Progress Towards Automated Grey Literature Public Health Intervention Summaries

This research was funded by the Robert Wood Johnson Foundation

Elizabeth D. Liddy, Center for Natural Language Processing, School of Information Studies, Syracuse University
Anne M. Turner, Oregon Health Science University
Jana Bradley, School of Library Science, Arizona State University

Grey Literature Conference

New York Academy of Medicine

December 6-7, 2004

Project Goals

- **Long Term Goal** - To provide Public Health professionals and policy makers with improved access to **Public Health Interventions** as reported in the **Grey Literature** by utilizing **Natural Language Processing** to provide a universally accessible web-site for searching, summarization, navigation, and visualization.

- **Intermediate Goal** - To generate and validate a model-based representation of **Public Health Interventions** to guide automatic NLP analysis and presentation of Public Health grey literature.

Public Health Intervention

An intervention is any strategy, procedure, therapy, approach, method or technique that changes, stops, deters or interacts with a problem, disorder, disease or disability of a patient, group, or community.

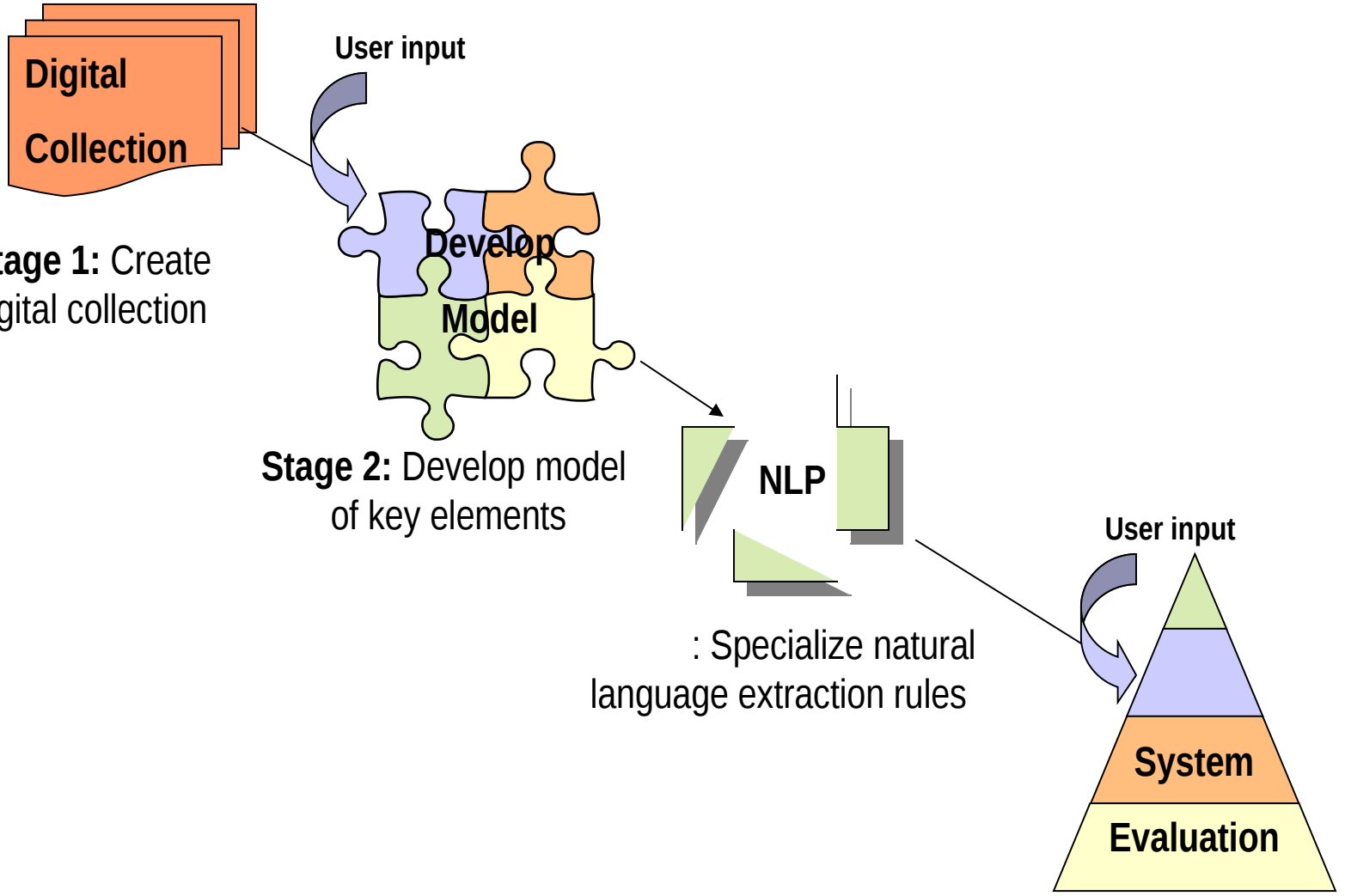
Community based programs that treat, prevent or educate about disease or health risks.

(Timmreck, 1997)

Typical Public Health Information

- Focused topically around public health problems and interventions to deal with them
- Broad domain with diverse formats, size, content, and intended audiences
- Available largely in grey literature, typically not available through traditional commercial publishing pathways
- Paucity of categorization and indexing, or web harvesting by popular search engines

Research Project Stages



Stage 1: Create digital collection

Stage 2: Develop model of key elements

: Specialize natural language extraction rules

Stage 4: Evaluate system by PH experts

STAGE 1:

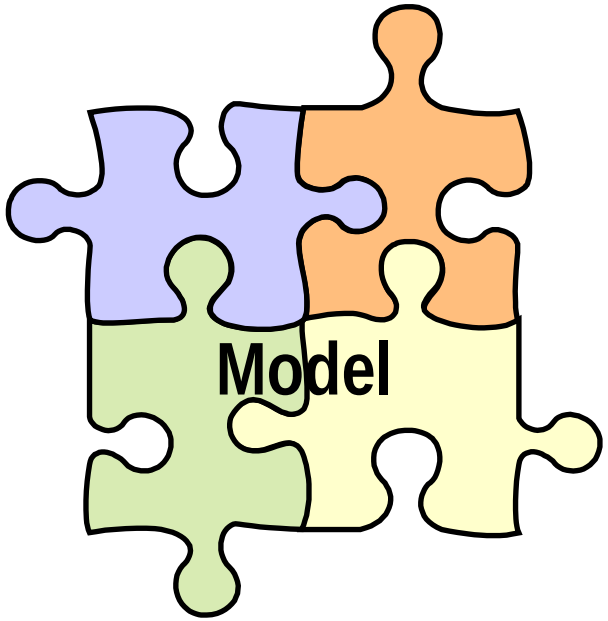
Create a training & testing digital collection of public health grey literature documents from county, state, and national public health sites.

Digital Collection Of Public Health Documents

	# Documents in Training Set	# Documents in Test Set	Total # of Documents
LAKE COUNTY	20	3*	23
HENNEPIN COUNTY	59	9	68
KENT COUNTY	21	3	24
ALL COUNTY DOCUMENTS	**100**	**15**	**115**
GEORGIA	27	5	32
NORTH CAROLINA	28	5	33
MINNESOTA	45	5	50
All STATE DOCUMENTS	**100**	**15**	**115**
NYAM *GL v. 3 n. 4 (Nov. 2001)	81	10	91
NYAM* GL v. 1 n. 1 (Aug.1999)	39	5	44
ALL NYAM * DOCUMENTS	**120**	**15**	**135**
ALL DOCUMENTS	**320**	**45**	**365**

** New York Academy of Medicine*

The research team would like to acknowledge the organizations listed above for their assistance in data collection and commend them for their efforts to promote access to Public Health Information.



STAGE 2:

Determine key content elements for extraction and representation based on input from public health professionals.

Model Development

1. Data-up analysis of this collection to identify commonly occurring intervention report elements across documents as candidates for the preliminary model.
2. Opinion of expert users – public health professionals - as to which report elements are important to include in a summary / surrogate of a PHI document.

Expert Subjects

Recruited 30 participants for web-based survey from 4 professional listservs:

- *PHNurses* - public health nurses
- *PH_SocialWork* - public health social workers
- *PH_Nut* - public health nutritionists
- *PH_Adm* - public health administrators

Participants in the user study were diverse educationally and academically, consistent with what is known about the public health workforce.

Document Collection

Collection of training documents presented broad and variable ranges of format, level of content & subject matter

- Newsletters, guidelines, annual reports, policy statements and data sets
- Documents ranged from a single page to over 100 pages
- 14% of reports consisted of multiple electronic files

Each document was reviewed by at least 3 subjects

Development Methodology

Participants were provided with copies of 4 Public Health reports and asked to:

- Rank a list of standard bibliographic elements
- Underline elements in the texts they thought would help PH professionals assess utility of a document
- Write an abstract of the length content necessary to determine if a document is useful in their work

Intervention Elements

PROBLEM

Description

Background Information

(Reports /Statistics /Guidelines/Protocols /Recommendations)

Description of Intervention

Organizations

Sponsoring /Funding /Affiliated

Governmental

- Federal
- State
- County
- Local

Non-governmental

- For-profit
- Non-profit

Intervention Type

- Education
- Prevention
- Treatment
- Surveillance

Methods

Date/Duration

Setting

- Individuals
- Practitioners
- Clinics
- Hospitals
- Institutions
- Community

Target Population

- Age
- Ethnicity
- Gender
- Employment
- Geographic Location
- Socio-Economic Status
- Insurance Status

Evaluation

Outcomes

- Results / Findings
- Knowledge Increase
- Behavioral Change
- Health Status Change
- Guidelines / Recommendations

Information Produced

Type of Information

- Guidelines
- Newsletter
- Program Reports
- Meeting notes
- Policy Brief
- Statistics/Data
- Fact Sheet

Bibliographic Elements

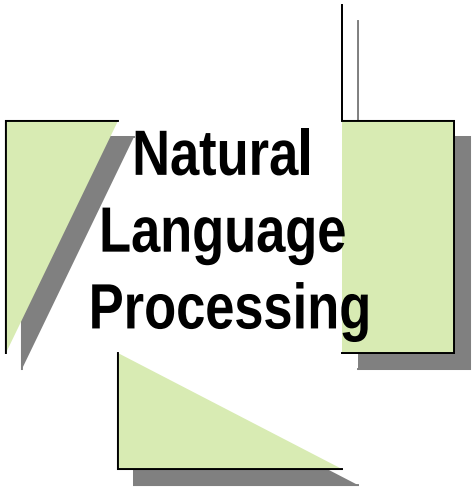
- Title
- Creator
- Publishing agency
- Publication date
- URL
- Length of document

Intervention Elements in Abstracts

Notable trends in abstracts:

- All included *a problem statement* with a description of the public health issue addressed.
- All provided a *description of the intervention* or purpose of the report.
- Most mentioned *document type*; such as policy brief, progress report or update.
- When articles included demographic parameters, such as *target population,* and when they included *results*, they were summarized in the abstract.

These guided the task of assigning priorities to task of automating element extraction



STAGE 3:

Specialize current NLP rules for extracting key elements from documents

Based on lexical, syntactic, semantic, and discourse information of entities themselves or context in which they occur

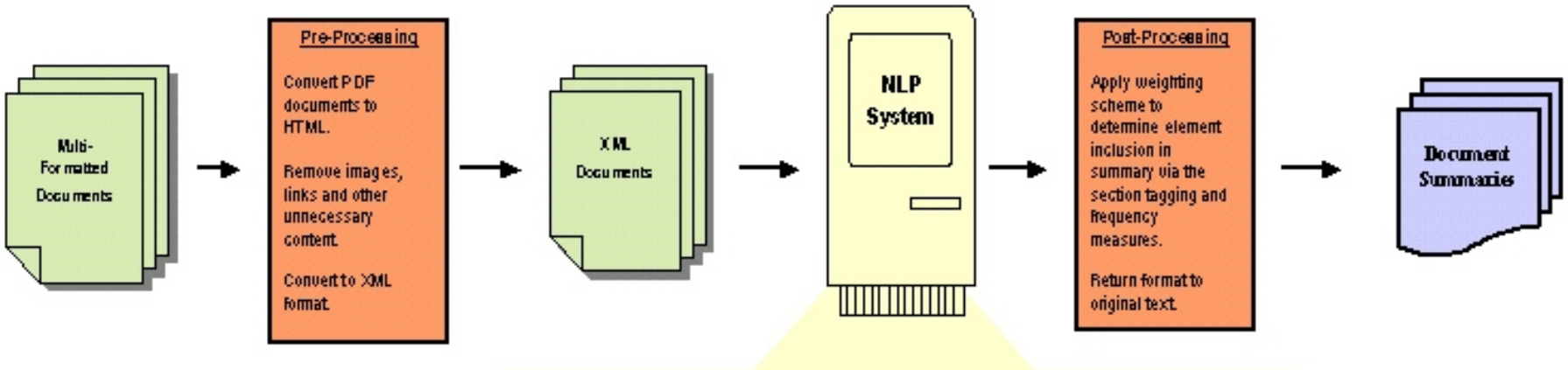
Literals, part-of-speech, context words, semantic word classes, genre clues

Metadata Element Generation

Used NLP to generate document summaries / surrogates comprised of the model elements, similar notion to metadata.

- Can distinguish between 2 kinds of metadata - “formal” metadata and metadata “*in situ*”
- Formal metadata are elements assigned by document creators and available in document header
- Metadata *in situ* are descriptive elements about the document’s contents found in the document itself for which NLP is essential in recognizing

System Diagram



The system utilizes the several levels of Natural Language Processing (NLP) to extract the intervention elements including POS Tagging, Entity Identification, Entity Categorization and a specialized extraction rules for public health intervention elements.

Example Text:

This report assesses many domains of senior health in Hennepin County including demography, quality of life, social and community support, morbidity, mortality, risk behavior, preventive care and screening utilization, and long term care.

Extraction Rules:

(inIIN) ($anywdINP) (CountyINP)
==> contextgeneric ($&, $context, 'entity', 'model-element', 'geo-location', sf($2,$3));

<S> ($RIIUthisIDT) ($RIIUinfotypeI$anypos) ($anywdI$anypos)* ($RIIUstudy_verbI$anypos) ($anywdI$anypos)* ($RIIUproblemI$anypos) ($anywdI$anypos)* </S>
==> contextgeneric ($&, $context, 'entity', 'model-element', 'description', sf($1,$2,$3,$4,$5,$6,$7), 'doc-type', $2);

Element Output:

Document Type: Report

Geographic Location: Hennepin County

Description: This report assesses many domains of senior health in Hennepin County including demography, quality of life, social and community support, morbidity, mortality, risk behavior, preventive care and screening utilization, and long term care.

Target Population: Seniors

Intervention Elements Initially Extracted by NLP System

Issue – the focus of the intervention; what health issue is being addressed.

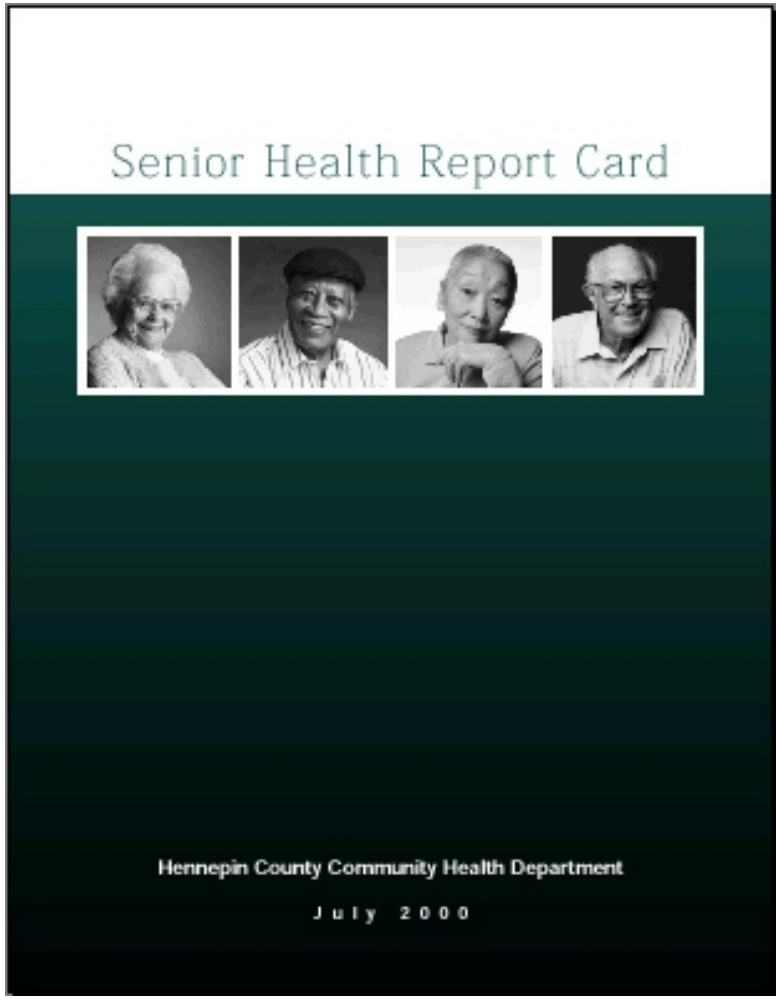
Description of Intervention – 1 sentence, high-level summary.

Target Population – target of the intervention, defined by specific demographic attributes, e.g. age, gender, ethnicity.

Geographic Location – specific locale of the target population.

Type of Information – genre / document type which embodies the intervention.

Example of Input: 45 Page Report



Example Output: 1 Page Summary

Senior Health Report Card

Description :

This report assesses many domains of senior health in Hennepin County including demography, quality of life, social and community support, morbidity, mortality, risk behavior, preventive care and screening utilization, and long term care.

Issue :

Because Hennepin County's senior citizen population (ages 65+years) is increasing, we felt it was timely to establish a set of indicators of the health of the senior population.

Document Type :

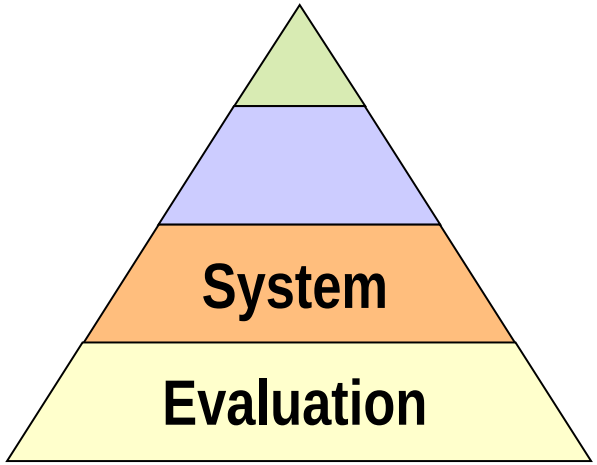
Report	Statistics-data

Target Population :

senior population	resident age 65 and older
Hennepin County resident	Senior

Geographic-Location :

Hennepin County, Minnesota, USA



STAGE 4:

Performed web-based user study with public health professionals to evaluate quality and value of the output.

Analyzed test documents and measured quality of the system.

RWJ Survey - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address Go Google

Search for Type search term(s) here Web Search Rank: 2,205

You are user 39. This is survey 1 out of 3. View Resource | Instructions | Definitions | Logout

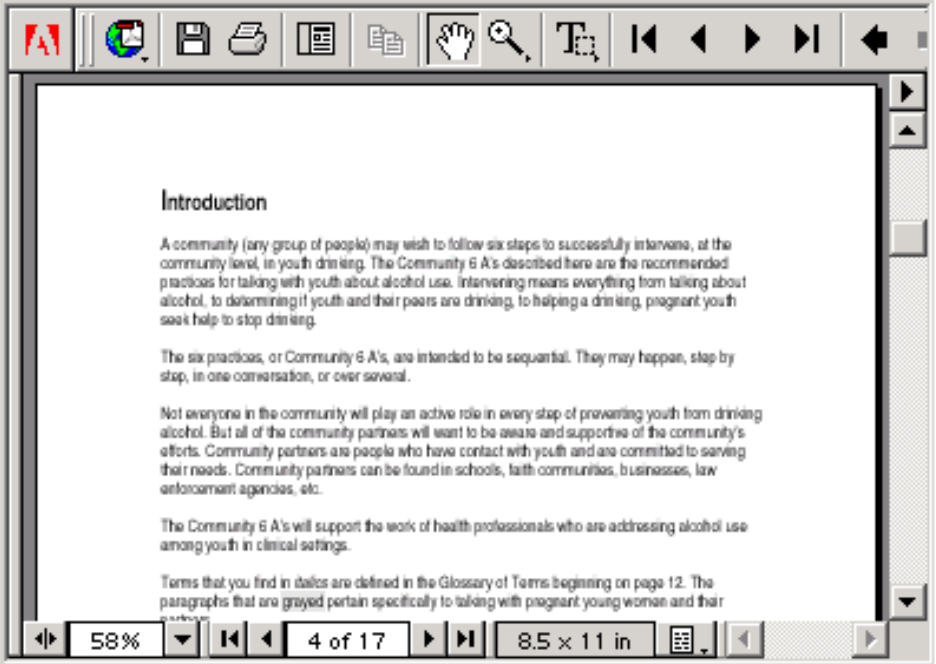
Summary

Title

The Community 6A's- Recommended Practice for Talking with Youth about Alcohol Use

Description

A. The Community 6 A's described here are the recommended practices for talking with youth about alcohol use .
B. The paragraphs that are grayed pertain specifically to talking with pregnant young wom



Survey--Summary & Full-Text Questions

Questions 4-22 out of 22

* indicates required fields.

4. * Please review the full article. (refer to Instructions for information about printing or altering window

Internet

Hide Wizard 1 17.500, 5.625 in. 23.500 x 1.750 in.

Start Gr... Ma... C:... Titl... Ne... Po... ne... R... 12:16 PM

User Survey Results

Element	Accuracy
Issue	87%
Description	83%
Target Population	73% *
Geographic-location	95%
Document type	76% *

Grey Literature Usage

Respondents were asked to name 2 documents used in the last month that were important to their work.

- Participants provided document titles and sources which we then located.
- **59% of documents listed were Grey Literature**.
- Many thought they could find all Grey Literature via traditional online services.

Study Conclusions

1. Although public health grey literature is diffuse in subject and format, a review of 300+ documents revealed that the literature can be represented by a single intervention model.

2. Key elements for extraction from the intervention model were confirmed by input from public health professionals.

3. Promising preliminary results suggest that Natural Language Processing can successfully extract these key elements based on an initial set of public health grey literature documents.

4. User input studies indicate initial extractions are sufficient and accurate for many elements. User input is being used to further refine rules.

Next Steps

Currently seeking funding to build on preliminary results and prototype technology for a system that will:

Next Steps

Currently seeking funding to build on preliminary results and prototype technology for a system that will:

1. Search web and recognize PHI grey literature reports

Next Steps

Currently seeking funding to build on preliminary results and prototype technology for a system that will:

1. Search web and recognize PHI grey literature reports
2. Harvest relevant web sites

Next Steps

Currently seeking funding to build on preliminary results and prototype technology for a system that will:

1. Search web and recognize PHI grey literature reports
2. Harvest relevant web sites
3. Use NLP to recognize PHI model elements in reports

Next Steps

Currently seeking funding to build on preliminary results and prototype technology for a system that will:

1. Search web and recognize PHI grey literature reports
2. Harvest relevant web sites
3. Use NLP to recognize PHI model elements in reports
4. Produce searchable metadata record/summary of report

Next Steps

Currently seeking funding to build on preliminary results and prototype technology for a system that will:

1. Search web and recognize PHI grey literature reports
2. Harvest relevant web sites
3. Use NLP to recognize PHI model elements in reports
4. Produce searchable metadata record/summary of report
5. Accept user query in either NL or model-based UI

Next Steps

Currently seeking funding to build on preliminary results and prototype technology for a system that will:

1. Search web and recognize PHI grey literature reports
2. Harvest relevant web sites
3. Use NLP to recognize PHI model elements in reports
4. Produce searchable metadata record/summary of report
5. Accept user query in either NL or model-based UI
6. Match query to PHI metadata record / summary

Next Steps

Currently seeking funding to build on preliminary results and prototype technology for a system that will:

1. Search web and recognize PHI grey literature reports
2. Harvest relevant web sites
3. Use NLP to recognize PHI model elements in reports
4. Produce searchable metadata record/summary of report
5. Accept user query in either NL or model-based UI
6. Match query to PHI metadata record / summary
7. Retrieve relevant PHI reports

Next Steps

Currently seeking funding to build on preliminary results and prototype technology for a system that will:

1. Search web and recognize PHI grey literature reports
2. Harvest relevant web sites
3. Use NLP to recognize PHI model elements in reports
4. Produce searchable metadata record/summary of report
5. Accept user query in either NL or model-based UI
6. Match query to PHI metadata record / summary
7. Retrieve relevant PHI reports
8. Display model-based summaries with links into full report for each metadata element

End Goals

1. Produce an NLP-based information **access** system for public health researchers, practitioners, and policy makers that provides high precision and high recall results when searching the grey literature of public health available on the web utilizing the tested model of the key data elements.

End Goals

1. Produce an NLP-based information **access** system for public health researchers, practitioners, and policy makers that provides high precision and high recall results when searching the grey literature of public health available on the web utilizing the tested model of the key data elements.

2. Provide a map of the work done in public health that shows the **"shape"** of the public health intervention domain._
   "Shape" is a meta-level overview of the problems that have been addressed with PHIs, the populations served, the types of interventions used, their success ratio, etc.

   Using automatic data-mining of model-based PHI reports.

Further Info

www.cnlp.org