Canadian Metadata Forum 2005

Metadata: A Reality Check

Auto-Categorization-In-A-Box

What's It All About?

Linda Farmer, Second Knowledge Solutions
Sean Murphy, Deloitte
Susan Thorne, Public Works & Government Services Canada
Clark Breyman, Interwoven

Workshop Agenda

1. **Technology, Value & Issues**
   Linda Farmer, Second Knowledge Solutions

2. **Government of Canada Case Study: Effective Metadata & Content Management**
   Sean Murphy, Deloitte
   Susan Thorne, Public Works & Government Services Canada

3. **Auto-Categorization Under the Hood**
   Clark Breyman, Interwoven

4. **Panel Discussion & Audience Questions**

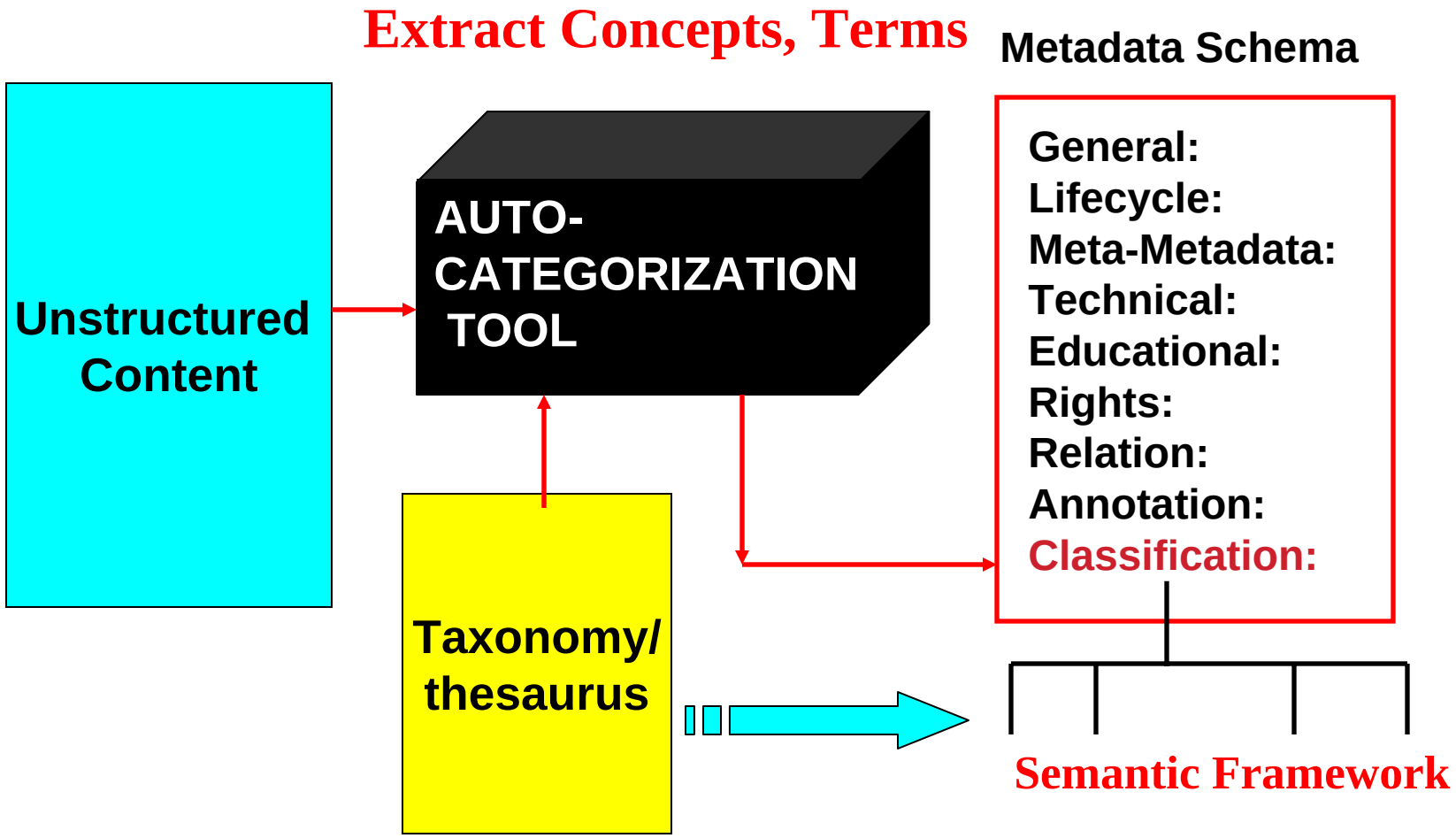
Canadian Metadata Forum, 2005

Auto-Categorization:

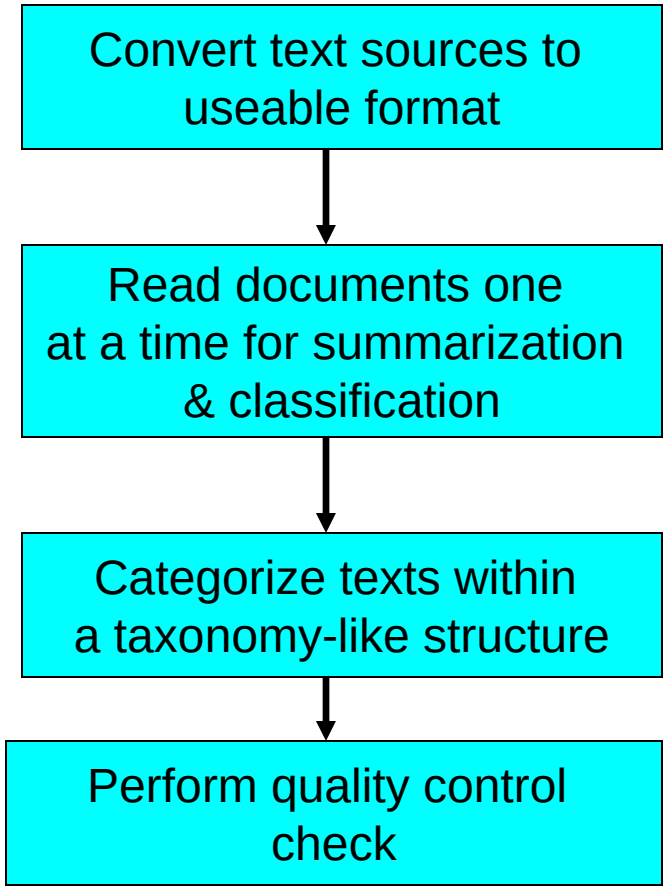
Technology, Value & Issues

Linda Farmer
Second Knowledge Solutions
lfarmer@k2s.ca
http://k2s.ca

Relationship to Metadata

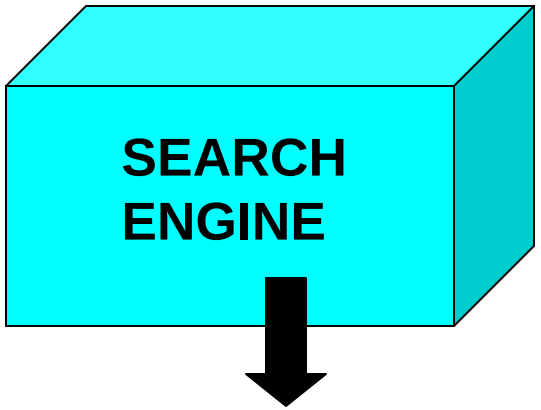


Information Processing Limits



- Lack of speed
- Lack of consistency
- Volume required not achievable
- Time-to-market becomes difficult

Search Technology Limits



- **Crawls for keywords Ignores stopwords**
- **Puts keywords into indexing database with occurrences & locations**
- **Applies Boolean logic for searching**
- **Stems words, ignores plurals**

- Relies primarily on keyword matching
- No relationships between keywords
- "Keyhole" view of content
- No context
- Largely indiscriminate retrieval of information

What's Needed

- All high volume, information-dependant industries are desperate for better content management and retrieval tools

- Tools that organize content, provide structure and serve up relevant information

Taxonomy & Classification Technology

- **Taxonomies** for giving semantic structure to content

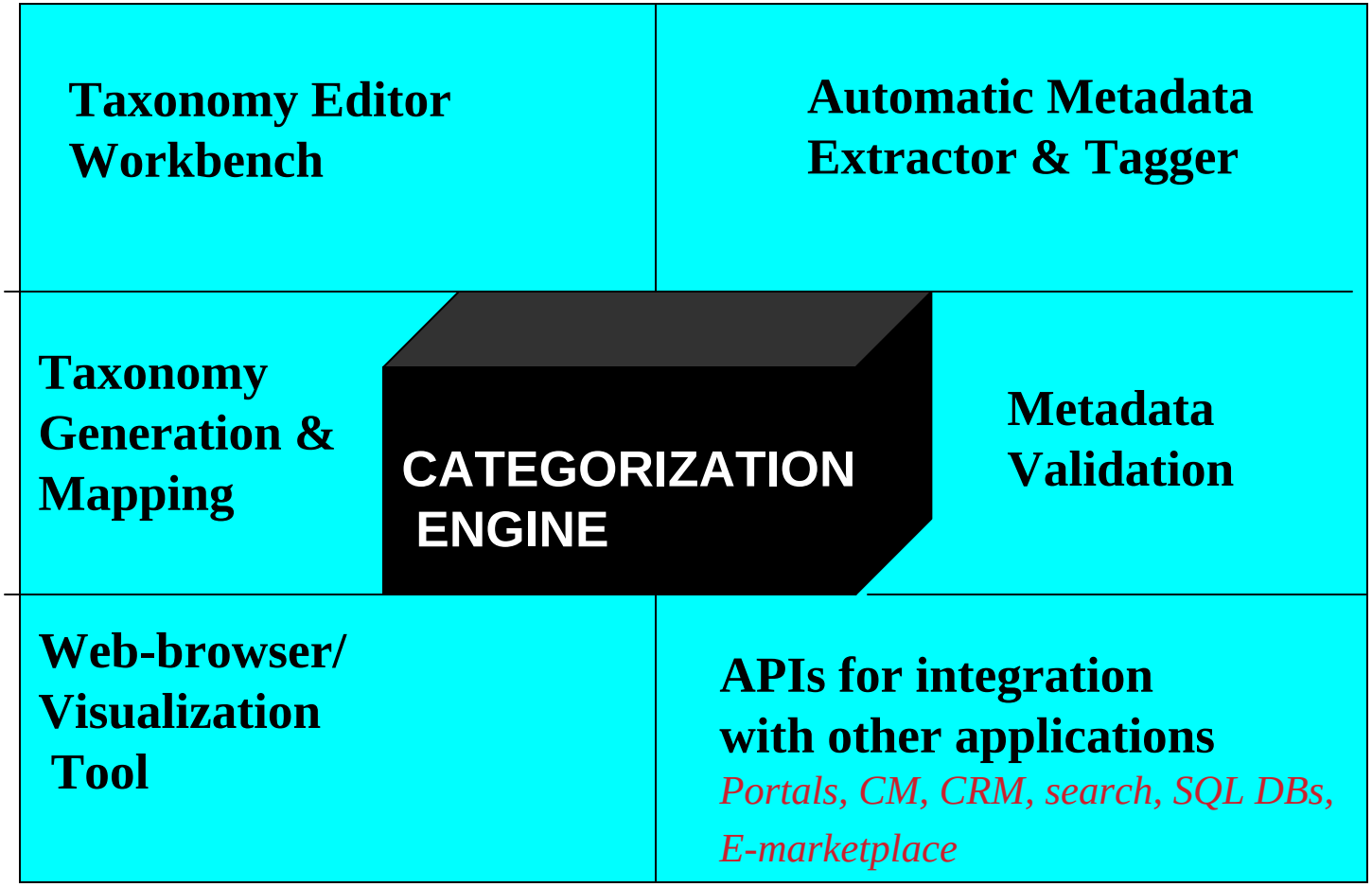
- **Auto-categorization tools**
  - Facilitate creation & maintenance of taxonomies
  - Classify/categorize content

Auto-Categorization Tools

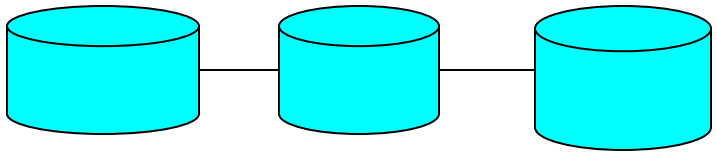
Lifeline for the enterprise swimming in unstructured information

1. Develop **taxonomy** structure
2. **Classify** existing collections of unstructured content
3. Apply **metadata** to content

Auto-categorization Components



Categorization Engine



Unstructured Networked Sources

Extraction of concepts, phrases, categories for taxonomy creation

CATEGORIZATION ENGINE

Black Box

- Rule-Based
- Statistical Analysis
- Semantic & Linguistic Clustering
  (Concept extraction)

Rule-Based Approach

- Precisely defines the criteria by which a document belongs to a specific category
- Matches terms in thesaurus to words in content
- Rules can also employ metadata values
- Experts organize concepts into categories using “If-Then” rules

If word=“shrub”, then assign to category=“bush”

If word=“Bush” and within 4 words of “President”, then assign to category=“nil”

If doc. type=email, then assign to category=“Internal Communication”

Rule-Based Approach

Upside

- Rules are powerful and flexible
- Most straightforward and user-controllable
- Can support complex operation & decision trees
- Very accurate

Downside

- Supports classification only
- Rules must be carefully articulated and made as unambiguous as possible
- Expensive human domain experts need to write and maintain rules
- Best for focused and stable subject domains

Statistical Analysis Approach

- Word frequency
- Relative placement of words, groupings
- Distance between words in a document
- Pattern analysis
- Co-occurrence of terms to find clumps or clusters of closely related documents

- Bayesian Probability
- Neural Networks
- Support Vector Machines

Assigns them a category according to a "training" set of documents

Statistical Analysis Approach

Training Set Requirements

- Collect & manually create subsets of 15- 30 documents representative of each topic or node of the taxonomy
- Sample content is analyzed and taxonomy is further refined and rules of classification established
- Rules used to automate the analysis of new documents and their classification into the taxonomy

Statistical Analysis Approach

Upside

- Supports first draft compilation of taxonomy & subsequent classification of content into taxonomy
- Common method used for concept extraction due to computational nature and its fit with computers

Downside

- Classification totally dependant on breadth & precision of manually defined training set
- Setting up and maintaining training set of documents is very time consuming & expensive
- Does not adapt well to changes in taxonomy
- Best used in tandem with linguistic processing

Semantic & Linguistic Clustering

- Language dependant
- Documents clustered or grouped depending on meaning of words using thesauri, parts-of-speech analyzers, rule-based & probabilistic grammar, etc.
- Analyzes structure of sentences

- **Morphological level**
  Analysis of words - prefixes, suffixes, roots
- **Lexical level**
  Word-level analysis incl. part of speech
- **Syntactical level**
  Analyzes structure & relationships between words in a sentence
- **Semantic level**
  Determine possible meanings of a sentence
  Enhanced by statistical analysis.

Semantic & Linguistic Clustering

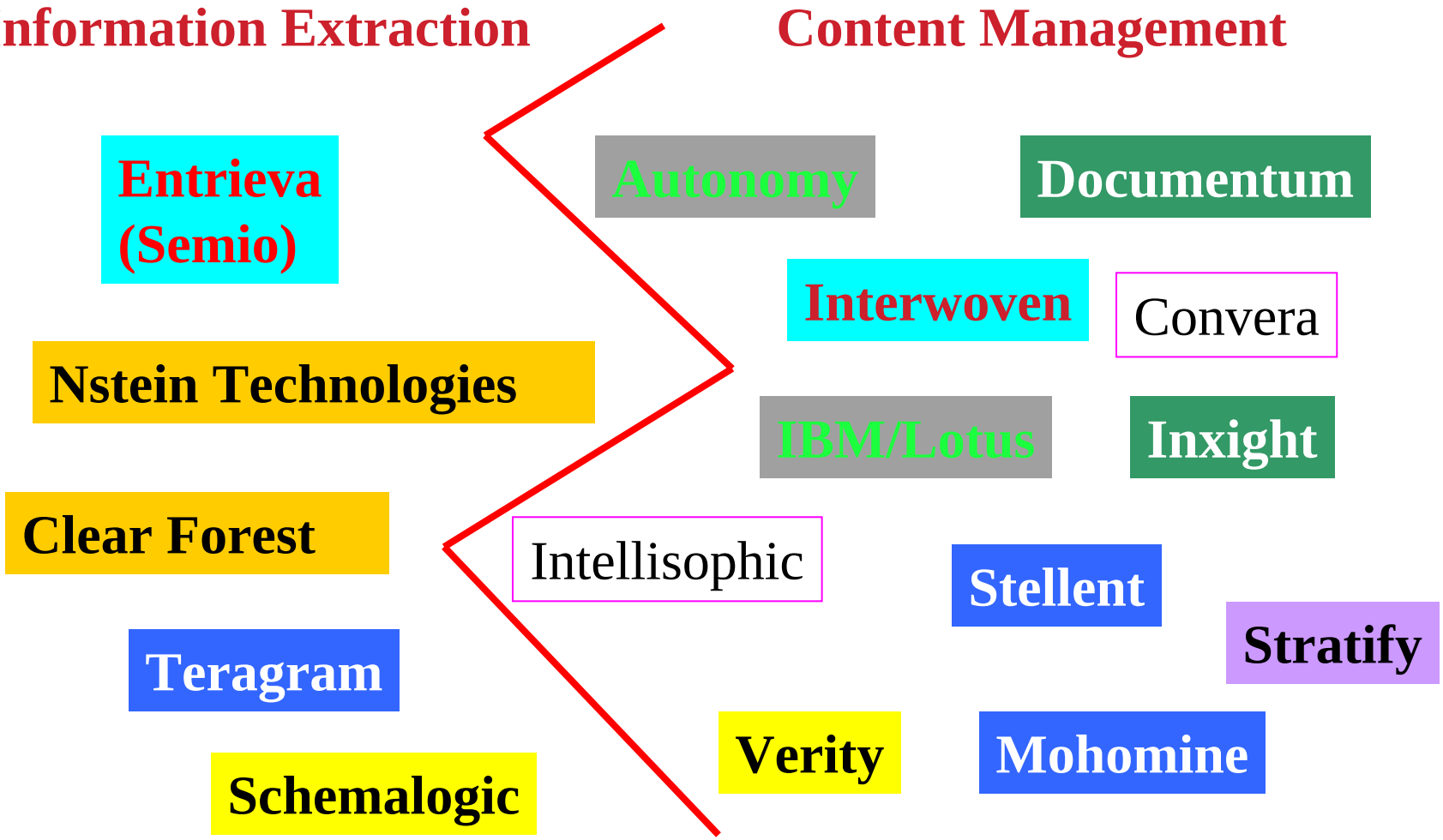
Upside

- Supports both taxonomy creation & classification
- No training set of documents required
- Supports automatic summarization of documents

Downside

- High degree of sophistication required to develop tool

Auto-Categorization Vendors



Auto-Categorization Products

Interwoven
MetaTagger

Inxight
SmartDiscovery
Analysis Server

Convera
- **RetrievalWare Knowledge Discovery Solution**
- **ExcaliburWeb Search**

Nstein
- **Global Intelligent Information Management**
- **Linguistic DNA**

Teragram
- **Categorizer**
- **Entity Extractor**

Documentum
Content Intelligence
Services

SchemaLogic
SchemaServer

Integrator

Entrieva
- **SemioTagger**
- **Semio Skyliner**
- **Knowledge Engineering Workbench**

Verity
- **Collaborative Classifier**
- **Verity Extractor**

Which One to Choose?

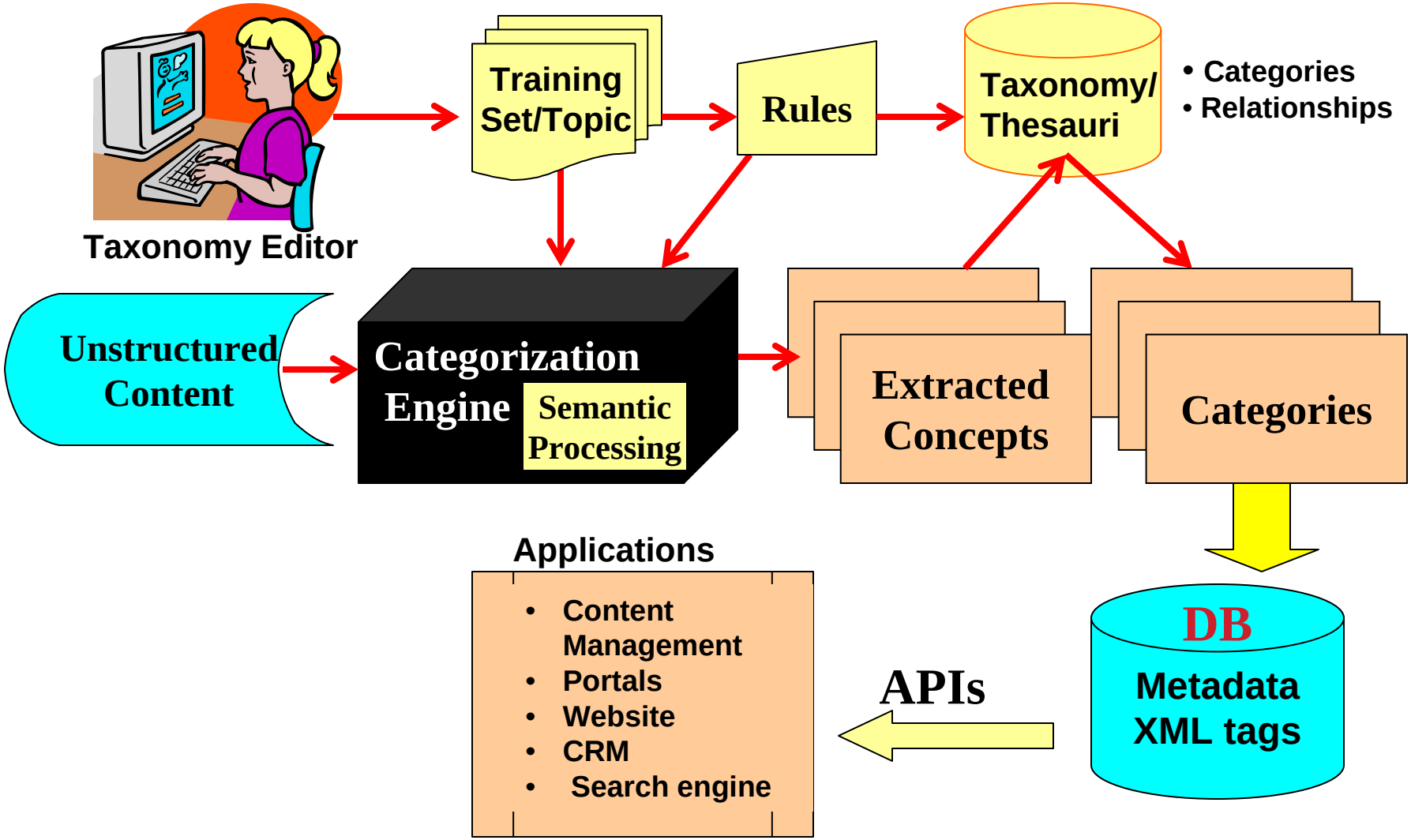
Condundrum

"..there is no universally accepted standard for evaluating the various algorithms or software configurations in regard to speed, accuracy, and scalability of taxonomy technology products." Delphi Group White Paper, 2004

Option

"..test the different solutions against a significant portion of your unstructured data, letting your users verify that the documents are categorized quickly and accurately and on a scale that meets your needs." Delphi Group White Paper, 2004

Categorization Process



Key Features of Tools

- Pre-defined taxonomy templates
- Multiple language support
- Confidence ratings for assignment of a document to a particular category
- Search/discovery tools
- Workflow management
- Entity extraction (people, places, company names, products.etc.) to automatically generate metadata
- Extraction of key sentences to generate text summaries/profiles
- Clustering/tagging on-the-fly
- Multiple taxonomy management

Value of Auto-Categorization Tools

1. **Speed**: Extremely large quantities of documents can be processed very quickly

2. **Superior results:** Generates highly accurate, highly granular categorization creating well indexed corpus of content

3. **Increased scalability:** Easily handles increases in users & documents without need for new products, infrastructure changes

4. **Control & flexibility:** Control over the way documents are categorized and ability to create multiple “views” into the content

Issues of Implementation

1. When does an auto-categorizer become essential?
2. How do you evaluate the performance of an auto-categorization tool?
3. What level of human involvement is desirable, required, or possible?
4. How do controlled vocabularies (CVs) contribute to performance of auto-categorizers?
5. How can categorizing tools help create CVs?

Linda Farmer
Second Knowledge Solutions

lfarmer@k2s.ca

http://k2s.ca

Auto-Categorization Under the Hood

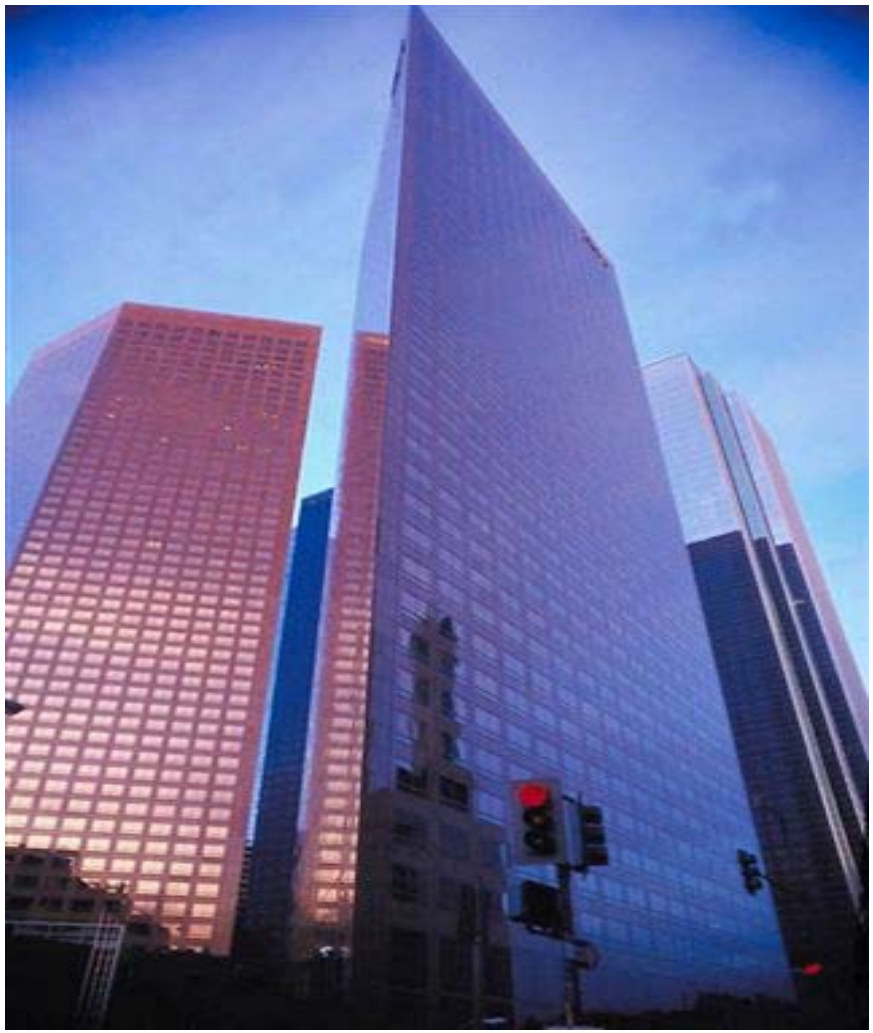
Clark Breyman
Director of Product Management, Interwoven

Agenda

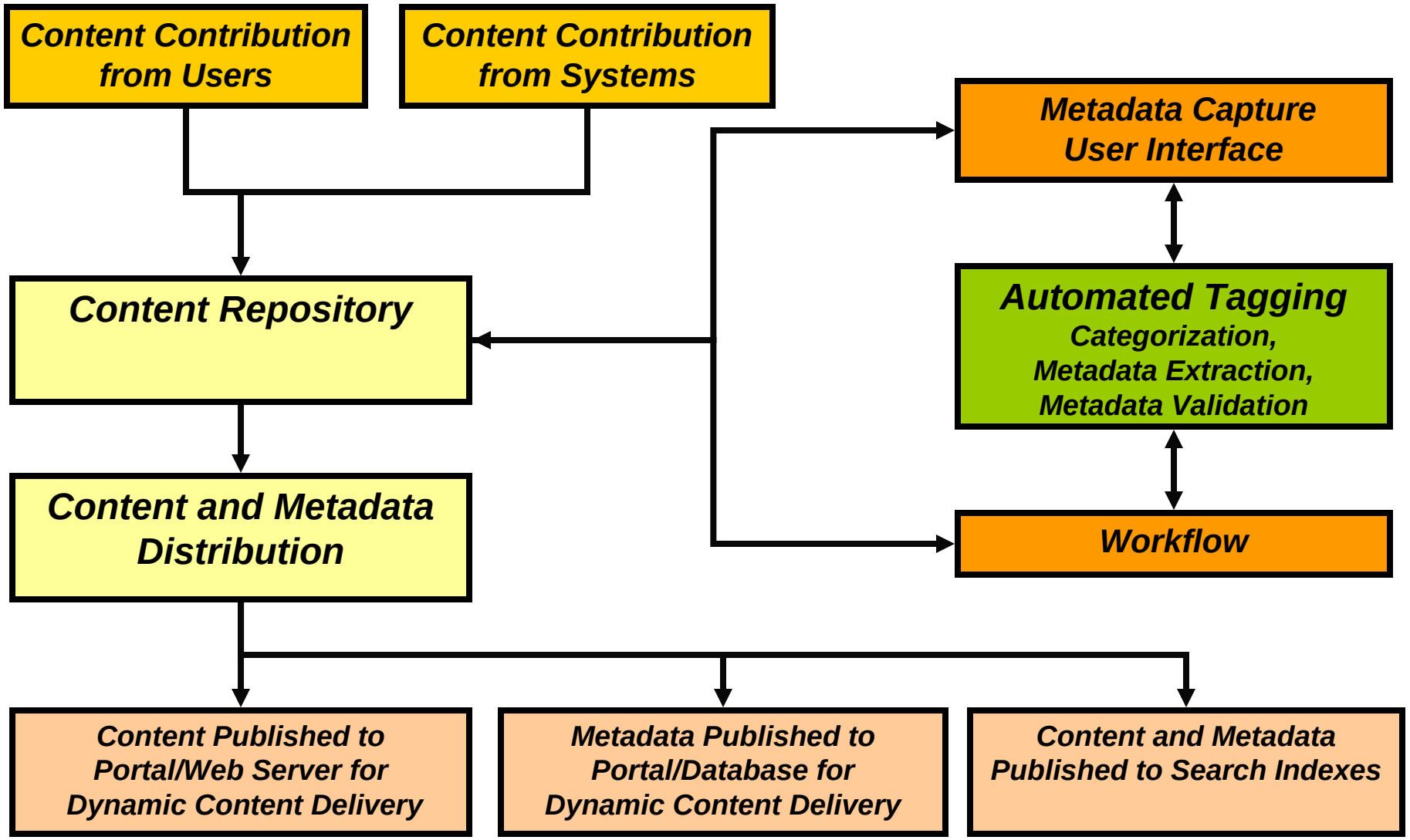
- **Overview**
- **Basic Assumptions**
- **Under-the-Hood:**
  - Content Analysis Stages
  - Contextual Recognition
  - Classification (K-NN)
- **Supporting Technologies**
  - Entity Extraction
  - Collection Profiling
- **Future Directions**

Basic Assumptions

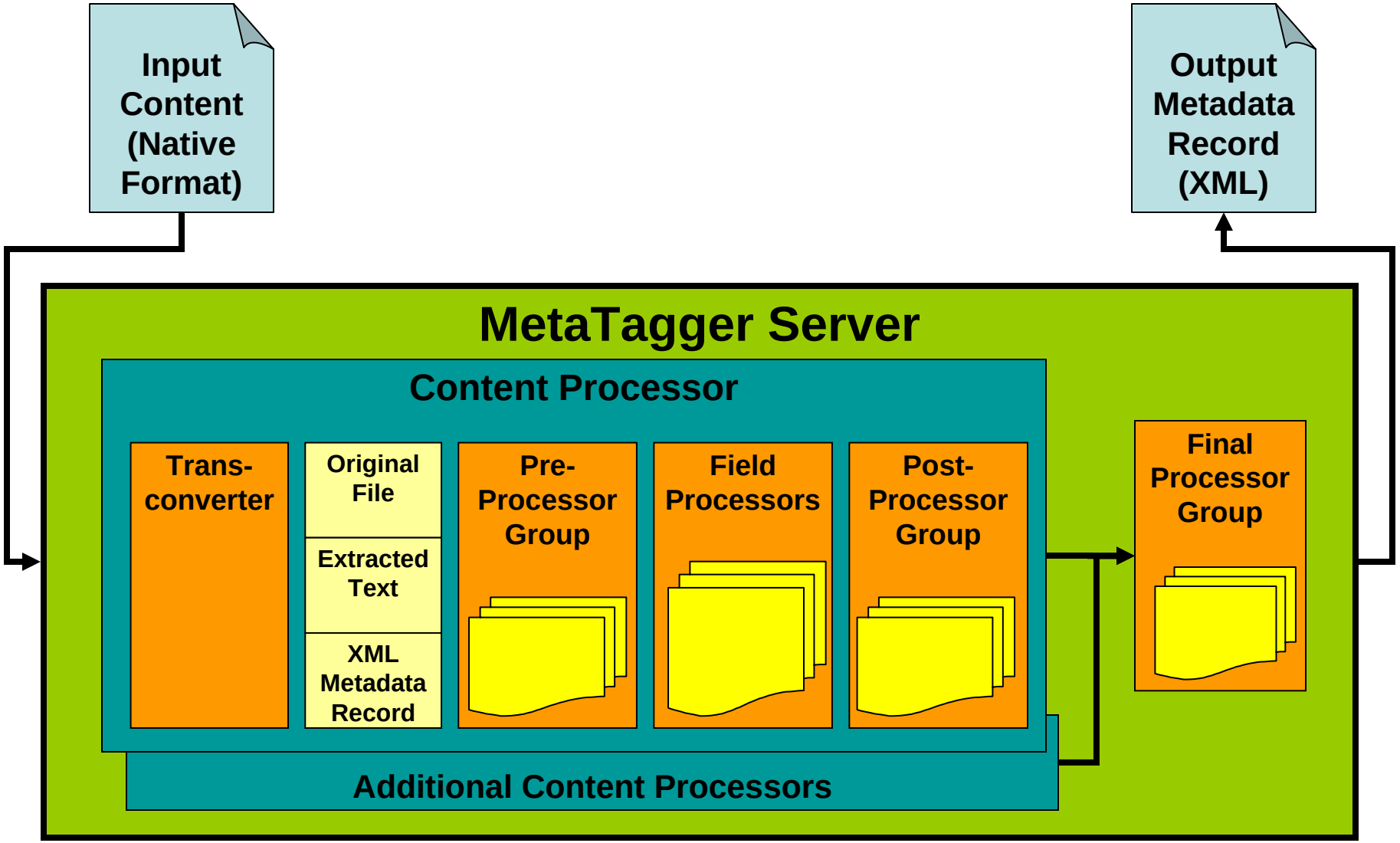
- **The Objective:**
  - A Scalable Content Architecture
- **The Method:**
  - Drive Content Presentation, Storage and Compliance from Metadata
- **Prerequisites:**
  - Metadata Standards: Defined Schemas and Taxonomies
  - Supporting Automation: Minimize Manual Document Review & Metadata Assignment



Metadata-Driven Presentation, Storage & Compliance

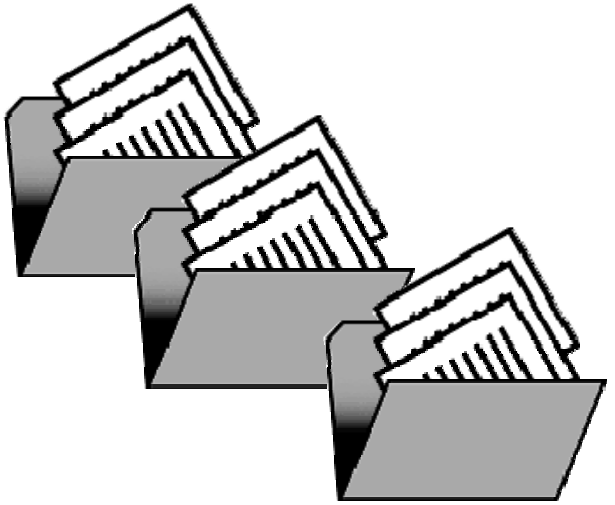


Automated Tagging Process

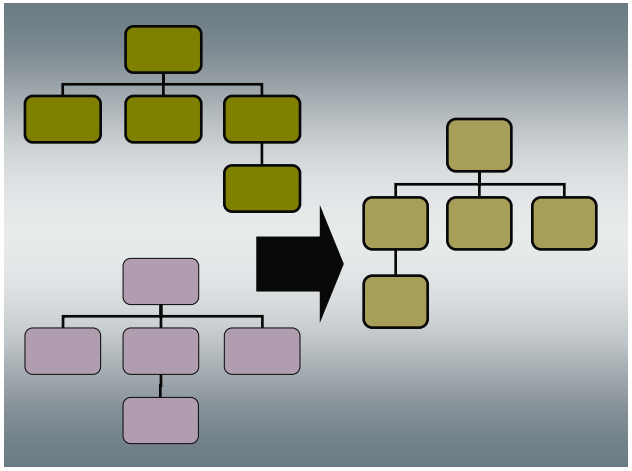


Field Processor Types – Categorization

- **Categorization by Recognition**
  - Categorize by matching words and phrases
  - Resolves ambiguous categories with contextual clues (e.g. financial bank vs river bank)



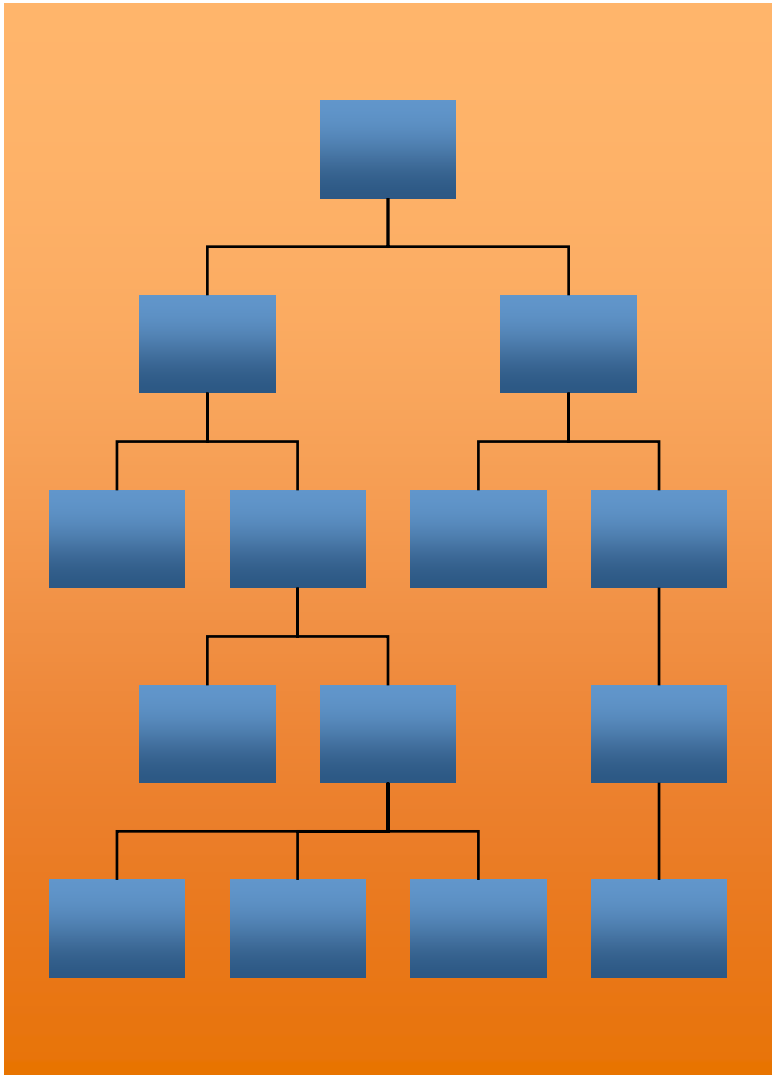
- **Categorization by Example**
  - Categorize using by comparison with expertly classified training documents.



- **Metadata Validation and Mapping**
  - Combine and Standardize Metadata using Business Rules
  - Convert Between Taxonomies

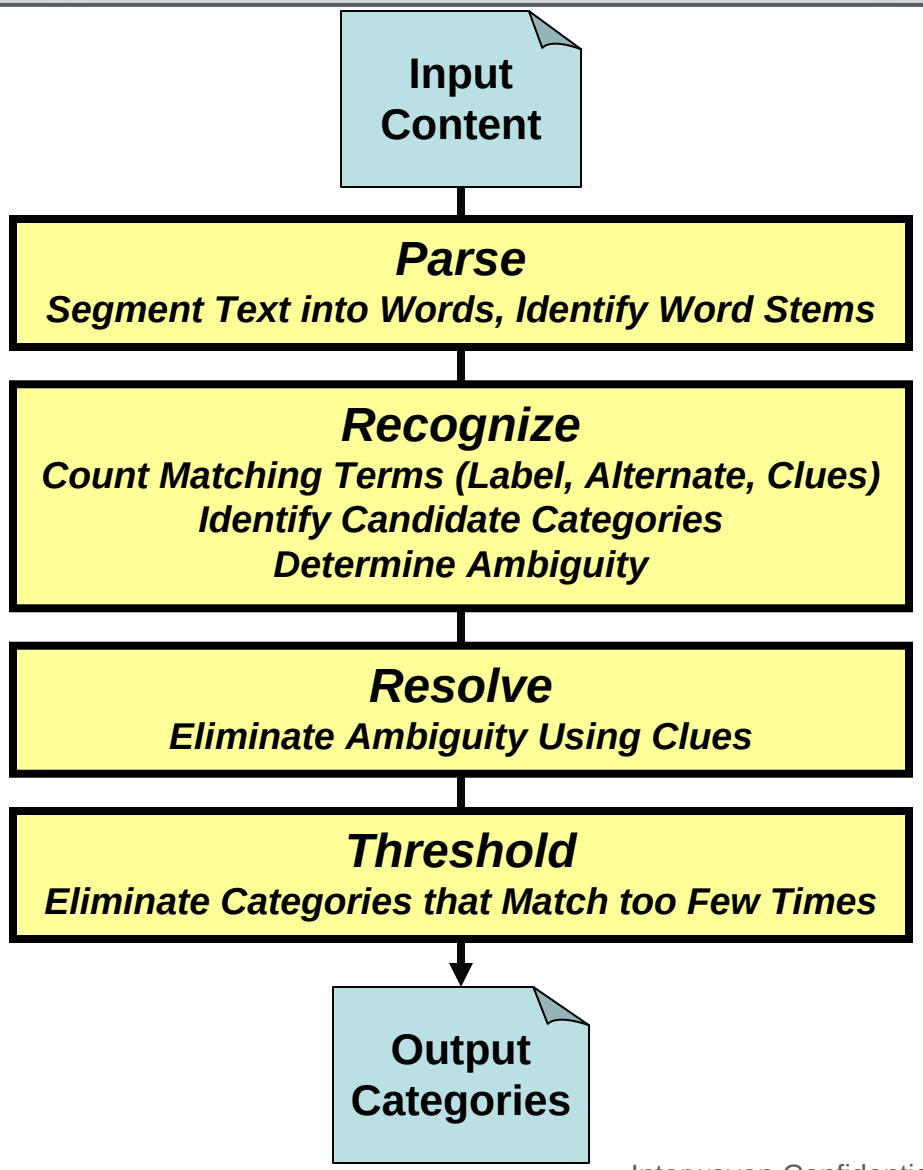
Implementing Taxonomies

Basic Information for All Categories	
UID (Universal Identifier)	A unique code that identifies a category, enabling label and other attributes to be changed and localized as necessary
Label	The display name for a category
Language	The language of the localized category attributes.
Definition	A plain-language description of the category and where it should be applied.
Parents	Parent (more general) categories
Child	Child (more specific) categories
Related	Additional related but distinct categories that may apply.



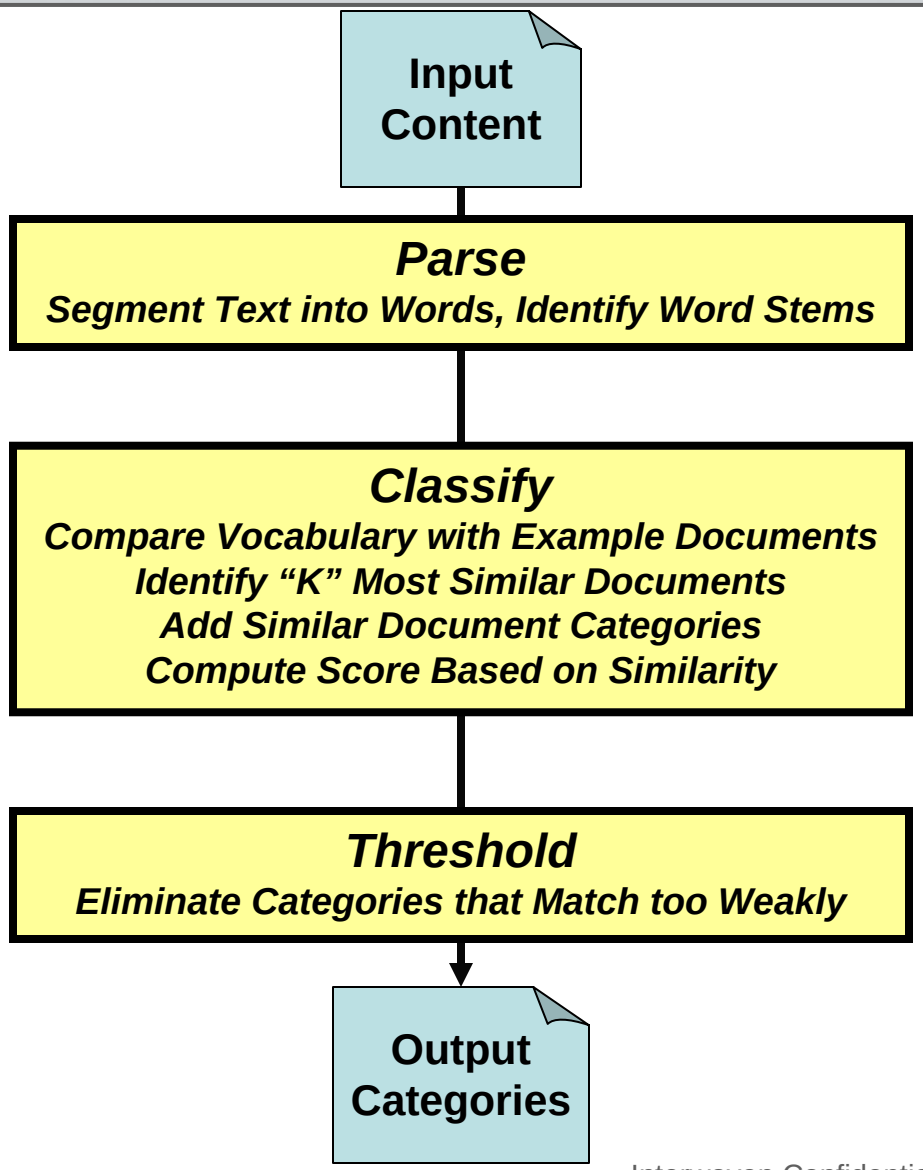
Adding Auto-Categorization: Contextual Recognition

Auto-Categorization Information for Contextual Recognizers	
Alternate Terms	Words and Phrases that indicate that a category applies.
Clue Terms	Words and Phrases used to resolve ambiguity.
Weak	Force label and alternate terms to be considered ambiguous.
Test Documents	Documents that should match a particular category.



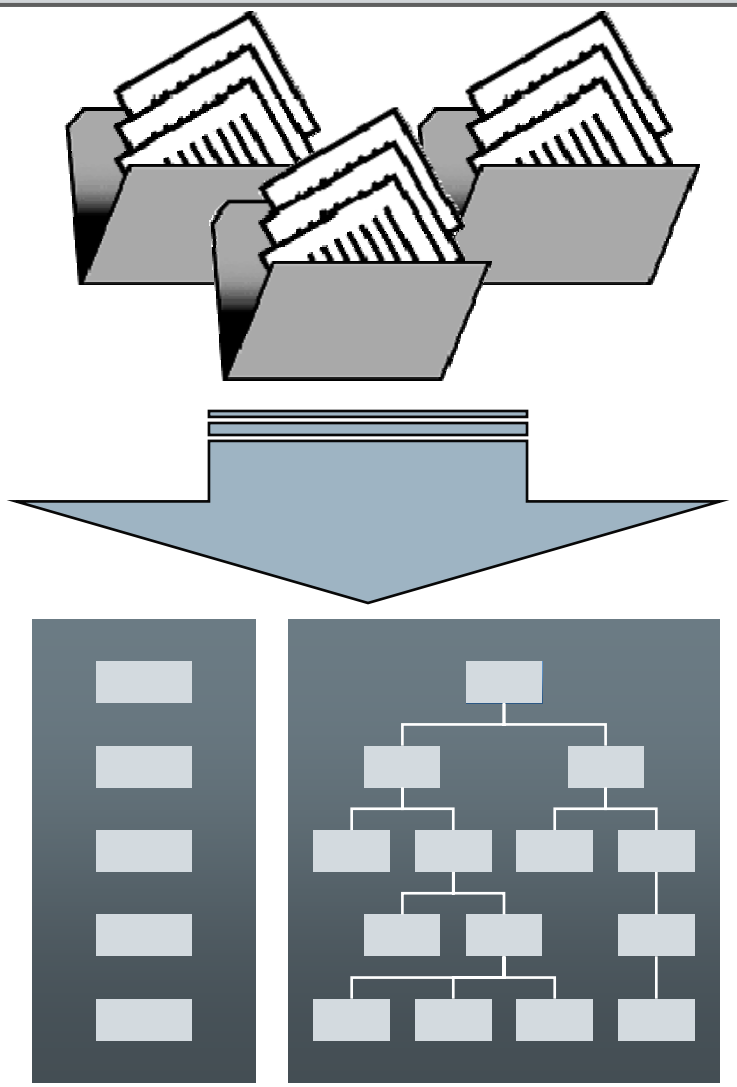
Adding Auto-Categorization: Classification (K-NN)

Auto-Categorization Information for Example-Based Classifiers	
Example Documents	Documents that define by example where a category applies.
Test Documents	Documents that should match a particular category.



Supporting Technologies

- **Entity Extraction**
  - Extract Names, Addresses and other Linguistic Patterns for content cataloging.
- **Content Profiling**
  - Identify Similar Groups in Document Collections (Clustering)
  - Identifying Co-Occurring Terms
- **Summarization**
  - Generate Summaries & Key Phrases



Implementing Entity Discovery & Extraction

- **Word Patterns to Identify Metadata**
- **Essential for Discovery Applications**
  - Social Networks
  - Related Concepts
  - Compliance Audit
- **Character Patterns**
  - URLs
  - Email Address
  - Part Numbers
  - Phone Numbers
- **Term-Type Patterns**
  - Person Names
  - Company Names
- **Hybrid Patterns**
  - Street Addresses
  - Events (e.g. Merger Announcement)

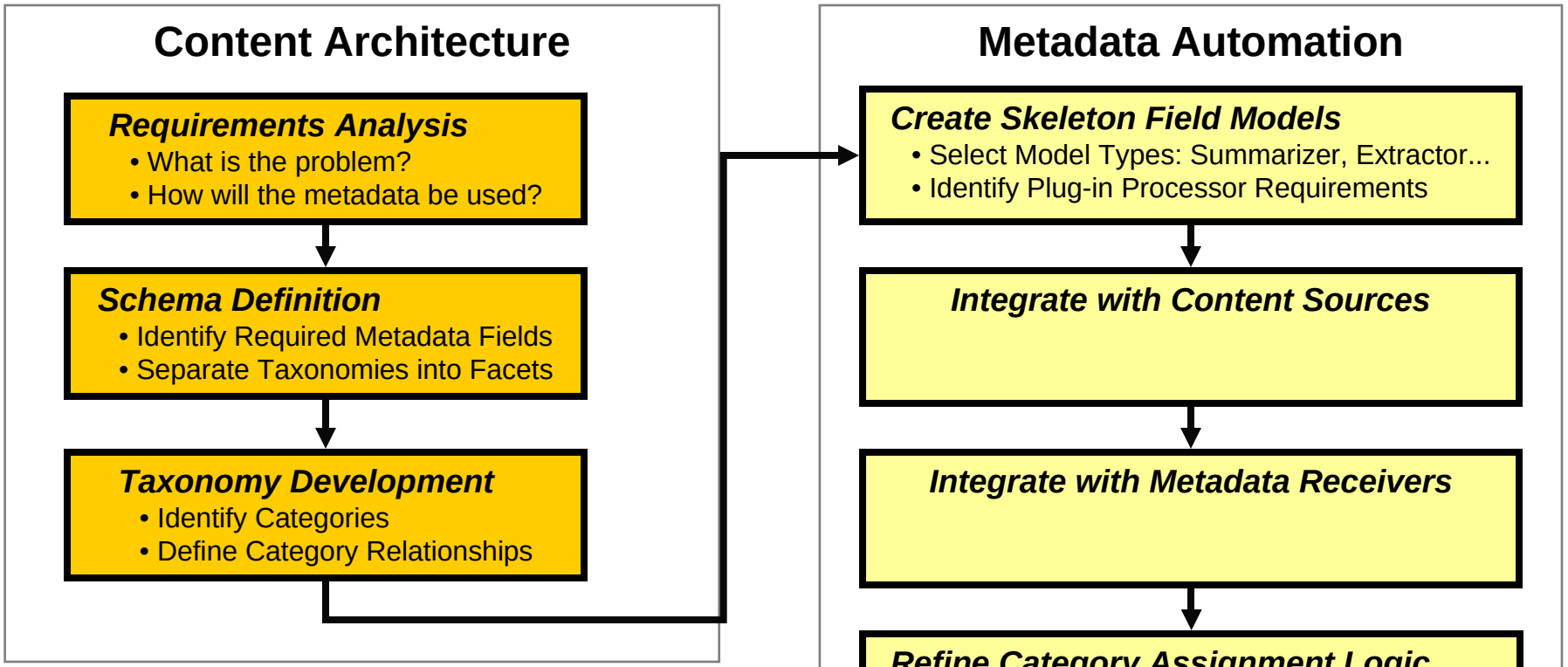
Examples:

```
<extract>
  <pattern>/http:\/\/[A-Za-z0-9\.\/]+/</pattern>
  <action report="true">
      specifier.url
  </action>
</extract>

<extract>
  <pattern>FIRSTNAME LASTNAME</pattern>
  <action report="true">
      name.person
  </action>
</extract>

<extract>
  <pattern>/[0-2]+/ INITCAP STREET</pattern>
  <action report="true">
      specifier.address
  </action>
</extract>
```

Implementation Methodology



Focus on the Business Goals

- Taxonomies are to USE not to HAVE
- Keep Navigational & Descriptive Elements Separate
- Leverage Existing Domain Resources

Future Directions

- **Easier Model Development**
  - Interactive Collection Profiling & Discovery
  - Collection-Driven Suggestions
- **More Powerful Hybrid Models**
  - Category Type Support in Rules Engine
  - Single-Point Authoring for Multiple Models
- **Better Feedback & Tuning Mechanisms**
  - Per-Category Thresholds
  - Trainable Feature Selection
  - Category Drift Analysis

Copyright 2005 Interwoven, Inc. All Rights Reserved

- **This confidential publication is the property of Interwoven, Inc.**

- **No part of this publication may be reproduced, translated into another language or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written consent of Interwoven, Inc. Some or all of the information contained herein may be protected by patent numbers: US # 6,505,212, EP / ATRA / BELG / DENM / FINL / FRAN / GBRI / GREC / IREL / ITAL / LUXE / NETH / PORT / SPAI / SWED / SWIT # 1053523, US # 6,480,944, US# 5,845,270, US #5,384,867, US #5,430,812, US #5,754,704, US #5,347,600, AUS #735365, GB #GB2333619, US #5,845,067, US #6,675,299, US #5,835,037, AUS #632333, BEL #480941, BRAZ #PI9007504-8, CAN #2,062,965, DENM / EPC / FRAN / GRBI / ITAL / LUXE / NETH / SPAI / SWED / SWIT #480941, GERM #69020564.3, JAPA #2968582, NORW #301860, US #5,065,447, US #6,609,184, US #6,141,017, US #5,990,950, US #5,821,999, US #5,805,217, US #5,838,832, US #5,867,221, US #5,923,376, US #6,434,273, US #5,867,603, US #4,941,193, US #5,822,721, US #5,845,270, US #5,923,785, US #5,982,938, US #5,790,131, US #5,721,543, US #5,982,441, US #5,857,036, GERM #69902752.7or other patents pending application for Interwoven, Inc. Misappropriation of the information contained in this publication may be a violation of applicable laws.**

- **Copyright 2005 Interwoven, Inc. All rights reserved. Interwoven, TeamSite, Content Networks, DataDeploy, DeskSite, iManage, LiveSite, FileSite, MediaBin, MetaCode, MetaFinder, MetaSource, MetaTagger, OpenDeploy, OpenTransform, Primera, TeamPortal, TeamXML, TeamXpress, VisualAnnotate, WorkKnowledge, WorkDocs, WorkPortal, WorkRoute, WorkTeam, the respective taglines, logos and service marks are trademarks of Interwoven, Inc., which may be registered in certain jurisdictions. All other trademarks are owned by their respective owners.**

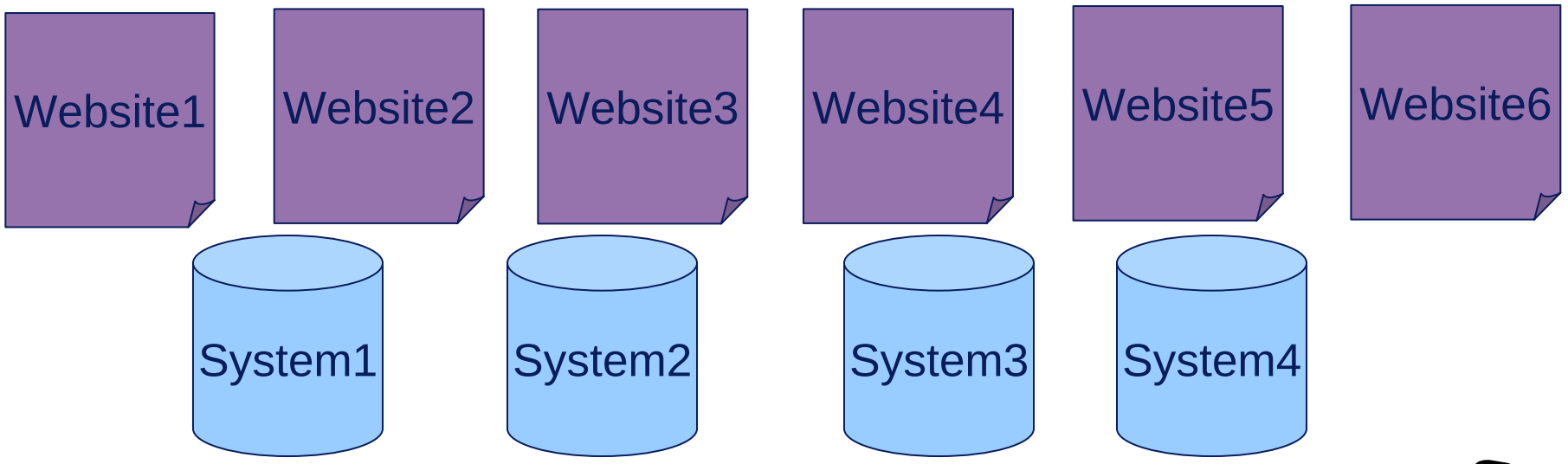
- **All other trademarks are owned by their respective owners.**

GoC Case Study (GoC CMS)
Our Definition of CMS

- Content Management Solutions (CMS) are the technologies, standards, metadata, business processes and people that are required to create, manage and deliver “content”

- “Content” encompasses documents, structured and unstructured data and other materials generally delivered through the Internet to citizens (external users) and to internal users via Intranets and Extranets

GoC Case Study (GoC CMS)
The Problem

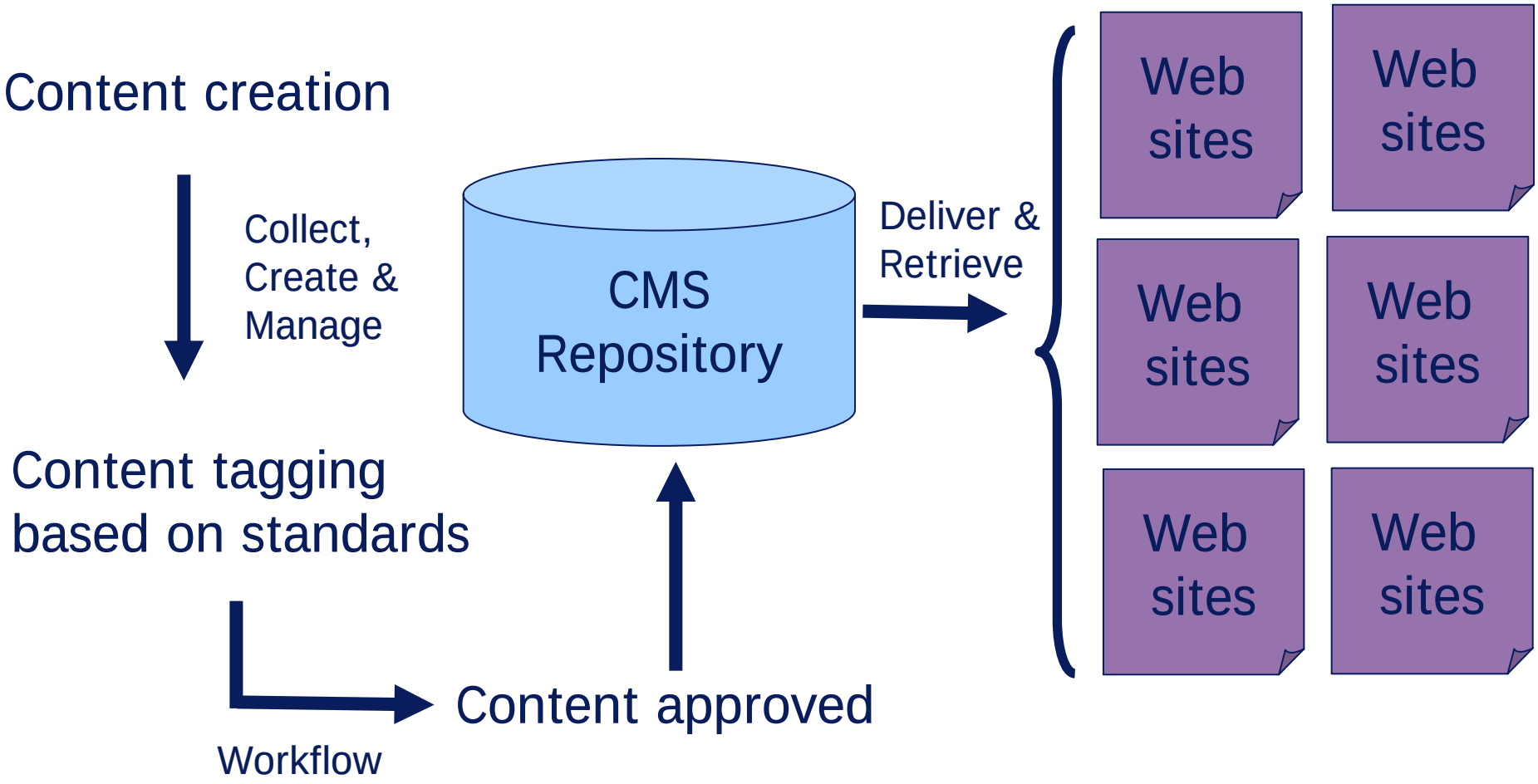


Lots of websites, fed by different systems filled with content written by a variety of groups.

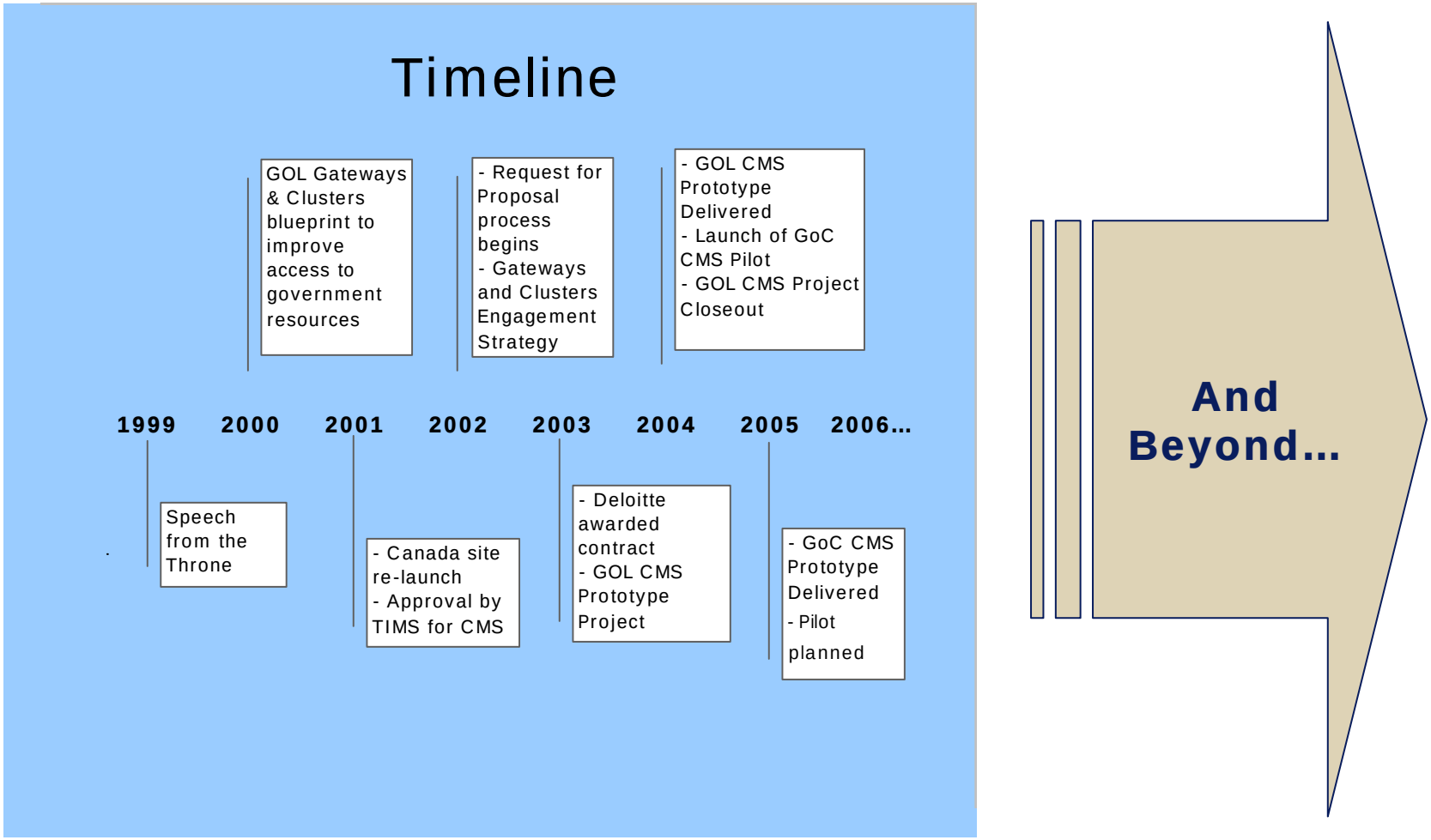
GoC Case Study (GoC CMS)

The Vision

GoC CMS enables stakeholders to manage, share and publish web resources and its metadata in a standard and rational way.



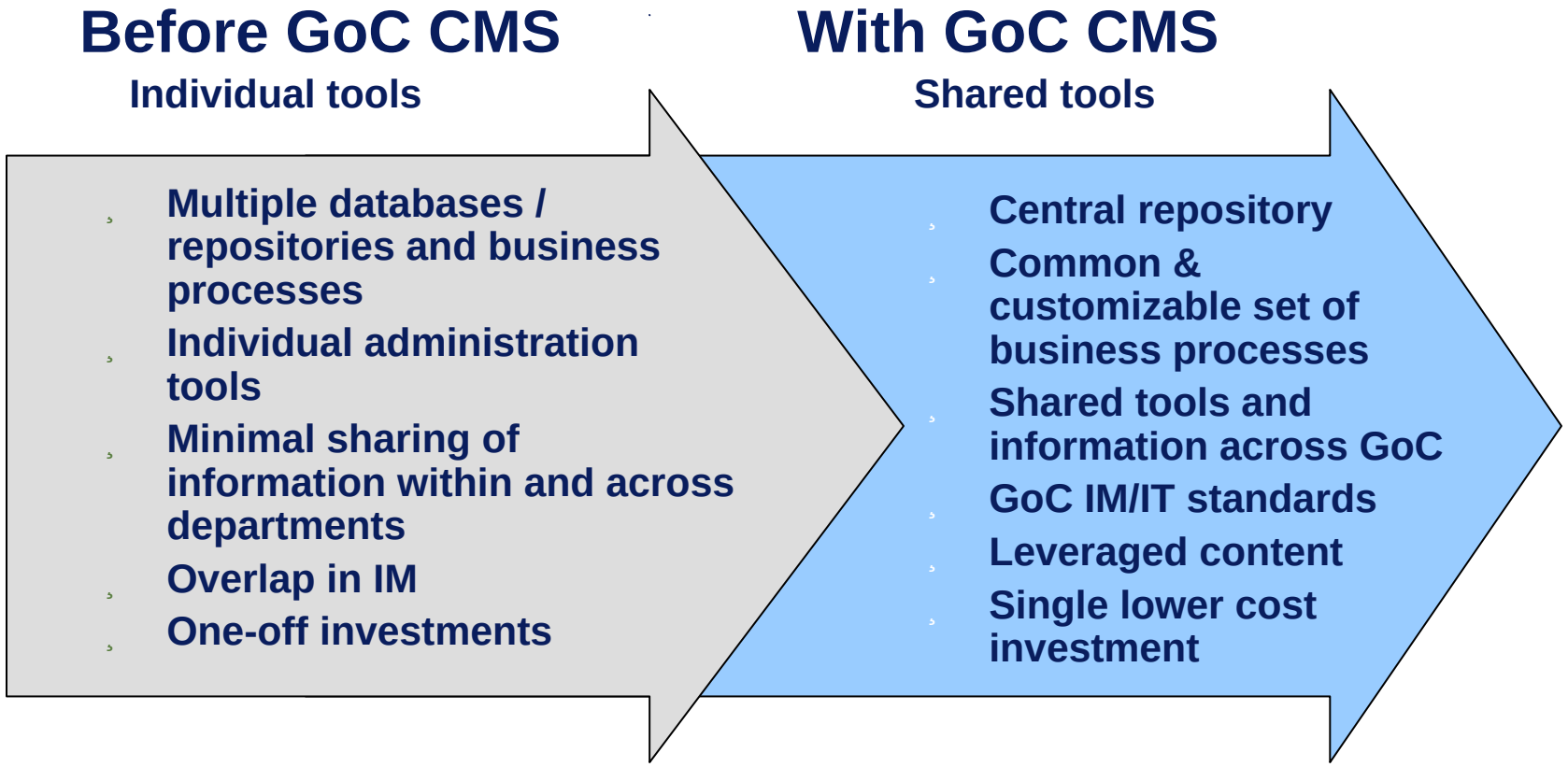
GoC Case Study (GoC CMS)
Background (1)



GoC Case Study (GoC CMS)

Background (2)

Need for an Enterprise Solution



Improved Operations through a Shared Solution!

GoC Case Study (GoC CMS)
The Components (1)

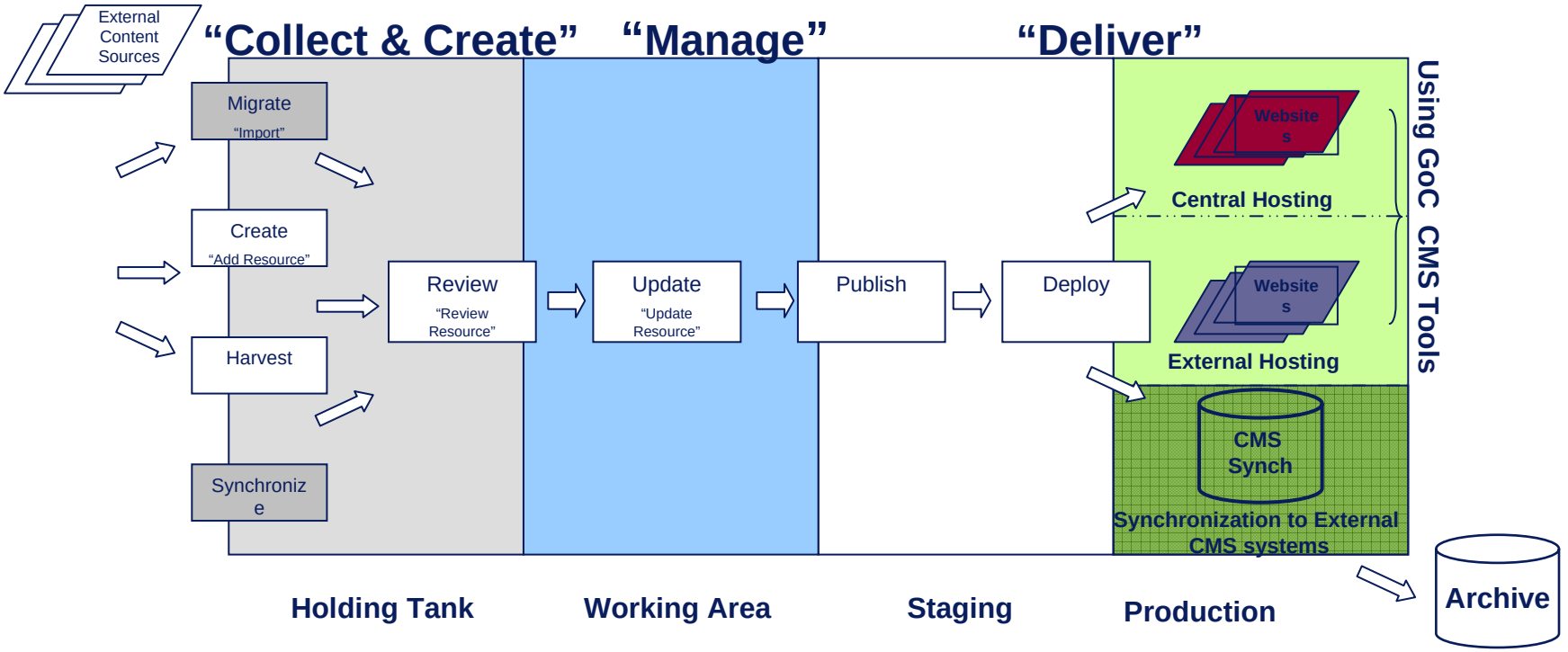
Technology (COTS products) have been integrated to develop the GoC CMS Prototype. The key technology components include:

- Interwoven: Product set includes: TeamSite for content management, MetaTagger for taxonomy management and automated keyword generation, and deployment tools (Open Deploy, Data Deploy).
- Verity K2: Search engine that provides content searching and indexing capability across the solution that can be adjusted to support the ranking of metadata in a search result.
- Cognos Impromptu: Business intelligence software used to generate usage and audit reports.
- BMC Patrol: Server monitoring software.

GoC Case Study (GoC CMS)
The Components (2)

GoC CMS enables stakeholders to manage, share and publish web resources and its metadata in a standard and rational way.

It includes features for the following key process components:



GoC Case Study (GoC CMS)
The Components (3)

Metadata standards are fundamental to the information management component of the CMS.

- Less about the technology, more about the standards which enable inter-operability across GoC and other levels of government
- We have moved beyond the "what is metadata and why is it important" phase
- Now we need to move beyond the phase of department-specific or application-specific metadata silos
  - Effective enterprise service to Canadians requires interoperability between content authoring, technical systems and processes, content repositories and end-user information needs
  - Facets of the GOC information holdings can be combined in virtual information and service "views" for client-centric or program-centric delivery
  - Information portability and reusability (write-once, use multiple times processes)
  - Connecting documents, publishing and archival systems

GoC Case Study (GoC CMS)
The Components (4)

Metadata Standards and Implementation Specifications

- CMS Metadata Sub-Group formed in April 2005
  - Reports to the GOL Metadata Working Group, led by TBS (Nancy Brodie)
  - Also acts as a sub-group of the CMS Functional Working Group
  - Role is to define, align and manage metadata frameworks and processes in support of the enterprise GOC CMS
  - Includes departmental and cluster representatives
- Objectives
  - CMS Metadata Element Set
  - CMS-Metadata Application Profile (MAP)
  - Align the Element Set with the Records Management Element Set by finding opportunities for interoperability and aligning metadata names

Metadata Element Sets and Application Profiles
The Components (5)

Metadata Standards and Implementation Specifications (continued)

- Metadata Elements Set
  - Name (dc.title, dc.coverage.spatial, dc.subject, gcms.caption, etc)
  - Label (Title, Subject, Caption, etc)
  - Definition (intended scope or purpose of the metadata element or sub-element)
  - Data type
  - The CMS Element Set will be a standard once completed
  - Is based upon Dublin Core, with GOC CMS extensions
- Metadata Application Profile
  - How the metatag value is populated and used within a CMS
  - Single or multiple values
  - Optional or mandatory
  - Schemes and vocabularies
  - Relationship to other metadata elements
  - Purpose and constraints

GoC Case Study (GoC CMS)
The Components (6)

Metadata standards and specifications development process: complex, costly and time consuming

- Designing your metadata for flexibility and extensibility
  - Working together with departments to define common metadata for the CMS
  - Departments will be able to extend the common metadata set to meet department-specific requirements
  - Design for flexibility … your metadata requirements will evolve
- Engaging your communities
  - Stakeholder community: If it doesn't come from them, they won't use it
  - Extended community: Share your experiences and challenges; it's unlikely that no one else has been developing solutions dealing with the same problems
- Keeping the end-goal in mind
  - To what end(s) do you expect to use the metadata element (facet browse, search filter, dynamic content feeds, etc.)?
  - Make sure everyone involved in the development process understands how the metadata element/vocabulary is intended to be used
  - Select representative sample of content and tag it to ensure that your element/vocabulary meets the requirements

GoC Case Study (GoC CMS)
The Components (7)

Metadata and Taxonomy Services

- Shared Metadata Services Unit as critical/hub to provide consistent metadata services as part of the central administration services offered around the CMS.
  - Taxonomy creation, integration and management services
  - Metadata quality assessment services
  - Metadata tagging services
  - Standards and guidelines development support services

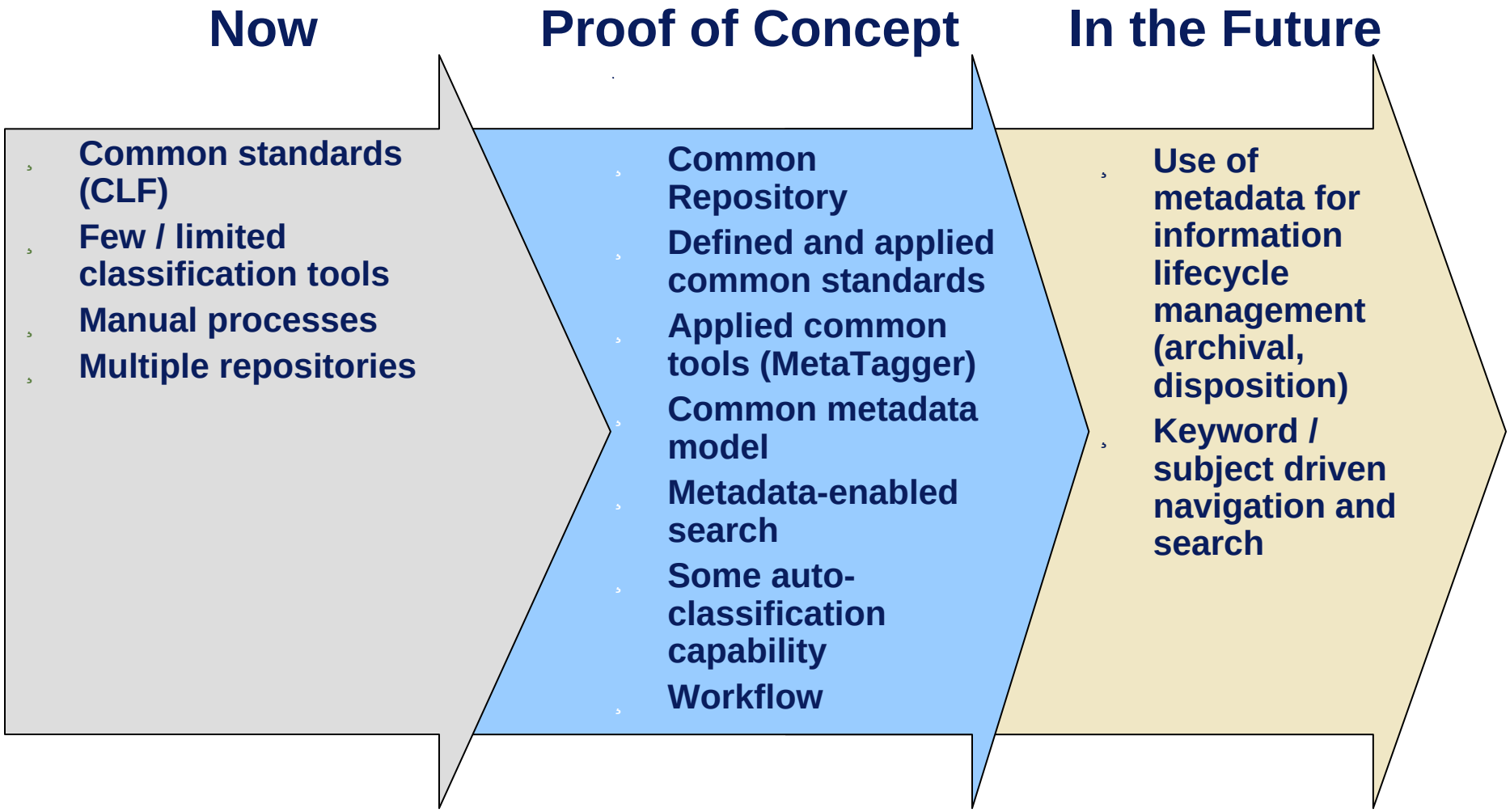
GoC Case Study (GoC CMS)
Secrets to Success

- Avoiding Scope Creep – focus project on content management solutions
- Crafting the Right Team – community leadership, collaborative effort …and lots of meetings and discussions/compromise
- Knowing the Products – they're a toolbox not an out-of-the-box packaged solution
- Clearly Defining the Requirements –metadata standards and taxonomy management are requirements that take time to develop, make the investment upfront
- Governance - it's all about the people, process and structure

…Evolving to a Shared Solution

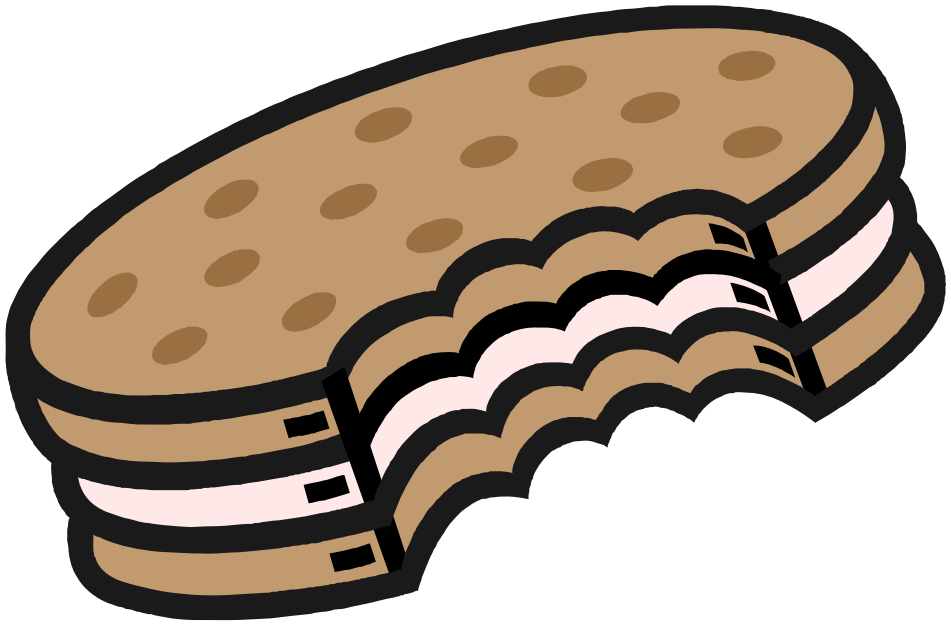
Proof-of-Concept Solution Evolution

An Auto-Classification Perspective



When all else fails…

Bring food



Questions?

Contact Information

Sean Murphy

Deloitte

(613) 786-7513

seamurphy@deloitte.ca

Susan Thorne

Public Works & Government Services Canada

(819) 956-5578

Susan.thorne@pwgsc.gc.ca