THE GOOD, THE BAD, AND THE UGLY OF USING THIRD PARTY AI

Kashif Riaz
Director, Data Engineering
Securian Financial

SOMETHING ABOUT ME

Passionate about human languages

Expertise in intelligent processing for

- non-Latin language e.g. Arabic, Chinese. Turkish, Hindi
- resource scarce languages e.g. Persian, Urdu
- Morphological Rich Languages (MRL) e.g. Arabic, Russian, German, Urdu

Enterprise Architecture of large scale systems for data processing

Doctorate in Computer Science specializing in Information Retrieval, NLP, Machine Learning

Adventure travelling and experiencing different cultures

WHAT IS A QUESTION ANSWERING (Q&A) SYSTEM?

Related to Information Retrieval (IR) systems

- Google, Bing, Yahoo are an Information Retrieval system

A Q&A system usually takes a query in a question format and provides a synthesized to the point **answer** that is ***not a document***

- *How tall is Stephen Colbert?*

LETS TALK ABOUT COMPLEXITY

What is an synthesized answer?

Does language paly a role?

- Components of a sentence   S → VP NP
- SVO – English *The cat chased a mouse*
- A language like Urdu has free word order.

What about the enabling technologies that are required to process text?

May be all of this is somewhat easy to do in English

- NER engines have supposedly 95% accuracy according to the marketing literature from the AI vendors

Lets look at a use case study to build a Question Answering system in a legal domain using IBM Watson in English

USE CASE: QUESTION ANSWERING SYSTEM FOR DATA PRIVACY

Regulatory Changes

CCPA, GDPR

Pain Points for A Data Privacy Professional

Data is geographically agnostic, so projects are:

- Multi-jurisdictional / global in scope

Rapidly changing legal & regulatory environment

High Stakes

Lack of reliable resources

SCENARIOS

Chief Counsel for a large, multinational co-operation was asked by Chief Privacy officer about the treatment of employee data

The CEO read in the Wall Street Journal that their main competitor's employee data was stolen in a recent data breach, she wants a complete review of policies and practices as they relate to data breach response and handling of employee data.

A client has asked the data privacy practitioner to provided advice and guidance on the requirements in Canada, Ireland, Germany, Australia, and India.

HOW CAN AI HELP?

Varying jurisdictional content

Common law vs. Coded law

Not all customers are trained lawyers who are able synthesis large legal documents

A Question Answer system with focused answers can help!

COGNITIVE SOLUTION FOR DATA PRIVACY ISSUES

Data Privacy Advisor has a component that is based on Watson Q&A capability (2018)

Watson Discovery Services is an advanced version of Watson

- Based on Machine Learning, Natural Language Processing, Search, and Knowledge Graph
- Domain adaptation on smaller set
  - Uses Statistical Information and Relation Extraction (SIRE)
- Train on large set of documents

Result is a trained Watson in the domain

Figure out IP

TRAINING DATA

Organization provided documents and data that was used by Watson to perform the following steps

Teach Watson the vocabulary of Data Privacy scenarios

- Need for data dictionaries, Acronyms (COPPA)

Teach Watson to understand documents from Data Privacy domain

- Global legislation, regulation, administrative decisions, know-how

Build ground truth

- Question Answers

Find answers and evidence to questions asked of Watson

MINIMAL COMPONENTS TRAINING DATA

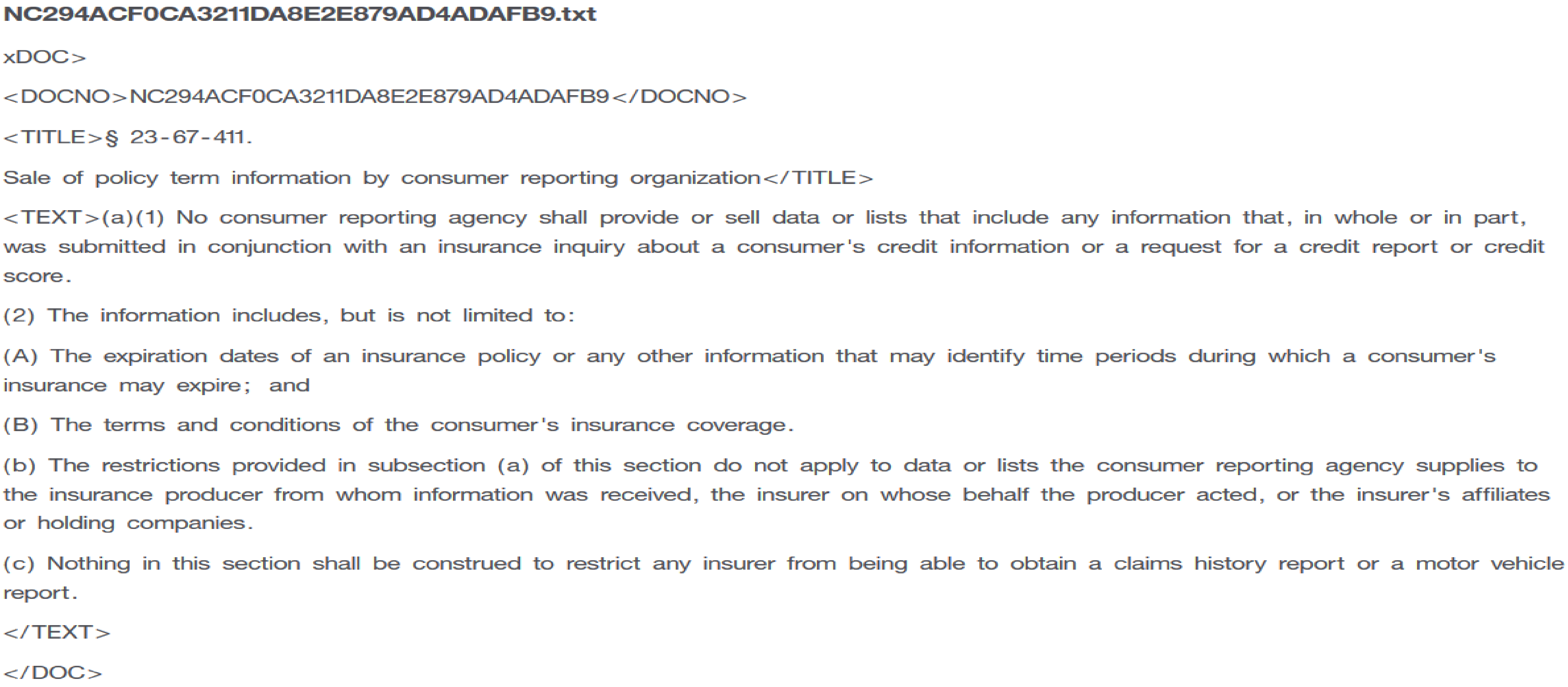
- Annotations of source text documents
  - Teaching Watson how to read and interpret legal and regulatory text
- Paired user questions with correct answers from source text
  - Teaching Watson how to answer questions

ANNOTATIONS OF SOURCE TEXT

Purpose: To break down legal text into components, and provide relationships between those components, to aid in Question Answer training.

- **Who:** Entities that are taking action, or of whom action is required, permitted or prohibited
  - Subject and Object
- **What:** Action that is described or defined
  - Verb/Predicate
- **How / When / Where:** Further parameters for the actions or entities – dates, numbers, processes

SAMPLE DOCUMENT TO ANNOTATE



NC294ACF0CA3211DA8E2E879AD4ADAFB9.txt

xDOC>

<DOCNO>NC294ACF0CA3211DA8E2E879AD4ADAFB9</DOCNO>

<TITLE>§ 23-67-411.

Sale of policy term information by consumer reporting organization</TITLE>

<TEXT>(a)(1) No consumer reporting agency shall provide or sell data or lists that include any information that, in whole or in part, was submitted in conjunction with an insurance inquiry about a consumer's credit information or a request for a credit report or credit score.

(2) The information includes, but is not limited to:

(A) The expiration dates of an insurance policy or any other information that may identify time periods during which a consumer's insurance may expire; and

(B) The terms and conditions of the consumer's insurance coverage.

(b) The restrictions provided in subsection (a) of this section do not apply to data or lists the consumer reporting agency supplies to the insurance producer from whom information was received, the insurer on whose behalf the producer acted, or the insurer's affiliates or holding companies.

(c) Nothing in this section shall be construed to restrict any insurer from being able to obtain a claims history report or a motor vehicle report.

</TEXT>

</DOC>

CHALLENGES TO ANNOTATIONS

Establishing a new classification system

- Continuing variations in content
- Refine definitions and revise elements
- Creating and defining elements that will be consistent across jurisdictions, content types

Obtaining agreement among xx lawyers

Goal: high 80%-90%+ agreement in marked annotations

Resourcing

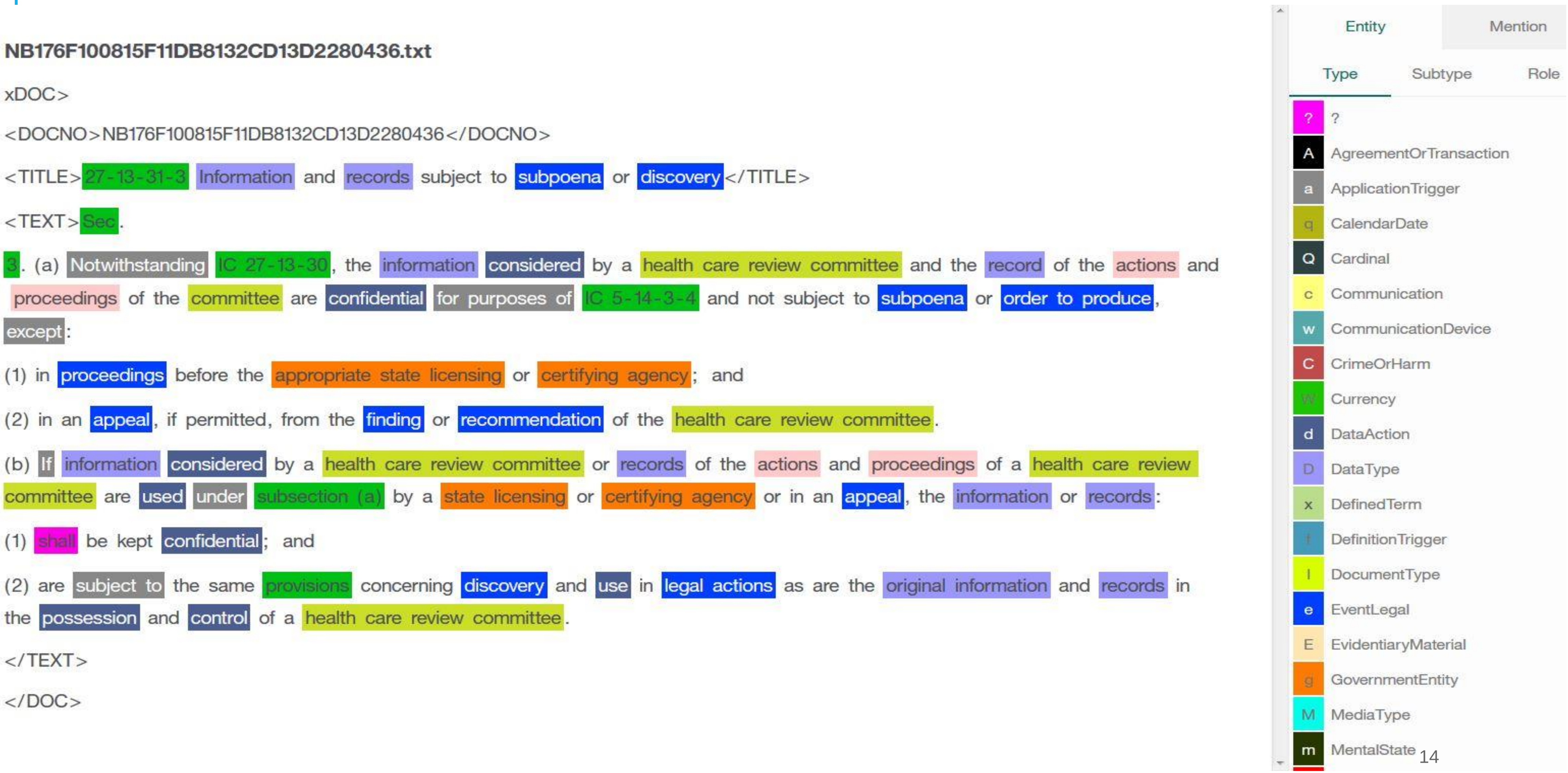
- Using legal editors with production responsibilities – not dedicated full-time to the project

Expectations

- Timeline defined before understanding how long it would take to produce the taxonomy or annotate a legal document

DATA PRIVACY USE CASE :WKS - PREDICATES

Lexical Analysis Type



NB176F100815F11DB8132CD13D2280436.txt

xDOC>

<DOCNO>NB176F100815F11DB8132CD13D2280436</DOCNO>

<TITLE>27-13-31-3 Information and records subject to subpoena or discovery</TITLE>

<TEXT>Sec.

3. (a) Notwithstanding IC 27-13-30, the information considered by a health care review committee and the record of the actions and proceedings of the committee are confidential for purposes of IC 5-14-3-4 and not subject to subpoena or order to produce, except:

(1) in proceedings before the appropriate state licensing or certifying agency; and

(2) in an appeal, if permitted, from the finding or recommendation of the health care review committee.

(b) If information considered by a health care review committee or records of the actions and proceedings of a health care review committee are used under subsection (a) by a state licensing or certifying agency or in an appeal, the information or records:

(1) shall be kept confidential; and

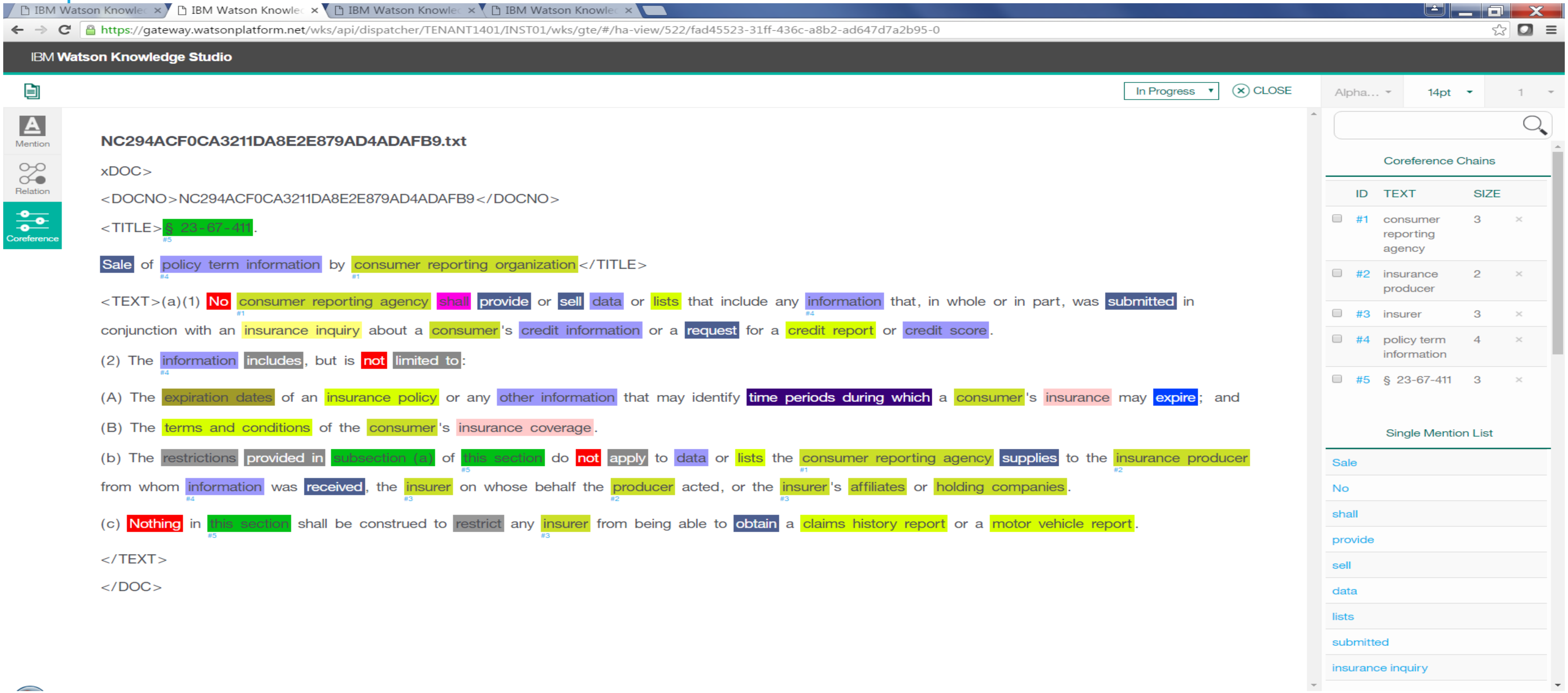
(2) are subject to the same provisions concerning discovery and use in legal actions as are the original information and records in the possession and control of a health care review committee.

</TEXT>

</DOC>

- ? ?
- A AgreementOrTransaction
- a ApplicationTrigger
- q CalendarDate
- Q Cardinal
- c Communication
- w CommunicationDevice
- C CrimeOrHarm
- W Currency
- d DataAction
- D DataType
- x DefinedTerm
- i DefinitionTrigger
- I DocumentType
- e EventLegal
- E EvidentiaryMaterial
- g GovernmentEntity
- M MediaType
- m MentalState

ANNOTATIONS – CO-REFERENCES


IBM Watson Knowle
IBM Watson Knowle
IBM Watson Knowle
IBM Watson Knowle
https://gateway.watsonplatform.net/wks/api/dispatcher/TENANT1401/INST01/wks/gte/#/ha-view/522/fad45523-31ff-436c-a8b2-ad647d7a2b95-0
IBM Watson Knowledge Studio
Mention
Relation
Coreference
In Progress
CLOSE
Alpha...
14pt
1

NC294ACF0CA3211DA8E2E879AD4ADAFB9.txt

xDOC>

<DOCNO>NC294ACF0CA3211DA8E2E879AD4ADAFB9</DOCNO>

<TITLE>§ 23-67-411.

Sale of policy term information by consumer reporting organization</TITLE>

<TEXT>(a)(1) No consumer reporting agency shall provide or sell data or lists that include any information that, in whole or in part, was submitted in conjunction with an insurance inquiry about a consumer's credit information or a request for a credit report or credit score.

(2) The information includes, but is not limited to:

(A) The expiration dates of an insurance policy or any other information that may identify time periods during which a consumer's insurance may expire; and

(B) The terms and conditions of the consumer's insurance coverage.

(b) The restrictions provided in subsection (a) of this section do not apply to data or lists the consumer reporting agency supplies to the insurance producer from whom information was received, the insurer on whose behalf the producer acted, or the insurer's affiliates or holding companies.

(c) Nothing in this section shall be construed to restrict any insurer from being able to obtain a claims history report or a motor vehicle report.

</TEXT>

</DOC>

Coreference Chains

	ID	TEXT	SIZE	
□	#1	consumer reporting agency	3	×
□	#2	insurance producer	2	×
□	#3	insurer	3	×
□	#4	policy term information	4	×
□	#5	§ 23-67-411	3	×

Single Mention List

- Sale
- No
- shall
- provide
- sell
- data
- lists
- submitted
- insurance inquiry

BEST PRACTICES – ANNOTATIONS

People

- Knowledgeable SMEs who are used to reading legal documents very closely
- Really picky about language, grammar, interpretation – strive for perfection
- Accustomed to production deadlines

Collaboration

- Daily meetings to walk through a shared example and agree on the correct markup for all elements
- Partnering 2-3 people together to further consistency
- Assigning x% of documents to more than one person, and resolving differences

Tools

- Tool that is easy to use – visual elements, keyboard shortcuts
- Choose short example documents to minimize unhelpful, repetitive work
- Written guidelines with lots of examples

QUESTION / ANSWER PAIRS

Purpose: To provide examples of how questions can be phrased, what types of information are responsive to those questions, and where that information can be found in text.

Q&A PAIRS – EXAMPLE

Q: It is unlawful for any person to use what kind of means to register for multiple electronic mail accounts or online user accounts from which to transmit to a protected computer a commercial electronic mail message that is unlawful under the CAN-SPAM Act?

A: scripts or other automated means

It is unlawful for any person to use ***scripts or other automated means*** *to register for multiple electronic mail accounts or online user accounts from which to transmit to a protected computer, or enable another person to transmit to a protected computer, a commercial electronic mail message that is unlawful under subsection (a) of this section.*

15 U.S.C.A. § 7704(b)(2)

CHALLENGES WITH QA PAIRS

Question format

- Atypical of real-world questions – specificity, governing law, Yes/No

Answer location

- Not all of the information for some questions can be found in a single source document

Watson’s original short-answer model

- 5-word maximum – does not fit the legal realm very well
- Training only on this question format would miss a large portion of use cases
- Had to work with IBM to convince them that we needed a new paradigm for legal data

Resourcing

- Using legal editors with production responsibilities

Expectations

- Timeline defined before we knew how long it would take to produce the taxonomy or annotate a legal document

BEST PRACTICES – Q&A PAIRS

Process

- Assign a content set to one person at a time to avoid overlap / duplication of work
- Review to ensure the questions fit at least one of the planned project use cases
- More than one reviewer – lots of documents to review per week
- Scheduled weekly submissions
  - Not daily or as work completed, as is done with annotations

People

- Easier to start a new person on Q&A pairs than on annotations
  - Less ramp-up time on requirements

Problem

- Define real-world challenge from the beginning, and get the dev team to listen

DATA: HIGH-LEVEL PROBLEM STATEMENT

Majority of the content already exists and in use in various Legal products

But

- Data pipelines for each jurisdiction and content sets based on product needs
- Initially Watson answers are served for 10 jurisdictions
- The goal is to reuse existing content instead of reacquiring content

Watson requires a **common data model** across jurisdictions and content sets

Watson answers for Data Privacy product must be sentences extracted automatically from data and and run through the Watson cognitive pipeline

COMMON DATA MODEL

Understand the jurisdiction and domain

The most important part of the answer is at different locations in a document

There are no pending legislations in UK. Everything is included in the legislation

Statutory Instruments vs. Regulations

Self Regulatory Organizations content's organization

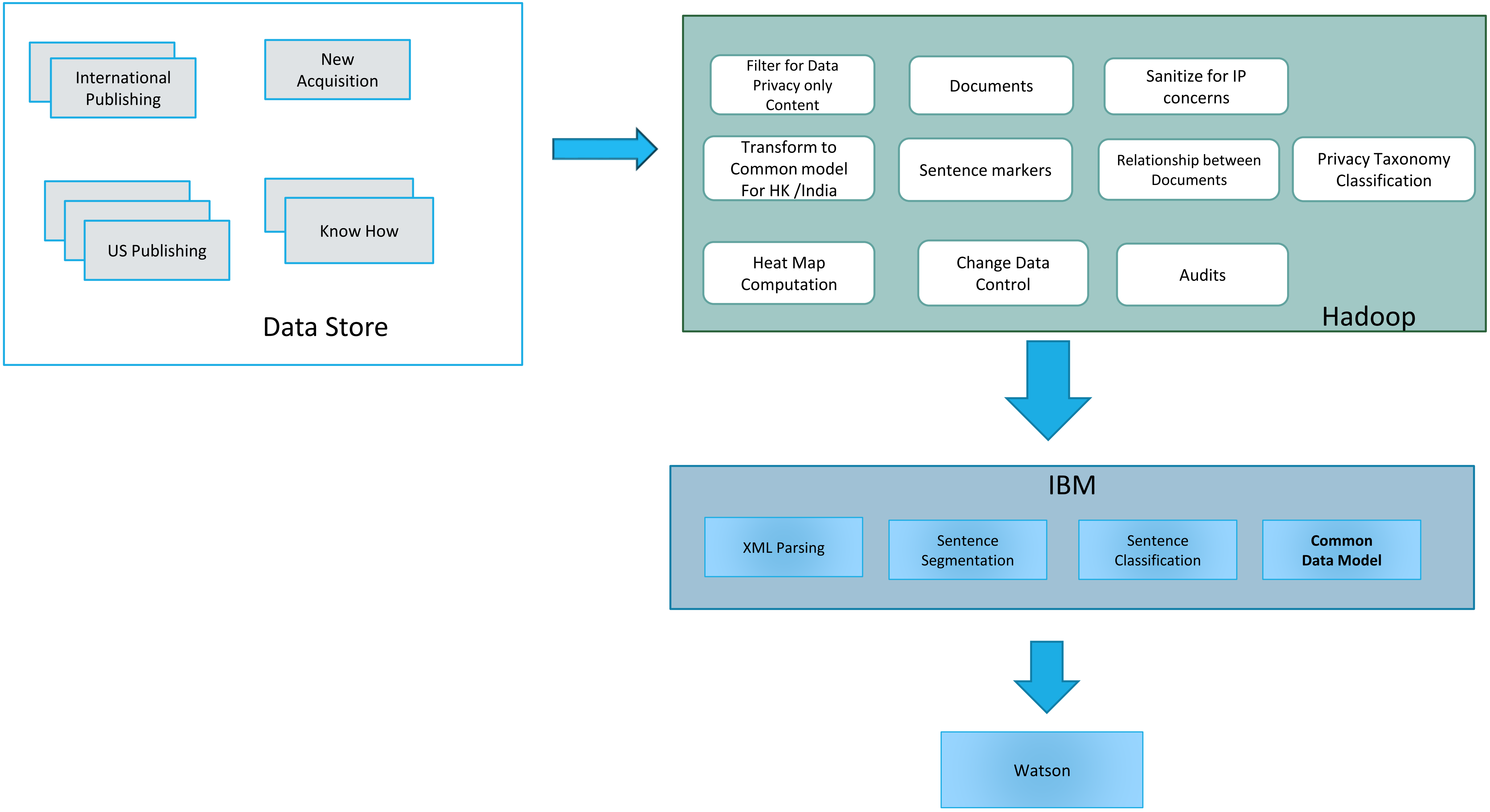
Define the schema for the Minimal Normalized Data Model (MNDM)

Each field is categorized as:

- Display
- Retrieval
- AI model

Process

- Meetings with data owners to map their content on the MNDM
- Who is going to transform the content?



International Publishing

New Acquisition

US Publishing

Know How

Data Store

Filter for Data Privacy only Content

Documents

Sanitize for IP concerns

Transform to Common model For HK /India

Sentence markers

Relationship between Documents

Privacy Taxonomy Classification

Heat Map Computation

Change Data Control

Audits

Hadoop

IBM

XML Parsing

Sentence Segmentation

Sentence Classification

Common Data Model

Watson

DATA CHALLENGES

Sentence identification

The data format is “XML”, ranges from valid to barely well formed.

Persistent Ids for the sentences → what if there is an update to a document

Repository of sentences and their IDs

Connecting them back to the training sets to up keep the validity of the training data.

Classification of sentences with the data privacy taxonomy

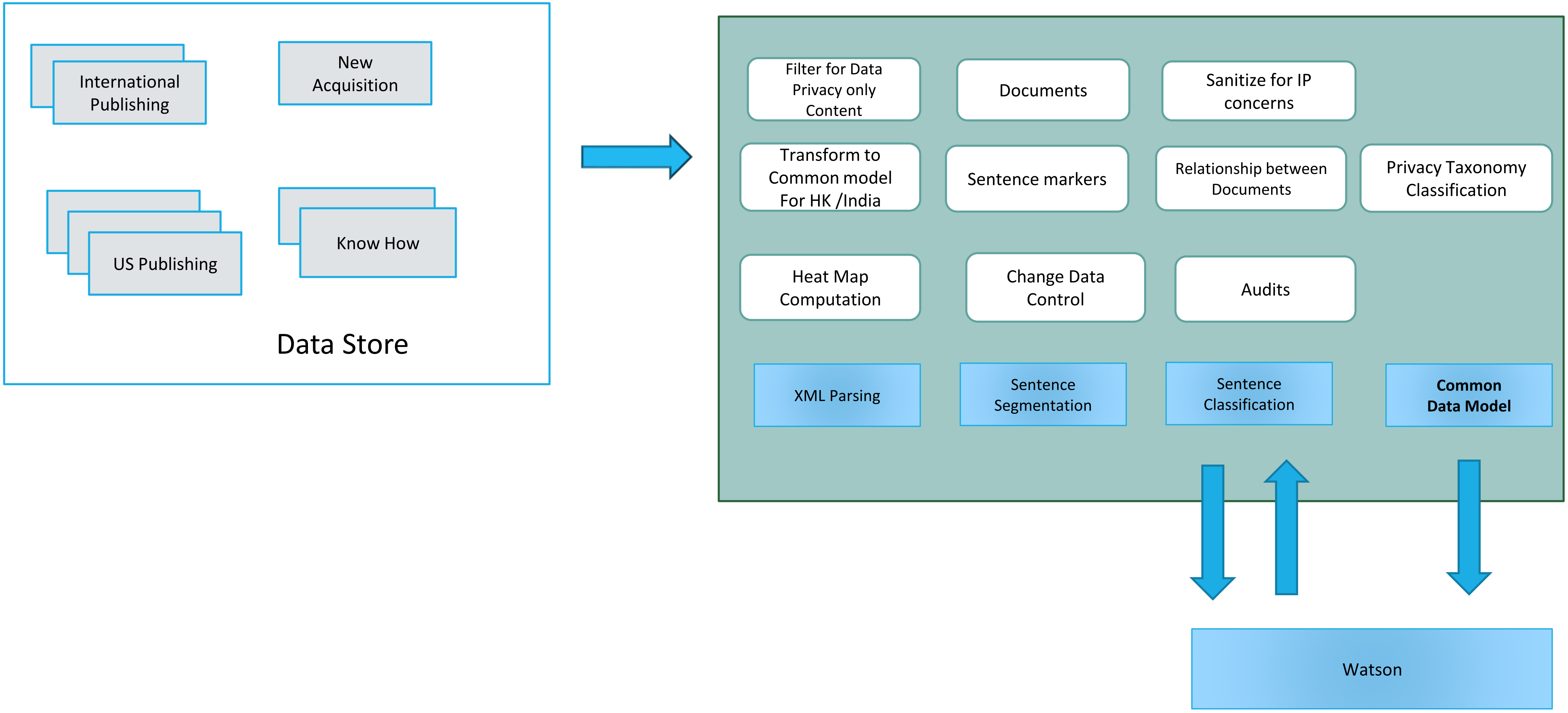
Building Change Data Control

Multiple updates of document in a day e.g. INSERT, DELETE, INSERT, UPDATE, DELETE

Auditing to check if data in house and a Watson is the same.

Requirement for deletion of document from the repository and training data and computed models → therefore all sentences

- Immediate delete of a document → therefore all sentence



International Publishing

New Acquisition

US Publishing

Know How

Data Store

Filter for Data Privacy only Content

Documents

Sanitize for IP concerns

Transform to Common model For HK /India

Sentence markers

Relationship between Documents

Privacy Taxonomy Classification

Heat Map Computation

Change Data Control

Audits

XML Parsing

Sentence Segmentation

Sentence Classification

Common Data Model

Watson

SUMMARY

Understand IP

Understand the reason for doing AI

- Automation or creating a differentiating product

Building an AI system using third party systems still requires significant work by internal resources

Quality of your organization’s data is paramount

AI data pipeline to refresh and keeping the training data relevant

QUESTIONS ?