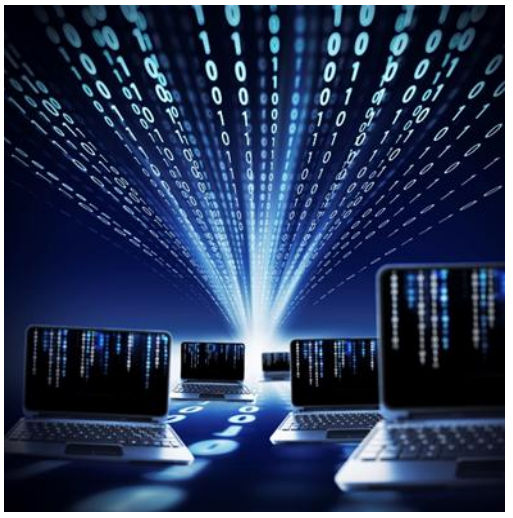
GALAXYCONSULTING
Information Management Solutions

DITA

White Paper
By Galaxy Consulting

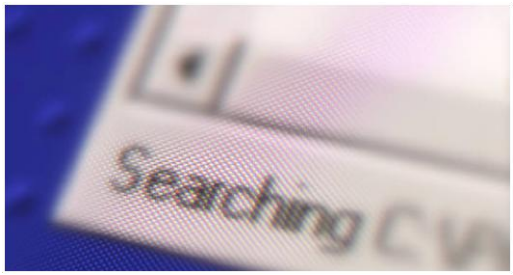
April

A



At Your Service…Today…Tomorrow
We Appreciate The Privilege Of Serving You!

May 2015



Abstract

Many organizations use component content management to create more sophisticated content, in more languages, and at lower cost. Component content management systems (CCMS) manage components at a fine granular level, in ways that allow the components to be easily used, reused, versioned, linked, assembled, and reassembled into different content products.

In order to support CCMS, content has to be in the eXtensible Markup Language (XML) format. Darwin Information Typing Architecture (DITA) format is widely used for this purpose. This white paper describes DITA, DITA architecture, and DITA support for metadata.

What is DITA?

The Darwin Information Typing Architecture (DITA) is an XML-based architecture for authoring, producing, and delivering information. Although its main applications have so far been in technical publications, DITA is also used for other types of documents such as policies and procedures. The DITA architecture and a related DTD and XML Schema were originally developed by IBM.

At the heart of DITA, representing the generic building block of topic-oriented information architecture is an XML document type definition (DTD) called the topic DTD. The point of the XML-based Darwin Information Typing Architecture (DITA) is to create modular technical documents that are easy to reuse with varied display and delivery mechanisms.

DITA Architecture

As the "Architecture" part of DITA's name suggests, DITA has unifying features that serve to organize and integrate information:

Topic Orientation. The highest standard structure in DITA is the topic. Any higher structure than a topic is usually part of the processing context for a topic, such as a print-organizing structure or the navigation for a set of topics.

DITA content is written as modular topics, as opposed to long "book-oriented" files. A DITA map contains links to topics, organized in the sequence (which may be hierarchical) in which they are intended to appear in finished documents. A DITA map defines the table of contents for deliverables. Relationship tables in DITA maps can also specify which topics link to each other.

Modular topics can be easily reused in different deliverables. However, the strict topic-orientation of DITA makes it an awkward fit for content that contains lengthy narratives that do not lend themselves to being broken into small, standalone chunks. Experts stress the importance of content analysis in the early stages of implementing structured authoring.

Fragments of content within topics (or less commonly, the topics themselves) can be reused through the use of content references (conref), a transclusion mechanism.

DITA includes extensive metadata elements and attributes, which make topics easier to find.

Topic Types. DITA specifies three basic topic types: Task, Concept and Reference. Each of the three basic topic types is a specialization of a generic Topic type, which contains a title element, a prolog element for metadata, and a body element. The body element contains paragraph, table, and list elements, similar to HTML.

- Task topic is intended for a procedure that describes how to accomplish a task. A Task topic lists a series of steps that users follow to produce an intended outcome. The steps are contained in a taskbody element, which is a specialization of the generic body element. The steps element is a specialization of an ordered list element.
- Concept topic is more objective, containing definitions, rules, and guidelines.
- Reference topic is for topics that describe command syntax, programming instructions, and other reference material, and usually contains detailed, factual material. DITA allows adding new elements and attributes through specialization of base DITA elements and attributes. Through specialization, DITA can accommodate new topic types, element types, and attributes as needed for specific industries or companies.

The extensibility of DITA permits organizations to specialize DITA by defining specific information structures and still use standard tools to work with them. The ability to define company-specific information architectures enables companies to use DITA to enrich content with metadata that is meaningful to them, and to enforce company-specific rules on document structure.

DITA map and topic documents are XML files. As with HTML, any images, video files, or other files which need to appear in output are inserted via reference. Any XML editor can therefore be used to write DITA content, with the exception of editors that support only a limited set of XML schemas (such as XHTML editors). Various editing tools have been developed that provide specific features to support DITA, such as visualization of conrefs.

Reuse. A principal goal for DITA has been to reduce the practice of copying content from one place to another as a way of reusing content. Reuse within DITA occurs on two levels:

Topic reuse. Because of the non-nesting structure of topics, a topic can be reused in any topic-like context.

Content reuse. DITA provides each element with a conref attribute that can point to any other equivalent element in the same or any other topic.

Specialization. Any DITA element can be extended into a new element.

Topic specialization. Applied to topic structures, specialization is a natural way to extend the generic topic into new information types (or infotypes), which in turn can be extended into more specific instantiations of information structures. For example, a recipe, a material safety data sheet, and an encyclopedia article are all potential derivations from a common reference topic.

Domain specialization. Using the same specialization principle, the element vocabulary within a generic topic can be extended by introducing elements that reflect a particular information domain served by those topics. For example, a keyword can be extended as a unit of weight in a recipe, as a part name in a hardware reference, or as a variable in a programming reference.

Property-Based Processing

The DITA model provides metadata and attributes that can be used to associate or filter the content of DITA topics with applications such as content management systems, search engines, etc.

Extensive metadata to make topics easier to find. The DITA model for metadata supports the standard categories for the Dublin Core Metadata Initiative. In addition, the DITA metadata enables many different content management approaches to be applied to its content.

Universal properties. Most elements in the topic DTD contain a set of universal attributes that enable the elements to be used as selectors, filters, content referencing infrastructure, and multi-language support.

Taking advantage of existing tags and tools. Rather than being a radical departure from the familiar, DITA builds on well-accepted sets of tags and can be used with standard XML tools.

Leveraging popular language subsets. The core elements in DITA's topic DTD borrow from HTML and XHTML, using familiar element names like p, ol, ul, and dl within an HTML-like topic structure. In fact, DITA topics can be written, like HTML for rendering directly in a browser.

Leveraging popular and well-supported tools. The XML processing model is widely supported by a number of vendors and translates well to the design features of the XSLT and CSS stylesheet languages defined by the World Wide Web Consortium and supported in many transformation tools, editors, and browsers.

Typed topics are easily managed within content management systems as reusable, stand-alone units of information. For example, selected topics can be gathered, arranged, and processed within a delivery context to provide a variety of deliverables to varied audiences. These deliverables might be a booklet, a web site, a specification, etc.

At the center of these content management systems are fundamental XML technologies for creating modular content, managing it as discrete chunks, and publishing it in an organized fashion. These are the basic technologies for "one source, one output" applications, sometimes referred to as Singe Source Publishing (SSP) systems.

The innermost ring contains capabilities that are needed even when using a dedicated word processor or layout tool, including editing, rendering, and some limited content storage capabilities. In the middle ring are the technologies that enable single-sourcing content components for reuse in multiple outputs. They include a more robust content management environment, often with workflow management tools, as well as multi-channel formatting and delivery capabilities and structured editing tools. The outermost ring includes the technologies for smart content applications.

It is good to note that smart content solutions rely on structured editing, component management, and multi-channel delivery as foundational capabilities, augmented with content enrichment, topic component assembly, and social publishing capabilities across a distributed network.

Content Enrichment

A descriptive metadata taxonomy is created or adopted and its use for content enrichment will depend on tools for analyzing and/or applying the metadata. These can be manual dialogs, automated scripts and crawlers, or a combination of approaches. Automated scripts can be created to interrogate the content to determine what it is about and to extract key information for use as metadata. Automated tools are efficient and scalable, but generally do not apply metadata with the same accuracy as manual processes. Manual processes, while ensuring better enrichment, are labor intensive and not scalable for large volumes of

content. A combination of manual and automated processes and tools is the most likely approach in a smart content environment. Taxonomies may be extensible over time and can require administrative tools for editorial control and term management.

Component Discovery/Assembly. Once data has been enriched, tools for searching and selecting content based on the enrichment criteria will enable more precise discovery and access. Search mechanisms can use metadata to improve search results compared to full text searching. Information architects and content managers can use search to discover what content exists, and what still needs to be developed to proactively manage and monitor the content. These same discovery and search capabilities can be used to automatically create delivery maps and dynamically assemble content organized using them.

Distributed Collaboration/Social Publishing. Componentized information lends itself to a more granular update and maintenance process, enabling several users to simultaneously access topics that may appear in a single deliverable form to reduce schedules. Subject matter experts, both remote and local, may be included in review and content creation processes at key steps. Users of the information may want to "self-organize" the content of greatest interest to them, and even augment or comment upon specific topics. A distributed social publishing capability will enable a broader range of contributors to participate in the creation, review and updating of content in new ways.

Federated Content Management/Access. Smart content solutions can integrate content without duplicating it in multiple places, rather accessing it across the network in the original storage repository. This federated content approach requires the repositories to have integration capabilities to access content stored in other systems, platforms, and environments. A federated system architecture will rely on interoperability standards (such as CMIS), system agnostic expressions of data models (such as XML Schemas), and a robust network infrastructure (such as the Internet).

These capabilities address a broader range of business activity and therefore fulfill more business requirements than single-source content solutions. Assessing your ability to implement these capabilities is essential in evaluating your organizations readiness for a smart content solution.

DITA Support for Metadata

Finding content in your file system or content repository is hard enough when you've got simple text documents to deal with. When you are using DITA and other component-oriented XML standards, you increase the difficulty by two or three orders of magnitude, because you're looking for smaller needles in bigger haystacks. Having thousands of media-independent content objects that can be shared and reused across multiple deliverables allows you to create more sophisticated knowledge products, but it definitely poses a challenge in findability for content authors.

Among its many features for content reuse, DITA provides content creators with a facility for tagging content objects with metadata. Metadata (data about the data) lets content authors and others who manage content describe what the content is about ("descriptive metadata"), as well as assign properties like who created the content, when, in what language, and for which audience ("administrative metadata").

A taxonomy is a hierarchical structure that organizes concepts and controls vocabulary. Taxonomies allow organizations to create and centrally manage important terms that can be applied to content as metadata. For example, a telecommunications manufacturer might have a taxonomy that includes concepts such as product categories (Mobile Phones, Wireless Routers, and so on), industries (Healthcare, Utilities, Transportation, and so on), or product models.

Once applied, this metadata and taxonomy can be leveraged by a search application to help users find and use content. Search engines can use taxonomy to organize search results in meaningful ways, such as refining search based upon certain properties ("faceted search") and suggesting related searches based upon relationships between search terms and other concepts in the taxonomy.

It is a natural fit — DITA and taxonomy. DITA creates a multitude of reusable components, and taxonomy helps describe and organize the components so that they may be readily found and reused by content authors and users.

Taxonomies and descriptive metadata is also a natural fit since metadata-based search would improve findability of content objects.

Compared to other XML standards, DITA provides a relatively rich and extensible framework for embedding metadata directly within the XML objects themselves. The embedded metadata can be used by processing tools like the publishing tools in the DITA Open Toolkit (DOTK) to conditionally publish content or to create metadata in the final outputs, like HTML.

DITA objects, both topics and maps, have a prolog section in which metadata can be specified. Within the prolog, the metadata section can define metadata about the topic itself such as the intended audience, the platform (for defining the applicability of the topic to specific hardware or operating systems), and so on. This metadata can be used for conditional publishing. For example, you can automate the production of a Linux version of your documentation by only outputting topics and maps that set platform to "Linux" in the metadata.

DITA objects can also embed administrative metadata about the author, copyright holder, source, publisher, and so on. Metadata can also contain descriptive keywords for the topic or map. Keywords or index terms are output to HTML or XHTML as metadata keywords to support search engines. Authors can also define index terms for the generation of back-of-book indices.

DITA also enables users to define custom metadata fields within the "othermeta" element. Like keywords, metadata defined as "othermeta" are output as HTML metadata elements but ignored for other types of output like PDF. Metadata is a powerful tool in helping to manage and publish DITA content.

Dynamic Publishing of Content

A major benefit of DITA is creating content that is media-independent. It also enables content objects to be organized by DITA maps, so that content can be recombined and re-sequenced into different deliverables. DITA maps provide flexibility.

Dynamic publishing lets content be chosen and presented to meet the unique needs of a user or situation. To best illustrate dynamic publishing, let's compare it with static publishing of a help system.

In a statically published help system, the hierarchy of topics is fixed by the author and the selection of content is limited to what is in the DITA map at publish time. All of the related topics are manually linked. If an author wants to add a related topic, the author needs to manually add the link (or update the related-links table) and republish. The publishing process creates a deliverable that—while interactive—is static with respect to its contents and the relationships among them.

To create the same help system with dynamic publishing, the author would publish his/her content to a server, but he/she would not create the structure and relationships between topics at publish-time. Instead, a taxonomy would specify the relationship between concepts and properties that are defined in metadata. The relationships among topics are generated at run-time, based upon metadata on the topics. The richer the metadata and the more complete the taxonomy, the more sophisticated the user experience.

If you have experienced faceted search on consumer web sites, where we can refine search results by selecting specific values for different attributes, such as the number of megapixels for a camera. This experience is driven by metadata. With rich metadata on DITA content, we can create very sophisticated electronic content browsers, where metadata-based search creates browser-like user experiences.

Final Recommendations

Start by identifying all your taxonomy use cases. You will be using taxonomy not only for authors to search

content objects for reuse but also potentially for serving up content to users dynamically or in a faceted interface. These perspectives will provide you with the framework for your taxonomy.

Reuse existing vocabulary. Many organizations already use controlled vocabularies for some metadata fields such as organization, audience, platform, and product. Look to existing sources for tagging your content such as hierarchical product or system models (from engineering), or hierarchical task models (from instructional/task analysis from the training organization) as places to start building hierarchical descriptive taxonomies.

Authors are the best people to apply descriptive metadata. After all, they do the analysis to determine what content was required in the first place, so they have the best context for classifying it. However, don't expect authors to tag a lot: automate tagging when possible, especially for administrative metadata (author, organization, creation date, language).

Leverage the technology. Many content management systems can integrate third-party classification servers for automating descriptive metadata. These servers can automatically apply metadata from a taxonomy or controlled vocabulary when content topics are checked-in, then automatically populate subject metadata fields in the CMS. The metadata can in turn be reviewed and manually adjusted by authors. This metadata can be embedded into your DITA content for use in conditional publishing or to generate HTML tags in the final output to support search or dynamic publishing.

The next frontier of DITA adoption is leveraging semantic technologies (taxonomies, ontologies and text analytics) to automate the delivery of targeted content. For example, a service incident from a customer is automatically matched with the appropriate response, which is authored and managed as a DITA topic.

About Galaxy Consulting

GALAXYCONSULTING
Information Management Solutions

Galaxy Consulting provides services in business analysis and usability, content and knowledge management, records management, information architecture, enterprise search, taxonomy development and management, document control, and information governance.

Galaxy Consulting was founded with the mission and vision of helping organizations to manage their valuable information assets. Many of our clients, both large and small, have dramatically improved efficiency and reduced unnecessary labor hours through efficient methods, processes, and solutions we created.

Galaxy Consulting believes in partnerships with our clients. We are committed to working with you and to helping you transform your business. We will increase efficiency and productivity, maintain regulatory and legal compliance, improve collaboration, enhance innovation, and reduce costs through effective information management!

Call us TODAY to schedule a free, no obligation consultation!

Contact Us

Office: 650-474-0955
Mobile: 650-716-3609
Info@galaxyconsulting.net
www.galaxyconsulting.net