Content Structure Models

Tom Reamy
Chief Knowledge Architect
KAPS Group
http://www.kapsgroup.com
Author: Deep Text

KAPS Group
Knowledge Architecture Professional Services

KAPS Group

Agenda

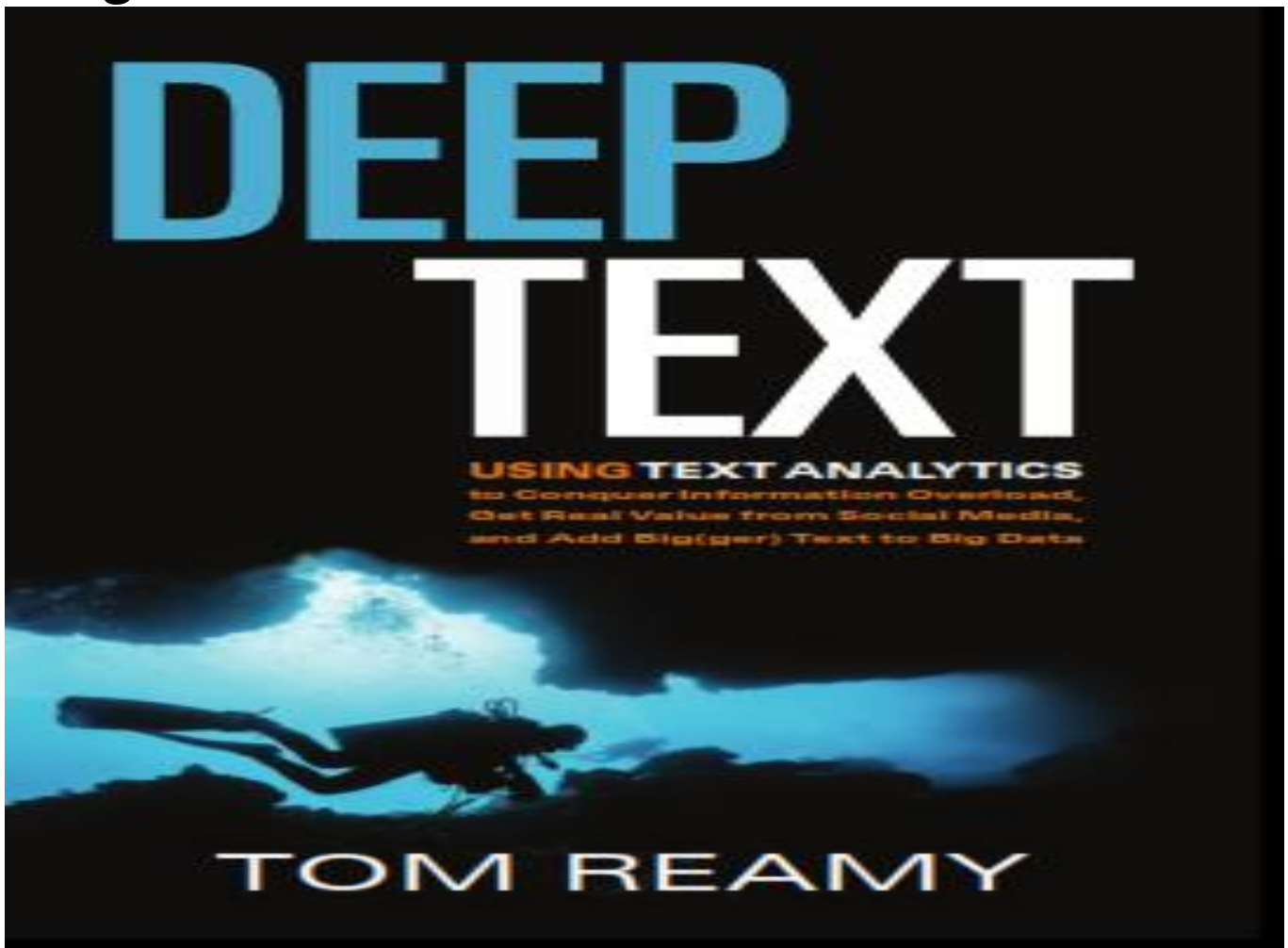
- Introduction
  - What is a Content Structure Model?
- Content Structure Models and Text Analytics
  - Auto-Categorization
  - Data Extraction
- Content Structure Models in Action
  - Search & Tagging
- Conclusions

Introduction: KAPS Group

- Network of Consultants and Partners - 2002
- Text analytics consulting: Strategy, Development-taxonomy, text analytics foundation & applications
- Mini-Projects – get started or take to next level
  - Strategy, Mini-POC - Categorization
- Partners –Synaptica, SAS, Smart Logic, Expert System, Clarabridge, Lexalytics, BA Insight, BiText
- Clients: Genentech, Novartis, Northwestern Mutual Life, Financial Times, Hyatt, Home Depot, Harvard, British Parliament, Battelle, Amdocs, FDA, GAO, World Bank, IMF, IFC, Dept. of Transportation, etc.
- Presentations, Articles, White Papers – www.kapsgroup.com
- Program Chair – Text Analytics Forum – Nov. 6-7 DC

A treasure trove of technical detail, likely to become a definitive source on text analytics – *Kirkus Reviews*

Book Sign / Meet the Author – Too late!



Content Structure Models

- Not Your Mama's Content Models!
  - Document types – nice to have, info management
- Content **Structure** Models
  - These change everything!
  - Combined with auto-categorization & data extraction
- Content **Structure** Models can:
  - Improve search by orders of magnitude
  - Improve auto-categorization by 30-50%
  - Automate entity and fact extraction
  - Build multiple analytical apps – from the other 80%

Content Structure Models
No Such Thing as Unstructured Text

- Documents are not unstructured – poly-structure
  - Words, Sentences, and Paragraphs
  - Sections and Clusters
- Text analytics rule can capture sections from the text or metadata
- Sections – Variety - “Abstract” to Function “Evidence”
  - Categorization – Title, Sub-title, Abstract, Executive Summary
  - Special - Results / Methods / Objectives
  - Systemic Text – Acknowledgements, References
  - Data Sections – Major and throughout – Tables, etc.
- Bag of Words = Bag of S**t

Content Structure Models

- Content Structure Model
  - Built on a content model – and a taxonomy
  - Foundation for applications, auto-categorization, data extraction
- Sections – types, sizes (words, sentences, paragraphs)
  - Text Indicators, start-end or size, position (first 100 words)
  - Flexible rules – use start-end if available, else use size
- Option – Organize by weight, not type
  - “Summary” – multiple text indicators
- Implementation
  - Store – spreadsheets, database, repository, text analytics software
  - Apply through text analytics rules

GrantWatch Report

Strategically Addressing The Nurse Shortage: A Closer Look At The Nurse Funders Collaborative

Geographic location and funders' goal areas influence the size of investments and where they are made.

by Denise A. Davis and Melanie D. Napier

ABSTRACT: Despite consistent public and private investments in nursing over several decades, nurse shortages persist, appearing more acute today than ever before. The Nurse Funders Collaborative, a group of foundations, government agencies, and corporations convened by the Robert Wood Johnson Foundation, has been meeting since 2003, seeking opportunities to address issues facing nursing and health care more strategically. This paper reports on a study conducted under the collaborative's auspices, which highlights the categorical and regional funding patterns of funders of nursing over five years. This information provides nursing funders with ways to craft new solutions to the nurse shortage.
[*Health Affairs* 27, no. 3 (2008): 876–881; 10.1377/hlthaff.27.3.876]

In 2003 the Robert Wood Johnson Foundation (RWJF) convened the Nurse Funders Collaborative in hopes of unifying the uncoordinated efforts to address the nurse shortage. Previously, the RWJF and trained workers, an aging workforce, and an imbalance between an increasingly diverse patient population and the nursing workforce).[1] Reports of medical errors within hospital settings continue to be prevalent, prompting

GrantWatch Report

Strategically Addressing The Nurse Shortage: A Closer Look At The Nurse Funders Collaborative

Geographic location and funders' goal areas influence the size of investments and where they are made.

by Denise A. Davis and Melanie D. Napier

ABSTRACT: Despite consistent public and private investments in nursing over several decades, nurse shortages persist, appearing more acute today than ever before. The Nurse Funders Collaborative, a group of foundations, government agencies, and corporations convened by the Robert Wood Johnson Foundation, has been meeting since 2003, seeking opportunities to address issues facing nursing and health care more strategically. This paper reports on a study conducted under the collaborative's auspices, which highlights the categorical and regional funding patterns of funders of nursing over five years. This information provides nursing funders with ways to craft new solutions to the nurse shortage. [*Health Affairs* 27, no. 3 (2008): 876–881; 10.1377/hlthaff.27.3.876]

In 2003 the Robert Wood Johnson Foundation (RWJF) convened the Nurse Funders Collaborative in hopes of unifying the uncoordinated efforts to address the nurse shortage. Previously, the RWJF and trained workers, an aging workforce, and an imbalance between an increasingly diverse patient population and the nursing workforce).[1] Reports of medical errors within hospital settings continue to be prevalent, prompting

GrantWatch Report

Strategically Addressing The Nurse Shortage: A Closer Look At The Nurse Funders Collaborative

Geographic location and funders' goal areas influence the size of investments and where they are made.

by Denise A. Davis and Melanie D. Napier

ABSTRACT: Despite consistent public and private investments in nursing over several decades, nurse shortages persist, appearing more acute today than ever before. The Nurse Funders Collaborative, a group of foundations, government agencies, and corporations convened by the Robert Wood Johnson Foundation, has been meeting since 2003, seeking opportunities to address issues facing nursing and health care more strategically. This paper reports on a study conducted under the collaborative's auspices, which highlights the categorical and regional funding patterns of funders of nursing over five years. This information provides nursing funders with ways to craft new solutions to the nurse shortage.
[*Health Affairs* 27, no. 3 (2008): 876–881; 10.1377/hlthaff.27.3.876]

In 2003 the Robert Wood Johnson Foundation (RWJF) convened the Nurse Funders Collaborative in hopes of unifying the uncoordinated efforts to address the nurse shortage. Previously, the RWJF and trained workers, an aging workforce, and an imbalance between an increasingly diverse patient population and the nursing workforce).[1] Reports of medical errors within hospital settings continue to be prevalent, prompting

GrantWatch Report

Strategically Addressing The Nurse Shortage: A Closer Look At The Nurse Funders Collaborative

Geographic location and funders' goal areas influence the size of investments and where they are made.

by Denise A. Davis and Melanie D. Napier

ABSTRACT: Despite consistent public and private investments in nursing over several decades, nurse shortages persist, appearing more acute today than ever before. The Nurse Funders Collaborative, a group of foundations, government agencies, and corporations convened by the Robert Wood Johnson Foundation, has been meeting since 2003, seeking opportunities to address issues facing nursing and health care more strategically. This paper reports on a study conducted under the collaborative's auspices, which highlights the categorical and regional funding patterns of funders of nursing over five years. This information provides nursing funders with ways to craft new solutions to the nurse shortage. [*Health Affairs* 27, no. 3 (2008): 876–881; 10.1377/hlthaff.27.3.876]

In 2003 the Robert Wood Johnson Foundation (RWJF) convened the Nurse Funders Collaborative in hopes of unifying the uncoordinated efforts to address the nurse shortage. Previously, the RWJF and trained workers, an aging workforce, and an imbalance between an increasingly diverse patient population and the nursing workforce).[1] Reports of medical errors within hospital settings continue to be prevalent, prompting

GrantWatch Report

Strategically Addressing The Nurse Shortage: A Closer Look At The Nurse Funders Collaborative

Geographic location and funders' goal areas influence the size of investments and where they are made.

by Denise A. Davis and Melanie D. Napier

ABSTRACT: Despite consistent public and private investments in nursing over several decades, nurse shortages persist, appearing more acute today than ever before. The Nurse Funders Collaborative, a group of foundations, government agencies, and corporations convened by the Robert Wood Johnson Foundation, has been meeting since 2003, seeking opportunities to address issues facing nursing and health care more strategically. This paper reports on a study conducted under the collaborative's auspices, which highlights the categorical and regional funding patterns of funders of nursing over five years. This information provides nursing funders with ways to craft new solutions to the nurse shortage.
[*Health Affairs* 27, no. 3 (2008): 876–881; 10.1377/hlthaff.27.3.876]

In 2003 the Robert Wood Johnson Foundation (RWJF) convened the Nurse Funders Collaborative in hopes of unifying the uncoordinated efforts to address the nurse shortage. Previously, the RWJF and trained workers, an aging workforce, and an imbalance between an increasingly diverse patient population and the nursing workforce).[1] Reports of medical errors within hospital settings continue to be prevalent, prompting

GrantWatch Report

Strategically Addressing The Nurse Shortage: A Closer Look At The Nurse Funders Collaborative

Geographic location and funders' goal areas influence the size of investments and where they are made.

by Denise A. Davis and Melanie D. Napier

ABSTRACT: Despite consistent public and private investments in nursing over several decades, nurse shortages persist, appearing more acute today than ever before. The Nurse Funders Collaborative, a group of foundations, government agencies, and corporations convened by the Robert Wood Johnson Foundation, has been meeting since 2003, seeking opportunities to address issues facing nursing and health care more strategically. This paper reports on a study conducted under the collaborative's auspices, which highlights the categorical and regional funding patterns of funders of nursing over five years. This information provides nursing funders with ways to craft new solutions to the nurse shortage.
[*Health Affairs* 27, no. 3 (2008): 876–881; 10.1377/hlthaff.27.3.876]

In 2003 the Robert Wood Johnson Foundation (RWJF) convened the Nurse Funders Collaborative in hopes of unifying the uncoordinated efforts to address the nurse shortage. Previously, the RWJF and trained workers, an aging workforce, and an imbalance between an increasingly diverse patient population and the nursing workforce).[1] Reports of medical errors within hospital settings continue to be prevalent, prompting

GrantWatch Report

Strategically Addressing The Nurse Shortage: A Closer Look At The Nurse Funders Collaborative

Geographic location and funders' goal areas influence the size of investments and where they are made.

by Denise A. Davis and Melanie D. Napier

ABSTRACT: Despite consistent public and private investments in nursing over several decades, nurse shortages persist, appearing more acute today than ever before. The Nurse Funders Collaborative, a group of foundations, government agencies, and corporations convened by the Robert Wood Johnson Foundation, has been meeting since 2003, seeking opportunities to address issues facing nursing and health care more strategically. This paper reports on a study conducted under the collaborative's auspices, which highlights the categorical and regional funding patterns of funders of nursing over five years. This information provides nursing funders with ways to craft new solutions to the nurse shortage.
[*Health Affairs* 27, no. 3 (2008): 876–881; 10.1377/hlthaff.27.3.876]

EXECUTIVE SUMMARY

The shortage of nursing faculty in the United States is a critical problem that directly affects the nation's nurse shortage, which is projected to worsen in future years. Short-term interventions to address the nursing shortage are inadequate given the increasing needs of a growing and aging population. A substantial increase in newly educated nurses will be needed to meet future demand; therefore, timely and sustainable interventions to reduce the nursing faculty shortage are required. This paper highlights solutions to the faculty shortage by:

- describing the current faculty shortage in relation to demand, supply, educational preparation and productivity;
- examining the factors that contribute to the faculty shortage;
- reviewing the array of interventions already undertaken; and
- outlining recommendations for further action.

The paper is based on a review of published literature and data, including surveys by government and professional organizations, studies by state task forces addressing the nursing shortage, foundation reports, and reviews by such groups as the National Conference of State Legislatures of activities at the state level, as well as author interviews with leaders in nursing education.

COMMONWEALTH OF VIRGINIA
DEPARTMENT OF TRANSPORTATION

WORK ORDER

Contract ID. No.: P00091296B00 FHWA No.: BH-BR03(259); BH-BR03(261) Work Order No.: 2

State Project No.: BRDG-041-718, B660; BRDG-041-719, B661 Category: MISC

Original Contract Value $ 646,308.25 Total of Other Work Orders $ 0

NOTE: If additional space is needed, use an additional sheet(s) and label as Supplemental Attachment #.

I. LOCATION AND DESCRIPTION OF PROPOSED WORK:

Time Extension

Dec. 22, 2010 to March 13, 2011 Suspension of work.
March 14, 2009 to April 15, 2011 Extension of 33days

50 days total time extension

One month additional Maintenance of Traffic

II. RESPONSIBLE CHARGE ENGINEER'S EXPLANATION OF NECESSITY FOR PROPOSED WORK:

This Work Order is needed to extend the contract time to allow the contractor to place the Asphalt Concrete TY. SM9.5A during warmer weather. Asphalt producers have shut down and will not be open until warmer weather returns. All remaining work to be completed at current contract prices.

"Burleigh Construction Company Inc. and VDOT agree that this Work Order fully resolves and settles all claims, demands or damages of any kind relating to or arising out of the work set forth in this Work Order, including but not limited to delay, impact and acceleration."

The additional Maintenance of Traffic cost are to cover the cost of rented traffic control equipment during the time when additional work was taking place.

III. FUNDING SOURCE/CHARGE Federal 80% / State 20%

IV. THE FIXED DATE TIME LIMIT FOR THIS CONTRACT PRIOR TO APPRVOAL OF THIS WORK ORDER IS Dec. 21, 2010

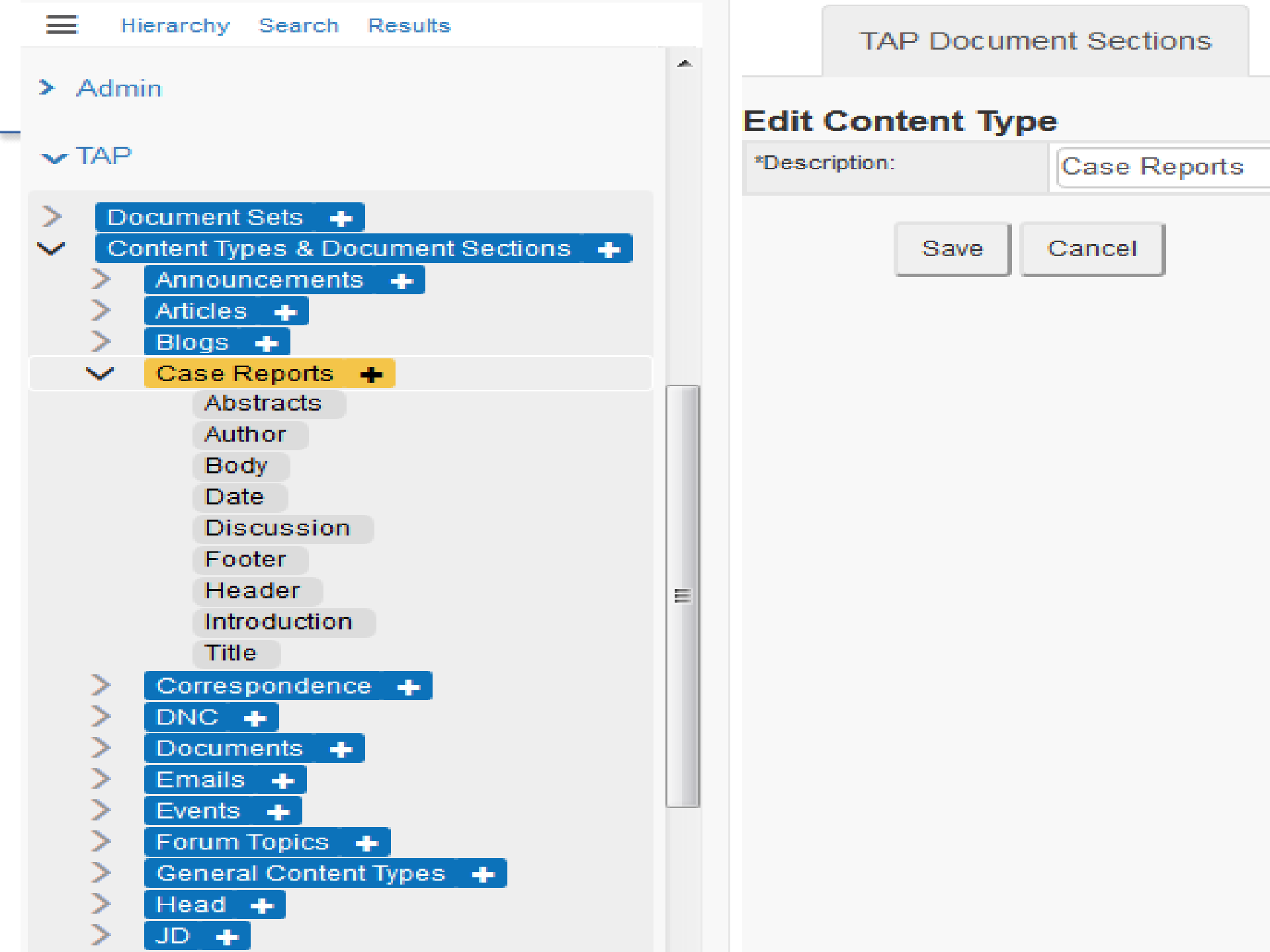
V. THE FIXED DATE TIME LIMIT FOR THIS CONTRACT UPON APPROVAL OF THIS WORK ORDER IS Apr. 15, 2011

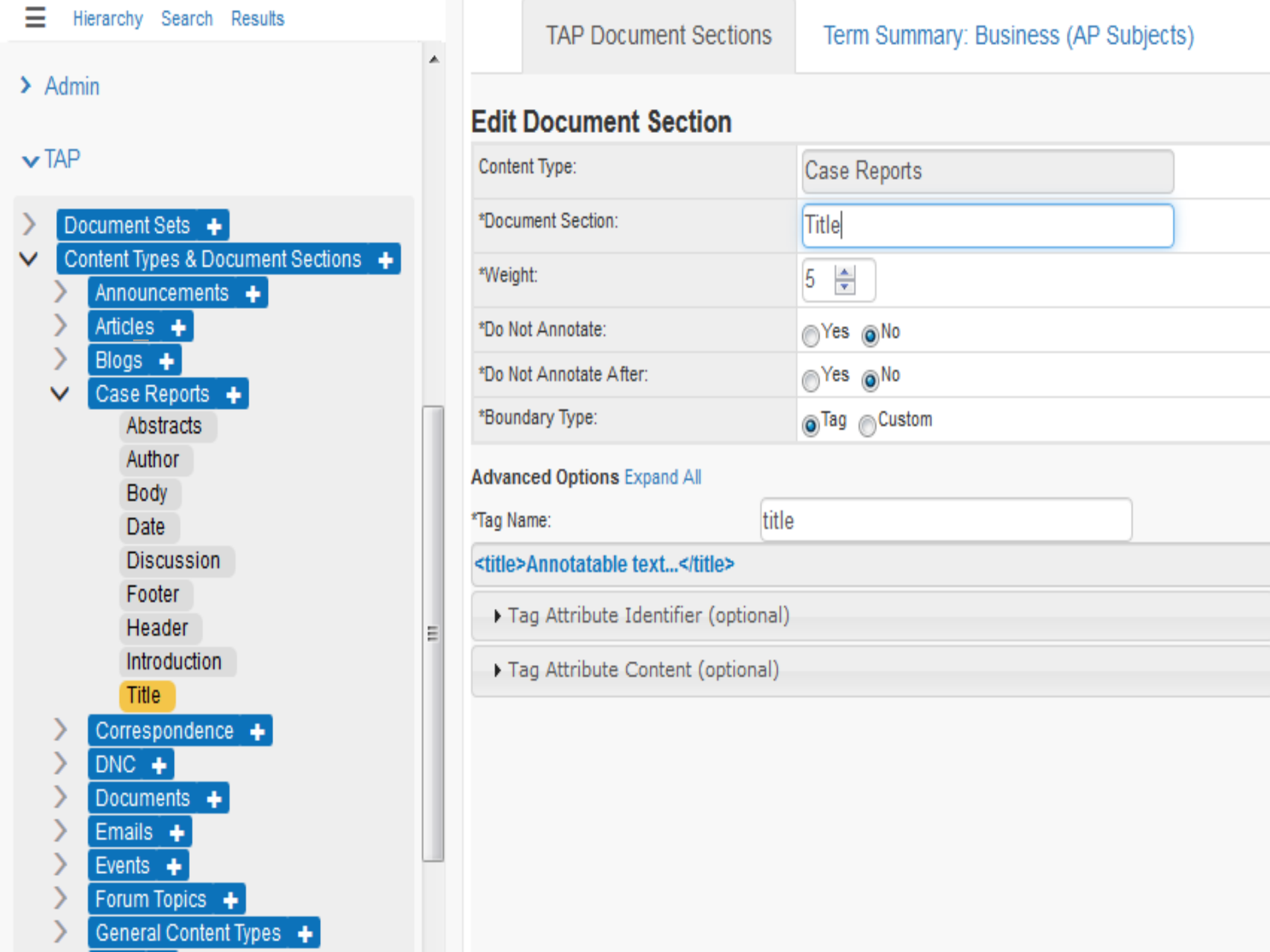
Content Structure Models

- Content Structure Models
  - And
  - Text Analytics
- Auto-Categorization & Data Extraction

Content Structure Models
What do you get for the effort?

- Use of human judgements about “aboutness”
  - Better than keywords
  - Can include variety of terms – journals, people, programs
- Less is more
  - Less text to process – easier to develop & maintain
  - Fewer terms in a categorization rule
  - More precise – not dependent on relevance algorithms
  - Relevance scores per section, not entire document
- Entity / fact extraction
  - Increase precision & speed up processing
  - Knowing where to look
  - Knowing what to ignore


Hierarchy
Search
Results
Admin
TAP
Document Sets
Content Types & Document Sections
Announcements
Articles
Blogs
Case Reports
Abstracts
Author
Body
Date
Discussion
Footer
Header
Introduction
Title
Correspondence
DNC
Documents
Emails
Events
Forum Topics
General Content Types
Head
JD
TAP Document Sections
Edit Content Type
*Description:
Case Reports
Save
Cancel


Hierarchy
Search
Results
Admin
TAP
Document Sets
Content Types & Document Sections
Announcements
Articles
Blogs
Case Reports
Abstracts
Author
Body
Date
Discussion
Footer
Header
Introduction
Title
Correspondence
DNC
Documents
Emails
Events
Forum Topics
General Content Types
TAP Document Sections
Term Summary: Business (AP Subjects)
Edit Document Section
Content Type:
Case Reports
*Document Section:
Title
*Weight:
5
*Do Not Annotate:
Yes
No
*Do Not Annotate After:
Yes
No
*Boundary Type:
Tag
Custom
Advanced Options
Expand All
*Tag Name:
title
<title>Annotatable text...</title>
Tag Attribute Identifier (optional)
Tag Attribute Content (optional)

w Tabular View

y: Require Same Sentence

Document Keywords

Document Keywords

Intelligence artificial intelligence Business

Positive Context Keywords

Add New Positive Context Keywords

AI deep learning machine learning

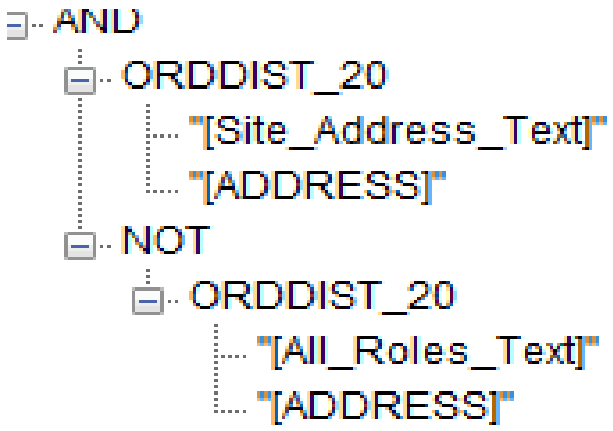
Negative Context Keywords

Add New Negative Context Keywords

No Keywords Found

Content Structure Models
Data Extraction

- Rich source of Metadata
- Facets need a lot of metadata
- Automated or semi-automated improves the quality of the tags and reduces the human tagging effort
- Resolve disambiguation – combine sections and context rules
  - Ford – company, car, person, stream crossing
  - Look at words around – sentence, distance, paragraph
- Fact extraction even more powerful
  - Not all people, entities
  - Distinguish entities – Site address not architect address
- Extract bulk data and analyze – combine internal & external
  - Example - financial, demographic, political, etc.


AND
ORDDIST_20
"[Site_Address_Text]"
"[ADDRESS]"
NOT
ORDDIST_20
"[All_Roles_Text]"
"[ADDRESS]"

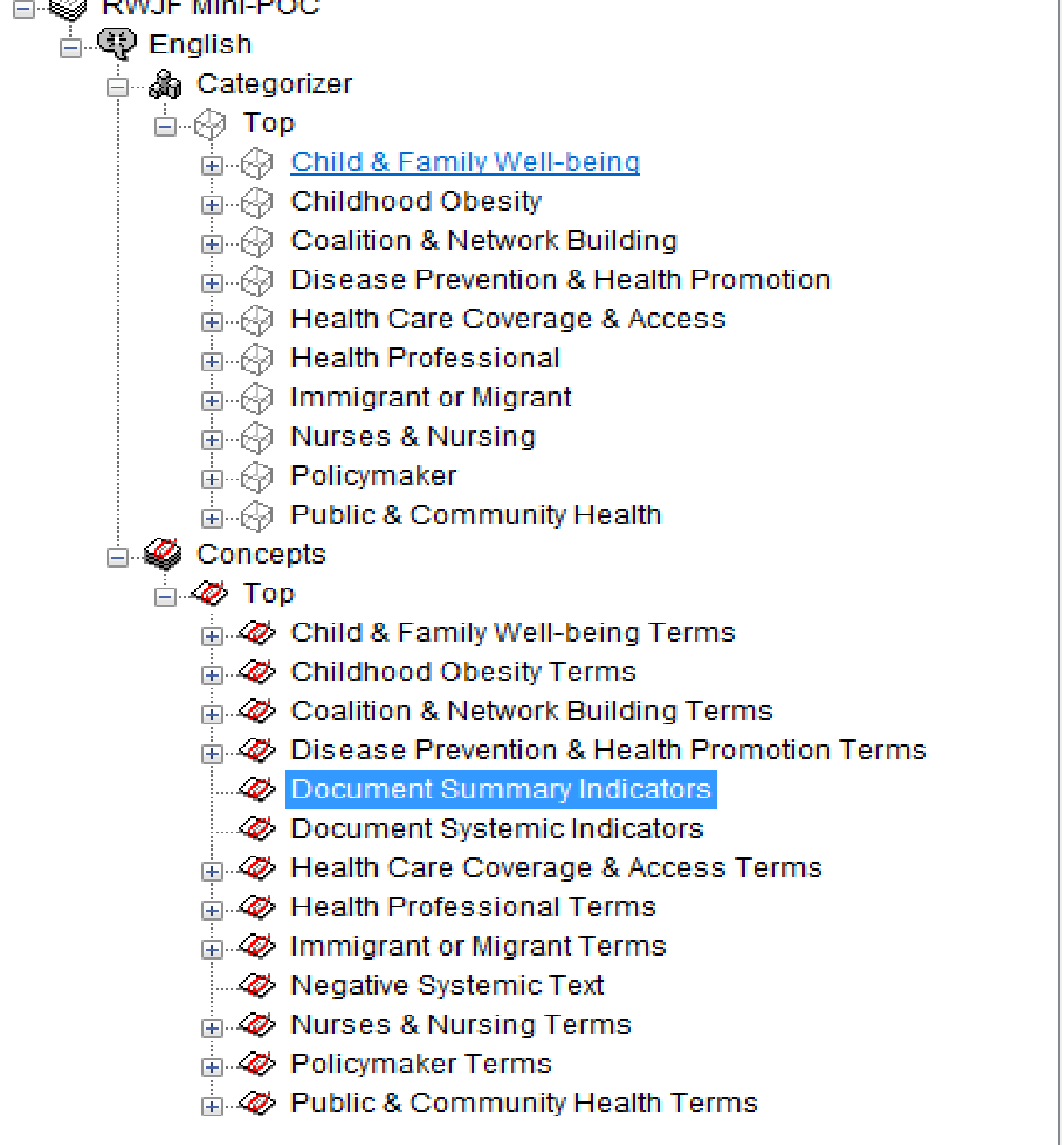
KAPS Group

Content Structure Models

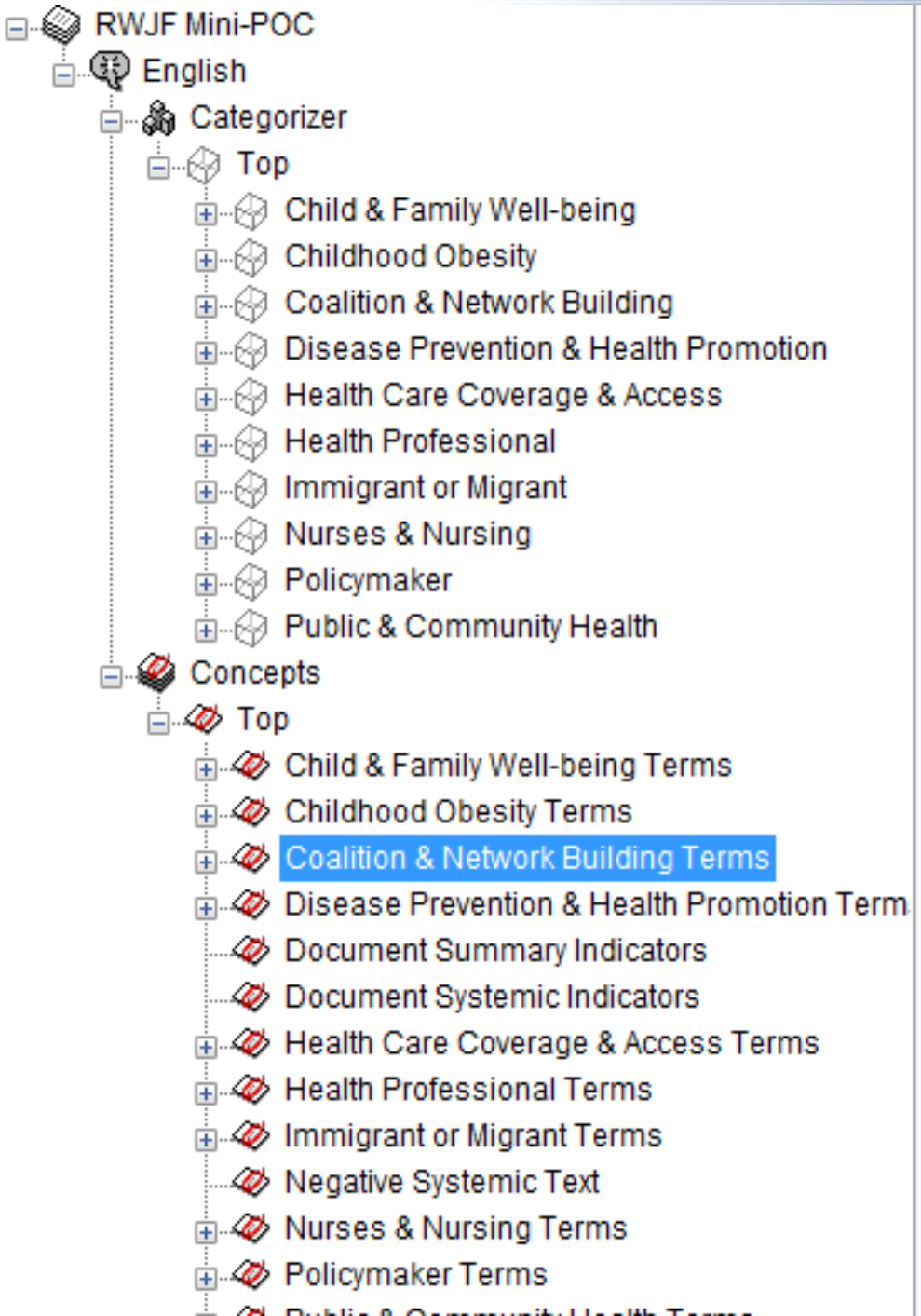
- Content Structure Models
  - In Action

Content Structure Models
Search & Tagging

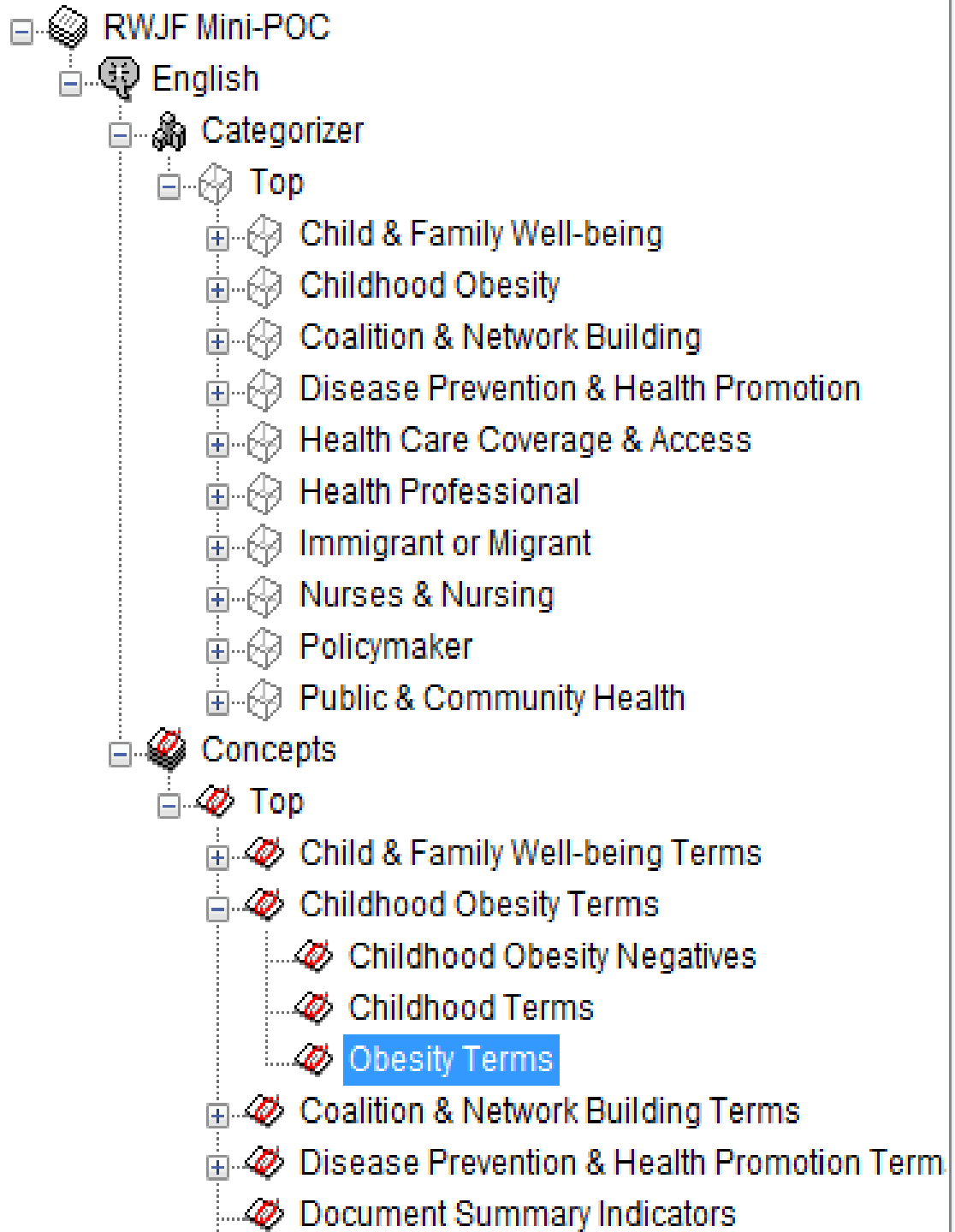
- Mini Categorization POC – 40 hours, 10 categories, 20 documents per category
  - Initial content structure model and rules
  - Develop terms – positive and negative until 90% +
- Scale to enterprise – range of approaches
  - Text Mining for terms, distribute tasks - SMEs
  - Organic – series of Mini-POCs – get important types done and use



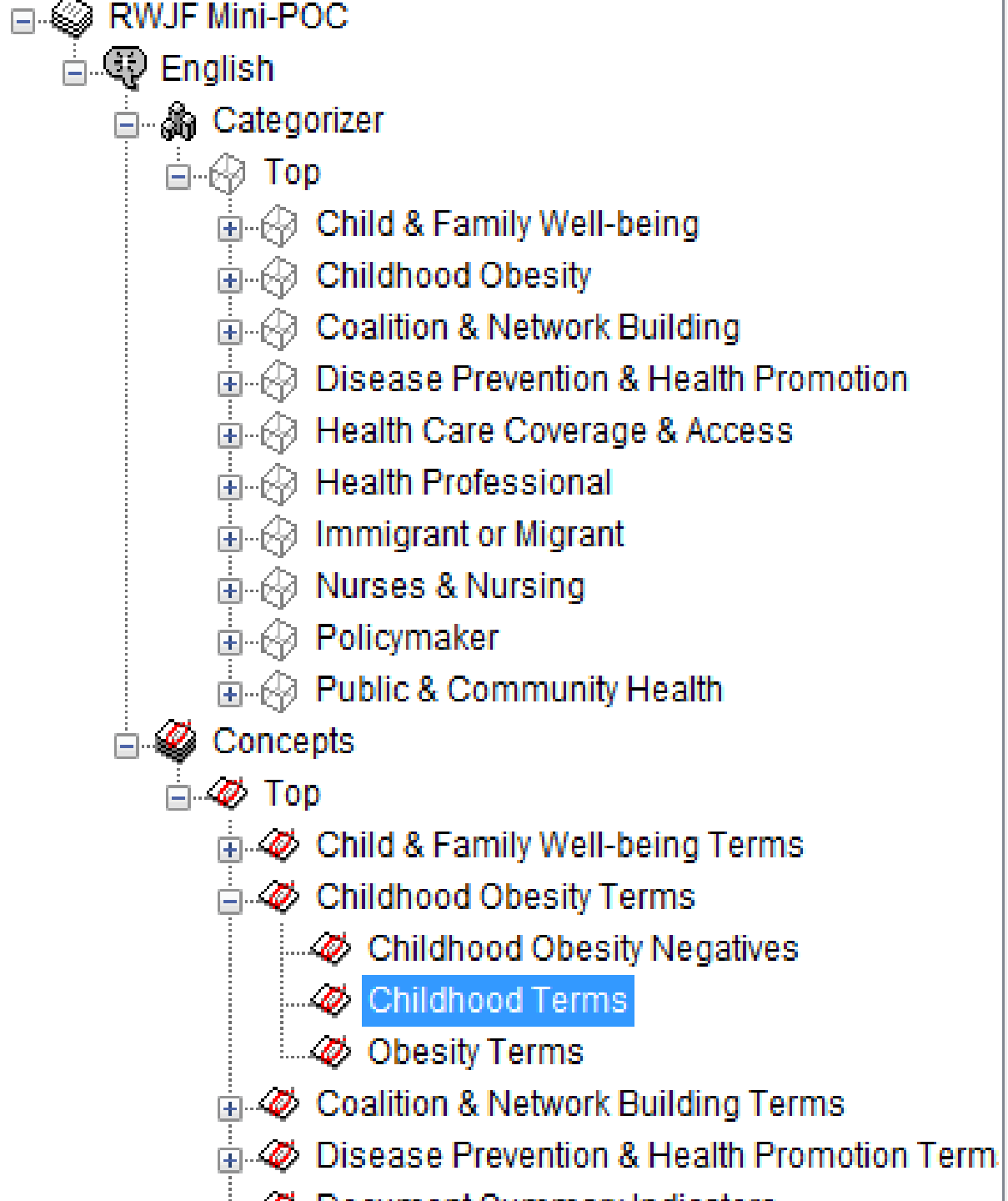
SUMMARY,
Summary,
Executive Summary,
executive summary,
Abstract,
Takeaways,
Overview,
overview,
Commentary,
Introduction,
introduction,
INTRODUCTION,
Key Meeting Themes,
Abstract,
ABSTRACT,
KEY WORDS,
KEYWORDS,
Overview,
issue brief,
Issue Brief,
Framing the issue,
Summary Brief,
Introd duction,
Contents,
SYNOPSIS,
Objective.,
eXeCUTive sUmmary,
EcEcutivESummary,
RESULTS,
Results,
Context,




```
Coalition Building,
coalition building,
Coalition,
coalition,
Coalitions,
coalitions,
Network Building,
network building,
collaboration,
Collaboration,
Collabortive,
collaborative,
Collaboratives,
collaboratives,
Partners,
partners,
Interprofessional Collaboration,
interprofessional teams,
Interprofessional,
interprofessional,
Partnership,
partnership,
Partnerships,
partnerships,
PARTNERSHIPS,
PARTNERSHIP,
building relationships,
NBCH,
National Business Coalition on Health,
```




```
obesity,
Obesity,
overweight,
obese,
Obese,
soft drink,
Snack Foods,
snack food,
snack foods,
Sugar-Sweetened Beverages,
Sugar-Sweetened,
sugar-sweetened,
obesity epidemic,
Nutrition,
nutrition,
```




```
child,
Child,
children,
Children,
youth,
Youth,
adolescents,
Adolescents,
school,
schools,
Schools,
School,
5th grade,
kindergarten,
```

```
(OR,(START_100,
(AND,
(OR,"[Child & Family Well-being Terms]",
(DIST_7, "[Child & Family Terms]", "[Well-being Terms]")),
(NOT,
(START_100,"[Child & Family Well-being Negatives]")
))),

(AND,
 (ORDDIST_500,
"[Document Summary Indicators]",
(OR,
"[Child & Family Well-being Terms]",
(DIST_7, "[Child & Family Terms]", "[Well-being Terms]"))),
(NOT,
(AND,

(DIST_25,
(AND,
 (ORDDIST_500,
"[Document Summary Indicators]",
(OR,
"[Child & Family Well-being Terms]",
(DIST_7, "[Child & Family Terms]", "[Well-being Terms]"))),
(AND,
 (ORDDIST_500,
"[Document Summary Indicators]",
 "[Child & Family Well-being Negatives]"))
))))))
```

Content Structure Models
Structure Rules Basic Logic

- Count terms that are in the list and in the first 100 words unless there are negative terms within 7 words
- Count terms that are in the list and that are within 500 words after a Document Summary Indicator unless there are negative terms within 7 words
  - Document Summary Indicators – 29 terms “Executive Summary”, “Issue Brief”, “Abstract”
- Terms in the list can be phrases or sets of terms within 7 words of each other
- Negative terms are ones that often show up but should belong to another category – they vary by category
  - Child & Family Well-being – “Coverage”, “Obesity”, “Nurses”

Content Structure Models

- Results

Score with Sections Category	Recall	Total Precision	Top 10 Precision	Notes
Child & Family Well-being	95%	100%	100%	
Childhood Obesity	100%	95%	100%	
Disease Prevention & Health Promotion	90%	85%	90%	
Health Care Coverage & Access	95%	95%	100%	
Nurses & Nursing	95%	95%	100%	
Public & Community Health	95%%	70%	100%	
Coalition & Network Building	93%	93%	100%	
Health Professional	85%	100%	100%	
Immigrant or Migrant	100%	94%	100%	
Policymaker	100%	91%	100%	
Average	**95%**	**92%**	**99%**	

Scores without Sections – Full Text	Recall	Total Precision	Top 10 Precision	Notes
Child & Family Well-being	75%	43%	80%	
Childhood Obesity	100%	67%	70%	
Disease Prevention & Health Promotion	50%	27%	40%	
Health Care Coverage & Access	80%	33%	90%	
Nurses & Nursing	40%	27%	80%	
Public & Community Health	45%%	17%	90%	
Coalition & Network Building	73%	48%	90%	
Health Professional	75%	31%	70%	
Immigrant or Migrant	100%	71%	100%	
Policymaker	75%	50%	100%	
Average	**71%**	**41%**	**81%**	

KAPS Group

RWJF Mini-POC
Overview Average Scores

	Recall	Precision	Precision Top 10
With Sections	95%	92%	99%
Full Text	71%	41%	81%
Difference	24%	51%	18%

Content Structure Models Conclusions

- No such thing as unstructured text
- Structure can be captured in a variety of ways
- Content structure models with text analytics provides a means to dramatically improve search and search-based applications
  - And build multiple analytical applications
  - Best is hybrid machine-human tagging
- CSM + TA – can use existing metadata and can create new metadata
- Best way to get value from your taxonomy = Add content structure models and auto-categorization & data extraction
- Don’t believe me? Try a Mini-POC for categorization on your content

Questions?

Tom Reamy
tomr@kapsgroup.com
KAPS Group
Knowledge Architecture Professional Services
http://www.kapsgroup.com

KAPS Group
Knowledge Architecture Professional Services