Topic Maps: backbone of Content Intelligence

Jean **Delahousse** <jean.delahousse@mondeca.com>

Abstract

Industrial enterprises, administrations and editors have a long history of organizing information so as to facilitate access and exchange. Developed for the world of print publication, these earlier solutions need to be incorporated now into the newer ones that are specifically adapted for digital content. The term “Content Intelligence” encompasses the different tools and methods that would allow companies to capitalize on their existing skills and resources even while offering fresh solutions for new stakes. At the heart of Content Intelligence, tools that are based on the Topic Maps standard are particularly well positioned as content aggregation solutions around industry-specific subject matters and their internal organization.

Table of Contents

1. Context 1
2. Which needs for which users? 2
3. What is Content Intelligence? 2
4. What are the steps in building an enterprise-wide solution? 3
    4.1. Manage the subjects 3
    4.2. Organize the content – organize the subjects 3
    4.3. Aggregate the content 4
5. What are the tools needed to build a solution? 4
    5.1. Building an enterprise repository of subjects 4
    5.2. Organize the subjects 5
    5.3. Referencing heterogeneous content 5
6. What are the tools used within the implemented solution? 6
    6.1. Search 6
    6.2. Navigate 6
7. Topic Maps as the backbone of Content Intelligence 6
    7.1. Solution for managing subjects 6
    7.2. Open and decentralized solutions for organizing subjects 6
    7.3. Personalization solutions 7
    7.4. Non-intrusive solutions for referencing contents 7
    7.5. Ability to describe all models of knowledge organization currently in use across the enterprise 7
    7.6. A repertory of subjects, organization and enterprise vocabulary at the service of other technical components 8
8. Conclusion 8
    8.1. Why Topic Maps tools for content intelligence solutions? 8
    8.2. What’s still needed to make content intelligence solutions a reality? 8
    8.3. Our convictions 9

1. Context

This presentation will focus on the specific context of information access within an enterprise, and will examine the particular case of large industrial enterprises.

- define the user needs,

- offer a working definition of Content Intelligence
- define the different components of a Content Intelligence solution
- and finally pose the question of the relevance of using the Topic Maps standard as a structuring and core component of Content Intelligence solutions.

Our secondary aim is to show that the technical discontinuity imposed by the management of digital content should not obscure the great continuity in the methodologies and in the conceptual tools used to organize information available in print and in digital form. This continuity allows companies that have a long tradition of document organization to transpose and enrich their tools within solutions adapted to the digital world.

2. Which needs for which users?

The end-users that we speak of are employees, technicians, engineers, salespeople, etc., working in industrial enterprises. Their enterprises already have a long experience in organizing technical documentation, standardization and conserving documents. These enterprises were among the first adopters of SGML and of content management tools like Documentum.

Today they are equipped with numerous intranets, databases, ERP, CRM and several document management systems. These industries have capitalized on the construction of internal thesauri or use external thesauri (Mesh, in the case of the pharmaceutical industry), cultivate centers of documentation, and are engaged in technological and competitive intelligence.

In short, these companies and their end-users have a wide experience in organizing and normalizing information.

We may define three categories of people around any system that allows efficient access to information:

- consumers of the content
- authors of the content
- organizers of the content

The needs of content consumers may be summarized as follows:

- ability to access from their workstation ALL and ANY useful and valuable information that is available throughout the enterprise
- ability to find at the least cost the information that corresponds to a given need, along with the context of this information
- ability to reuse the contents so identified in new publications

For the authors, the need is to ensure that the document created will be available and identifiable for the end-users when needed.

The principal needs of the organizers of content are:

- ability to reuse the organizational elements of existing information within the context of a digital publication
- ability to complete and enrich the content organization in response to new technical and user needs
- to minimize the cost of content organization within the enterprise

3. What is Content Intelligence?

The term “Content Intelligence” for us encompasses all the methods, tools and knowledge organization bases that make it possible to implement efficient content access solutions within an enterprise context.

4. What are the steps in building an enterprise-wide solution?

4.1. Manage the subjects

The subjects handled within an enterprise are listed through classificatory schemes, document indexes, thesauri, classifications of document centers, Lotus Notes databases, keywords in document management systems, domain-specific XML standards, etc.

The core component of any document organization scheme, the subject offers a reference language within the enterprise. The exact definition of the subject, managing its expression in different languages, using different words to talk about it depending on the departments, are problems that are never wholly resolved but which demonstrate the importance of the subject and its definition.

The subject may be a general concept, a technique, a product, the name of a process, a profession, a person, a project, a company that is a client or supplier. The totality of subjects may be divided among:

- general subjects (country, language, currency, geographical locations, etc.)
- subjects linked to an industrial domain (pharmacy, telecom, automobile, etc.) that are indispensable to all enterprises (names of molecules, diseases, technologies, regulations, etc.)
- subjects linked to a particular enterprise (jobs, departments, clients, products, etc.)
- subjects linked to a particular department (projects, business processes, persons, etc.)

The subject is a basic tool for:

- the content consumer to express her requests
- the author, whose documentary production must be published with appropriate subjects as keywords
- organizers and documenters, who use subjects as the fundamental tool for organizing content

4.2. Organize the content – organize the subjects

Just as in the case of ordered and systematic access to books in a library or bookshop, content organization relies on the organization of subjects in the enterprise. This organization has existed since a long time in the enterprise and is found scattered across different systems of document management:

- Classificatory scheme used by documentation centers
- Lotus Notes bases
- Enterprise thesaurus or industry thesaurus (Mesh, for example)
- Bibliographies
- Index of technical documentation
- Product classification, nomenclature, etc.
- etc.

These different resources for organizing subjects use differing methods:

- classification (documentation center, product classification)
- simple semantic relations (thesaurus)
- non-hierarchical relations (bibliographies, book index, etc.)

- hierarchical relations (break-down of a machine, technical documentation)

Part and parcel of the culture of the enterprise and its employees, these different elements of organizing subjects must be federated and incorporated within the Content Intelligence system of which they form the infrastructure

4.3. Aggregate the content

We are considering here what happens within a single enterprise, but the enterprise is global, composed of several entities, departments. The content management systems are numerous, dispersed and specialized:

- several installations of document management software for technical documentation and patents in XML and SGML
- dozens of intranets
- numerous installations of Lotus Notes
- databases
- management applications
- tools for competitive and technical intelligence
- etc.

The aggregation of content is necessary in order to meet the content consumer's need to access all the relevant documents throughout the enterprise.

For the responses to precise subjects to be pertinent and exhaustive, there must be:

- an aggregation of contents around the subjects of the enterprise
- the ability to access all the contents of the enterprise

5. What are the tools needed to build a solution?

5.1. Building an enterprise repository of subjects

Building an enterprise-wide subject repository is first of all a work of reusing whatever subjects already exist:

- at the level of an industry (industry thesaurus, standardization of keywords, etc.)
- at the enterprise level (document classification, classification used for document management, inventory of products, jobs, tools, etc.)
- at the department level (projects, people, etc.)

Automatic taxonomy creation tools that mine a corpus of documents may help enrich and control the list of subjects.

Each subject must be defined, along with its different synonyms and orthographies, so as to take into account the different languages used in the enterprise.

The management of subjects requires software tools that are adapted for administering these subjects, for updating and deleting them, and so on. The software solutions for managing subjects must include functions for collaborative work, workflow, and the automatic importation of legacy data. They must likewise be sufficiently standardized and open so as to be used both by humans and through automated handling by other software components of a Content Intelligence solution.

5.2. Organize the subjects

Subject organization relies on different tools depending on the nature of the subjects to be organized:

- the reuse of document classifications already in use within the enterprise or common to the whole industry allows the most general organization of subjects
- the reuse of database content allows the building of a subject organization appropriate for the specific enterprise (organization of products, projects, departments, etc.)
- the reuse of process descriptions makes it possible to organize subjects having to do with job processes
- the incorporation of local databases allows reuse of subject organizations such as projects, people, etc.

The creation of one or several systems of subject organization may thus rely on the reuse of elements already existing in the enterprise.

Managing the organization of subjects requires software tools adapted for this specific purpose. These tools must be capable of administering hundreds of thousands of relations that organize subjects, of reusing pre-existing organizations that derive from ontologies and thesauri, of mapping the relationships described in the databases.

These software solutions must be open enough to serve as the basis for building tools that handle requests and for transacting with the various other components of the content intelligence solution: natural language searches, inference engine, tool for graphically representing the organization of content, etc.

Software solutions of the inference engine type may likewise be implemented so as to enrich the organization of subjects from already existing organizational elements.

5.3. Referencing heterogeneous content

The referencing of content relies on different techniques depending on the nature of the subject:

- reuse of the indexing scheme of documentation centers
- use of keywords in technical documentation
- automatic categorization
- text mining
- use of database information
- manual referencing and automatic verification of referencing
- use of full text search tools based on a repertory of subjects and their organization

The solutions will be different depending on the nature of the documents:

- organized document production (technical documents, patents, marketing documents) may be automatically referenced if it is based on the enterprise-wide subject repository
- structured external documentation may be likewise automatically referenced by relying on its metadata
- internal production deriving from office tools may be referenced if assistance is provided to the authors in the choice of document keywords (automatic categorization under author control)
- unstructured documents coming from the outside may be referenced by combining the categorization and text mining tools.

For all such contents referenced in a stable manner to subjects, an adequate software solution should allow the enterprise to manage these relations between the subjects and the contents, and to place these relations at the disposal of the other components of the Content Intelligence solution.

6. What are the tools used within the implemented solution?

6.1. Search

Document search using search functions should combine:

- search against the different subject identifiers (product name in different languages, reference number, etc.)
- solutions that use natural language search tools to allow the subjects of the enterprise to be explored without knowing beforehand their exact identification

Natural language search, based on a defined professional vocabulary and against subjects and their organization, seems to us an indispensable tool as soon as the end-user attempts to search within domains with which she is the least familiar.

6.2. Navigate

Navigation functions between the subjects (broader or narrower subjects, different parts of a machine, patents linked to a molecule, successive steps of a process, technologies used for a project, etc.) are of the utmost importance for finding the correct information when the search has made it possible to identify the subjects.

These navigation functions may be wholly manual by using the hyperlinks on the navigation pages, but may also be automated.

A simple automation would allow all the documents of a certain type for a product to be found along with its hierarchically linked components. By relying on navigation automations or on inference engines, a more complex automation of the navigation would allow all semantic and logical relations between the subjects to be explored.

Components for graphically representing subject organization should likewise be included within a Content Intelligence solution, whether to represent relations between the product components or a manufacturing process.

7. Topic Maps as the backbone of Content Intelligence

7.1. Solution for managing subjects

The extensions of the Topic Maps standard regarding the identification of subjects (Published Subject / OASIS technical group) allow to manage, in a centralized or decentralized manner, the stable and unique identifiers of the subjects.

This basic component is of utmost importance to ensure:

- the coherence of the subject identifiers used in all the content systems throughout the enterprise
- the interoperability of the different software tools that use these subjects (text mining, categorization, natural language requests)

7.2. Open and decentralized solutions for organizing subjects

The separation between:

- the enterprise-wide subject repository
- the organization of subjects among themselves within the Topic Maps based management software for knowledge organization

- the contents managed in the various content management tools of the enterprise

allows to decentralize the solutions for managing subject organization and those for managing contents, and thereby ensures their mutual independence.

One could have a general organization of the enterprise subjects so as to aggregate the content of generalist intranets and also, for example, a subject organization relating rather to the technical documentation of products The ability to manage these two systems independently, in an autonomous manner, is in fact indispensable.

7.3. Personalization solutions

Any system of referencing subjects, organizing subjects and content access should include personalization mechanisms that ensure the following functions:

- personalization that conforms to the language of the end-user
- personalization that conforms to the end-user's authorizations to information access
- personalization that conforms to the specific job needs of the end-user

The Topic Maps standard along with its derived tools allow for these personalization mechanisms.

7.4. Non-intrusive solutions for referencing contents

Enterprises already have various sorts of content management solutions in place, with a multiplicity of systems, and they need to federate these diverse contents. This is why implementing solutions that would require a modification of these existing systems is out of the question.

The principles underlying the Topic Maps standard and the software that implements this standard are based on the referencing of contents through subjects, without there being any need to modify the content management.

The Topic Maps solutions moreover allow to manage a set of associated metadata for each document without modifying the content management solutions already in place.

7.5. Ability to describe all models of knowledge organization currently in use across the enterprise

Deriving from the world of documents and having its origin in the problems posed by the reuse of pre-existing document organizations, the Topic Maps standard does not impose any particular model of organizing documents. The association mechanisms support all possible models of knowledge organization by taking the current needs and existing state of affairs into account, and enabling the coexistence of these different models:

- simple repertory of subjects
- hierarchical organization: classifications and taxonomies
- thesaurus
- relational non-hierarchical organization with or without semantics to the relations
- process description
- inter-document organization
- etc.

thus allowing just as well the description, within a single system, of the document classification, the organization of a complex thesaurus, the breakdown of a machine, the relationships between the participants in a project, etc.

7.6. A repertory of subjects, organization and enterprise vocabulary at the service of other technical components

We've seen that a Content Intelligence solution puts to work several technologies:

- Text mining
- Automatic categorization
- Natural language search
- Inference
- Full text search tools
- Collaborative tools

These tools, taken as a whole, use:

- repertories of subjects
- the organization of subjects among themselves
- the referencing relations between the subjects and their contents

Though each of these different tools uses its own techniques and handling algorithms, a close coordination must be maintained between the tools and the Topic Maps solutions so as to ensure:

- coordination in referencing the subject repository
- coordination in organizing the subjects
- close coordination between the subject repertory, subject organization and these tools during handling

This coordination requires a normalization of transactions through standardized XML files, API standards, and Web Services.

8. Conclusion

8.1. Why Topic Maps tools for content intelligence solutions?

Topic Maps tools are ideally suited for providing the infrastructure to content intelligence solutions because of:

- the separation between subject management, subject organization and contents
- the ability to implement decentralized interoperable solutions
- the ability to reuse and exploit the existing information organization within the enterprises

These strengths explain the proliferation of operational projects based on the Topic Maps standard and tools within various industries, the publishing world and in administration.

8.2. What's still needed to make content intelligence solutions a reality?

Several elements are still missing in the Topic Maps standard and its derivative software for the solutions implemented to be complete and interoperable:

- the finalization of the PSI standard

- a standardization, within the Topic Maps framework, of the most commonly used models of knowledge organization so as to ensure interoperability of the solutions implemented
- the definition of generic APIs, Web services and standards that would together facilitate the emergence of a dialogue between the different software components of a Content Intelligence solution

But, as of today, these different challenges have already been identified by the different actors and are progressively finding solutions within the framework of standardization committees, operational projects for clients and industrial collaborations among the different providers of technical solutions.

8.3. Our convictions

At the core of the enterprise, Content Intelligence tools will make it possible to:

- build solid, generic and universal solutions
- place the human knower and actor at the center of solutions for content organization and access
- reuse and exploit existing document organization solutions within enterprises
- federate the different technologies that will help humans in the work of organizing contents and knowledge and of facilitating their access.

Biography

Jean **Delahousse**
Mondeca
Paris
France
jean.delahousse@mondeca.com

Jean Delahousse, born in 1958, graduated from ESCP (France). He worked for Andersen Consulting, the Paris Stock Exchange and Diagram (a financial software company). He created BDB, a financial software company, in 1993. Since 1999, he has been CEO and co-founder of MONDECA, a software company dedicated to content organization and content access solutions based on Topic Maps. Jean Delahousse has a wide experience of software design and marketing. He was one of the cofounders of the XTM authoring group and belongs to the board of XML France. Merging knowledge engineering, graph theory and XML knowledge representation standards into useful tools for end-users within the entreprise is his main concern today