / 05_1D_Motion / AccelerationHist.problem
Add	msu-prob44.problem	physics/mechanics	/ msu / physicslib / msuphysicslib / 08_2D_Motion_and_Motion_in_a_Circle / msu-prob44.problem
Add	msu-prob08.problem	physics	/ msu / physicslib / msuphysicslib / 07_Vector_Calculus / msu-prob08.problem

Next

$$\text{Quality} = \frac{\text{Number of associations}}{\text{Maximum number of associations in list}} + \frac{\text{Number of uses}}{\text{Maximum number of uses in list}} \quad (1)$$

where the respective “Max.Number” values refer to the highest numbers appearing in the list, and thus the quality function varies between 0 and 2. Items with a higher quality number appear higher on the list of recommendations.

## 5 Evaluation

In order to evaluate the quality of the taxonomy assignments and the recommender tool, users were asked to use the prototype (installed on one server in the network) instead of the normal import and search functionality that is built into the current release of LON-CAPA. They were also asked to approve or modify the taxonomy assignments of assets.

### 5.1 Taxonomy

Users evaluated 348 taxonomy assignments, which is about a tenth of a percent of the total assignments (see Fig. 8), and were able to indicate agreement or suggest modifications. In particular, choices were:

Agree: Accept assignment as given.

Extend: Accept the current assignment as correct, but add one or two additional taxonomy terms, thus making the assignment more specific.

Change: Change the current assignment at any level. Here, a change at the first level is the most sweeping change, indicating a wrong subject assignment, while a change at the third level merely corresponds to the item being sorted into the wrong week of a course.

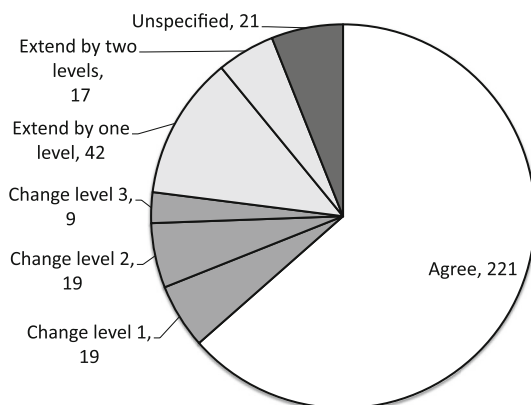
Table 4 shows examples of taxonomy changes made by users of the prototype (first versus second column), as well as how we classified this change. Some changes have obvious reasons: for example, the change from “geometry” to “mathematics:calculus:series” is due to the fact that the asset dealt with the geometric series and was thus mislabeled by the automated assignment for rather trivial reasons. Some changes are more surprising at first, for example “physics” to “geology:waterresources”. As it turns out, the affected asset was part of a course on renewable energy and sustainability that was offered by a physicist.

Figure 10 shows the result of this user evaluation. 221 of the taxonomy assignments were accepted by the users, and 59 extended—meaning, in this user evaluation, 280 of 348 (80 %) of the assets were correctly classified at some level. 47 of the taxonomy assignments were modified at a particular level, where a completely wrong assignment was indicated for 19 resources, which had a wrong first-level subject assignment (5.4 % of the assets). 21 assets underwent more sweeping but possibly accidental changes in their taxonomy assignment; for example, users eliminated taxonomy assignments where more than one set of terms was assigned or they submitted an empty new taxonomy assignment (indicating disagreement with the current assignment, yet confused about how to submit a new one).



**Table 4** Examples of user corrections and extensions to generated taxonomy assignments

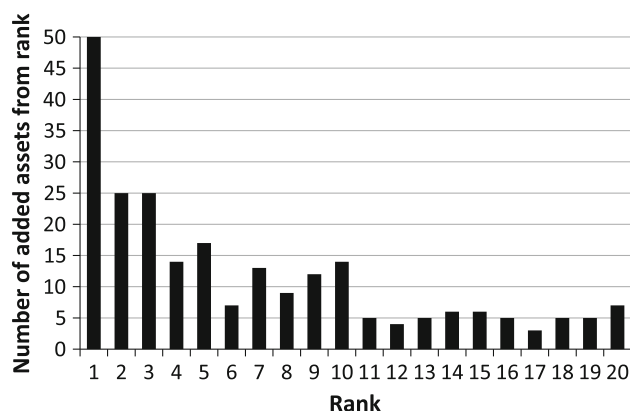
Old assignment	New assignment	Type
Biology	Biology:cells:metabolism	Extend by two levels
Biology	Biology:cells:photosynthesis	Extend by two levels
Chemistry	Chemistry:bonding	Extend by one level
Chemistry	Chemistry:introduction	Extend by one level
Chemistry	Chemistry:reactions:electrochem	Extend by two levels
Chemistry:introduction:states	Chemistry:introduction:compounds	Change level three
Chemistry:introduction:states	Chemistry:introduction:massconservation	Change level three
Chemistry:matter	Chemistry:introduction:periodictable	Change level two
Chemistry:matter:nuclear	Chemistry:introduction:measurement	Change level two
Chemistry:matter:shell	Chemistry:bonding	Change level two
Chemistry:matter:shell	Physics:modern:quantum	Change level one
Chemistry:organic	Chemistry:organic:alkenes	Extend by one level
Chemistry:organic:alkanes	Chemistry:organic	Change level three
Design	Design:colors	Extend by one level
Geometry	Mathematics:calculus:integrals	Change level one
Geometry	Mathematics:calculus:series	Change level one
Mathematics:calculus	Mathematics:calculus:derivatives	Extend by one level
Mathematics:calculus:derivatives	Mathematics:functions:trigonometric	Change level two
Mathematics:numbers:classes	Mathematics:calculus	Change level two
Physics	Biology:cells:photosynthesis	Change level one
Physics	Chemistry:reactions:redox	Change level one
Physics	Geology:waterresources	Change level one
Physics	Physics:electromagnetism:emwaves	Extend by two levels
Physics	Physics:modern:nuclear	Extend by two levels
Physics	Physics:modern:relativity	Extend by two levels
Physics:electromagnetism	Physics:electromagnetism:capacitance	Extend by one level
Physics:electromagnetism	Physics:electromagnetism:potentials	Extend by one level
Physics:mechanics	Physics:mechanics:lineardynamics	Extend by one level
Physics:mechanics	Physics:mechanics:linearmomentum	Extend by one level
Physics:mechanics	Physics:mechanics:rotationaldynamics	Extend by one level
Physics:mechanics	Physics:mechanics:rotationalkinematics	Extend by one level
Physics:mechanics:rotationalkinematics	Physics:mechanics:linearkinematics	Change level three
Physics:modern	Physics:modern:quantum	Extend by one level

**Fig. 10** User evaluation of generated taxonomy assignments

## 5.2 Recommendations

Users added 269 assets to courses using the Recommender prototype. An important indicator of the quality of recommendations is how far down the ranked list of suggested assets the items appeared that ended up getting picked. Figure 11 shows a histogram of picked assets versus their ranked position. 100 of the 269 picked assets (37 %) appeared in the first three positions of the list of recommended assets (see Fig. 9), while 32 assets (12 %) appeared much further down the list beyond position 20 or during “browse” of user directories (which appear as unranked directory listings).

In written feedback, a user indicated that the Recommender “does a very good job of linking similar topics



**Fig. 11** Histogram of number of added assets versus rank at which they appeared in the recommender listing

together”. Another user stated that he found the “browse by topic” (i.e., ranked recommendations based on taxonomy terms) a lot more useful than keyword searches (however, this user gave “energy” as an example for a less-than-useful keyword search, which is not surprising, given the broad applicability of the energy concept across STEM courses). While we did not state how the Recommender decides on ranking, yet another user figured it out, stating “it was apparently looking at the folder I was in, the topics that were in there and already making suggestions based on the content in my folder”. Overall, we feel encouraged that context-based recommendation, looking at folder-by-folder usage by other instructors, is a promising way to manage large educational assets pools.

## 6 Limitations

This case study was limited to getting to know the assets and relied on 10 years worth of usage data. An area of concern is the incorporation of new assets into the library: as initially nothing would be known about these items, a recommender system like the one presented would systematically favor older items. How does the user discover new items, which may be better than existing ones?

A possible answer would be a “New for You” feature, which would present an instructor with potentially interesting new assets. These recommendations would be based on the instructor, not the course context. Indeed, earlier studies found that there are distinct communities of practice among the instructors, which can be identified from usage data [22]. Thus, the system can also “get to know”, the users and make recommendations based on the user.

## 7 Outlook

Future studies will need to be based on broader usage, as the tool is integrated into the LON-CAPA releases. Particularly

an ongoing evaluation of the data presented in Fig. 11 can lead to an optimization of Eq. 1, where the currently used or additional terms may enter with different weighted coefficients. User suggestions on improved taxonomy assignments need to be evaluated and assignments recalculated based on corrections.

The dynamic metadata presented in this case study is for the most part constructed based on harvested data, i.e., data gathered at one particular point in time and then asynchronously analyzed. This limited approach is due in part to the exploratory nature of our investigation, but also reflects the fact that the LON-CAPA system was never built to accommodate a constant flow of usage data and its continuous analysis. In a next generation version of such a system, one would choose a data model which from the start facilitates a dynamic and near-realtime provision of such quality and selection information for all assets.

The techniques presented in this case study should be combined with the results of earlier studies to construct a compressive recommendation system, which would make the resulting educational library clearly far more attractive than searches on the “open web” and traditional course management solutions. We believe that a fully integrated system like the one presented in Fig. 2, i.e., a comprehensive cross-institutional learning content management system, will eventually replace scattered isolated digital libraries, insular course management systems, and the current e-text platforms. On the base of this technology, an economy for granular learning content can be established, in which faculty can make guided choices when assembling dynamic online course packs for their students.

## 8 Conclusion

In this case study, we found that the integrated nature of systems like LON-CAPA allows automatically constructing large amounts of useful metadata based on latent organizational features and asset usage. Over time, due to usage in courses and particular modules, chapters, and assignments, the library “gets to know” the assets, even if static metadata is sparse, incomplete, or erroneous. Based on this metadata, we were able to construct a prototype of a recommender system, which allows multiple access routes to the materials beyond basic searches and browsing. Initial user testing of the dynamic metadata and recommendation ranking yielded promising results.

**Acknowledgments** We are grateful that this work was supported by NSF grant DUE-1044294, which is not responsible for the content and conclusions of this study. S. D. has been partially supported by BMBF grants 01PL11059 and 01PL11066H. We are grateful to Jonathan Abbott, Daniel Seaton, and Peter Riegler for assistance with this work. We are also grateful to the LON-CAPA users who ventured

out and tested the recommender prototype. Finally, we would like to thank the anonymous reviewers of this manuscript for their helpful and constructive criticism.

## References

1. Recker, M.M., Wiley, D.A.: A non-authoritative educational metadata ontology for filtering and recommending learning objects. *J. Interact. Learn. Environ.* **9**(3), 255–271 (2001)
2. Kortemeyer, G.: Ten years later: why open educational resources have not noticeably affected higher education, and why we should care. *EDUCAUSE Review*, Online, 02/26 (2013)
3. Park, N., Roman, R., Lee, S., Chung, J.E.: User acceptance of a digital library system in developing countries: an application of the technology acceptance model. *Int. J. Inf. Manag.* **29**(3), 196–209 (2009)
4. Turner, M., Kitchenham, B., Brereton, P., Charters, S., Budgen, D.: Does the technology acceptance model predict actual use? A systematic literature review. *Inf. Softw. Technol.* **52**(5), 463–479 (2010)
5. Jeng, J.: What is usability in the context of the digital library and how can it be measured? *Inf. Technol. Libr.* **24**(2) (2005)
6. Buchanan, S., Salako, A.: Evaluating the usability and usefulness of a digital library. *Libr. Rev.* **58**(9), 638–651 (2009)
7. Bethard, S., Wetzler, P., Butcher, K., Martin, J.H., Sumner, T.: Automatically characterizing resource quality for educational digital libraries. *JCDL '09: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*. pp. 221–230 (2009)
8. Strohmaier, M., Helic, D., Benz, D., Körner, C., Kern, R.: Evaluation of folksonomy induction algorithms. *ACM Trans. Intell. Syst. Technol.* doi:[10.1145/2337542.2337559](https://doi.org/10.1145/2337542.2337559) (2012)
9. Lau, S.B. -Y., Lee, C.-S., Singh, Y.P.: A folksonomy-based lightweight resource annotation metadata schema for personalized hypermedia learning resource delivery. *Inter. Learn. Environ.* doi:[10.1080/10494820.2012.745429](https://doi.org/10.1080/10494820.2012.745429) (2012)
10. Zervas, P., Sampson, D.G.: Computers in human behavior. doi:[10.1016/j.chb.2013.06.026](https://doi.org/10.1016/j.chb.2013.06.026) (2013)
11. Mimi, R., Leary, H., Walker, A., Diekema, A., Wetzler, P., Sumner, Y., Martin, J.: Modeling teacher ratings of online resources: a human-machine approach to quality. Paper presented at the American Educational Research Association annual meeting, New Orleans (2011)
12. Klebl, M., Krämer, B.: Distributed repositories for educational content-Part 1: Information Management for Educational Content. eled, vol. 7, 2010 (urn:nbn:de:0009–5-27716)
13. Klebl, M., Krämer, B., Annett, Z., Matthias, H., Christian, L.: Distributed repositories for educational content-Part 2: Technology. eled, vol. 7, 2010 (urn:nbn:de:0009–5-27748)
14. Kortemeyer, G., Kashy, E., Benenson, W., Bauer, W.: Experiences using the open-source learning content management and assessment system LON-CAPA in introductory physics courses. *Am. J. Phys.* **76**, 438–444 (2008)
15. Duval, E.: Standardized metadata for education: a status report. In: *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications (EDMEDIA)* (2001)
16. Duval, E., Hodgins, W., Stuart, S., Stuart, W.: Metadata principles and practicalities. *D-Lib Magazine* **8**(4), (2002)
17. Koutsomitropoulos, D.A., Alexopoulos, A.D., Solomou, G.D., Papatheodorou, T.S.: The use of metadata for educational resources in digital repositories: practices and perspectives. *D-Lib Magazine* **16**(1/2) (2010)
18. <http://dublincore.org/documents/dces/>. Accessed November 2012
19. Bergner, Y., Dröschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., Pritchard, D.: Model-based collaborative filtering analysis of student response data: machine-learning item response theory. In: *Proceedings of the 5th International Conference on Educational Data Mining* pp. 95–102 (2012)
20. Kortemeyer, G.: Gender differences in the use of an online homework system in an introductory physics course. *Phys. Rev. ST Phys. Educ. Res.* **5**, 010107 (8 pages) (2009)
21. Palazzo, D.J., Lee, Y.-J., Warnakulasooriya, R., Pritchard, D.E.: Patterns, correlates, and reduction of homework copying. *Phys. Rev. ST Phys. Educ. Res.* **6**, 010104 (12 pages) (2010)
22. Han, P., Kortemeyer, G., Krämer, B.J., von Prümmer, C.: Exposure and support of latent social networks among learning object repository users. *J. Univ. Comput. Sci. (J.UCS)*, **14**(10) (2008)