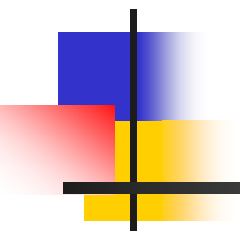
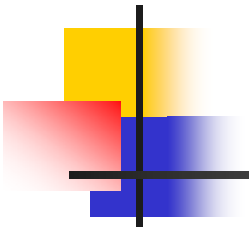


Towards a General-Purpose XML Search Engine

Michael Wollowski Robert Signorelli

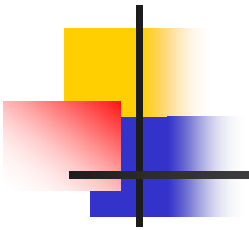
Computer Science and Software Engineering Department

Rose-Hulman Institute of Technology



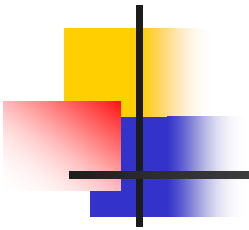
Objective of Work

- Avi Rappoport: *Many people say that XML will "solve the search problem".*
- It is not clear how
- Prior work focused on enterprise solutions



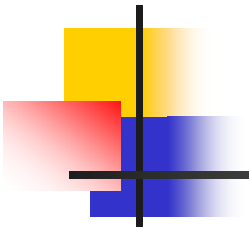
Objective of Work (cont'd)

- To develop a search engine for a broad range of XML documents
- Take advantage of the structure inherent in XML documents



Overview of Presentation

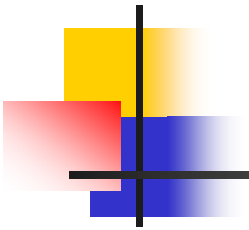
- Environment of our search engines
- Custom-fit search engines
- General-purpose search engine
- Conclusions



Environment

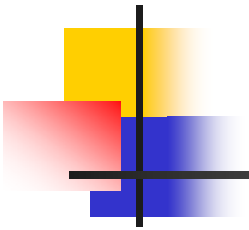
- **Domain:** XML documents with course related information
- We use XSL to produce HTML
- Transformation performed on client side

Custom-Fit Search Engine Definition



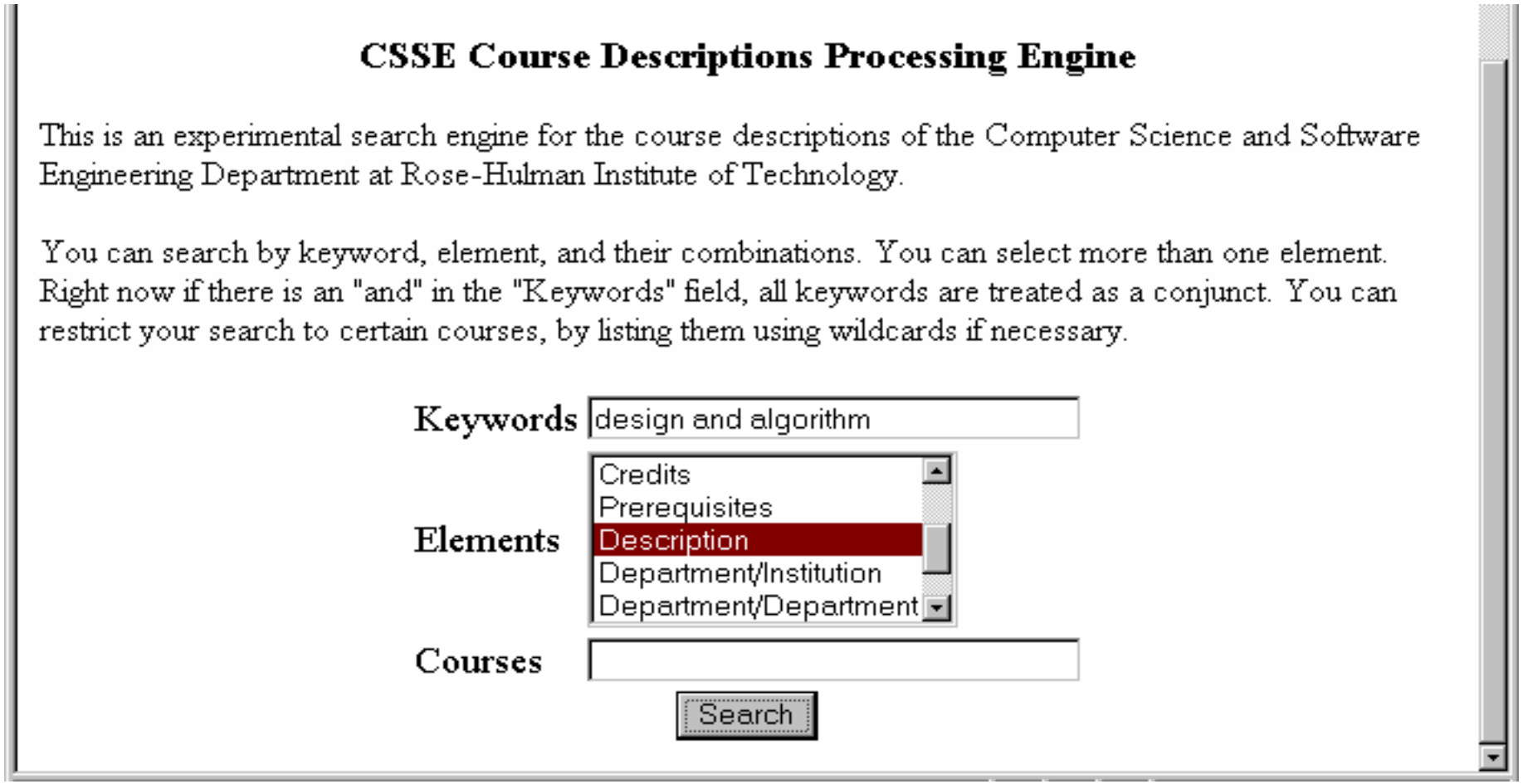
- A *custom-fit search engine* is designed for a single class of documents.

Custom-Fit Search Engine Notes



- It searches for keywords inside of specified elements
- Designer of search engine knows about elements in use
- Tailor search engine to those elements

Custom-Fit Search Engine Interface



Custom-Fit Search Engine Sample Document

<course_description>

<id> CSSE 120**</id>**

<title> Fundamentals of Software Development I **</title>**

<credits> 3R-3L-4C F,W,S **</credits>**

<description> (…) Students complete a series of projects requiring the choice of appropriate algorithms and the use of procedural abstraction, control constructs, and elementary data structures. (…) The use of a disciplined design process is emphasized in each of the projects, including good programming style and thorough testing. This course presumes no prior programming experience.

</description>

</course_description>

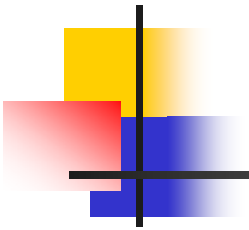
Custom-Fit Search Engine Initial Results Screen

Search Results

Your search resulted in the following hits. The ordering of the files is immaterial.

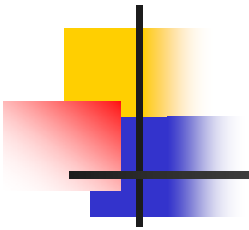
1. csse120.xml
   description: ...ing the choice of appropriate algorithms and the use of procedural abs...ect. The use of a disciplined design process is emphasized in each...
2. csse230.xml
   description: ...nd their use of a disciplined design process to include formal ana...roduces the use of randomized algorithms. Students design and implement software indivi...
3. csse445.xml
   description: ...Students study techniques for designing algorithmsalgorithms. The algorithm desig...de divide-and-conquer, greedy algorithms, dynamic programming, randomi...rithms and parallel algorithms. The algorithm analysis inclu...

Custom-Fit Search Engine Evaluation



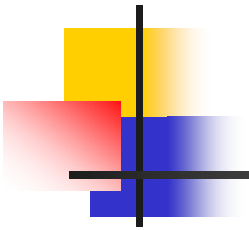
- Works extremely well for associated class of documents
- Can easily be customized to classes of documents
- Fails miserably as general-purpose search engine

General-Purpose Search Engine Search Paradigm



- A custom-fit search engine assumes that the user has a sense of which classes of documents contain desired information
- This is not a good assumption to make
- To this extent we developed a general-purpose search engine

General-Purpose Search Engine Search Paradigm (cont'd)

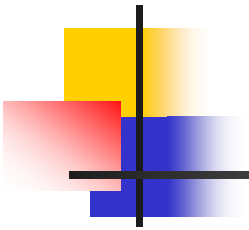


- Goal of general-purpose search engine: to determine classes of documents for subsequent searching
- Classes of documents are specified by DTDs/Schemas
- Similar to the idea of a *facet* [Marti Hearst: *Flexible Search and Navigation using Faceted Metadata*, Search Engine Meeting 2002]

General-Purpose Search Engine Search Paradigm (cont'd)

- Classes form a basis for further searching
- Embrace work that suggest several iterations in search [Amanda Spink: *A Human Behavior Approach to the Future of Search: IS THE SINGLE SEARCH DEAD?*, Search Engine Meeting 2002]

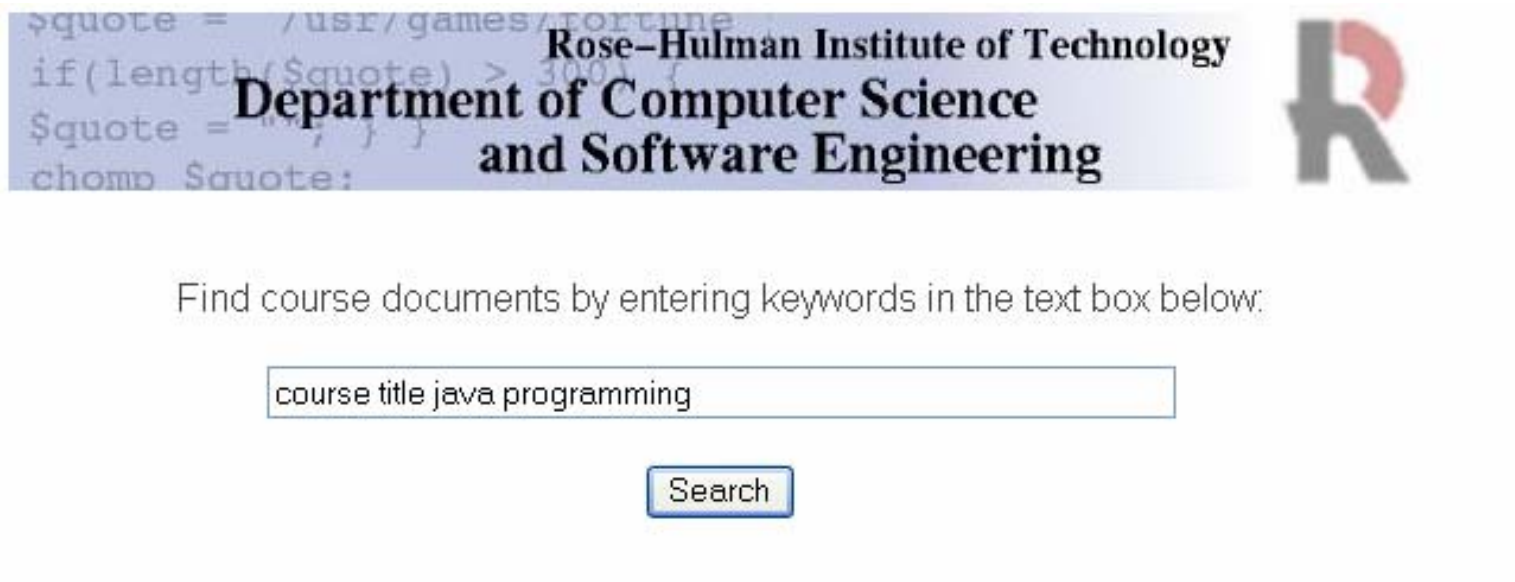
General-Purpose Search Engine Setup



- Developed a robust, independent XML search API
- API searches URLs and stores various statistics used in ranking metrics
- For testing purposes, developed various user interfaces that utilize the API's features in different ways

General-Purpose Search Engine Workings

- Start page has single text box:



General-Purpose Search Engine Workings

- First results screen displays the TYPES of documents found based on DTD/Schema
- Results presented in alphabetical order
- 2 paths of control

Below are possible contexts for the following query:
"course title java programming"

Results have been color-coded as follows:
- Structure matches
- Document matches

Course Descriptions	2 Results	33 Results	Focused Search
Course Syllabi	3 Results	5 Results	Focused Search
Instructor Bios	0 Results	11 Results	Focused Search

General-Purpose Search Engine Red Dot Results

- Red dots: Structure match

Below are the results in Course Descriptions for the following query:
"course title java programming"

http://donte.zapto.org/xml/xml-data/descriptions/csse304.xml	
title	**Programming** Language Concepts
description	Syntax and semantics of **programming** languages. Grammars, ..., binding times, functional **programming** and procedural abstraction, ...

http://donte.zapto.org/xml/xml-data/descriptions/csse100.xml	
title	Introduction to **Programming** and Problem Solving
description	...design, and elementary computer **programming**.

General-Purpose Search Engine Yellow Dot Results

- Yellow dots: Document match

Below are the results in Course Descriptions for the following query:
"course title java programming"

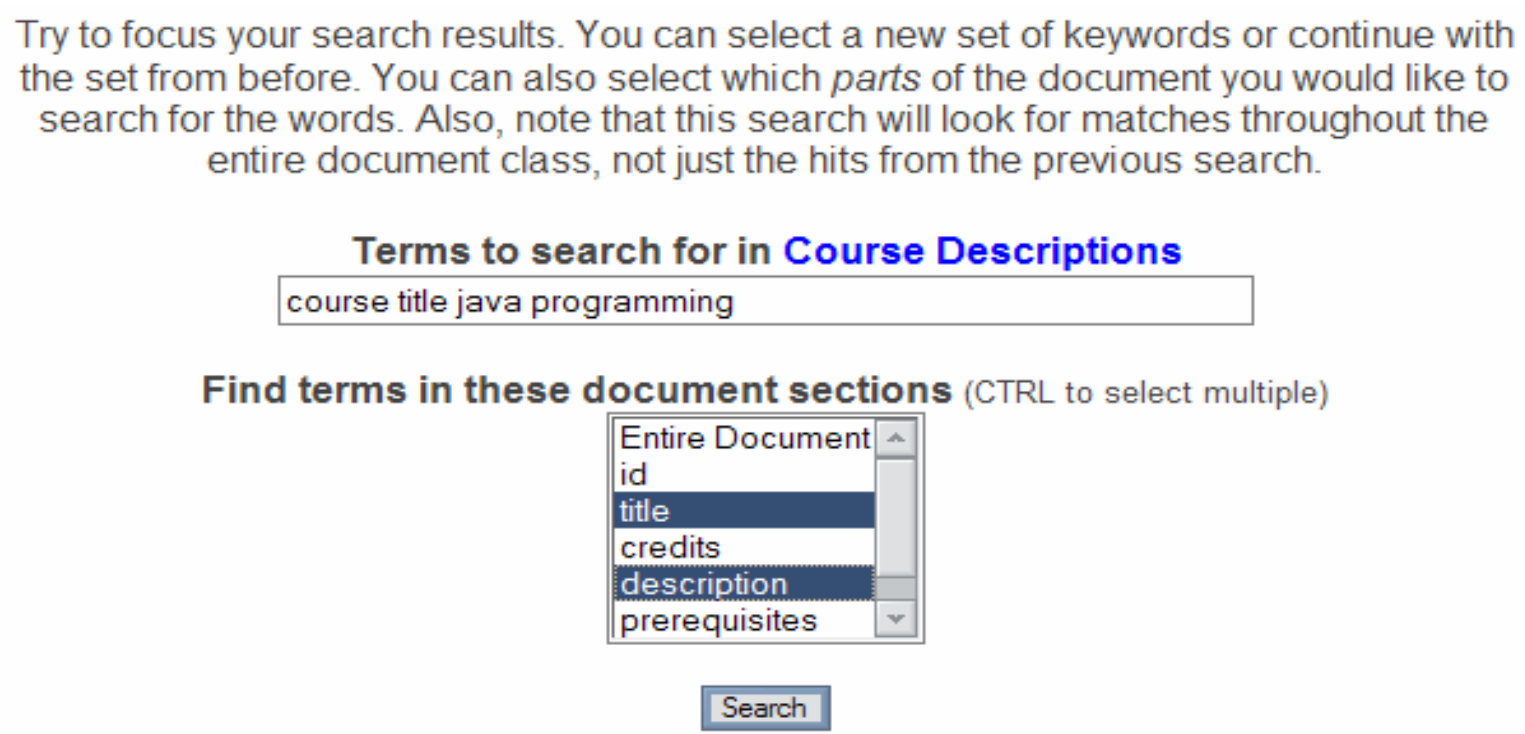
http://donte.zapto.org/xml/xml-data/descriptions/csse120.xml	
description	This **course** develops problem solving skills... multi-threaded event-driven **programming**, the development of ...
title	Fundamentals of Software Development I

http://donte.zapto.org/xml/xml-data/descriptions/csse220.xml	
title	Fundamentals of Software Development II
description	This **course** reinforces and extends students...in software maintenance. This **course** introduces networking and database... an example of functional **programming**. Students design and...

http://donte.zapto.org/xml/xml-data/descriptions/csse230.xml	
title	Fundamentals of Software Development III
description	This **course** reinforces and extends students... data structures. This **course** introduces the use of...

General-Purpose Search Engine Focused Search

- Dynamically generates a custom-fit search engine



General-Purpose Search Engine

Results of a Dynamic Custom-Fit Search

Below are the results in Course Descriptions for the following query:
"*course title java programming*"

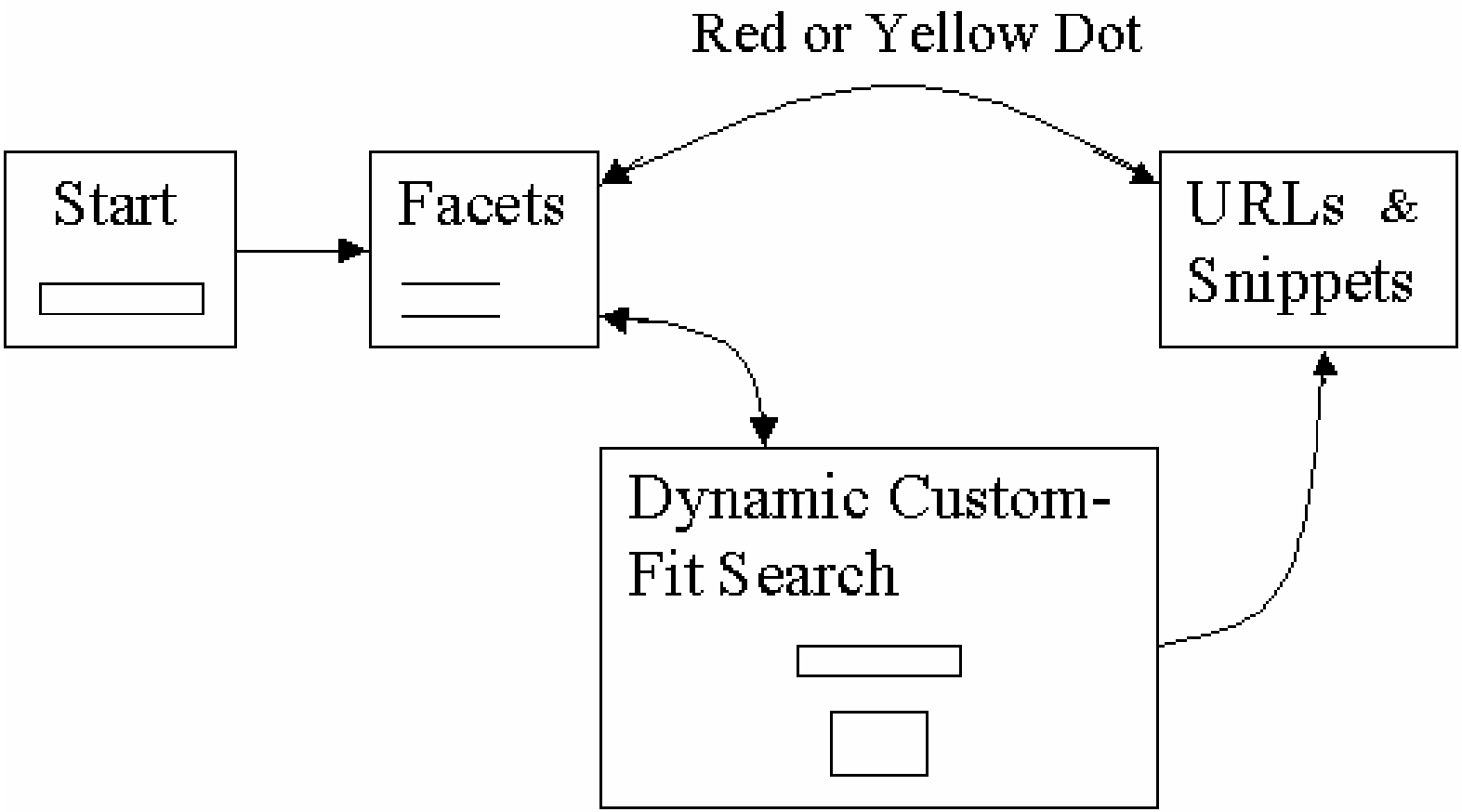
in the following set of data elements:
[title, description]

http://donte.zapto.org/xml/xml-data/descriptions/csse304.xml	
description	Syntax and semantics of **programming** languages. Grammars, ..., binding times, functional **programming** and procedural abstraction, ...
title	**Programming** Language Concepts

http://donte.zapto.org/xml/xml-data/descriptions/csse220.xml	
description	This **course** reinforces and extends students...in software maintenance. This **course** introduces networking and database... an example of functional **programming**. Students design and...

http://donte.zapto.org/xml/xml-data/descriptions/csse120.xml	
description	This **course** develops problem solving skills... multi-threaded event-driven **programming**, the development of ...

General-Purpose Search Engine Search Process



General-Purpose Search Engine User Testing: Methodology

- Developed a list of questions based on data found in some of our documents
- Users were given a very brief introduction to the functionality of the search engine
- Users then attempted to answer the questions using only the search engine
- Users supplied feedback about types of queries used

General-Purpose Search Engine User Testing: Result

- *What is "Cary Laxer's" email address?*
  - Answer determined from result screen

http://donte.zapto.org/xml/xml-data/instructors/laxer.xml	
email	**Cary.Laxer**@Rose-Hulman.edu
name	**Cary Laxer**
username	**laxer**

General-Purpose Search Engine User Testing: Result

- *Which courses use a textbook authored by "Weiss"?*
  - Complex question which was easily answered

http://donte.zapto.org/xml/xml-data/syllabi/csse220.xml	
author	**Weiss**, Mark Allen
author	Gittleman, Art
author	Stein, Lynn Andrea

General-Purpose Search Engine User Testing: Result

- *What is the weight of the final exam in CSSE 220?*
  - Correct document to open was obvious

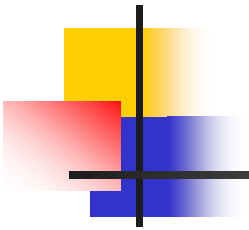
Below are the results in Course Syllabi for the following query:
"*weight final csse220*"

http://donte.zapto.org/xml/xml-data/syllabi/csse220.xml	
footer	... You must complete the **final** group project to pass... A very strong/weak **final** exam may raise/lower...
item	**Final** examination

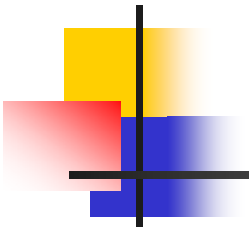
http://donte.zapto.org/xml/xml-data/syllabi/csse120.xml	
header	...two evening exams, a **final** group project, and a **final** exam.
item	**Final** group project
item	**Final** exam

http://donte.zapto.org/xml/xml-data/syllabi/csse304.xml	
footer	... your score on the **Final** Exam is outstanding, ... If your **Final** Exam score is disastrous...
item	**Final** Exam

General-Purpose Search Engine User Testing: Evaluation



- Small learning curve for interface
- After that, users liked the interface
- Users have an intuition about contents for kinds of documents
- Providing document classes as intermediate results worked well



Future Work

- Developed a ranking of classes of documents to replace the “red” and “yellow” dots
- Started work on improving the selection of element names for dynamically generated custom-fit search engine
- To make it scale:
  - Deal with classes of documents with same contents but different element names and structure
  - Develop synonym searches (surface and real)

Numeric “First Results” Page

Below are possible contexts for the following query:
"course title java programming"

Course Syllabi	4 Results	Top Score: 176	Focused Search
Course Descriptions	33 Results	Top Score: 62	Focused Search
Instructor Bios	11 Results	Top Score: 10	Focused Search

Numeric “Results” Page

Below are the results in Course Syllabi for the following query:
"course title java programming"

http://localhost/xml/xml-data/syllabi/csse220.xml	Score: 176
title	In reality, the **course** will involve two parallel
title	... & Problem Solving using **Java**, Second Edition
title	Advanced **Java**: Internet Applications, Second
title	Introduction to Interactive **Programming**
footer	... project to pass the **course**. You... exams to pass the **course**. A...

http://localhost/xml/xml-data/syllabi/csse120.xml	Score: 124
title	Interactive **Programming** in **Java**
title	Object Oriented Design
title	Coordinating Multiple Processes
title	Software Engineering Skills
attendance_policy	...-class time in this **course** constitutes an important learning... discuss continuation of the **course** with the instructor. ... automatic failure of the **course**.

http://localhost/xml/xml-data/syllabi/csse304.xml	Score: 48
goal	Syntax and semantics of **programming** languages, grammars, parsing..., binding times, functional **programming**, object-oriented languages,...
footer	... Regardless of your overall **course** average, you will not pass the **course** unless you have a...