Daddy, How do I Find a Book on Purple Frogs?
Representation Issues for Children and Youth

Abstract
Subject access and controlled vocabularies used for representing children’s resources are examined. Wittgenstein’s Language Games theory is presented as a possible framework for determining sources to use for controlled vocabulary construction. Preliminary results regarding existing controlled vocabulary use in databases and digital libraries designed for children are presented, as are extent of match findings comparing users’ search terms with LCSH, Sears, and a subject-specific proprietary controlled vocabulary are reported.

Introduction
A key challenge to retrieval in any type of system is how to represent the resources appropriately so that the user(s) can find what they are looking for. In systems being used by or designed for children and youth\(^1\), there exist two fundamental representation problems: (1) the metadata or representation scheme of the system may not be designed with this specific user group in mind, and (2) few age-appropriate controlled vocabularies exist for use in creating metadata.

Research has shown that children and youth have unique information needs and information seeking strategies (for an overview see Abbas (2005a); Abbas, (in press); Walter (1994)). In today’s increasingly digital world, children have access to a variety of global resources in many different formats. Children have access via OPACs in libraries, proprietary databases, and specialized collections of resources in digital libraries, subject directories and web portals that are designed specifically for their use. As the Web continues to develop and more Web 2.0 applications are introduced, children will access information in many differing contexts and formats, for example, social sharing sites like YouTube or MySpace where they can search through dynamic content and formats like videos, photos, blogs and wikis. Each of these different systems and information ecologies are organized by different underlying structures and standards, or lack thereof.

In order to maximize the success of children’s information seeking, the system design and the resources contained within the system must be represented at a level appropriate for this particular group of users. The metadata scheme and the metadata describing the system resources should reflect an understanding of how children access, organize, and use information (Abbas, 2005a; Abbas, 2005b). In social sharing sites, users control the content and the basic structures, though they design their “spaces” using pre-determined sets of available features provided by system designers/programmers.

Representation of resources is a “complex web of attributes of disparate objects and concepts, idiosyncratic and socially constructed codes and agreements, and neurological abilities” (O’Connor, 1996 p. 11) and has been defined using many lenses such as library and information science, cognitive science, and linguistics, among others. Representation \textit{for children} is a complex socio-cognitive process in which many variables come into play: 1) user’s developmental and cognitive state, 2) user’s domain and system knowledge, 3) indexer’s knowledge of the user’s information seeking habits, 4) indexer’s understanding of the intended purpose(s) for the objects, and 5) the controlled vocabu-

\(^1\) For the ease of the reader, the term children will be used throughout to represent all younger users between ages 0–15.
lary used to create representations. The representation process is further complicated by: 1) our incomplete picture of this group of users information seeking and system use, 2) metadata schemes designed for use by adults and not children, 3) the guidelines and tools to construct metadata, and 4) the differing cognitive abilities and system knowledge of children.

Blair (1990, 2006) sees the problem of representation and information retrieval as linguistic in nature. How effectively we utilize language to represent an object, determines the success or failure of the information retrieval process. Blair also posits that the language that we use to represent both our information needs or questions and that are used to index documents is learned in a social context or community. Blair explains the theory of “language games”, as first developed by the early twenty century philosopher Ludwig Wittgenstein, and the process in which we learn language and meaning. We do not acquire language purely by learning the word and its definition, but instead learn its use and appropriateness within the context of our “forms of life” or everyday experiences. Children engage in “language games” as they go through their daily “forms of life” or interactions. Direct influences on their learning are their parents, teachers, the documents with which they engage (books and textbooks, web resources, popular and trade media in many formats such as magazines, movies, social sharing sites (YouTube, MySpace, etc.)), and the information systems they interact with. Learning the appropriate “language” or terminology to use within these contexts is vital to their success both in information retrieval and content understanding.

The tools provided the cataloger/metadata creator obviously also can affect successful retrieval by children. A cataloger must have an age-appropriate controlled vocabulary to use for representing subject content of children’s resources. The most prevalent controlled vocabulary in use in library systems, the Library of Congress Subject Headings (LCSH) contains words that are at mostly above a sixth grade level, thereby making their use inappropriate for younger users. The Annotated Card (AC) Program of the Library of Congress is a subset of LCSH that has been altered to fit more appropriately for use in cataloging children’s resources. The AC also provides guidelines to follow to standardize cataloging for children. The Sears List of Subject Headings is another controlled vocabulary that can be used to create metadata for children’s resources.

Few researchers have explored the issue of developing more age-appropriate controlled vocabularies for children. Abbas (2005b) explored the idea of using children’s frequently used search terms as a source for controlled vocabulary terms. Jannson (as reported by Lundgren, 1998) developed a special thesaurus for children consisting of about 800 simple, concrete words within 21 areas of interest. Pejtersen’s Book House project developed an icon-based interface for an OPAC. What is most interesting about the Book House project is that it does not use a traditional metadata scheme such as MARC. The bibliographic records include additional information such as the emotional effect the book may produce and descriptions and evaluative comments written by 11 and 12 year old children. (Lundgren, 1998)

Of particular interest in the study being reported here is the application of Wittgenstein’s “Language Games” (LG) as a means to develop an age-appropriate controlled vocabulary for children and youth. Using LG as a framework, the study examined: 1) existing controlled vocabularies used to catalog children’s resources; 2) the usefulness of existing controlled vocabularies, 3) identified a set of sources present in LG “forms of life” encountered by children as they learn to use language for interacting and searching;
and 4) compared existing controlled vocabularies with the terminology used within LG sources.

**Methodology**

**Research Questions**

To guide this study a set of research questions were formulated:

1. Which controlled vocabularies are used by catalogers of children’s resources?
2. Which controlled vocabularies are used by systems used by/design for children?
3. Is Wittgenstein’s “Language Games” (LG) theory a useful framework for determining sources to construct a controlled vocabulary for children?
4. Using LG theory, which resources can inform controlled vocabulary development?
   a. Do users’ search terms match controlled vocabulary terms used by the system?
   b. Do users’ search terms match terms used in LG sources?

This paper will report preliminary findings of questions 1, 2, 4, 4a and 4b. The exploration to date serves as a “proof of concept” for examining question 3.

**RQ 1:** Standards for controlled vocabulary use in library catalogs are determined by the library itself, the country’s library governance body, and by the discipline. For example, in the U.S., the Association for Library Collections & Technical Services (ALCTS), Cataloging and Classification Section, Cataloging of Children’s Materials Committee, division of the American Library Association developed a set of guidelines catalogers can follow when cataloging children’s resources. These guidelines establish principles for applying Library of Congress’s Annotated Card (AC) Program headings. They also provide guidance for using other age-appropriate controlled vocabularies, such as Sears List of Subject Headings, or curriculum-specific headings. The guidelines were developed to solve the issue of non-standardized cataloging of children’s resources and to dissuade catalogers who contribute to shared cataloging from adding local subject headings into their records. (ALCTS; Inttner, Fountain, & Gilchrist, 2006)

Unfortunately no complete answer to this research question could be determined by reviewing the literature related to children’s cataloging. Much of the literature is anecdotal in nature and discusses current or best practices or is focused on subject access in general, not specifically for children. Weihs (2004) reports on her “unscientific survey” of the use of Sears List of Subject Headings. She emailed friends and contacts given by friends in the U.S. and Canada (no sampling frame or size given), and asked them: 1) if they knew the difference between Sears and LCSH; and 2) whether or not they used Sears List of Subject Headings or some other controlled vocabulary. The results revealed: 1) most school librarians did not know which controlled vocabulary was used in their library catalog, 2) many used CIP (Cataloging in Publication) information only, 3) some reported not knowing the difference between Sears and LCSH; others thought LCSH was the only controlled vocabulary; and 4) the librarian does not always make the decision regarding the source of cataloging information, this is often determined by the school board. Weihs also notes a personal communication with the editor of Sears, Joseph Miller, who reported that Sears is used “in countries where there is some library collection or education in English” (2004, p. 6). While the Weihs “unscientific survey” reveals a bit of the picture of U.S. and Canadian understanding and use of Sears and LCSH, it is really not a definitive answer to the research question posed.
Other useful international efforts related to subject access issues in general have been conducted or are in process. These studies, however, do not focus on subject access for children’s works. Heiner-Freiling (2000) presents findings of an IFLA Section on Classification and Indexing survey. The goal of the study was to determine subject heading languages used in national bibliographies or catalogs of national libraries. Responses from 88 national libraries or national bibliographic agencies were received. The responses were from 37 European countries; 18 African countries; 15 Asian countries; 15 countries in the Americas; and from 3 other countries (Australia, New Zealand, Papua New Guinea). The results show: 1) twenty-four national libraries use LCSH and twelve countries use translations or adaptations of LCSH; 2) other subject heading lists including RAMEAU and the German subject heading system RSWK (Regeln für den Schlagwortkatalog) are gaining prominence in the countries sharing French or German as the common language; and 3) Sears List of Subject Headings as a second English language subject heading system was also shown. It is being used in ten countries, mostly those with English as the first or second language (e.g., the Bermuda Islands, Belize, Barbados, Ethiopia, Zambia, and Bangladesh) and it is being used for cataloging children’s works (Heiner-Freiling, 2000, pp. 191–192). Landry (2007) reports on efforts currently underway by the IFLA Working Group on Guidelines for Subject Access by National Bibliographic Agencies to establish guidelines for subject access by national bibliographic agencies and to develop an indexing policy for use in national bibliographic catalogs. The work of this group may give us more of a complete picture of controlled vocabularies used internationally for subject access, however, it is not noted whether it will examine those used specifically for children’s resources.

Next steps include: 1) review IFLA findings when available, and 2) develop and conduct survey of school and public librarians to determine controlled vocabularies being used.

**RQ 2:** Controlled vocabulary use in online and commercial systems varies. It is difficult to determine which vocabulary a system is using because of the limited documentation on the systems’ web sites and/or the lack of literature containing this information. In many cases Commercial online databases such as Gale Group, EBSCOHost and Proquest use content specific metadata schemes and subject specific controlled vocabularies, such as the ERIC thesauri for educational resources in the ERIC database. Web metadata creators use either generic metadata schemes such as the Dublin Core (DC) to represent resources or they adapt existing metadata schemes such as MARC and DC by adding additional elements (or database fields) that are subject or audience specific. Additionally, specialized metadata schemes are being designed with a particular group of user(s), resources, and use in mind. For example the Gateway to Educational Materials (GEM) metadata scheme has been developed using the DC as a base set of elements. Additional elements were then added to make the scheme more useful to the users of the collection, K-12 educators. Multiple communities with specific users and needs are currently developing metadata schemes for systems and resources designed for adults but few have children as their focus. (Abbas 2005a)

To explore this question, a set of online commercial databases (Ebsco’s Science Reference Center, and Ebsco’s General Science Collection) and one national digital library (National Science Digital Library) NSDL were reviewed to determine which controlled vocabularies were used to represent resources for children. The research at this point
chose to focus on science-related systems in order to narrow the subject areas being explored. In future explorations, other subject areas will be included as well. The following steps were conducted:

1. Emails and phone calls to representatives of the companies were sent but none returned to date.
2. Each site was reviewed for information about the controlled vocabulary used. The proprietary databases included FAQ and search help pages. The site explained the subject authority file as “a controlled vocabulary” [being] “drawn from the subjects assigned to each record” which would lead one to assume it is a controlled vocabulary constructed by terms derived from the resources themselves (natural language). However this assumption could not be verified. When the NSDL was reviewed, it was noted that multiple subject-specific controlled vocabularies are in use. Due to the number of contributors to the NSDL and the scope of the collections included, NSDL determined that it is not feasible to have one all encompassing controlled vocabulary for use in all collections. The Library of Congress Subject Headings was noted but contributors to NSDL continue to use their own controlled vocabularies. NSDL instead provides cataloging best practices for contributors to follow. (NSDL)
3. A sample of 33 search strings by 5th–8th grade users\(^2\) were run in the Ebsco General Science Collection online database. These same terms were also compared to the Library of Congress Subject Headings list (both standard list and Annotated Card) and to Sears List of Subject Headings. The findings are presented in Table 1.

<table>
<thead>
<tr>
<th>Match form</th>
<th>LCSH</th>
<th>Annotated Card(^3)</th>
<th>Sears</th>
<th>Gen. Sci. Col.</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>21 (63%)</td>
<td>2 (6%)</td>
<td>21 (63.6%)</td>
<td>22 (66.6%)</td>
</tr>
<tr>
<td>Derived</td>
<td>5 (15%)</td>
<td>0</td>
<td>2 (6%)</td>
<td>5 (15%)</td>
</tr>
<tr>
<td>Paren. Qual.</td>
<td>3 (9%)</td>
<td>0</td>
<td>2 (6%)</td>
<td>2 (6%)</td>
</tr>
<tr>
<td>USE ref.</td>
<td>2 (6%)</td>
<td>1 (3%)</td>
<td>1 (3%)</td>
<td>1 (3%)</td>
</tr>
<tr>
<td>No term</td>
<td>1 (3%)</td>
<td>30 (90%)</td>
<td>1 (3%)</td>
<td>1 (3%)</td>
</tr>
</tbody>
</table>

It is interesting to note that in all controlled vocabularies used by the systems, with the exception of AC, the majority of the users’ search terms exactly matched those in the controlled vocabulary and only 1 (3%) of the terms did not match any of the controlled vocabularies. In other cases, the search term matched with a derived term in the controlled vocabulary (plural form or extended form) or contained a parenthetical qualifier. Few of the terms (between 3–6%) were unauthorized terms. These findings indicate that for this sample of terms, the controlled vocabulary of the system actually matches to a good degree with those terms used by the searchers. This is an important and positive finding for the existing controlled vocabularies being used by systems (LCSH, Sears, and Ebsco).

\(^2\) The terms used were chosen from a sample of searches run over one school semester in the GoKnow digital library. The terms were the most frequently used search terms used by this set of users, having been used in searches in a range of 50–519 times. For a list of the terms and table of results, please email author at abbasjm@buffalo.edu.

\(^3\) While the print and online versions of the AC did not include many of the terms in the sample, AC guidelines have provisions for either using existing LCSH terms or for adding terms.
Next steps will include: 1) expanding the sample of terms and systems to include non-science related terms, 2) run the queries in the non-science related systems, and 3) follow up with all database vendors for more information on controlled vocabularies.

**RQ 4:** Identifying LG sources to use is the heart of the further study. As mentioned above children encounter many sources in their daily “forms of life” that could inform us about controlled vocabulary development. This preliminary study examined one of those sources, science textbooks used by 8th graders, but further exploration is warranted and more potential sources need to be identified and reviewed. To test the efficacy of this idea, one 8th grade science textbook (used by students ages approximately 10–12) was reviewed. The glossary, index, text of the chapters, and chapter review word banks were examined to determine if the search terms used in RQ #3 appeared as prominent terms in the textbook. Of the 33 search terms, 14 or 42% of the terms appeared in the textbooks in one or many of the areas: 1) 6 terms or 18% were in glossary entries; 2) 12 terms or 36% were in index entries; 3) 4 terms or 12% were in word banks; and the sample terms appeared prominently 65 times in the textbook.

Next steps will include: 1) expanding this idea to more textbooks and new set of non-science terms; and 2) identifying and examining other LG sources to use.

**Conclusion**
The study presents some intriguing findings but also shows areas that need further exploration. Each of the research questions proves to be useful focus areas to explore the issue of children’s representation further. The Language Game theory as a framework for the development of a more-age appropriate controlled vocabulary was shown to be useful, though more sources need to be identified and examined. While this study adds to the knowledge we have regarding subject access in systems designed specifically for use by children and the use of existing controlled vocabularies to catalog children’s resources within databases and OPACs, it also shows the large gaps in the research and literature concerned with this issue.

**References**


