

# **Breathing Life into Digital Archives: Use of Natural Language Processing Technology to Revitalize the Grey Literature of Public Health**

**Funded by the Robert Wood Johnson Foundation**

**Anne M. Turner**

Oregon Health & Science  
University  
Portland, OR

**Elizabeth D. Liddy**

Center for Natural Language Processing  
Syracuse University  
Syracuse, NY

**Jana Bradley**

Syracuse University  
Syracuse, NY

# Purpose:

Improve access to difficult-to-find literature  
of public health

# Definitions:

- Grey literature - documents that are fugitive, ephemeral, produced along non-traditional commercial pathways such as meeting notes, think tank reports, legislative documents, pamphlets, pre-prints and annual summaries.
- Intervention - any strategy, procedure, therapy, approach, method, or technique that changes, stops, deters, or interacts with a problem, disorder, disease, or disability or patient, group, or community. (Timmreck, 1997)

# Public Health Information

- Focused topically around public health problems and interventions
- Broad domain with diverse formats, content, and audiences
- Largely grey literature, not available through traditional commercial publishing pathways
- Paucity of categorization and indexing

# Public Health Grey Literature

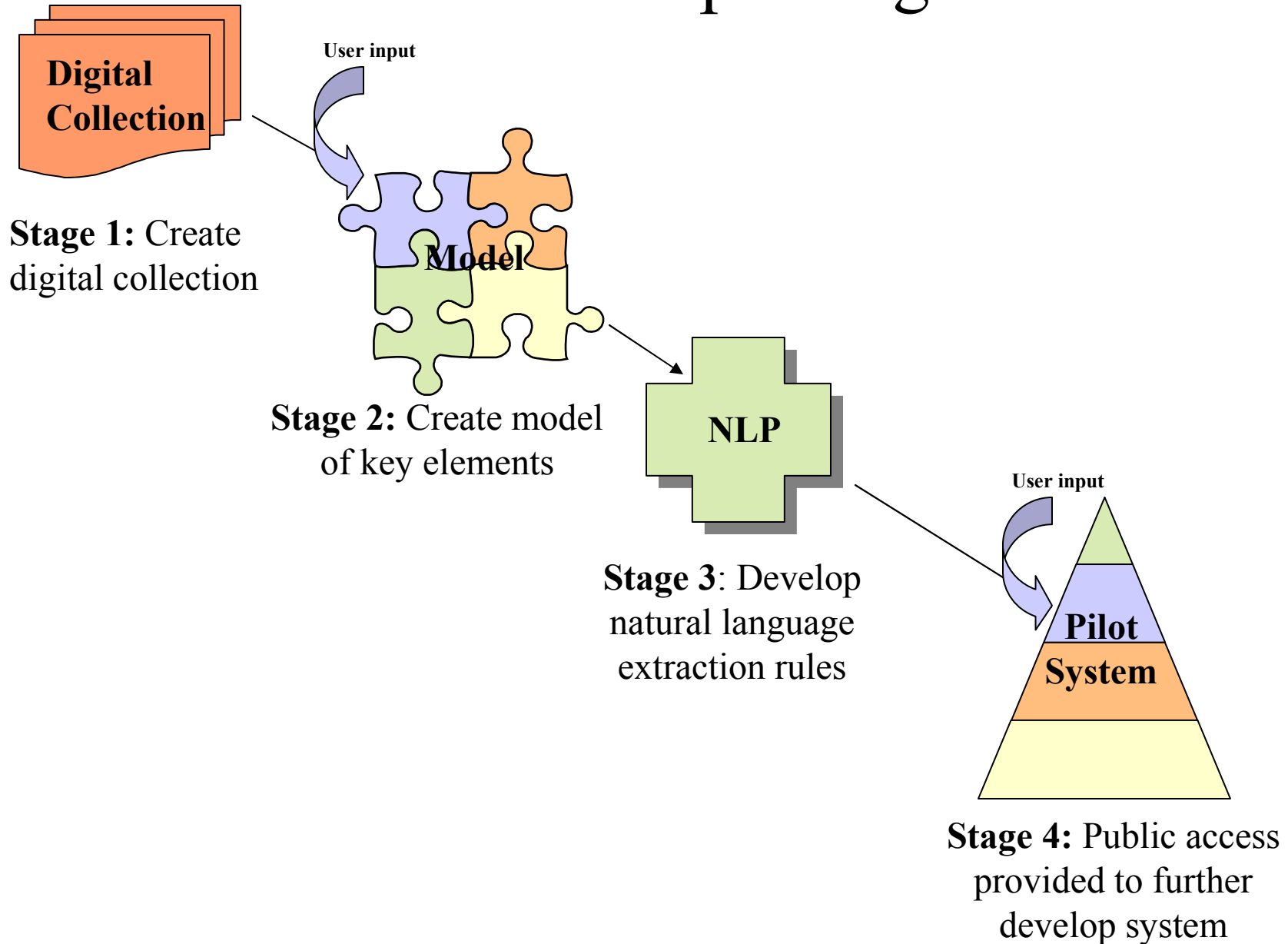
- Meeting notes
- Think-tank reports
- Legislative documents
- Data sets
- Pamphlets
- Guidelines
- Toolkits



# Project Stages

1. Create a digital collection from existing public health grey literature documents.
2. Implement an information access system which automatically recognizes these components.
3. Identify manually the components of information important for inclusion in reports of PH interventions.
4. Provide public access to the grey literature, hosted on NYAM's web site for PH practitioners and policy makers for a limited trial period.

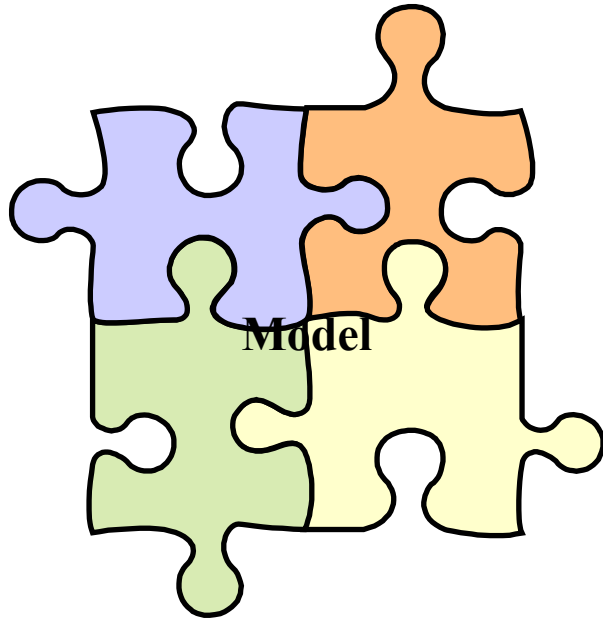
# Process of Improving Access





**Digital  
Collection**

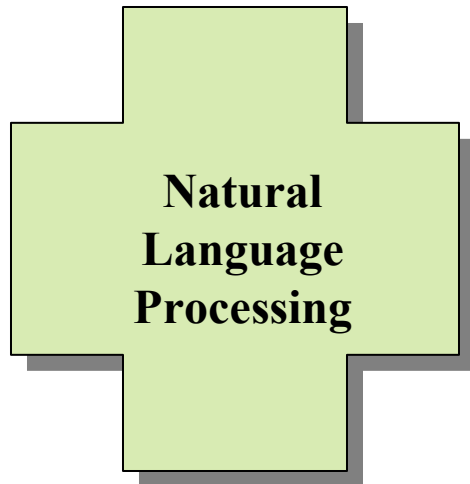
**STAGE 1:**  
**Create test digital collection  
of public health grey  
literature documents from  
county, state and national  
public health sites.**



## **STAGE 2:**

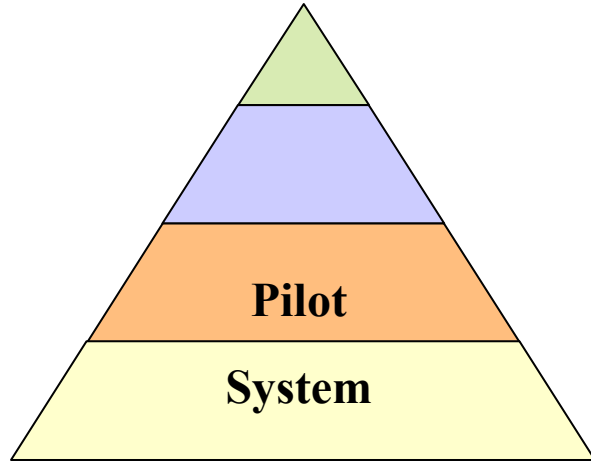
- a. Organize collection based on model of public health interventions**
- b. Determine key content elements for extraction based on input from public health professionals**

## **STAGE 3:**



**Write rules for extracting key elements from documents**

- **Based on lexical, syntactic, semantic, and discourse information of entities themselves or the context in which they occur**
- **Literals, Part-of-Speech, Context Words, Semantic Word Classes, Genres**



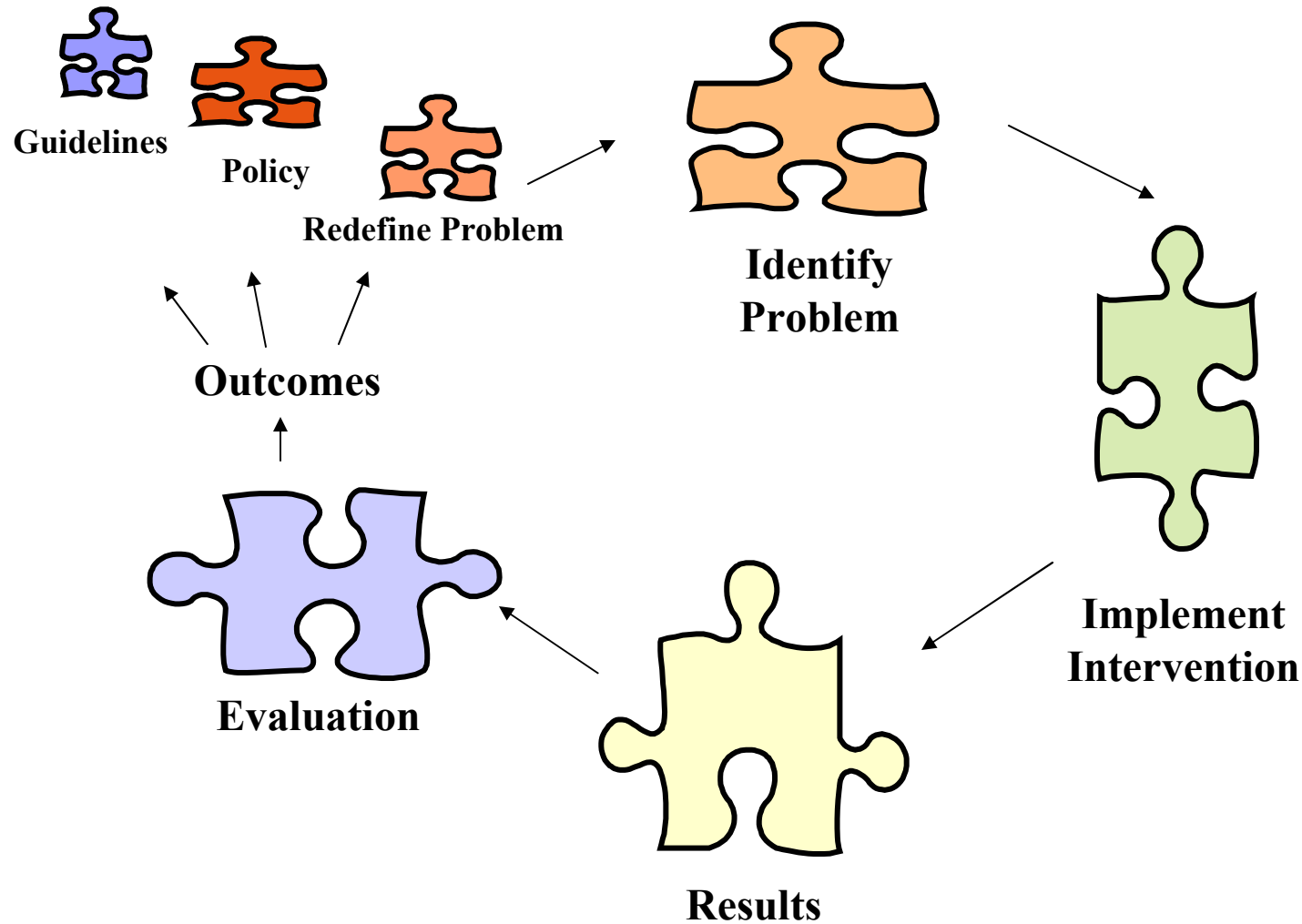
#### **STAGE 4:**

**Provide public access to the New York Academy of Medicine's Grey Literature website for a limited period of time. User input will enable developers and funders to determine whether the new capabilities are useful, efficient and productive.**

# Determine Key Content Elements Based on Input of Public Health Professionals

- Select a sample of representative documents from digital collection
- Recruit Public Health professionals through listserves (PH\_Nurses, PH\_SW, PH\_Nut, PH\_Adm)
- Ask participants to identify key content elements of documents
- Analyze input and create a set of key elements for extraction

# Organized Documents in the Context of a Model of Public Health Interventions



# Public Health Professionals Input

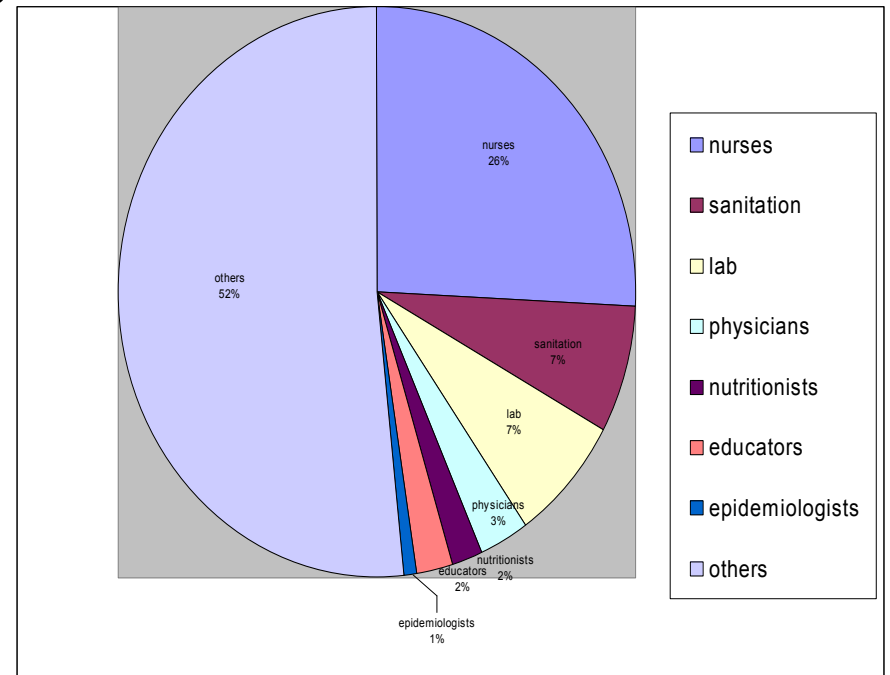
**Analyze a subset of documents with PH professionals to identify key content elements.**

## **Tasks:**

- Read sample of 3-5 documents
- Identify key elements to be included in a summary of the document
- Write an abstract of desired length and content to summarize document

# User Input: Response and Participant Selection

- 145 respondents, 11.4 % response rate
- 25 participants selected based on professional background and work experience
- Participants reflect the makeup of the public health workforce



# Intervention Elements

- **Problem:** Teenage Smoking
- **Sponsor:** Tobacco Free Kids
- **Investigators:** Jane Doe, PhD., Univ. of Virginia
- **Type of Intervention:** Media Campaign
- **Stages of Intervention:** Program Implementation, Results
- **Dates of Intervention:** 1990-1993
- **Demographic focus:** National, Ages 12-18 years
- **Results :** 25% reduction in smoking among teenagers exposed to media campaign
- **Significance:** First quantitative study to show reduction of teenage smoking from focused media campaign
- **Document address:** <http://stopteensmoking.pdf>
- **Contact information:** Jane Doe, JD@uv.edu

# Natural Language Processing

- Write rules for system to recognize & extract components.
- Integrate new rules into generic NLP system.
- Iteratively run & improve rules.
- Gather test set of PHIs and run through system.
- Have experts identify components in parallel.
- Compute Precision and Recall.

# Implement Information Access System

- Similar to automatic NSDL Meta-Data Generation task
  - Intervention reports contain components of information useful for searching, browsing, and summarizing
  - Need to determine how to:
    - Recognize, label, and extract these components of information
    - Index and store this information
    - Provide best access & display of PHI records
  - Can do extraction task using symbolic rules or Machine Learning based on linguistic features

# Conduct Focused Test

- Index all documents in selected problem area.
- Provide access for a selected group of PH practitioners.
- Conduct focus groups with practitioners to learn if the representation helps in selecting useful documents when planning an intervention.
- Explore which parameters are useful for aggregate presentations of interventions.