Search options and content tagging

NKOS 2008
Aarhus Denmark
September 19

Marjorie M. K. Hlava
Access Innovations, Inc – Data Harmony

DataHarmony

Access Innovations.

In the olden days.....

- Online from the 70's
  - Dialog
  - Data Star
  - Many others
- Secondary publishers
  - Mead – Lexis
  - CAS
  - NASA & DOE & many others

ONLINE SEARCH

- Worked very well
  - Focused
  - Controlled
  - Specialized
- Content analysis
  - Database design - context
  - Extensive markup
  - Proprietary formats (Dialog format b)

BACK AT THE LAB

- Computer science
  - Full text
  - Isolated
  - Content without context
- Developing shortcuts became critical
  - Relevance
  - Weighting
  - Probabilities

Natural Language Processing

- Since the early 1970's
- Replicate human intelligent processes
- HUGE body of research
- Extract information
- Increasing mountains of textual information
- Holy Grail

Natural Language Processing

- Artificial intelligence
- Computational linguistics.
- Problems of automated generation and understanding of natural human languages.
- Convert samples of human language into more formal representations that are easier for computer programs to manipulate
- Nine major areas (or so…)

Natural Language Processing

- Linguistic Study of language
- Semantic - study of meaning in communication
  - Literal and connotation
  - Lexical, Applied, Structural
- Syntactic principles and rules for constructing sentences
- Morphological - structure and content of word forms
- Phraseological - peculiar form of words
- Grammatical -the rules governing the use of any given natural language
- Stemming – lemmatization reducing inflected word to stem, base or root
- Synonyms - semantically equivalent
- Pragmatics - Common sense - indexicality - use and effects of language

Other Techniques - Sample

- Vector calculus (*vector analysis*) quaternion analysis
  - Latent Semantic
- Statistical - multivariate analysis
- Bayesian probability uses probability as 'a measure of a state of knowledge'
  - Objectivist school
  - Subjectivist school
- Neural networks - connectionism
  - Statistical learning theory
- **SMART (System for the Mechanical Analysis and Retrieval of Text) Information Retrieval System**
  - Cornell University Gerard Salton
  - Vector space model
  - Relevance feedback Rule based

They don’t work well

- Search is broken
- Google stole the show
- Precision and recall went out the window
- Relevance became the buzzword

The Potential

- To access content directly
- Find it
- Tag it
- Know what the user will ask for
  - And the next user, And the next user
- Not all people search the same way
- Persistent Clustering – find it again!

DataHarmony

Access Innovations.

Use term control - applied

- At the input end
- On the search (query) side as well
- Accommodate all learning styles
- ……
- High relevance
- Total recall
- Excellent precision
- Happy users

DataHarmony

Access Innovations.

Look at The Weather Channel: 15 synonyms for “rain”

- Rain
- Shower
- Sprinkles
- Downpour
- Cloudburst
- Gully washer
- Monsoon
- Deluge
- Thunderstorm
- Thundershower
- Drizzle
- Mist
- Liquid precip.
- Torrent
- Virga

The Weather Channel®

vir•ga \ ‘vərgə \ *n* –*s*

Precipitation (usually rain or snow) that evaporates before it reaches the ground, often seen as gray streaks in the sky near the base of the cloud.

DataHarmony

Access Innovations.

"Hammered!"

Hammered, Hit, Slammed, Buffeted, Slapped, Sprayed, Pushed, Pummeled, Drenched, Buried, Blasted, Blown, Abused, or otherwise manhandled by the elements.

DataHarmony

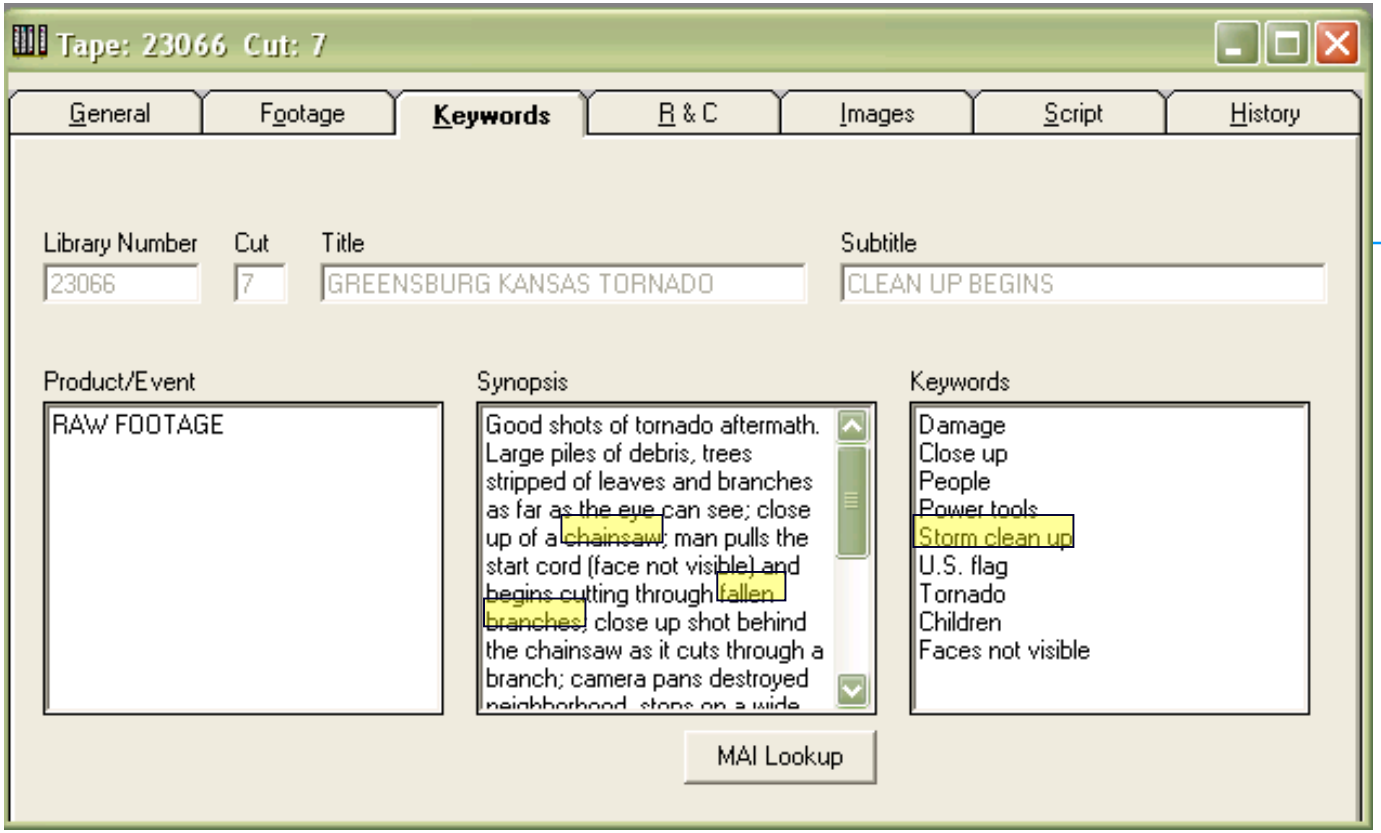
The Weather Channel®

Access Innovations.

Adding the taxonomy terms to the content

- Time of creation
- Adding to the corpus in the System
  - Content Management
  - Digital Asset Management System
  - Repository

Attach to the record or information object



The Weather Channel®

Automatic term suggestion – VERY rich in synonyms for search

DataHarmony

Access Innovations.

The Weather Channel®

weather.com

Bringing weather to life

“Using the MAI has cut our search time by 50%”

Jay Tellock, Weather Channel

DataHarmony

Access Innovations.

Then pull it out again

- Search software
- Inverted index – fast look up
- Display records – show the user
- Accommodate different learning styles
  - Browse (taxonomy)
  - Search (the box)
  - Advanced search (faceted navigation)
  - Follow a thread (ontology)

Use term control - applied

- At the input end
- On the search (query) side as well
- Accommodate all learning styles
- ……
- High relevance
- Total recall
- Excellent precision
- Happy users

DataHarmony

Access Innovations.

Rules of thumb - general

- Index to the most specific level
- Role up the terms for presentation
- Add lots of synonyms
- Review the search logs
- Add candidate terms

Rules of thumb - metrics

- Hit Miss and Noise
  - 85 % accuracy to launch
- 4 hours per month to maintain
  - With candidate term feeds
  - With search log data
- 5 minutes per term – rule and record
- 1 hour per training term

DataHarmony

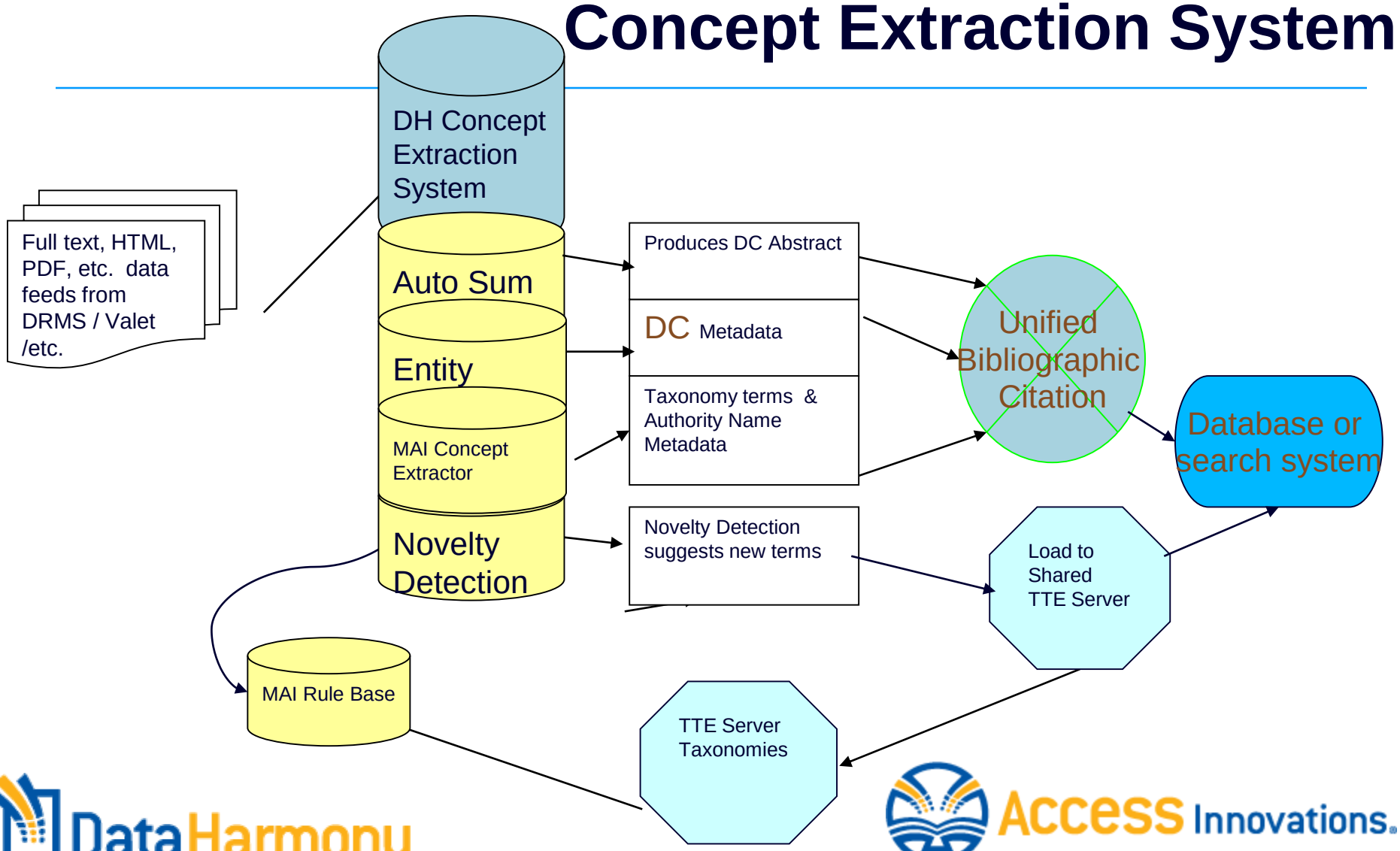
Access Innovations.

Justification – the ROI

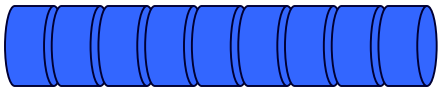
The Pain of Search

Mission critical	Percent	Number of Employees 1000	Search & Use Timel Per Week Hours	Time Searching Per Week Hours	Time Analysing Per Week Hours	Average Loaded Salary $ Per Hour	Annual Cost of Looking	Search Time Reduction 10%	Difference
High	10	100	14	8.4	5.6	200	8,736,000	7,862,400	873,600
Medium	80	800	12	7.2	4.8	150	44,928,000	40,435,200	4,492,800
Low	10	100	10	6	4	100	3,120,000	2,808,000	312,000
							$56,784,000	$51,105,600	$5,678,400

Concept Extraction System

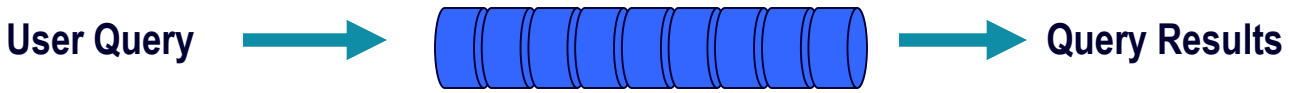


DH M.A.I.™ Process

User Taxonomy →  → Subject term indexing

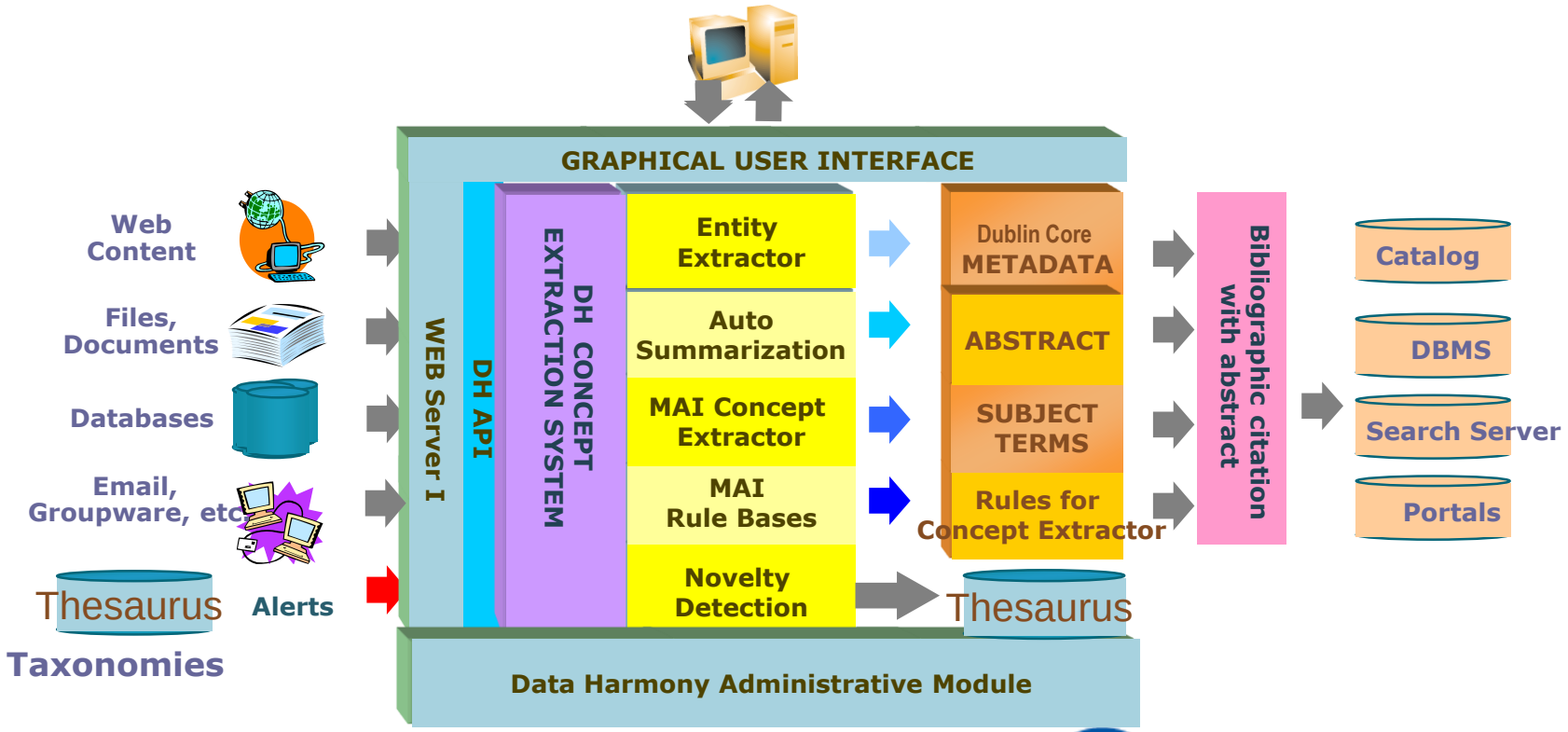
Data Harmony MAI Concept Extractor Module		
Formulation	**Term Suggestions**	**Term Selection**
Query formulations Use NLP to parse query Expand query term to all factors in rule base	Pass text through rule bases Concept Extraction Provide suggested term list	Categorization of results by frequency Convert frequency to weights Present results

DH MAI Query Process



Data Harmony MAI Query Module		
Formulation	**Query revolver**	**Reporting**
Query formulations Use NLP to parse query Expand query term to all factors in rule base	Pass query to Search Concept Extraction Analyze reply	Categorization of results by frequency Group results Present results

Data Harmony Architecture



DataHarmony

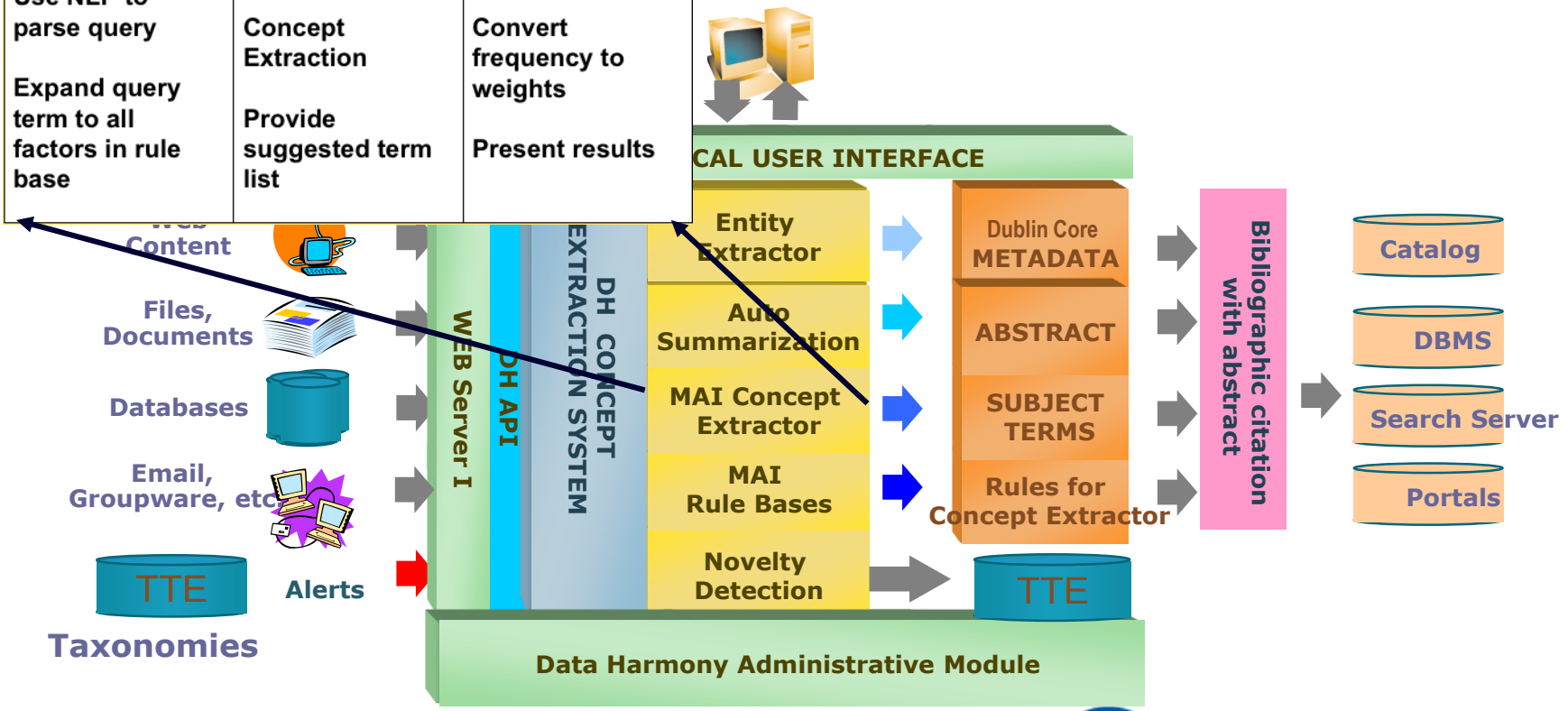
Access Innovations.

Architecture

DH MAI Process

User Taxonomy → Subject term indexing

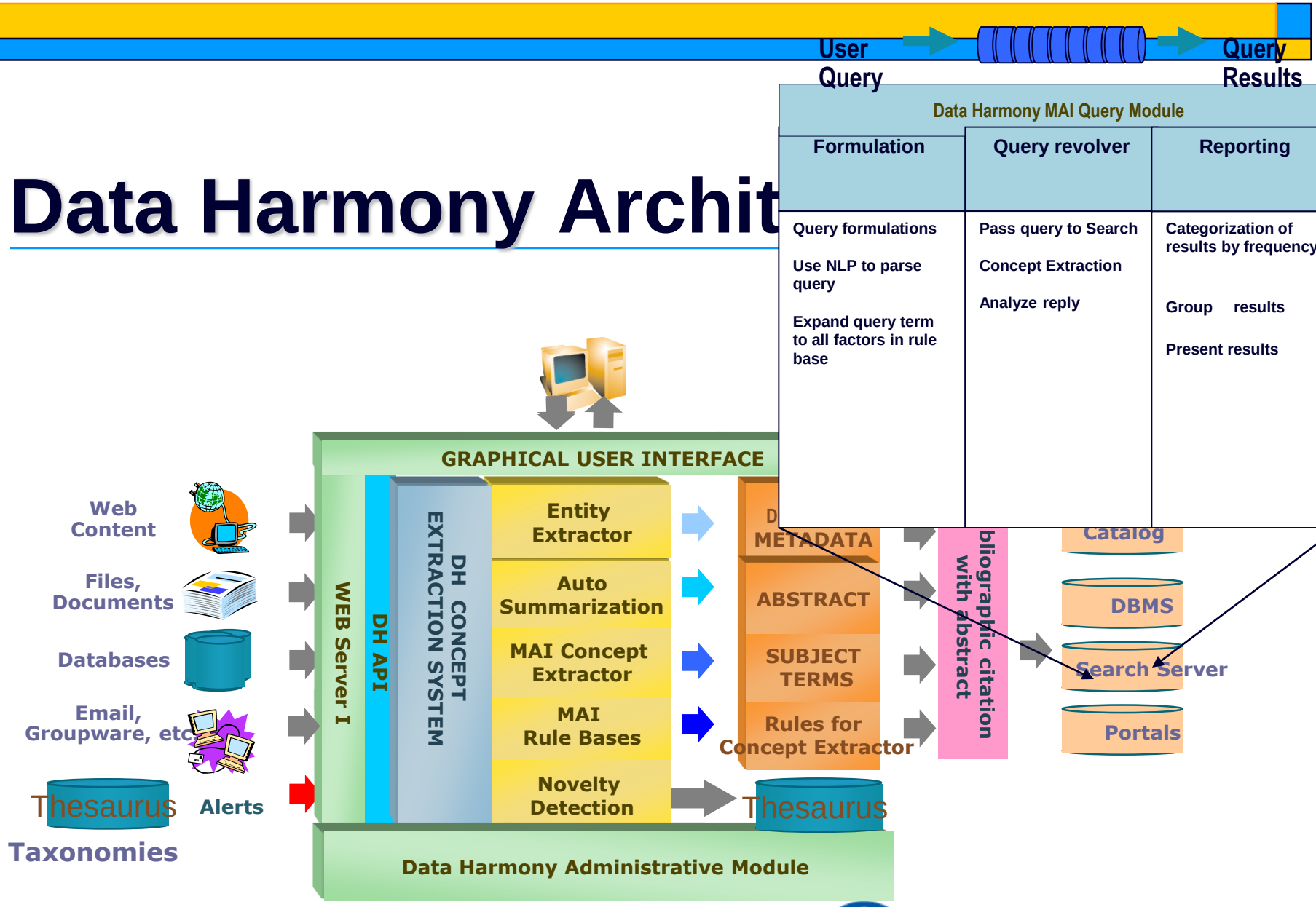
Data Harmony MAI Concept Extractor Module		
Formulation	**Term Suggestions**	**Term Selection**
Query formulations Use NLP to parse query Expand query term to all factors in rule base	Pass text through rule bases Concept Extraction Provide suggested term list	Categorization of results by frequency Convert frequency to weights Present results



DataHarmony

Access Innovations.

Data Harmony Archit



DataHarmony

Access Innovations.

Thank you for attention!

Marjorie M. K. Hlava

Access Innovations / Data Harmony

mhlava@accessinn.com

+1-505-998-0800

DataHarmony

Access Innovations.