An OASIS White Paper

DITA 1.2 Glossary and Terminology Specialization Feature Description

Kara Warburton, IBM Terminology

21 October 2009



OASIS (Organization for the Advancement of Structured Information Standards) is a not-for-profit, international consortium that drives the development, convergence, and adoption of e-business standards. Members themselves set the OASIS technical agenda, using a lightweight, open process expressly designed to promote industry consensus and unite disparate efforts. The consortium produces open standards for Web services, security, e-business, and standardization efforts in the public sector and for application-specific markets. OASIS was founded in 1993. More information can be found on the OASIS website at http://www.oasis-open.org.

The OASIS DITA Adoption Technical Committee members collaborate to provide expertise and resources to educate the marketplace on the value of the DITA OASIS standard. By raising awareness of the benefits offered by DITA, the DITA Adoption Technical Committee expects the demand for, and availability of, DITA conforming products and services to increase, resulting in a greater choice of tools and platforms and an expanded DITA community of users, suppliers, and consultants.

DISCLAIMER: All examples presented in this article were produced using one or more tools chosen at the author's discretion and in no way reflect endorsement of the tools by the OASIS DITA Adoption Technical Committee.

Table of Contents

Table of Contents	3
1. Statement of the problem	4
2. Terms and definitions	5
3. Fundamental principles	6
4. Explanation of the elements	7 8 8 9
4.5 Miscellaneous elements	
5.1 Glossary entries as DITA files 5.1.1 Each glossary entry in a separate DITA file 5.1.2 Multiple glossary entries in one DITA file 5.2 Referencing glossary entries 5.2.1 Declaring a key for a glossary entry topic 5.2.2 Referencing a glossary entry from the text 5.2.3 Creating a glossary deliverable 5.3 Examples of glossary entries 5.3.1 A single defined term 5.3.2 Preferred and alternate forms 5.3.3 Full forms, acronyms, and surface forms 5.3.4 An entry for a published glossary 5.3.5 An entry destined for populating a terminology database 5.3.6 An entry with a cross reference	111213141515161617
6. Processing considerations	
6.1 Referring to a glossary entry from the text	

1. Statement of the problem

Markup for glossaries and glossary-type data has not evolved since the origins of HTML. The focus has been on formatting publishable glossaries instead of managing terms and their meaning. Today, even the most sophisticated authoring markup languages mirror the simple two element definition-list style of markup found in HTML, providing just one element for the term and one element for the definition. There are no additional elements available to handle other types of metadata, such as part of speech values, relationships between abbreviations and full forms, comments, usage notes, sources, and so forth. The existing model does not even allow for terms that have multiple meanings. The lack of granularity in markup for glossaries has presented significant barriers to using information about terms that is recorded in glossaries for down-stream processes, such as inclusion in a terminology database, extraction for pre-translation, and reuse in controlled authoring applications.

In technical writing, synonyms and variants should be used judiciously and often avoided altogether. The use of one term consistently to express a given concept is preferred so that communication is clear and so that translation costs are minimized. For this reason, when synonyms and variants do exist in popular usage, it is common practice in commercial environments to choose one of the terms as the "preferred term." This indicator of preferred usage needs to be documented in glossaries. Due to the limitations of markup languages for creating glossaries, usually the so-called preferred term is identified simply by making it the headword in a glossary entry and providing a definition in this glossary entry. Synonyms and variants are also recorded as headwords in their own glossary entries, but instead of being accompanied by a definition, a reference is provided to the preferred term entry, such as by using the formulation "See X." where "X" is the preferred term. By virtue of the fact that a term has a "see" reference instead of a definition, it is deemed to be a non-preferred term. However, this type of unstructured, prose-based reference is only usable by humans; machine processes cannot interpret this information in order to establish logical links between terms, such as synonyms, that could be usable for other automated processes.

Also due to the limitations of markup languages, in the past synonyms were not clearly identified in glossaries and often not identified at all. If they existed in glossaries, the reader could only find them by following the "see" references. Among content developers, there are increasing requirements for knowledge about synonyms. The DITA glossary specialization records all synonyms clearly in the same entry structure.

To address these requirements, DITA 1.2 adds a glossary specialization with markup that supports terminological information.

2. Terms and definitions

For the purposes of clarity, the meanings of the following terms, which are used in this document, are provided below:

glossary entry

A <glossentry> element and all the information contained within it.

NOTE: A glossary entry is not only used to record information for glossaries. It can be used for other purposes such as to record acronyms and abbreviations that are referred to in documents or to record information that may be later retrieved for terminology databases or other lexical resources.

sense

The concept or meaning denoted by a term.

NOTE: Although more common in the field of terminology, the term "concept" has been avoided due to the conflict with the concept topic type in DITA.

term

Any of the designations documented in a glossary entry. Terms include the main entry term (<glossterm>) and all variants (<glossAbbreviation>, <glossAcronym>, <glossShortForm>, and <glossSynonym>). Terms can comprise single words, such as "satellite" and multi-word expressions, such as "cascading style sheet."

3. Fundamental principles

Each glossary entry (<glossentry>) is a container element for all other elements defined in this glossary specialization best practices document. Each glossary entry describes and documents one sense (meaning, or *concept*). In the field of terminology management, the principle whereby a glossary entry (usually called a terminological entry) describes a concept is referred to as "concept orientation." However, because the term "concept" is a topic type in DITA, to avoid confusion, in this document, the term "sense" is used to refer to the concept denoted by a term.

Multiple terms can denote the same sense, and these are known as synonyms. For instance, "laptop" and "portable computer" could be considered synonyms. These types of synonyms are sometimes referred to as "lexical synonyms" because the form of the words is entirely different. In this DITA specialization, the term "synonym" and the element <glossSynonym> shall be reserved for this type. Other types of synonyms exist, often called "variants" because their lexical forms are related, such as abbreviations, acronyms, and other short forms. In this DITA specialization, the elements <glossAbbreviation>, <glossAcronym>, and <glossShortForm> are provided for these variants. DITA also introduces a new element, <glossSurfaceForm>, to record an alternate expanded surface form for glossary entries that contain abbreviations and acronyms.

With the DITA glossary specialization, dedicated elements are available to uniquely mark each of these types of terms in a machine-readable fashion. Because synonyms and variants all have the same meaning, they are all recorded in the same glossary entry as the preferred term. If there is no preferred term marker, the term in the <glossterm> element is assumed to be the preferred term.

With the exception of <glossSurfaceForm>, each variant and synonym is contained within a <glossAlt> element along with other elements that describe it.

4. Explanation of the elements

4.1 Types of terms

The following elements each contain a term.

<glossterm>

The term being described in the glossary entry. When there is more than one term being described in the glossary entry, this element shall contain the preferred term. In cases where a term has an abbreviated form, this element shall contain the expanded form of the term.

<glossAbbreviation>

An abbreviated form formed by omitting letters from a longer form.

Example:

glossterm: adjective glossAbbreviation: adj

<glossAcronym>

An abbreviated form made up of the initial letters of the components of the full form or from the syllables of the full form.

Example:

glossterm: United Nations

glossAcronym: UN

<glossShortForm>

An abbreviated form that includes fewer words than the full form.

Example:

glossterm: Intergovernmental Group of Twenty-four on Monetary Affairs

glossShortForm: Group of Twenty-four

<glossSynonym>

A term having the same meaning as the term in the <glossterm> element, and which is not an abbreviated form.

Example:

glossterm: automobile glossSynonym: car

<glossSurfaceForm>

An unambiguous presentation of the term that may combine multiple forms to introduce the term in the first occurrence in text. For instance, in an entry that contains an acronym, the <glossSurfaceForm> might provide the full form as well as the acronym in parentheses. Refer to the OASIS document *Best Practice for Managing Acronyms and Abbreviations in DITA* for more details.

4.2 Elements for free-text descriptions

The following elements contain descriptions or other free text.

<glossdef>

A definition of the meaning of the term.

<glossUsage>

Information about how the term is used in context.

<glossScopeNote>

A clarification of the sense designated by the terms such as examples of included or excluded companies or products. For instance, a scope note for "Linux" might explain that the term doesn't apply to UNIX products and give some examples of Linux products that are included as well as UNIX products that are excluded.

<glossProperty>

Any other linguistic or semantic property of the term, such as the gender of the term.

<note>

Any other note about the term or sense.

4.3 Limited value elements

The following two elements express a limited set of values, through the *value* attribute. These are empty elements.

Note: The list of values is extensible or replaceable.

<glossPartOfSpeech>

Values: noun, properNoun, verb, adjective, adverb

Identifies the part of speech for the preferred and alternate terms. Alternate terms must have the same part of speech as the preferred term because all terms in the glossary entry designate the same sense. If the part of speech is not specified, the default is noun.

<glossStatus>

Values: preferred, restricted, prohibited, obsolete

Identifies the usage status of a preferred or alternate term. If the status is not specified, the <glossterm> represents the preferred term and the remaining term elements described in section 3.1 document allowed terms.

4.4 Structural elements

The following elements are nesting elements that create the structure of the glossary entry.

<glossentry>

The root element of the glossary entry. It contains:

- one <glossterm>
- zero or one <glossdef>
- zero or one <prolog>
- zero or one <glossBody>
- zero or one <related-links>

<glossBody>

Contains elements that describe the main or preferred term in <glossterm> and a <glossAlt> nesting element to group alternate terms. It contains:

- zero or one <glossPartOfSpeech>
- zero or one <glossStatus>
- zero or more <glossProperty>
- zero or one <glossSurfaceForm>
- zero or one <glossUsage>
- zero or one <glossScopeNote>
- zero or more <glossSymbol>
- zero or more <note>
- zero or more <glossAlt>

<glossAlt>

Contains one alternate term and additional elements to describe that alternate term. The existence of an alternate term is of course specific to a language, so the translated version of the glossary entry may result in empty elements in this section. It contains:

- one <glossAbbreviation>, <glossAcronym>, <glossShortForm>, or <glossSynonym>
- zero or one <glossStatus>
- zero or more <glossProperty>
- zero or one <glossUsage>
- zero or more <note>
- zero or more <glossAlternateFor>

Note: all the elements within the <glossAlt> element describe the alternate term, not the preferred term.

4.5 Miscellaneous elements

olog>

The standard DITA prolog element, here optionally occurring at the glossary entry (<glossentry> or glossary group (<glossgroup>) level. It can be used to provide administrative information about the glossary entry.

<glossAlternateFor>

A type of <xref> that indicates when a variant term has a relationship with another variant term as well as with the preferred term.

<related-links>

The standard DITA related-links element. It can be used to associate a glossary topic with other DITA topics or resources on the web.

<glossSymbol>

A graphical symbol or image (that is, a visual rather than a linguistic designator) of the sense being described by the glossary entry.

5. Best practices

This section describes the recommended ways to mark up glossary entries by means of examples.

The only mandatory element in a glossary entry is the term (<glossterm>). This practice permits the creation of glossary entries in stages, the first stage being the identification of the terms to document. It would be highly unusual for a completed glossary entry to contain only a term.

Since glossary entries are sense-oriented (i.e., concept-oriented), each glossary entry shall have only one definition (<glossdef>). For terms that have multiple meanings, create a separate glossary entry for each sense or meaning of the term.

Since a particular sense of a word can only express one grammatical function (noun, verb, etc.), only one part of speech value is permitted. All the variants and synonyms inherit the part of speech of the so-called main term (<glossterm>). Even though, historically, recording the part of speech has not been not common in glossaries, it is now highly recommended to record this data particularly if it is anticipated that the glossary may be required for purposes dependent on machine-readability of the data, such as for creating lexical resources for controlled authoring applications, for importing the glossary data into a terminological database, or for populating linguistic tables for text analysis applications. The part of speech is also valuable information appreciated by translators if the glossary is to be translated, to help them select the appropriate translation. If desired, the part of speech can be automatically excluded from the published glossary by the processing software.

Each term and variant can only have one status value (<glossStatus>), since, for example, a particular term with a given sense cannot be preferred and prohibited at the same time.

5.1 Glossary entries as DITA files

A glossary entry can be recorded alone in a single DITA file or grouped with other glossary entries in a DITA file.

5.1.1 Each glossary entry in a separate DITA file

It is important to point out the value of the best practice of putting each glossary entry in a single DITA topic file. This practice enables reuse of glossary entries because the different parts of the

glossary can be assembled differently for different deliverables. Instead of the writer authoring a series of glossary terms and their definitions (or "see" references to preferred terms) in alphabetical order, DITA processors are expected to sort the glossary entries alphabetically, insert the definition or "see" reference based on the term's status, and assemble all definitions of the same term under one headword. This arrangement can differ depending on the organization's publishing rules and transforms.

Besides being essential for reuse, the separation of glossary entries into individual DITA files also supports translation. The decision on which term in an entry is "preferred" is language-dependent. If the writer authors the glossary as if it was the formatted deliverable, the translator has to rework the format. But if a processor generates the formatted glossary, the translator only has to rerun the processor.

A typical DITA glossary entry file that contains just one entry would look like the following, (some attributes and optional elements have been omitted for simplicity purposes):

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE concept PUBLIC "-//OASIS//DTD DITA 1.1 Concept//EN" "concept.dtd"
<glossentry id="abs">
<glossterm>anti-lock braking system</glossterm>
...
</glossentry>
```

5.1.2 Multiple glossary entries in one DITA file

There are two ways to create a DITA glossary entry file that contains multiple glossary entries: using the <glossgroup> element and using the <dita> element.

Sometimes it is desirable to group glossary entries with convenient titles, such as by letter of the alphabet or by sub-domain. For this purpose, the <glossgroup> element is provided. It contains:

- one <title>
- zero or one <prolog>
- zero or more <glossgroup> or <glossentry> topics

The following example shows the method using the <glossgroup> element.

```
<glossgroup>
<title>Glossary entry file containing multiple entries</title>
<glossentry id="abs">
<glossterm>Anti-lock Braking System</glossterm>
...
</glossentry>
<glossentry id="wms">
<glossterm>Weapons of Mass Destruction</glossterm>
...
</glossentry>
...
</glossentry>
...
</glossgroup>
```

The following example shows the method using the ditabasefor assembling multiple topics:

```
<dita>
<glossentry id="abs">
<glossterm>anti-lock braking system</glossterm>
...
</glossentry>
<glossentry d="wms">
<glossterm>weapons of mass destruction</glossterm>
...
</glossentry>
...
</dita>
```

5.2 Referencing glossary entries

Using the keyref mechanism, references to glossary entry files are created for the purpose of inserting terms in text, associating terms used in the text with their glossary definitions, creating a glossary deliverable, recording information for non-publishing purposes (such as for a terminology database,) and so forth.

5.2.1 Declaring a key for a glossary entry topic

To declare a key for a glossary entry topic, use the <glossref> element. The <glossref> element requires the keys and href attributes and defaults the print, search, and toc attributes to the "no" value and the linking attribute to the "none" value. The <glossref> element is always empty.

```
<map>
...
  <topicref href="printers.dita"/>
...
  <glossref keys="rasterp" href="rasterp.dita"/>
... key declarations for other referenced terms ...
</map>
```

Note that <glossref> defines the key without actually inserting the glossary entry topic at the position within the map. If the glossary entry topic should actually appear at that position, use an ordinary <topicref> with the keys attribute.

A best practice is to create a map with a list of <glossref> elements and then use a <mapref> element to pull this term map into any deliverable requiring the glossary terms.

NOTE: The keyref value does not need to match the term itself. In fact, using a more specific value for the keyref will reduce conflicts in situations where the term may occur in more than one glossary entry. For example, for the acronym ABS, an information set could use "cars.abs"

as the key for Anti-lock Braking System and "ship.abs" to refer to the American Bureau of Shipping.

5.2.2 Referencing a glossary entry from the text

You can use the <term> element with a keyref attribute that points to the declared key of the glossary entry to insert a term into the text. For example:

```
A digitized <term keyref="rasterp"/> is an array of bits.
```

If the term you are referencing has an abbreviated form, you can use the <abbreviated-form> element with a keyref attribute that points to the id of the <glossentry> in order to insert the abbreviation into the text. For example:

```
The <abbreviated-form keyref="abs"/> will prevent the car from skid-ding.
```

Whether the abbreviation itself, the surface form, or the expanded form is inserted into the text depends on certain contextual conditions. See the section "Processing considerations" for more information.

Also, instead of using an empty <term../> element as shown above, you can supply a local surface form, as in the following example:

```
The <term keyref="abs">anti-lock braking system</term> will prevent the car from skidding.
```

The keyref pointer in the above example allows the writer to associate the term enclosed in the <term> tag with a definition documented in the associated glossary entry. Processors can then display the definition as hover text in online deliverables. Because the term is inserted by the writer in the local file as the content of the tag, it can take an entirely different surface form, if desired, than the terms formally documented in the glossary entry.

5.2.3 Creating a glossary deliverable

The following example shows how to use a <map> element with <topicref> elements pointing to individual glossary entry DITA files, to assemble a glossary deliverable. Manually arranging a glossary with a <map> element is standard for assembling deliverables but required only when a processor is not available that can automatically sort and group the glossary entries referenced in topics.

Alternatively, you can create a glossary deliverable by using the <glossgroup> element as described in section 5.1.

You can also combine <topicref> elements with <glossref> elements, as shown in the following example. Note that the topicref both defines the key and inserts the topic while the glossref only defines the key.

5.3 Examples of glossary entries

5.3.1 A single defined term

The following is a basic glossary entry for a term and its definition.

```
<glossentry id="rasterpattern">
<glossterm>raster pattern</glossterm>
<glossdef>A series of picture elements (pels) arranged in scan lines to form an image.
</glossdef>
</glossdef>
</glossentry>
```

5.3.2 Preferred and alternate forms

The following example shows a detailed glossary entry specifying the usage for the preferred and alternate terms:

```
<glossentry id="usbfd">
    <glossterm>USB flash drive</glossterm>
    <glossdef>A small portable drive.</glossdef>
    <glossBody>
        <glossPartOfSpeech value="noun"/>
            <glossUsage>Do not provide in upper case (as in "USB Flash Drive") because that suggests a trademark.</glossUsage>
            <glossAlt>
```

```
<glossAcronym>UFD</glossAcronym>
      <glossUsage>Explain the acronym on first occurrence.</glossUsage>
    </glossAlt>
    <glossAlt id="memoryStick">
      <glossSynonym>memory stick</glossSynonym>
      <qlossUsage>This is a colloquial term.
    </alossAlt>
    <glossAlt>
      <glossAbbreviation>stick</glossAbbreviation>
      <glossStatus value="prohibited"/>
      <glossUsage>This is too colloquial.</glossUsage>
      <glossAlternateFor href="#memoryStick"/>
    </glossAlt>
    <glossAlt>
      <glossAbbreviation>flash</glossAbbreviation>
      <glossStatus value="prohibited"/>
      <glossUsage>This short form is ambiguous.</glossUsage>
    </glossAlt>
 </glossBody>
</glossentry>
```

5.3.3 Full forms, acronyms, and surface forms

The following example shows a full form, an acronym, and a surface form:

The surface form may differ, and an acronym may not even exist in other languages. For information about how to handle abbreviations in glossary entries for translated languages, see the *Best Practice for Managing Acronyms and Abbreviations in DITA*.

5.3.4 An entry for a published glossary

For traditional glossary publishing, you can explain one sense of a term with a glossary entry similar to the following example:

```
<glossentry id="abs">
  <glossterm>Anti-lock Braking System</glossterm>
  <glossdef>A brake technology that minimizes skids.</glossdef>
  <glossBody>
    <glossSurfaceForm>Anti-lock Braking System (ABS)</glossSurfaceForm>
  <glossAlt>
```

```
<glossAcronym>ABS</glossAcronym>
    </glossAlt>
</glossentry>
```

5.3.5 An entry destined for populating a terminology database

Published glossaries provide a practical source for terminology to populate terminology databases.

If you are interested in acronym resolution, glossary publishing, and populating a terminology database, you can create a glossary entry similar to the following:

```
<glossentry id="abs">
 <glossterm>Anti-lock Braking System
 <glossdef>A brake technology that minimizes skids./glossdef>
 <qlossBody>
    <glossPartOfSpeech value="noun"/>
    <qlossSurfaceForm>Anti-lock Braking System (ABS)/qlossSurfaceForm>
    <qlossAlt>
     <glossAcronym>ABS</glossAcronym>
      <qlossStatus value="preferred"/>
      <glossUsage>Recommended because more readers are familiar with the
acronym than the term.</glossUsage>
   </glossAlt>
    <qlossAlt>
     <glossSynonym>Anti-skid Brakes/glossSynonym>
      <qlossStatus value="restricted"/>
      <glossUsage>Allowed in legacy content but not in new
content.</glossUsage>
   </glossAlt>
  </glossBody>
</glossentry>
```

Note that the above example also demonstrates the relatively rare case when an abbreviated form is preferred instead of the full form. This preference may occur when the abbreviated form is more recognizable to the target audience than the full form.

5.3.6 An entry with a cross reference

The id attribute on the elements used for terms and variants allows you to create cross references between terms, using the <related-links> element.

The following example shows the syntax when the glossary entries are in the same DITA file:

```
<glossentry id="table">
    <glossterm>table</glossterm>
    <glossdef>In a relational database, a database object that consists of a specific number of columns and is used to store an unordered set of rows.</glossdef>
```

If the glossary entries are in different DITA files, use the following syntax:

```
<link href="db2gloss.dita#table"/>
```

You can also use a keyref to identify the target of the link if the target has a keys definition in a map that refers to both topics:

```
<link keyref="rasterp"/>
```

6. Processing considerations

6.1 Referring to a glossary entry from the text

When a writer uses the <abbreviated-form> element with a keyref attribute to refer to a glossentry topic from the text, the processor should perform the following checks in the attempt to find an abbreviated or surface form with text for the reference, skipping all subsequent checks once the text has been found:

- 1. If the context for the term reference requires the surface form (for instance, it is the first occurrence of the term, or it is in a tooltip) and the glossary entry contains a <glossSurfaceForm> element that contains text, insert the surface form.
- 2. If the context doesn't require the surface form and the glossary entry has at least one <glossAcronym> or <glossAbbreviation> alternate form that contains text and does not have a <glossStatus> of prohibited or obsolete, insert the abbreviated form. (If the glossary entry has multiple abbreviated forms that qualify and one has a <glossStatus> of preferred, insert that abbreviated form; otherwise, insert the first qualified abbreviated form.)
- 3. Otherwise, insert the <glossterm> text for the term.

Writers can also use the <term> element with a keyref attribute to refer to a glossy entry. Processing inserts text from the glossary entry topic only when the referencing <term> element does not contain text. As a result, writers can use the <term> element to delimit terms within content while identifying the corresponding glossary entry. That is, the <term> element can provide a context-specific surface form as its content where appropriate.

6.2 Publishing a glossary

To simulate traditional published glossaries, the publishing process should create published entries according to the particular style of the implementer. For instance, a published glossary may have the following properties:

- Each term including synonyms and variants occurs as a headword in its own entry
- If a term (glossterm) has multiple senses (i.e., it occurs in multiple <glossentry> elements), the definition from each glossentry should be assembled below the headword. Usually they are separated by bullets or numbers.
- The definition appears only with the preferred term
- All the non-preferred terms have a "see" reference to the entry of the preferred term
- All other data from glossary entries, such as usage notes, part of speech values, status values, and so forth, are omitted.

Note that organizing glossary entries in separate DITA files promotes flexibility in the final publishing styles, as mentioned in the section 5.1.1 *Each glossary entry in a separate DITA file*.

The following examples show the format of several entries from a PDF glossary from the IBM Corporation. These examples show the formatting styles of the following types of glossary data:

- An entry for an abbreviation, with a reference to its full form.
- An entry for a full form, showing the abbreviation in parentheses and a definition.
- An entry for a term that has two senses.
- Cross references to related terms.

ASYNC. See asynchronous.

asynchronous (ASYNC). Pertaining to events that are not synchronized in time or do not occur in regular or predictable time intervals.

client. (1) A system or process that is dependent on another system or process (usually called the server) to provide it with access to data, services, programs, or resources. See also server, host. (2) A classifier that requests a service from another classifier. See also association.