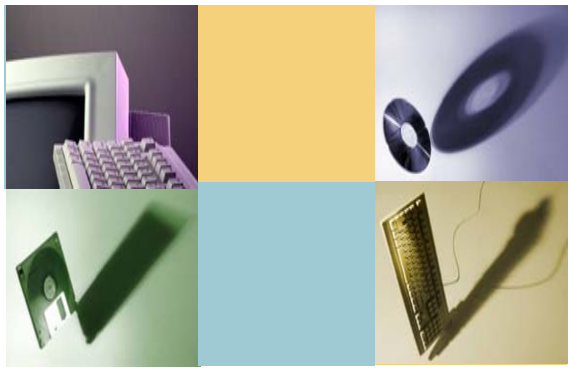
Use of Learning Object Vocabulary in GEM Queries



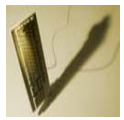
Jian Qin
School of Information Studies
Syracuse University
Syracuse, NY 13244
USA
jqin@syr.edu

F. Javier Calzada Prado
Departamento de Biblioteconomía y Documentación
Universidad Carlos III de Madrid
SPAIN
fcalzada@bib.uc3m.es

Outline

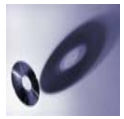
- Research questions
- Data
- Results
  - Term frequencies
  - Query term patterns
  - Similarity to ERIC Thesaurus terms
- What the results imply
- Further study in the near future





Why query log analysis?

- Provide information about the use of DLs to a much larger scale with lower costs
- Understand what queries the user issued, and how many results the system returned according to user's query
- Collect terms through a bottom-up approach for building a vocabulary


Research questions

- To what extent did users use controlled vocabulary in resource discovery?
- What non-controlled vocabulary was used in their resource discovery?
- How can we integrate user query terms into a learning object vocabulary for improving learning object representation and discovery?

Data

- Source: Gateway to Educational Materials (GEM)
- Covers four months (March-May, August) in 2003
- Total number of queries: 411,899
- Preprocessing:
  - Used SQL programs to:
    - Parse query components into separate text strings
    - Cleanse the data
    - Define codes for query fields
    - Code query components
    - Calculate frequencies
  - Created subsets of data for certain query terms:
    - Cleanse, normalize the data (low/upper case, singular/plural, etc.)
    - Categorize query terms
    - Generalize patterns



Data cleansing

Before

10	-pricecode:1 -pricecode:2
10	grade:10 grade:all
10	grade:11 grade:all
10	grade:12 grade:all
10	+"Spanish teaching"
10	+"Spanish Language"
10	grade:9 grade:all
11	|
11	grade:preschool grade:all
11	Writing (composition)
11	"phonics"
11	"writing"
12	Alphabet
13	|
13	grade:preschool grade:all
13	Writing (composition)
13	"alphabet"
13	"writing"

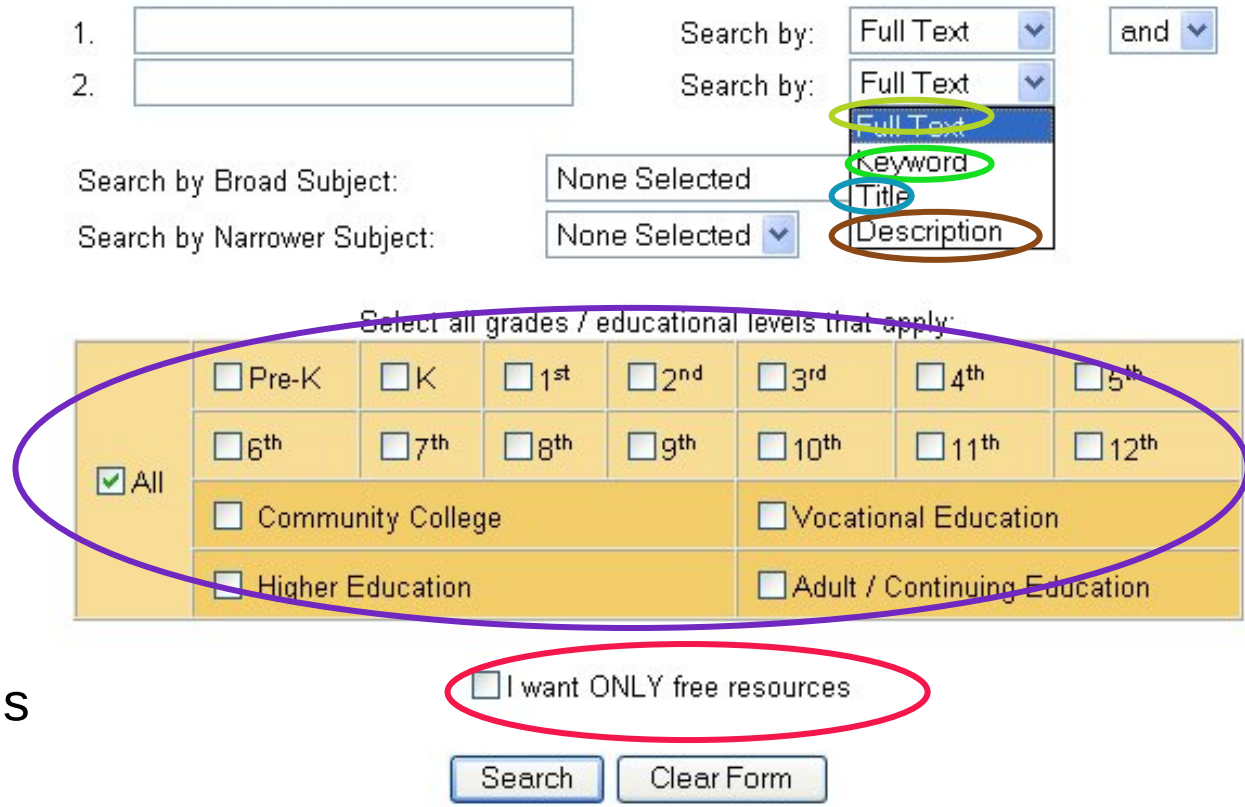
After

10	-pricecode:1 -pricecode:2	235
10	grade:10 grade:all	235
10	grade:11 grade:all	235
10	grade:12 grade:all	235
10	**Spanish teaching**	235
10	**Spanish Language**	235
10	grade:9 grade:all	235
11	grade:preschool grade:all	12347
11	**Writing (composition)**	12347
11	**Phonics**	12347
11	**writing**	12347
12	**Alphabet**	7
13	grade:preschool grade:all	12347
13	**Writing (composition)**	12347
13	**Alphabet**	12347
13	**writing**	12347

1 = Description, 2 = Full Text, 3 = Grade, 4 = Keywords,
5 = Pricecode, 6 = Title, 7 = Topic

Data coding (1)

- Original Query fields
  - 1 = Description
  - 2 = Full Text
  - 3 = Grade
  - 4 = Keywords
  - 5 = Pricecode
  - 6 = Title
  - 7 = Topic
- Combination of any single fields
- Aggregated query fields
  - Any query with field:
    - DE, FT, KW, TI, TO



Data coding (2)

SN	Field	Query component	Field pattern
5	FullText	|	12347
5	Grade	grade:preschool grade:all	12347
5	topic	Writing (composition)	12347
5	keywords	"phonics"	12347
5	description	"writing"	12347
6	Grade	grade:12 grade:all	237
6	Grade	grade:9 grade:all	237
6	Grade	grade:10 grade:all	237
6	Grade	grade:11 grade:all	237
6	topic	Vocabulary	237
6	FullText	+"Spanish teaching"	237
6	FullText	+"Spanish Language"	237
7	Grade	grade:12 grade:all	1347
7	Grade	grade:9 grade:all	1347
7	Grade	grade:10 grade:all	1347
7	Grade	grade:11 grade:all	1347
7	topic	Vocabulary	1347
7	keywords	"Spanish teaching"	1347
7	description	"Spanish Language"	1347

Results

- General description of data
  - Overall distribution of query fields
  - Distribution of number of hits
- Top query fields and terms
  - Original and aggregated query fields
  - Top query fields
  - Top terms
  - Patterns of highly occurring terms
- Similarity to ERIC Thesaurus terms



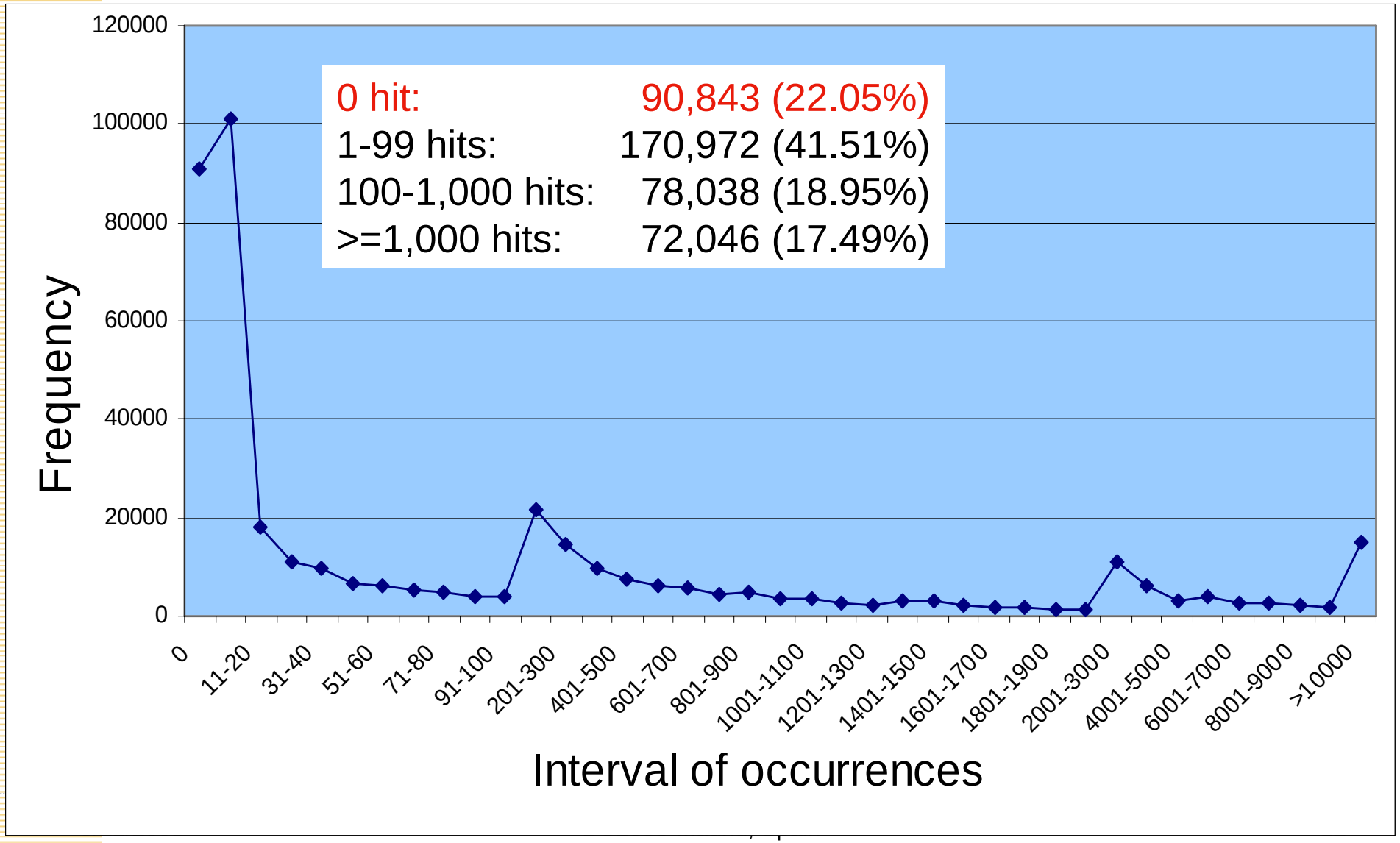
Query term occurrences by search field

Field name	Occurrences	% of total occurrences
Grade	444017	42.53
Fulltext	223579	21.41
Topic	167203	16.01
Pricecode	108136	10.36
Keywords	87330	8.36
Title	7028	0.67
Description	6750	0.65
Total	1044043	100

Distribution of query fields

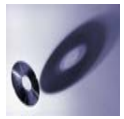
#	Query Field	Occurrences	Percent	Cum Percent
1	Full text/Grade	68781	16.70	16.70
2	Full Text/Keyword	66266	16.09	32.79
3	Topic	52097	12.65	45.44
4	Full text	50267	12.20	57.64
5	Grade/Price/Topic	31268	7.59	65.23
6	Grade/Topic	20759	5.04	70.27
7	Full text/Grade/Price	20101	4.88	75.15
8	Full text/Grade/Price/Topic	19752	4.80	79.95
9	Full text/Grade/Topic	15236	3.70	83.65
10	Grade	8784	2.13	85.78
11	Grade/Price	8541	2.07	87.85
12	Full text/Price	6821	1.66	89.51
13	Full text/Topic	6449	1.57	91.07
14	Full text/Grade/Keyword/Price/Topic	4410	1.07	92.14
15	Price/Topic	3231	0.78	92.93
16	Full text/Grade/Keyword/Topic	3101	0.75	93.68
17	Full text/Price/Topic	2995	0.73	94.41
18	Full text/Grade/Keyword/Price	2785	0.68	95.08

Distribution of number of hits



Top 20 GEM controlled terms

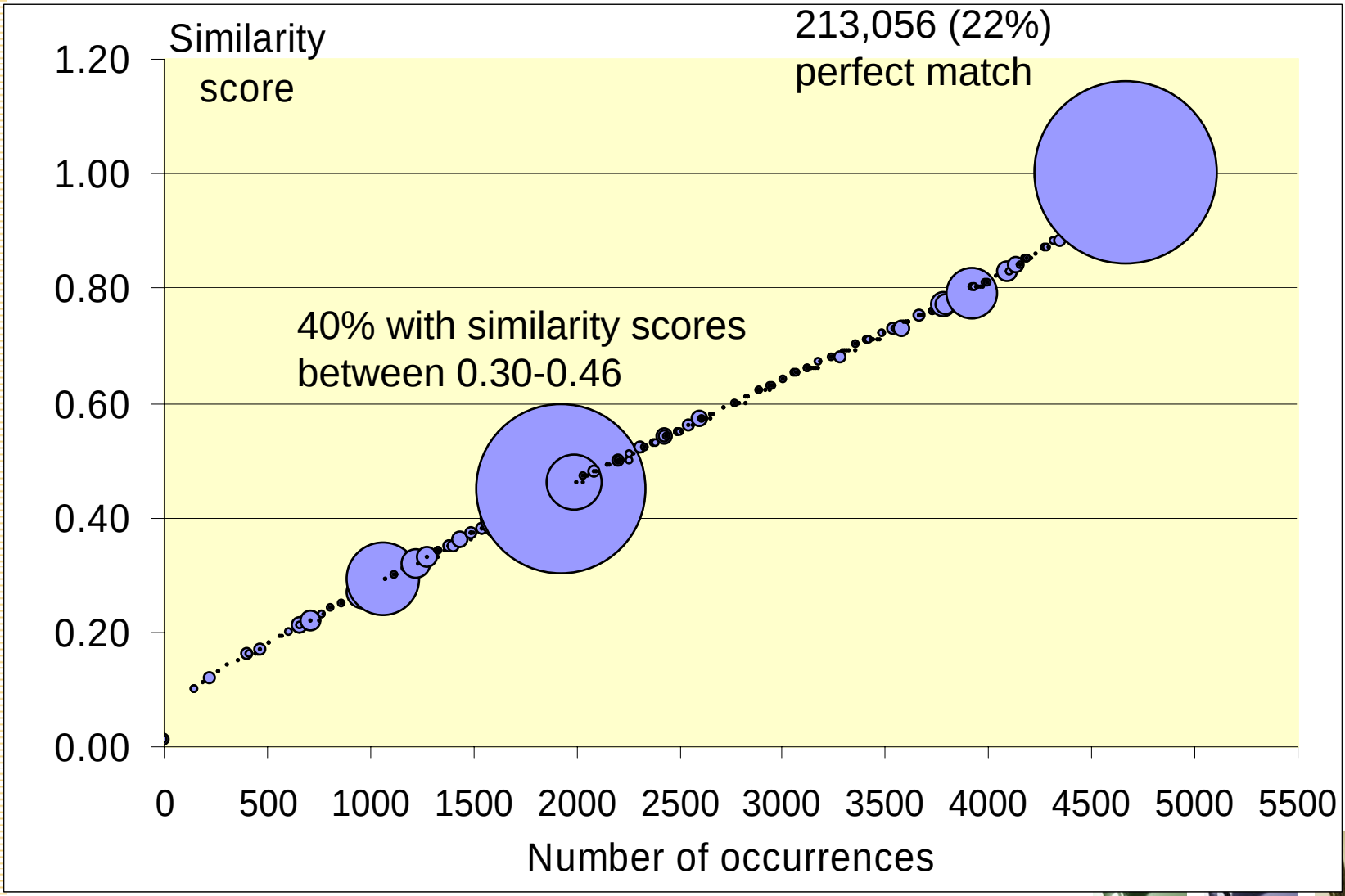
GEM subject category	Freq.	GEM subject category	Freq.
Language arts	13131	Arts	3544
Science	10384	Grammar	3252
Mathematics	9581	Writing (composition)	2822
Social studies	8567	Algebra	2788
Reading	5704	Nutrition	2788
Literature	5395	Process skills	2599
History	5346	Biology	2537
Educational Technology	4470	Vocational Education	2537
Technology	4344	Instructional issues	2482
Health	3613	Geography	2429


Top 20 non-controlled terms

Non-controlled vocabulary	Freq.	Non-controlled vocabulary	Freq.
careers	2841	civil war	538
foreign languages	2575	School to work	522
math	1898	ESL	458
lesson plans	1887	Water	397
Propaganda	875	computer	394
poetry	853	dinosaurs	389
fractions	784	curriculum	377
LESSON PLAN	713	money	369
weather	643	Internet	354
spanish	636	Library	351

Similarity to ERIC Thesaurus terms



Top matches from fulltext searches

Query terms	ERIC Thesaurus terms	Score	Occurrences
math	Mathophobia	0.36	1676
lesson plans	Lesson Plans	1	1574
reading	Reading	1	1492
science	Science	1	1147
propaganda	Propaganda	1	763
poetry	Poetry	1	732
fractions	Fractions	1	679
music	Music	1	679
language arts	Language Arts	1	614
social studies	Social Studies	1	584
Weather	Weather	1	573
Lesson Plan	Lesson Plans	0.9	568
technology	Technology	1	563
writing	Writing (1966 1980)	0.75	542

Top terms by query field

Full text

lesson plans	2019
math	1387
reading	1231
Science	947
propaganda	714

Keyword

english (second language)	443
lesson plans	194
math	156
reading	122
Spanish	93
Propaganda	91
lesson plan	87
writing	83
poetry	81
plants	80

Topic

Language Arts	12440
Science	9122
Mathematics	9036
Social studies	7901
Literature	5033
History	4851
Educational Technology	4437
Reading	4020
Technology	3684
Arts	3513
Health	3068
Grammar	2905
Writing (composition)	2820
Careers	2595
Process skills	2593

Occurrences of pedagogical terms

Pedagogical Keywords	Number of Occurrences	Pedagogical Keywords	Number of Occurrences
Process	3135	Application	180
Comprehension	1666	Exercises	140
Assessment	625	Principle	140
Project	491	Content	116
Creative	286	Facts	114
Analysis	282	Objectives	112
Concept	278	Context	108
Practice	231	Procedure	92

Note: These terms occurred alone or in a phrase



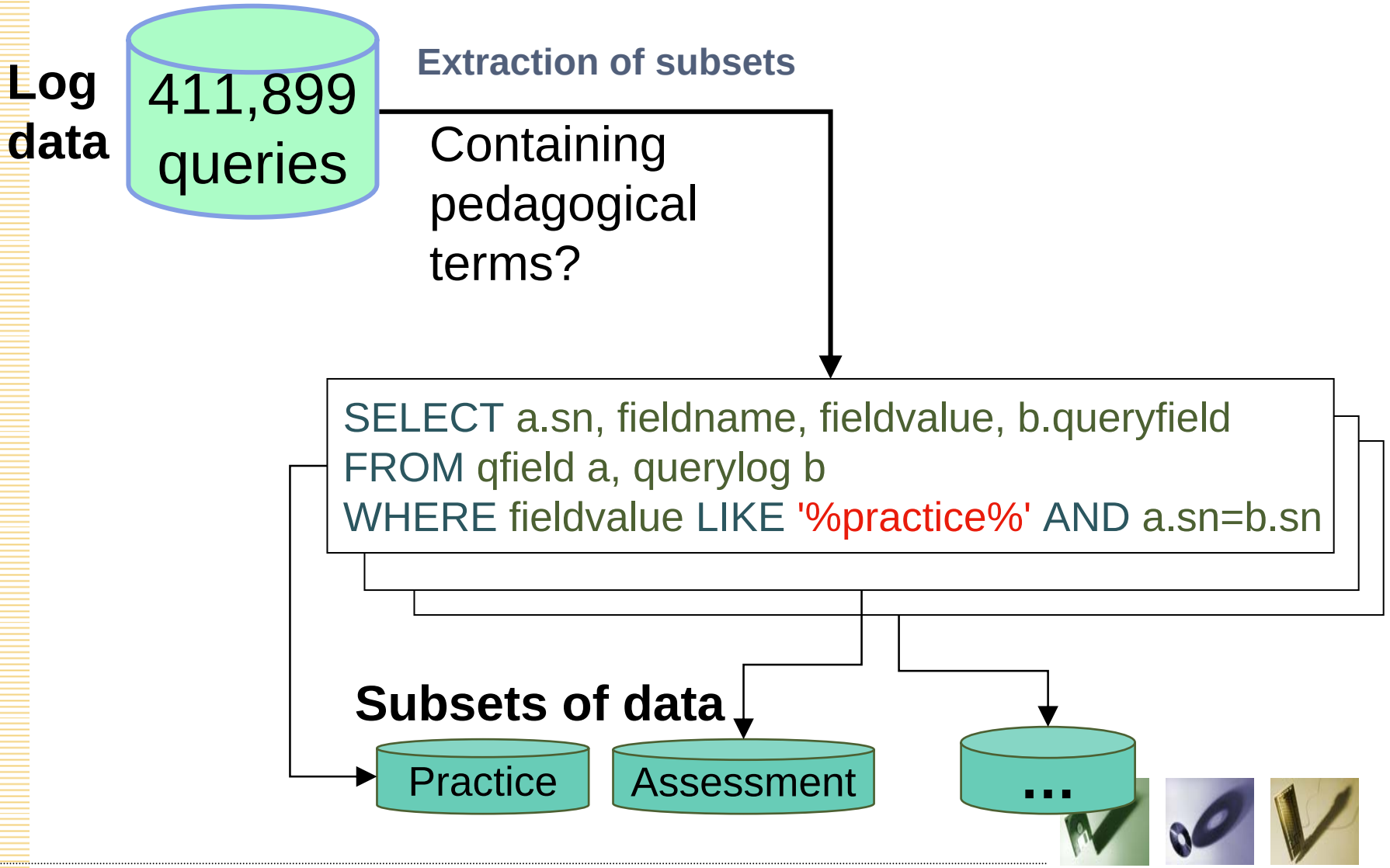
What the results imply

- Pattern terms
  - Linguistic
  - Semantic
- Associations
  - Facets of Concepts
  - Types of relationships
- Middle and long-tail term groups

- Example of pedagogical terms: A majority of pedagogical terms are in the top 1% of total occurrences



Example of pedagogical terms



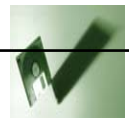
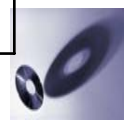
Concept facets extracted (1)

- Each of the subsets was examined by the researchers
- Facets for each concept were inducted from manual examination

Analysis **Areas of analysis** **Methods of analysis**	**Application** **Areas of application** **Application for jobs**
Concept **Discipline** **Assessment** **Instruction**	**Content** **Discipline**
Facts **Discipline**	**Practice** **Assessment** **Subject areas** **Best practice**
Project **Community** **Academic** **Methods**	**Process** **Operation** **Application**

Concept facets extracted (2)

Assessment Assessment areas Assessment methods Assessment tools	Comprehension Language Activities Skills Assessment
Context Clues Vocabulary	Creativity Action Thinking Teaching
Principle Disciplinary Learning	Procedure Practice Learning Instruction


Summary of findings (1)

- Terms occurred most frequently come from controlled vocabulary
- Long-tail terms: mostly semantic rich, non-controlled terms
- Facets of concepts may be used as subject categories for metadata



Summary of findings (2)

- Query log mining is useful for understanding:
  - Use of search terms and query fields
  - Uncovering term patterns
  - Generalizing concepts and facets of concepts
  - And more
- A large amount of data preprocessing before the data became usable
- Main findings:
  - Full text, keyword, and topic searches were at the top
  - Controlled vocabulary was used extensively in queries
  - Using subsets of data for in-depth examination as a way of uncovering concepts for enhancing controlled vocabulary



Next step

- Query patterns: What combination of query fields and terms are used for what purposes
  - Disciplinary/topical
  - Instructional methods
  - Educational research
  - General information
- Query terms
  - What kinds of queries generated zero hits?
  - What terms resulted in the largest numbers of hits?
  - How will we use the pattern concepts and terms to enhance the knowledge structure and controlled vocabulary?
- Applying data mining techniques