

GETTING MORE VALUE OUT OF CONTENT

Key success factors in auto-classification

APQC April 2018

KM – the downside of success

The “runaway train”
of content overload



How do we re-establish control and deliver true value to users?

The content explosion... and the Google effect

- Content is easy to create and revise
 - Most of it is unstructured
 - It's going to get worse
- “Digital natives” are now firmly in the working world...
 - ... and now everyone expects to find everything in a few clicks



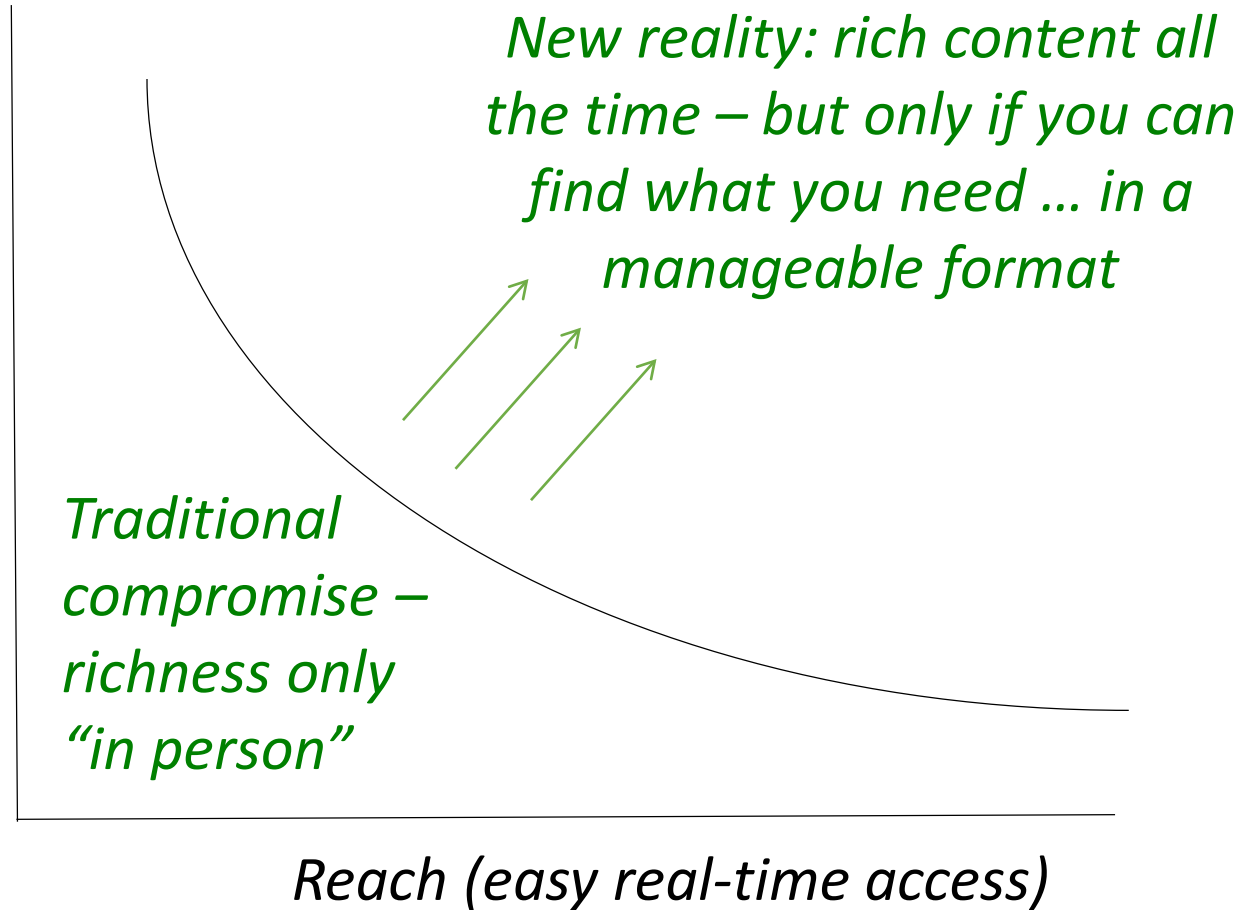
Why can't our intranet be like Google?

- Nearly everyone hates their enterprise search!
 - “I use the public “Google” to find our content, because I can't find any of it using our internal search”
 - “We still rely on emailing documents because we can't find it on the intranet”
- Google isn't set up to effectively search enterprise content
- SharePoint search is limited, and also often not tuned properly
- Users expect to find answers across different systems – “an integrated view of what we know”



The promise: information richness and reach

*Richness
(in depth
knowledge)*



But “Cognitive Computing” is not yet reality in most enterprises...

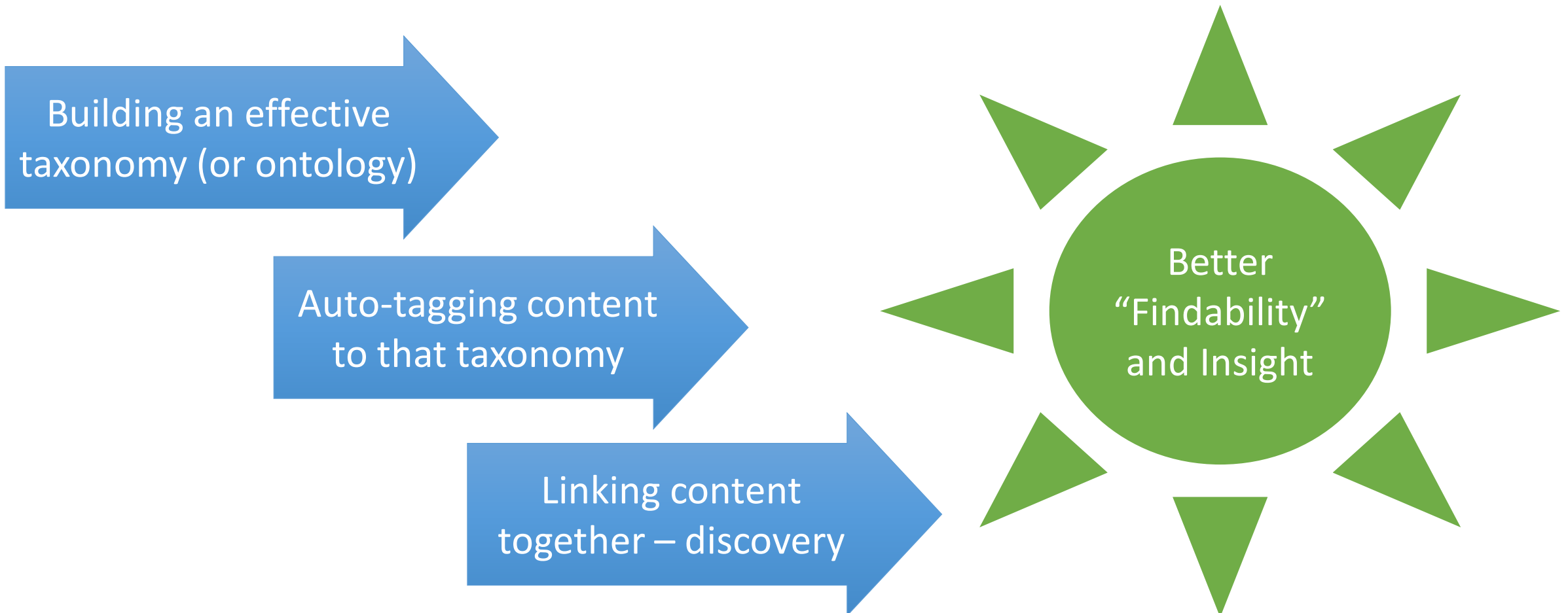
- Tools that recognize and extract content themes, and route them to the right people in context
- Practical applications are costly, complex, and remain elusive



However...

- Auto-classification is a practical, immediate step forward for almost everyone -- immediate value for managing content overload

What is auto-classification?



Taxonomy remains the fundamental building block of content curation and findability

Taxonomy means...

A common language for key concepts



Increased search precision
Enhanced content discovery

Ontology Example:

Houston in Wikipedia

<p>n, but slave dealers were near the city worked on Houston was granted ame the county seat of unity established a Bayou.^[20]</p> <p>the Texas inland vil War, Houston served of Galveston.^[21] After ld accept more as.</p> <p>port were accelerated.^[22] Texas petroleum hip Channel. By 1910, ed a large part of the</p> <p>gan. By 1930, Houston sus Bureau reported</p> <p>however, the war did</p>	<table><tr><td>Country</td><td>United States</td></tr><tr><td>State</td><td>Texas</td></tr><tr><td>Counties</td><td>Harris, Fort Bend, Montgomery</td></tr><tr><td>Incorporated</td><td>June 5, 1837</td></tr><tr><td>Named for</td><td>Sam Houston</td></tr><tr><td>Government</td><td></td></tr><tr><td>• Type</td><td>Mayor–council</td></tr><tr><td>• Body</td><td>Houston City Council</td></tr><tr><td>• Mayor</td><td>Sylvester Turner (D)</td></tr><tr><td>Area</td><td></td></tr><tr><td>• City</td><td>627 sq mi (1,552.9 km²)</td></tr><tr><td>• Land</td><td>599.59 sq mi (1,552.9 km²)</td></tr><tr><td>• Metro</td><td>10,062 sq mi (26,060 km²)</td></tr><tr><td>Elevation</td><td>80 ft (32 m)</td></tr><tr><td>Population (2010)^[3]</td><td></td></tr><tr><td>• City</td><td>2,099,451</td></tr><tr><td>• Estimate (2016)</td><td>2,303,482^[2]</td></tr><tr><td>• Rank</td><td>US: 4th</td></tr><tr><td>• Density</td><td>3,660/sq mi (1,414/km²)</td></tr><tr><td>• Urban</td><td>4,944,332 (7th U.S.)</td></tr><tr><td>• Metro</td><td>6,313,158 (5th U.S.)</td></tr><tr><td>• Demonym</td><td>Houstonian^[1]</td></tr><tr><td>Time zone</td><td>CST (UTC -6)</td></tr></table>	Country	United States	State	Texas	Counties	Harris, Fort Bend, Montgomery	Incorporated	June 5, 1837	Named for	Sam Houston	Government		• Type	Mayor–council	• Body	Houston City Council	• Mayor	Sylvester Turner (D)	Area		• City	627 sq mi (1,552.9 km ²)	• Land	599.59 sq mi (1,552.9 km ²)	• Metro	10,062 sq mi (26,060 km ²)	Elevation	80 ft (32 m)	Population (2010)^[3]		• City	2,099,451	• Estimate (2016)	2,303,482 ^[2]	• Rank	US: 4th	• Density	3,660/sq mi (1,414/km ²)	• Urban	4,944,332 (7th U.S.)	• Metro	6,313,158 (5th U.S.)	• Demonym	Houstonian ^[1]	Time zone	CST (UTC -6)
Country	United States																																														
State	Texas																																														
Counties	Harris, Fort Bend, Montgomery																																														
Incorporated	June 5, 1837																																														
Named for	Sam Houston																																														
Government																																															
• Type	Mayor–council																																														
• Body	Houston City Council																																														
• Mayor	Sylvester Turner (D)																																														
Area																																															
• City	627 sq mi (1,552.9 km ²)																																														
• Land	599.59 sq mi (1,552.9 km ²)																																														
• Metro	10,062 sq mi (26,060 km ²)																																														
Elevation	80 ft (32 m)																																														
Population (2010)^[3]																																															
• City	2,099,451																																														
• Estimate (2016)	2,303,482 ^[2]																																														
• Rank	US: 4th																																														
• Density	3,660/sq mi (1,414/km ²)																																														
• Urban	4,944,332 (7th U.S.)																																														
• Metro	6,313,158 (5th U.S.)																																														
• Demonym	Houstonian ^[1]																																														
Time zone	CST (UTC -6)																																														

Auto-classification is very different than traditional manual tagging

	Taxonomy - what are we going to call things?	Rules – what determines tagging? →	Result
Manual	Scope limited by tagging workload and the setup of CMS tools	Judgment of tagging staff (or end users) →	High-level categories for browsing, with questionable quality
Automated	Deep ontology with alternative terms, lateral relationships	Programmed rules based on occurrence, frequency, restrictions, etc. →	Content drill-down, enhanced search, related terms

Huge benefits from getting this right

Cost and quality

- Manual classification - time consuming
- Quality/consistency challenges

Curation and records management

- Expose content by topic - “spot the ROT”
- Identify sensitive/compliance content

Content discovery and integration

- Business value from combining structured and unstructured data across sources

Search effectiveness

- Tagging helps with relevance and search refinement
- Expanded browsing to related topics

Example: Making sense of a huge repository

Electric power utility - multiple facilities/locations



> 1 million documents across intranet and records management

Enable access for key business use cases

- Business areas
- Equipment and technical standards
- Locations
- Types of content (drawings, etc.)
- Vendors and customers

Understand what content exists and start to curate it

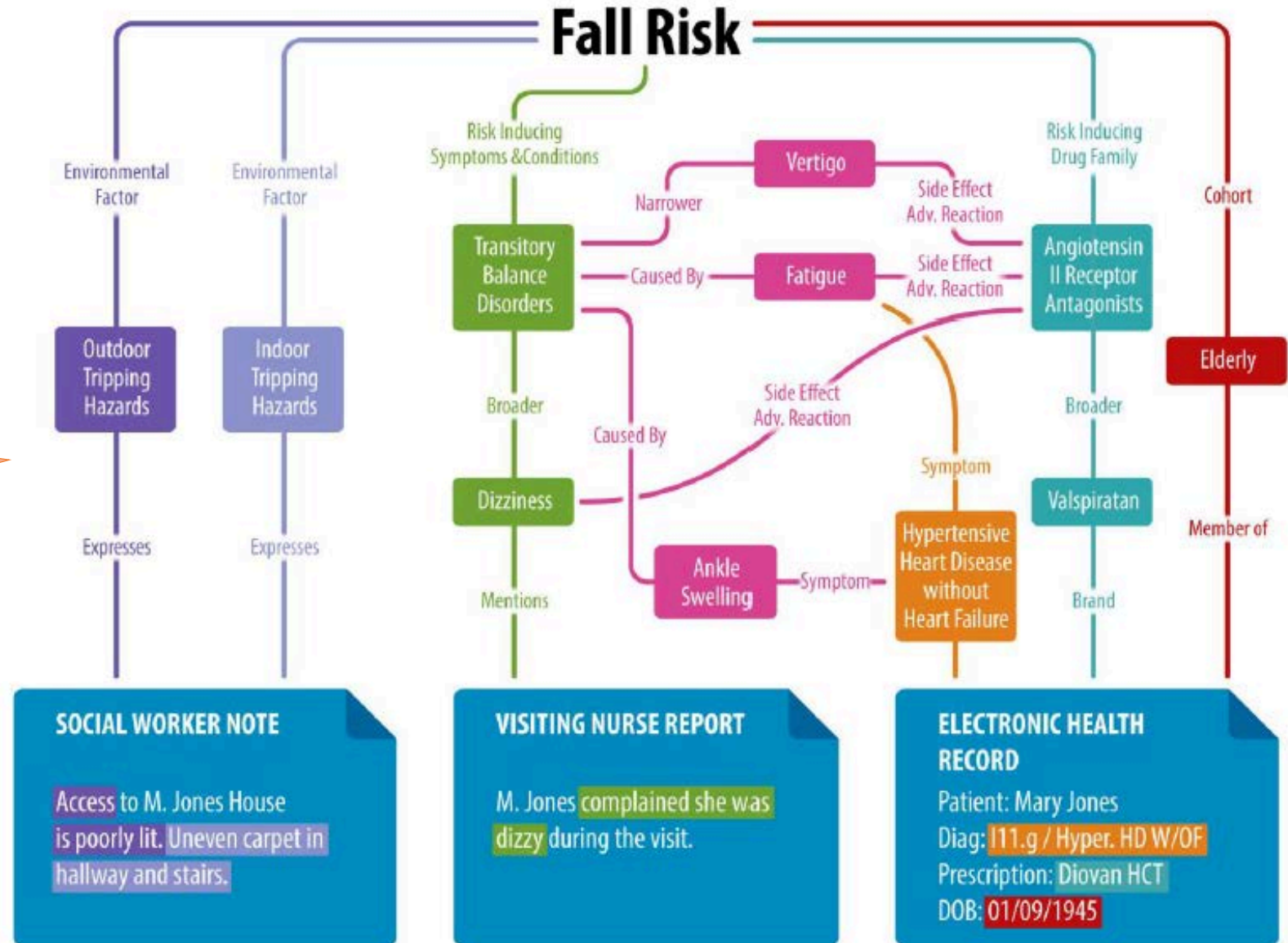
Leverage the taxonomy for useful business applications (adding structure)

Example: Improving patient care for seniors

Identify individuals at risk and prevent injuries

Content auto-classified against the ontology for "Fall Risk"

Multiple unstructured information sources



So why isn't everyone doing it?

- Enterprise taxonomy is “sliding” on the Gartner hype cycle (implementation effort?)
- Economics – number of documents, number of people to manage
- It is challenging to implement initially – so what are the lessons learned?

Key challenges along the way



Taxonomy development doesn't happen on its own...

The taxonomy will be larger than you think...

Classification rules require iterative development...

Classification runs may be time consuming

You will need active ongoing governance...

Taxonomy development requires investment



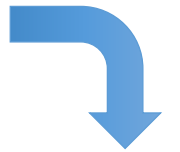
Existing systems
+
Search log review
+
External models
+
“Concept mining”



Human
judgment
- what's
important?

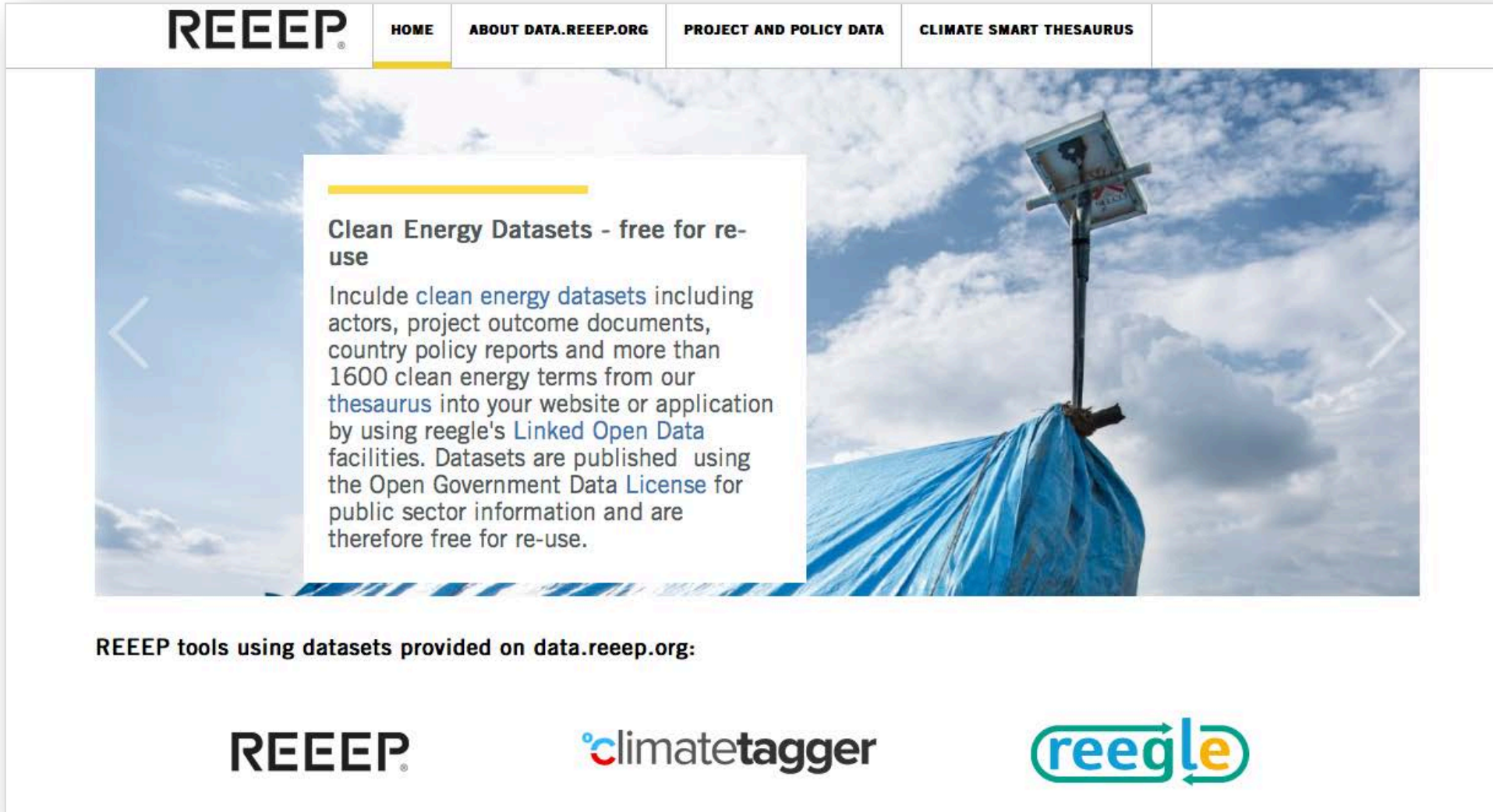


SME review



Business buy-in

Open source ontologies can be helpful



The screenshot shows the REEEP website with a navigation menu at the top containing: REEEP, HOME, ABOUT DATA.REEEP.ORG, PROJECT AND POLICY DATA, and CLIMATE SMART THESAURUS. Below the menu is a large image of a solar panel on a pole. A white text box is overlaid on the left side of the image, containing the following text:

Clean Energy Datasets - free for re-use

Include [clean energy datasets](#) including actors, project outcome documents, country policy reports and more than 1600 clean energy terms from our [thesaurus](#) into your website or application by using reegle's [Linked Open Data](#) facilities. Datasets are published using the [Open Government Data License](#) for public sector information and are therefore free for re-use.

Below the image, the text reads: REEEP tools using datasets provided on [data.reeep.org](#):

At the bottom of the page, there are three logos: REEEP, climatetagger, and reegle.

REEEP Climate Smart Thesaurus:

- Links to synonyms and related terms
- Multi-language
- Multi sector
- Tagging tools

Taxonomy will be deeper than you think

With manual tagging

- Business Functions
 - Finance
 - Human Resources...



Quick to tag
Easy to browse, but
non-specific

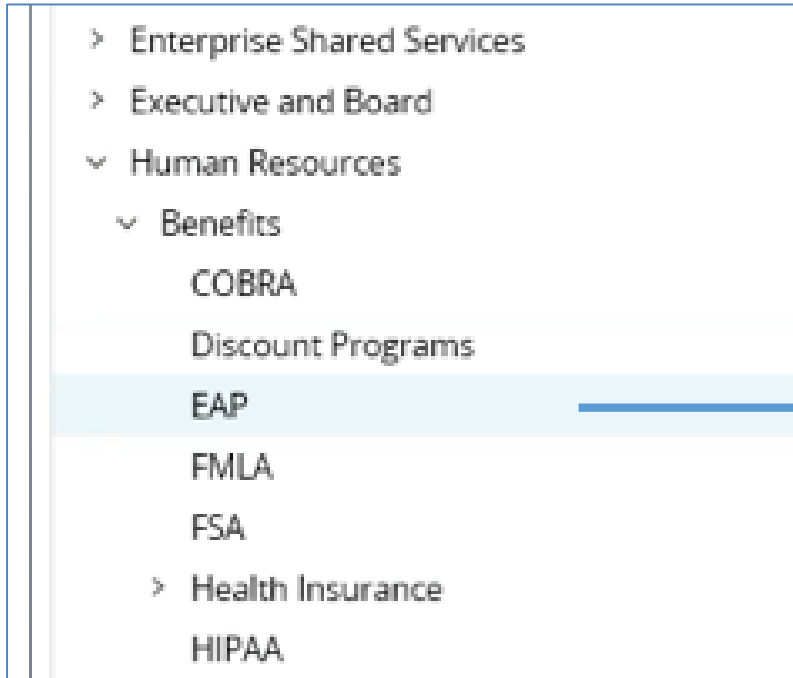
With auto-tagging

- Business Functions
 - Human Resources
 - Benefits (..only in HR context...)
 - 401K (Retirement plan, 457 plan, ...)
 - Health Plan (Health insurance, ...Cigna,...Medical insurance,...)



Depth creates more
“evidence” for auto-
tagging

“Alternative labels” for taxonomy terms drive the rules



EAP

Concept Class

Business Group

+ Preferred Labels

[Create a preferred label](#)

EAP en

+ Alternative Labels

[Create an alternative label](#)

Content Server Classification > 7-05-16 en

Negative evidence > Emergency Action Plan en

Negative evidence > Expenditure Authorization Procedure en

Negative evidence > Project EAP en

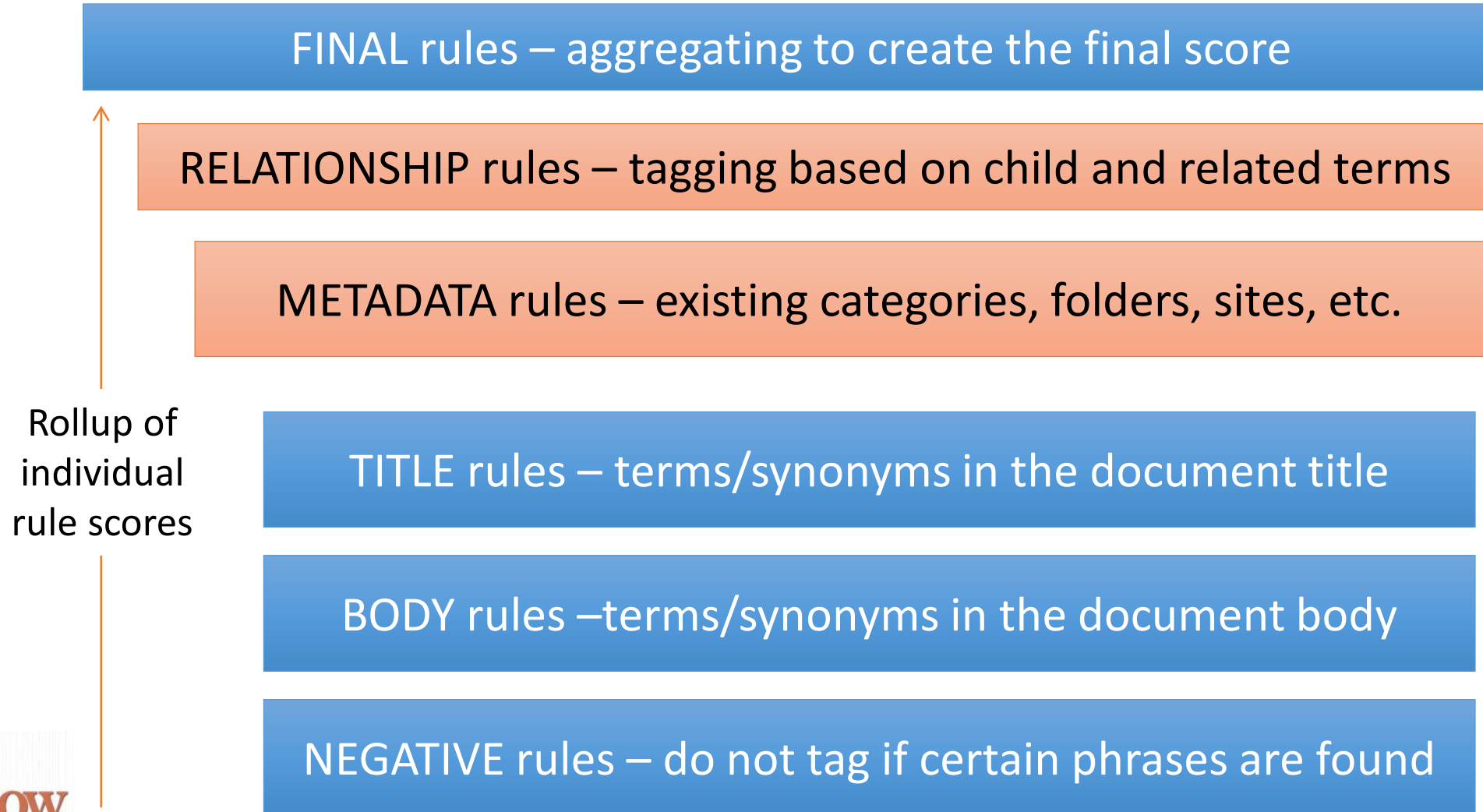
Required Site > */hr/* en

SharePoint Site URL > */hr/bene/Pages/BenefitsVendorLinks.aspx en

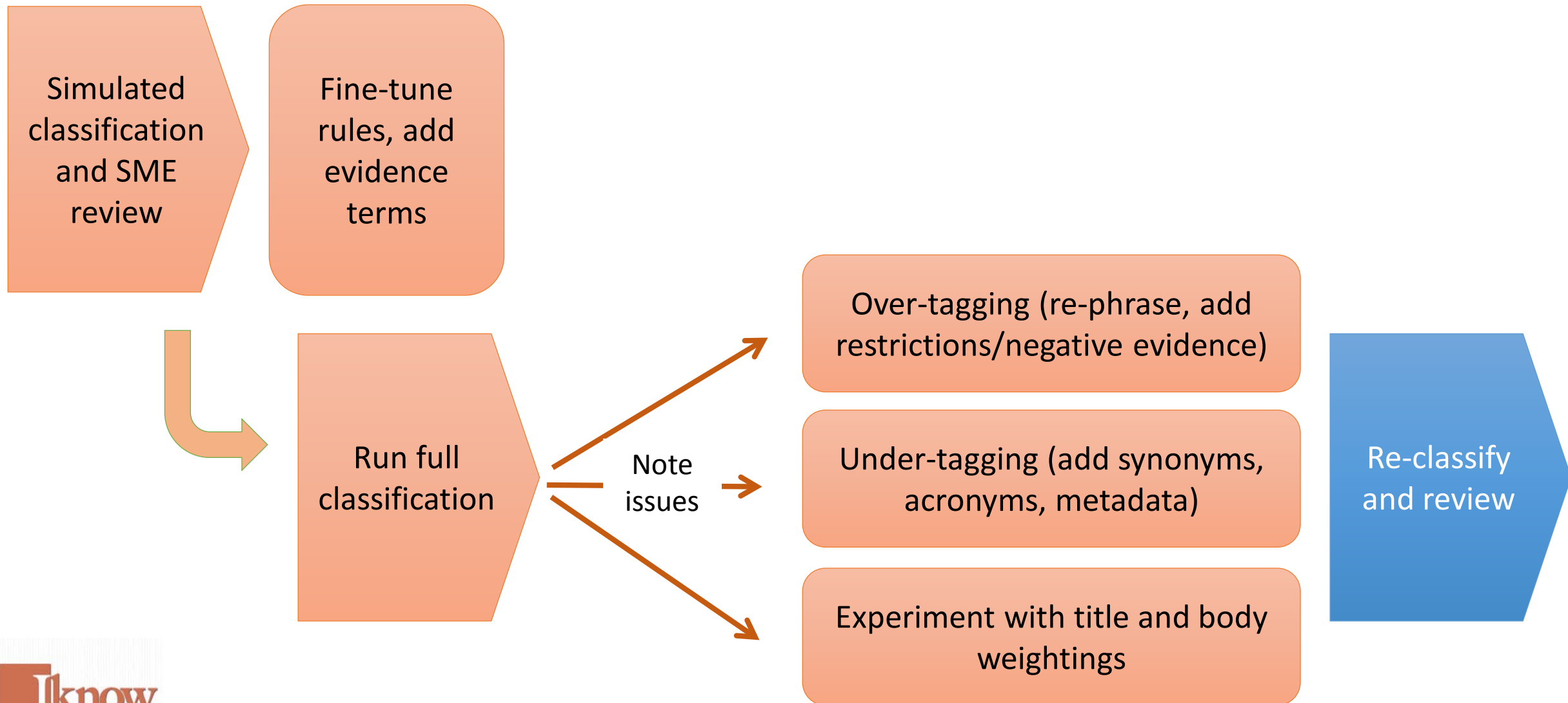
Synonym > Employee Assistance en

Synonym > Employee Assistance Program en

Classification rules drive auto-tagging



Rules development will require iteration



Classification process may be time-intensive

Alternative approaches to classification processing

During indexing only

Tags are used in the search index only – and not added back to the source documents

- Simple, quick
- But does not affect underlying content

“At source” or “write-back”

Tags are added to the underlying content repositories

- Many benefits, such as structured browsing and other UI features
- But, more complex to build and manage

Full write-back classification takes time
(~20K documents per hour)
Need to plan updates carefully!

Implies strong governance process...

Ongoing monitoring

Action planning

User feedback

Search logs

Tag frequency



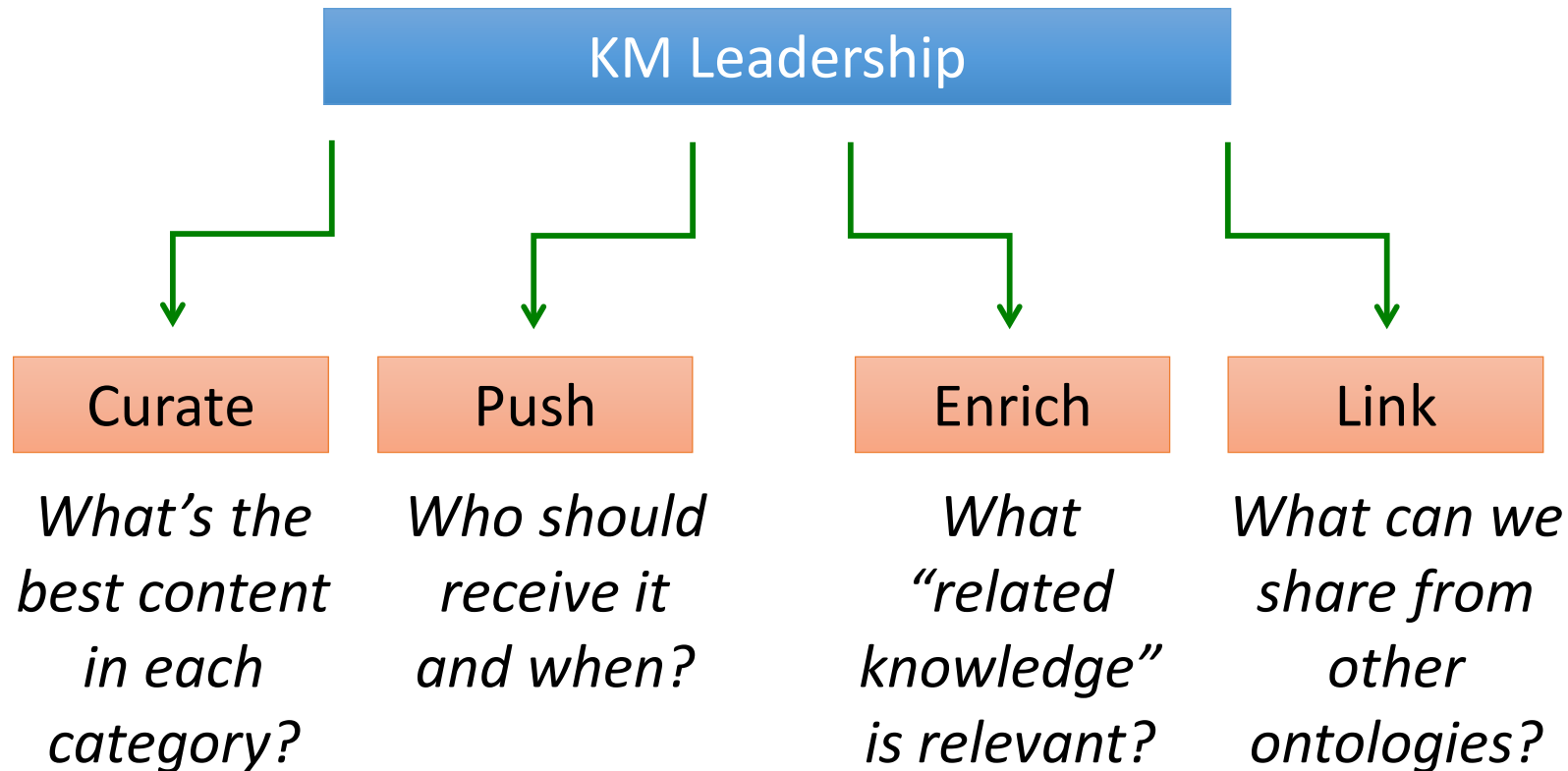
Suggested changes to
taxonomy and rules



Re-classification
schedule and scope

... and the right organizational follow-up

New and enhanced KM staff roles



Planning an auto-classification project



Business case

- Volume
- People
- Use cases

Current state

- Taxonomies
- User surveys
- Search logs

How it will be used (business scenarios)

Main taxonomy facets

Specific search functionality



Plan for ~ 6 months for a large enterprise

Closing thoughts

- This does require investment and commitment, but...
- Challenges can be overcome, and there are many proof points
- Huge benefits from getting this right – for users and for KM professionals