



# Breaking the Metadata Generation Bottleneck

Elizabeth D. Liddy, **Syracuse University**

Stuart Sutton, **University of Washington**

Woojin Paik, **[solutions-united.com](http://solutions-united.com)**

# Project Overview

**Status:** A National Science Digital Library Project

**Goal:** Demonstrate feasibility of automatically generating metadata for a digital library through Natural Language Processing (NLP) and Machine Learning (ML)

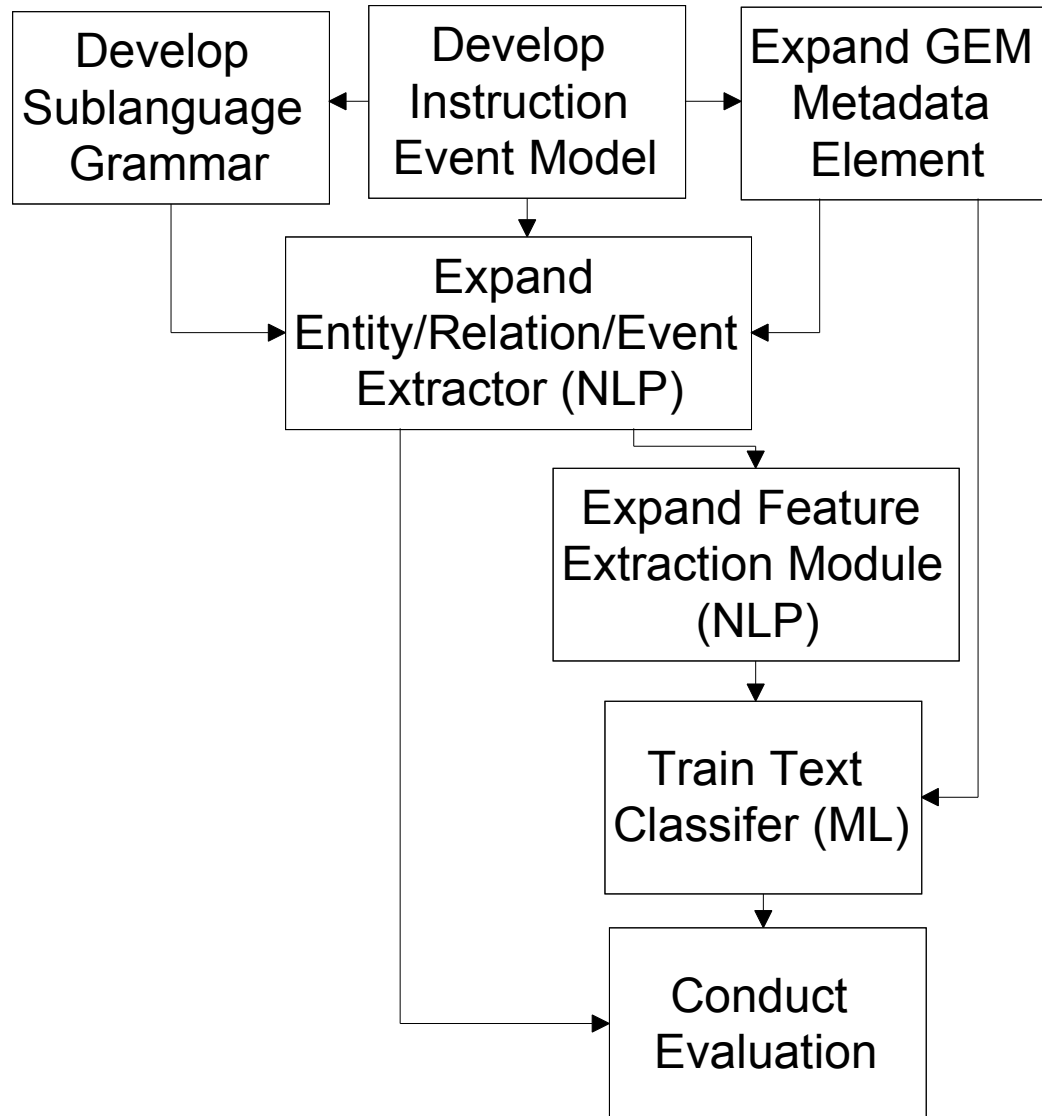
**Target Data:** Full-text collections from Eisenhower National Clearinghouse on Science & Mathematics

**Target Metadata Schema:** Enhanced Gateway for Educational Materials (GEM) metadata repository

# Research Aims

1. Develop sublanguage & discourse model for science & mathematics education materials
2. Extend generic information extraction technology to extract domain-dependent **concepts & relations**
3. Extend an automatic metatagger, <!metaMarker>, using:
  - **Machine learning**
  - **Gateway for Educational Material (GEM) metatags**
  - **Extended metatag sets**
  - **Heuristics based on the sublanguage & discourse model**
4. Compare automatic and manual metatagging in both quantitative & qualitative experiments

# Research Process Overview



# Information Extraction via **NLP**

## Natural Language Processing

- An approach which enables a system to accomplish human-like access to valuable information
- Extracts both explicit and implicit meaning
- Utilizes all levels of human language understanding when representing the content of text

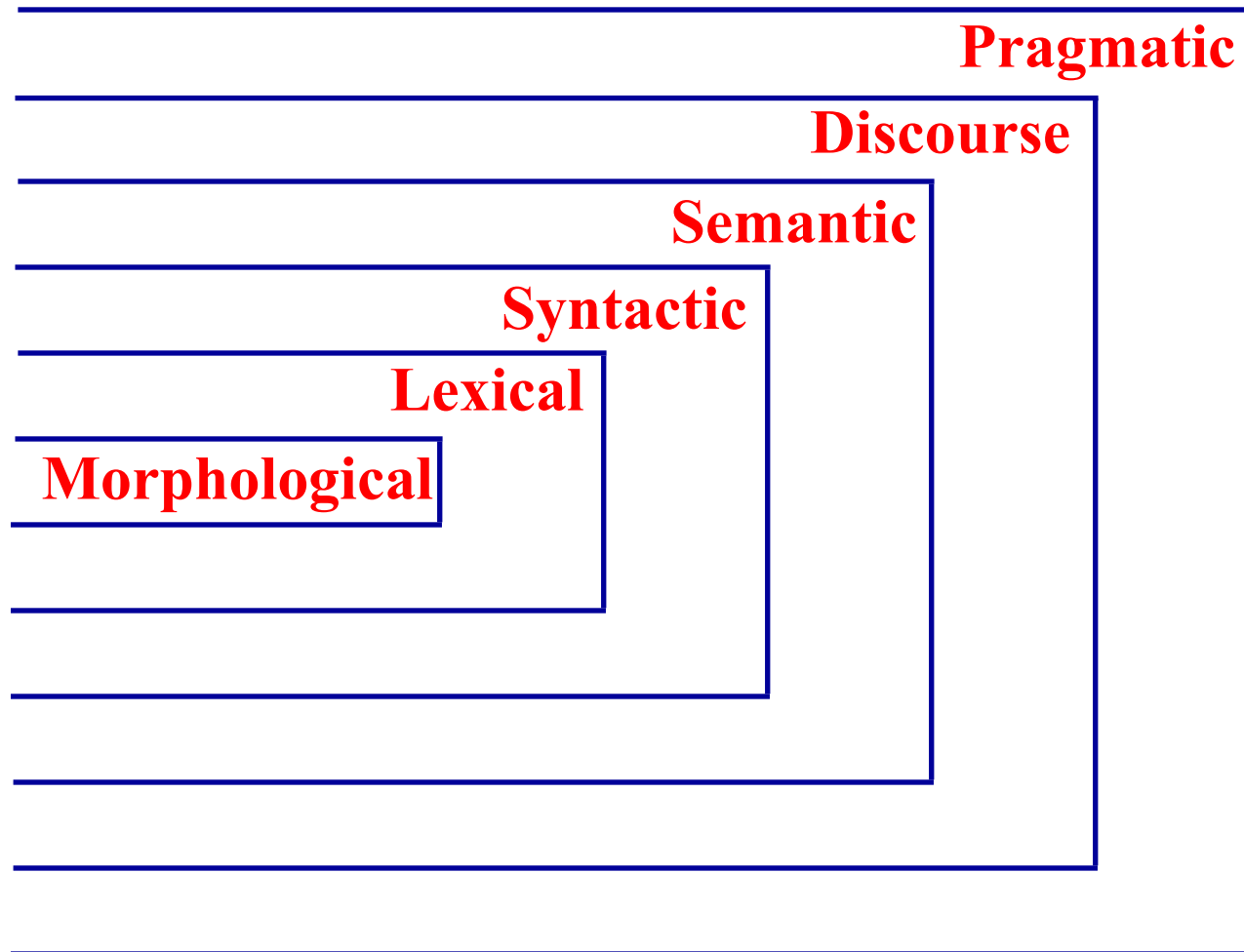
## Sublanguage Analysis

- Captures regularities of a text type in a sublanguage grammar

## Discourse Model Development

- Based on conceptual model of the process being reported in text

# Levels of Language Processing



# <!metaMarker> Machine Learning Tool

**Utilizes a readily extendable metadata framework developed for enterprise communications (e.g., email):**

- Traditional descriptive, citation-like features
  - Email header, Sender's Biographical Information
- Descriptive features unique to enterprise communications
  - Subject/Topic
  - Items mentioned
- Additional situational or use aspects which provide critical contextual information
  - Intention
  - Mood
  - Urgency

# Target Metadata Schema

## GEM Metadata Elements

- GEM Audience
- GEM Cataloging
- GEM Duration
- GEM EssentialResources
- GEM Grade
- GEM Pedagogy
- GEM Quality
- GEM Standards

## Dublin Core Metadata Elements

- DC Contributor
- DC Coverage
- DC Creator
- DC Date
- DC Description
- DC Format
- DC Identifier
- DC Language
- DC Publisher
- DC Relation
- DC Rights
- DC Source
- DC Subject
- DC Title
- DC Type

# Educational Resource Example

## ***Stream Channel Erosion Activity***

### **Student/Teacher Background Information:**

Rivers and streams form the channels in which they flow. A river channel is formed by the quantity of water and debris that is carried by the water in it. The water carves and maintains the conduit containing it. Thus, the channel is self-adjusting. If the volume of water, or amount of debris is changed, the channel adjusts to the new set of conditions.

...

### **Student Objectives:**

The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.

...

# GEM Metadata - **manually** generated

**Title:** Grand Canyon: Flood!  
- Stream Channel Erosion Activity

**Grade Levels:** 6 7 8

**GEM Subjects:** Science--Geology  
Mathematics--Geometry  
Mathematics--Measurement  
Science--Process Skills  
Science--Instructional Issues

**Tool For:** Teachers

**Resource Type:** Lesson Plan

**Format:** text/HTML

**Placed Online:** 1998-09-02

**Name:** PBS Online

**Role:** onlineProvider

**Homepage:** <http://www.pbs.org>

# NLP Processing Example

## Input:

The student will discuss stream sedimentation that occurred in the Grand Canyon as a result of the controlled release from Glen Canyon Dam.

## Morphological Analysis:

The student will discuss stream sedimentation that occurred<sup>ed</sup> in the Grand Canyon as a result of the controlled<sup>ed</sup> release from Glen Canyon Dam.

## Lexical Analysis (part-of-speech Tagging):

The|DT student|NN will|MD discuss|VB stream|NN sedimentation|NN that|WDT occurred|VBD in|IN the|DT Grand|NP Canyon|NP as|IN a|DT result|NN of|IN the|DT controlled|JJ release|NN from|IN Glen|NP Canyon|NP Dam|NP .|.

## Syntactic Analysis (phrase identification):

The|DT student|NN will|MD discuss|VB <CN> stream|NN sedimentation|NN </CN> that|WDT occurred|VBD in|IN the|DT <PN> Grand|NP Canyon|NP </PN> as|IN a|DT result|NN of|IN the|DT <CN> controlled|JJ release|NN </CN> from|IN <PN> Glen|NP Canyon|NP Dam|NP </PN> .|.

## Semantic Analysis Phase One (proper name interpretation)

The|DT student|NN will|MD discuss|VB <CN> stream|NN sedimentation|NN </CN> that|WDT occurred|VBD in|IN the|DT <PN cat=geography/location> Grand|NP Canyon|NP </PN> as|IN a|DT result|NN of|IN the|DT <CN> controlled|JJ release|NN </CN> from|IN <PN cat=buildings&structures> Glen|NP Canyon|NP Dam|NP </PN> .|.

## Semantic Analysis Phase Two (domain-independent extraction)

extraction: discuss	<b>agent of:</b>	student
	<b>object:</b>	stream sedimentation
extraction: occur	<b>object:</b>	stream sedimentation
	<b>location:</b>	Grand Canyon
extraction: controlled release	<b>location:</b>	Glen Canyon Dam
	<b>associated:</b>	stream sedimentation

---

## Semantic Analysis Phase Three (event-specific extraction)

event: discuss	<b>actor:</b>	student
	<b>topic:</b>	stream sedimentation
event: occur	<b>occurred:</b>	stream sedimentation
	<b>location:</b>	Grand Canyon
event: controlled release	<b>source:</b>	Glen Canyon Dam
	<b>result:</b>	stream sedimentation

## Semantic Analysis Phase Two (domain-independent extraction)

extraction: discuss	<b>agent of:</b>	student
	<b>object:</b>	stream sedimentation
extraction: occur	<b>object:</b>	stream sedimentation
	<b>location:</b>	Grand Canyon
extraction: controlled release	<b>location:</b>	Glen Canyon Dam
	<b>associated:</b>	stream sedimentation

---

## Semantic Analysis Phase Three (event-specific extraction)

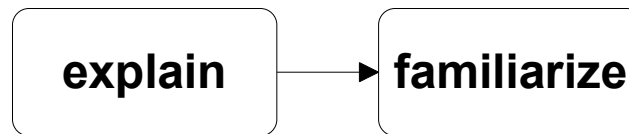
event: discuss	<b>actor:</b>	student
	<b>topic:</b>	stream sedimentation=id1
event: stream sedimentation=id1	<b>location:</b>	Grand Canyon
	<b>cause:</b>	controlled release
	<b>source:</b>	Glen Canyon Dam
	<b>result:</b>	stream sedimentation=id1

# Model Generation

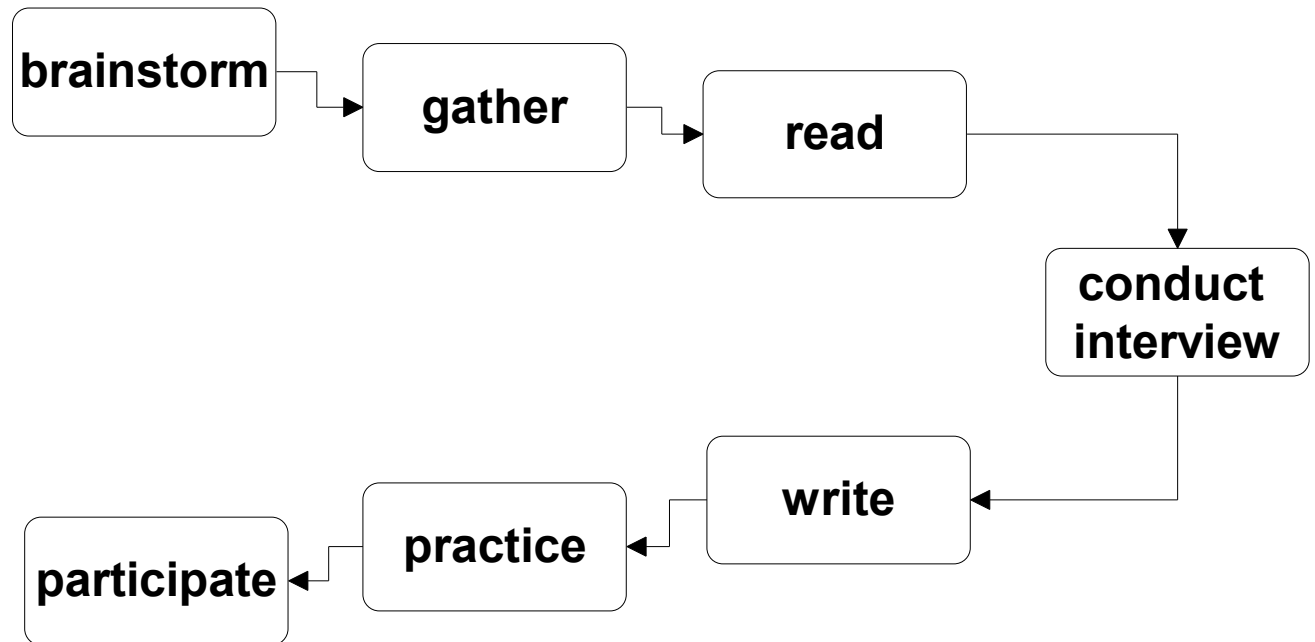
Discourse Analysis (model derivation from extracted events)

Teaching Process

Teacher



Student



# Feature Extraction & ML-based Text Classification

## Types of Features:

- Non-linguistic
  - Length of a document
- Linguistic
  - Root forms of a words
  - Part-of-speech tags:
  - Noun, Verb, Proper Noun, and Numeric Concept phrases
  - Proper Name & Numeric Concept categories
  - Semantic Relations
  - Discourse Analysis Results
  - Concepts (sense disambiguated words/phrases)

## ML-based Text Classification Techniques:

- Regression model
- Nearest neighbor classifier
- Bayesian probabilistic classifier
- Decision trees

# GOAL - GEM Metadata - **automatic** generation

**Title (SIE):** Grand Canyon: Flood!  
- Stream Channel Erosion Activity

**Grade Levels (SIE):** 6, 7, 8

**GEM Subjects (TC):** Science--Geology  
Mathematics--Geometry  
Mathematics--Measurement  
Science--Process Skills  
Science--Instructional Issues

**Keywords (TIE):**

Proper Names: **Colorado River** (river), **Grand Canyon** (geography / location), **Glen Canyon Dam** (buildings&structures)

Subject Keywords: **channels, conduit, controlled release, dam, reservoir, rivers, sediment, streams, volume of flow**

Material Keywords: **cookie sheet, roasting pan, cup, sand, clayboard, water, paper towel, pencil, paper**

Procedure Keywords: **poke a hole, divide, take, hold, pour, make drawing, identify areas, diagram, compare**

<b>Pedagogy (TC)</b>	Collaborative learning Hands on learning
<b>Tool For (SIE):</b>	Teachers
<b>Resource Type (TC):</b>	Lesson Plan
<b>Format (SIE):</b>	text/HTML
<b>Placed Online (SIE):</b>	1998-09-02
<b>Name (SIE):</b>	PBS Online
<b>Role (SIE):</b>	onlineProvider
<b>Homepage (SIE):</b>	<a href="http://www.pbs.org">http://www.pbs.org</a>

#### **Metadata Generation Methods:**

**Structured Information Extraction (SIE)**

**Textual Information Extraction (TIE)**

**Text Categorization (TC)**

# Future Directions

1. Automatically assign metatags identifying the educational standards and benchmarks for which the item would serve as a learning resource
  - As new resources are added to the digital library
  - Based on the *Compendium of Standards & Benchmarks*
  - To support standards-based discovery & retrieval
  - Enable teachers from any state to easily locate front-line teaching resources to assist their students to achieve a particular state competency

# Future Directions

2. Study a broad group of digital library users in the educational community to understand:
  - Which of the metadata elements are really used for locating and assessing resources
  - Whether new metadata elements need to be added
  - Vocabulary issues involved