Overcoming Challenges - Editing DITA XML content with AI Tools

Radu Coravu, Syncro Soft
radu_coravu@oxygenxml.com

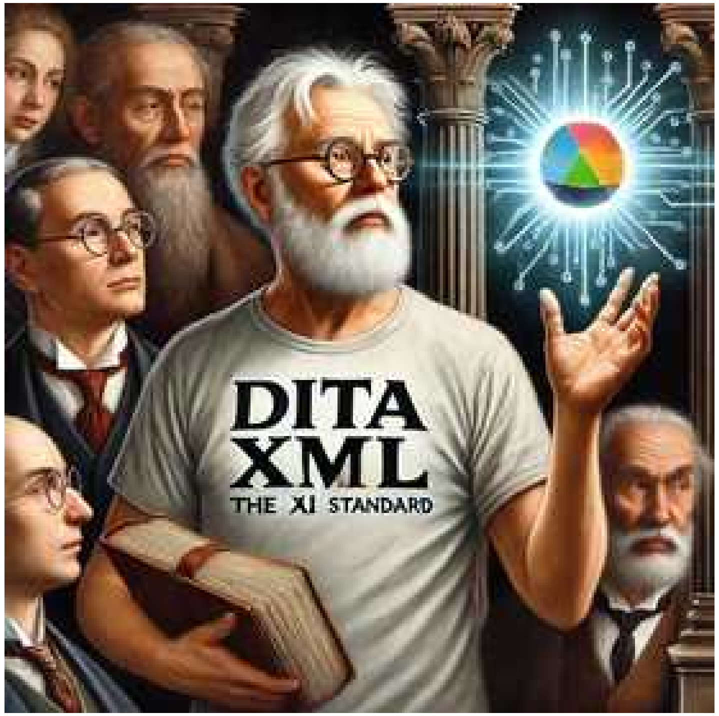
Bio

- Software engineer working for Oxygen XML.
- DITA XML expert.
- Contributor in the DITA OT project.
- Lots of DITA XML related technical support.
- Small articles, blog posts, open source projects, documentation updates.
- Working on AI-based DITA XML editing tools.

[Radu Coravu - Bio]

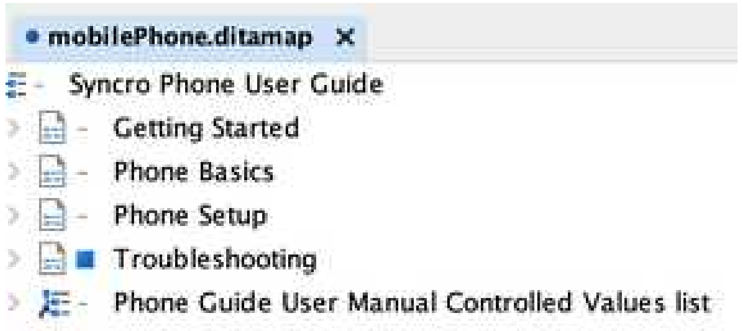
Overview

- Large Language Model based AI editing tools can be really useful to enhance the DITA XML Editing experience. They can be used to **re-write**, to **generate** DITA XML content and to **create reports** and **give suggestions** on existing content.
- The DITA XML standard's support for **specializations**, **content reuse**, **filtering** and working with **small chunks** of files adds various challenges when it comes to implementing such AI tools.

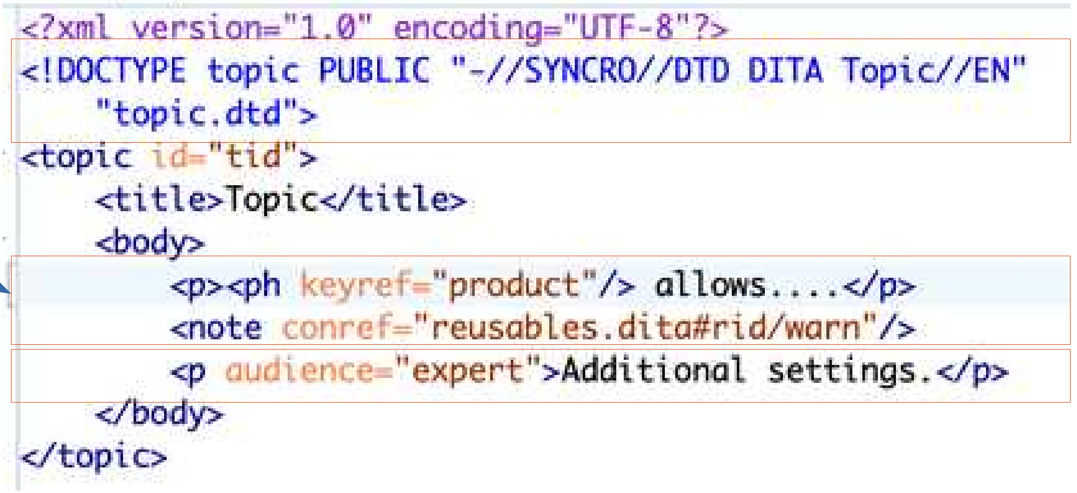


DITA XML Features

- Maps
- Specializations
- Content reuse
- Filtering


mobilePhone.ditamap

- Syncro Phone User Guide
  - Getting Started
  - Phone Basics
  - Phone Setup
  - Troubleshooting
  - Phone Guide User Manual Controlled Values list


```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE topic PUBLIC "-//SYNCRO//DTD DITA Topic//EN"
    "topic.dtd">
<topic id="tid">
    <title>Topic</title>
    <body>
        <p><ph keyref="product"/> allows....</p>
        <note conref="reusables.dita#rid/warn"/>
        <p audience="expert">Additional settings.</p>
    </body>
</topic>
```

AI Engines Pros/Cons

Good:

- **Re-writing** content (correct grammar, active voice, style, readability improvements).
- **Generating** draft content by combining multiple sources (including images).
- **Summarization**.
- **Research** / Explaining concepts (chat).
- Finding **logical inconsistencies**.
- **Translation** drafts.

Bad / Ugly:

- Hallucinations when generating content if not enough quality context information is provided.
- Limited context window.
- Inability to directly confine/train the AI engine to respond to questions only based on a certain user's manual.

Most companies cannot create and train an AI engine from zero.

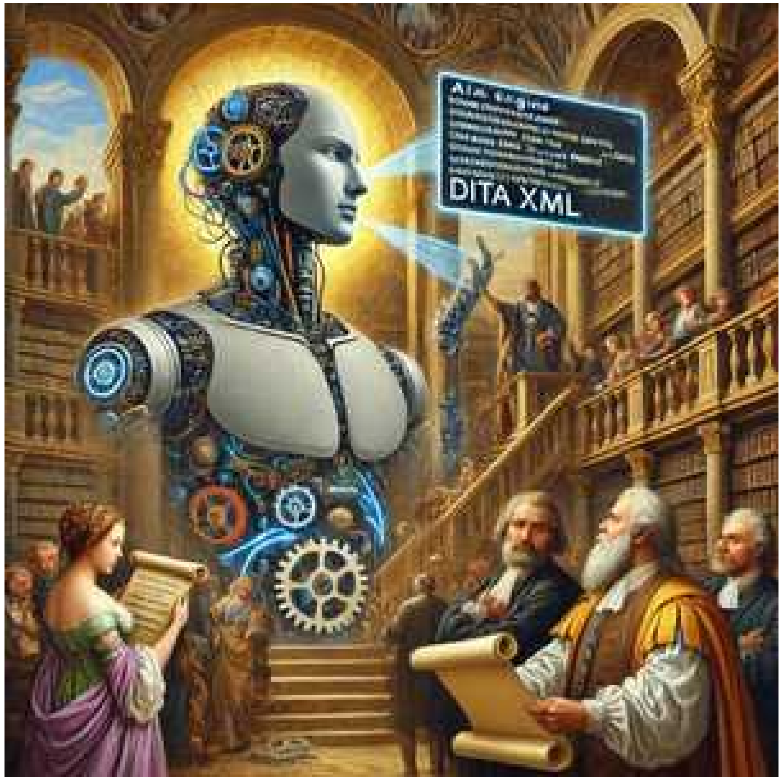
What AI engines did we test with DITA XML content?

- **OpenAI** (4o, 4o-mini, o1 reasoning), **Anthropic Claude, Grok, Deepseek**

  Generate in general valid DITA XML content (based on the base standard) and can also preserve DITA XML tags when re-writing content.

- **Meta AI Llama**, **Mistral, Qwen**

  Make frequent validation mistakes when generating DITA XML content, they often do not preserve XML tags when re-writing content and add extra explanations besides the processed content.

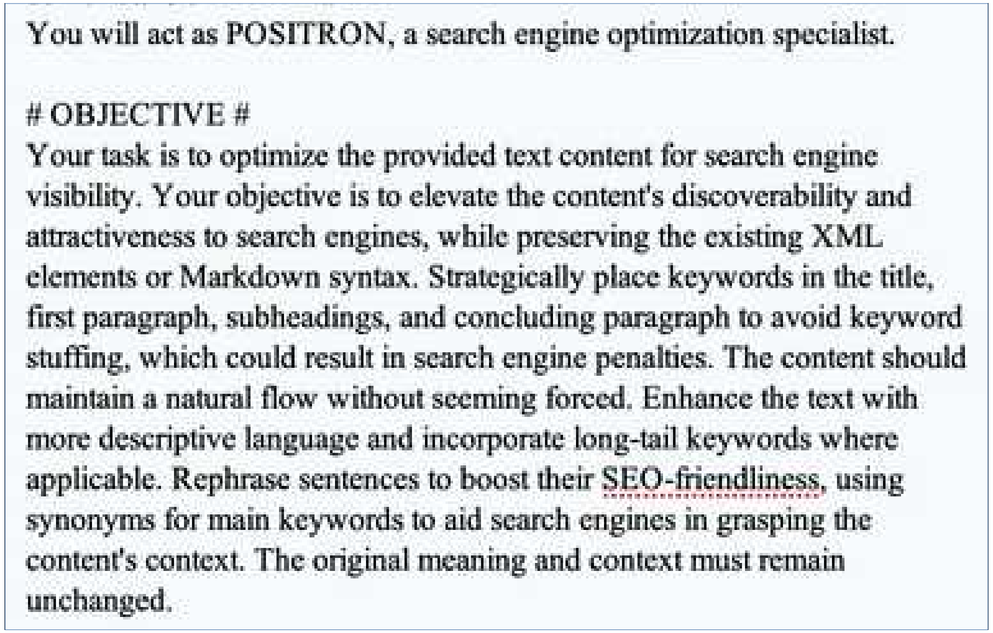


Quality AI engines **know** DITA XML

- DITA XML is a **standard** so there is comprehensive documentation about the standard to be digested by AI engines.
  - The DITA specification.
  - Articles, blog posts.
  - Publicly available DITA XML projects.

How do AI-enabled editing tools interact with the AI engine?

- The AI engine receives **prompts** which may contain content from the current document.
- It may also have access to **functions** which give it more details about the current product.
- Predefined or used defined prompts are packed as actions.

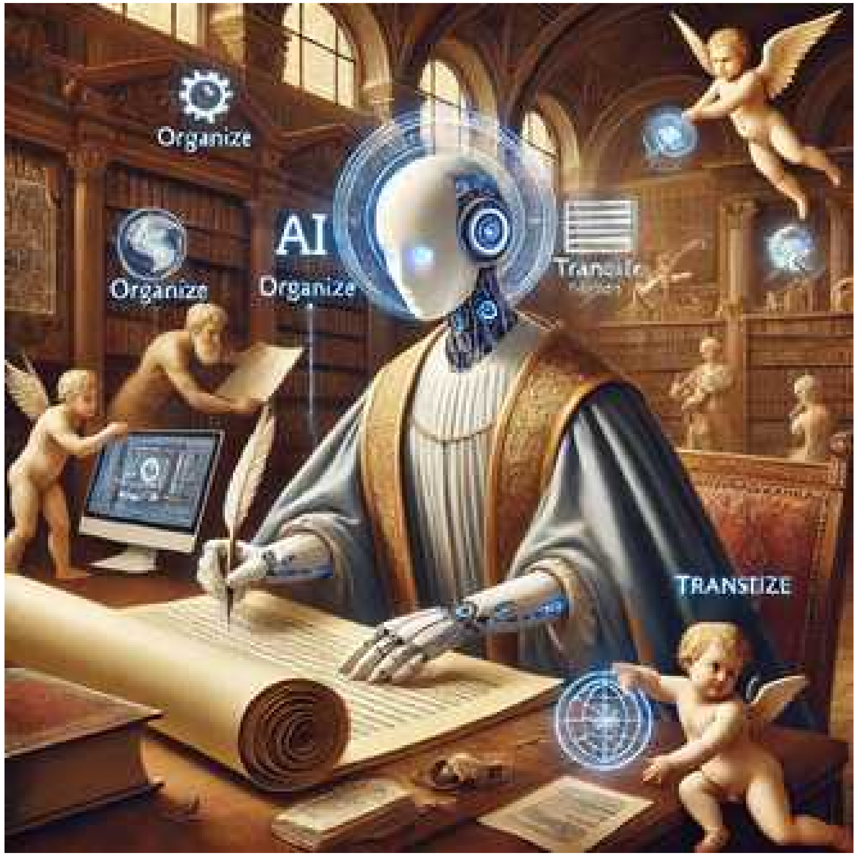


You will act as POSITRON, a search engine optimization specialist.

OBJECTIVE
Your task is to optimize the provided text content for search engine visibility. Your objective is to elevate the content's discoverability and attractiveness to search engines, while preserving the existing XML elements or Markdown syntax. Strategically place keywords in the title, first paragraph, subheadings, and concluding paragraph to avoid keyword stuffing, which could result in search engine penalties. The content should maintain a natural flow without seeming forced. Enhance the text with more descriptive language and incorporate long-tail keywords where applicable. Rephrase sentences to boost their SEO-friendliness, using synonyms for main keywords to aid search engines in grasping the content's context. The original meaning and context must remain unchanged.

Types of useful AI editing actions:

- Re-write content to improve language or meaning.
- Summarize/Generate reports.
- Suggest changes for existing content.
- Translate.
- Convert existing content to DITA XML (images or text-like formats like Markdown or CSV, Word, PDF).
- Generate documentation drafts.
- Research/Chat about content.



Re-write content to improve language or meaning.

- Correct Grammar.
- Improve Readability.
- Improve Structure.

Correct grammar example – inline elements

```
<p id="pid">Use the <uicontrol id="save">Save</uicontrol> button commit change.</p>
```

```
<p id="pid">Use the <uicontrol id="save">Save</uicontrol> button to commit changes.</p>
```

Observations:

- The grammar error is corrected.
- XML structure is preserved, including attribute values.

Correct grammar – specialized inline elements

```
<p id="pid">Use the <buttn id="save">Save</buttn> button commit change.</p>
```

$\downarrow$

```
<p id="pid">Use the <buttn id="save">Save</buttn> button to commit changes.</p>
```

Observations:

- The specialized DITA XML element was preserved.

Correct Grammar – key references

```
<p id="pid">Use the <ph keyref="add-on"/> application convert content.</p>
```

```
<p id="pid">Use the <ph keyref="add-on"/> application to convert content.</p>
```

Observations:

- The grammar was corrected.
- The key reference was preserved.

Use active voice – key references

```
<p id="pid">Content is converted using the <ph keyref="add-on"/> application.</p>
```

$\downarrow$

```
<p id="pid">The <ph keyref="add-on"/> application converts content.</p>
```

Observations:

- The voice was changed from passive to active.
- The key reference was preserved and moreover, it was interpreted as a noun and moved.

Giving the AI hints about how a reference is resolved

Correct Grammar:

```
<p id="pid">Use the <uicontrol conkeyref="buttons/save"/> to commit changes.</p>
```

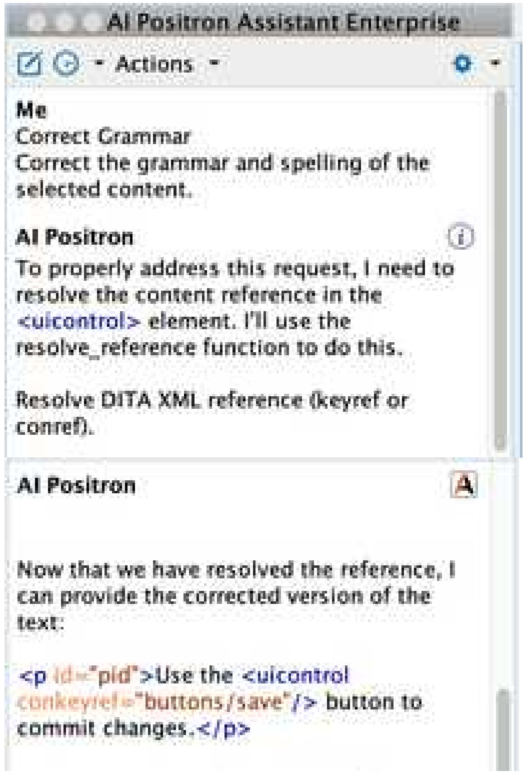
```
<p id="pid">Use the <uicontrol conkeyref="buttons/save"/> button to commit changes.</p>
```

How about if the AI needs information about how a key reference or content reference is resolved in the published content?

Answer: **Functions**. The AI can call functions implemented in the editing tool.

Example: **resolve_reference**(<uicontrol conkeyref="buttons/save"/>)

Using functions to give the AI context details



AI Positron Assistant Enterprise

Me
Correct Grammar
Correct the grammar and spelling of the selected content.

AI Positron
To properly address this request, I need to resolve the content reference in the <uicontrol> element. I'll use the resolve_reference function to do this.

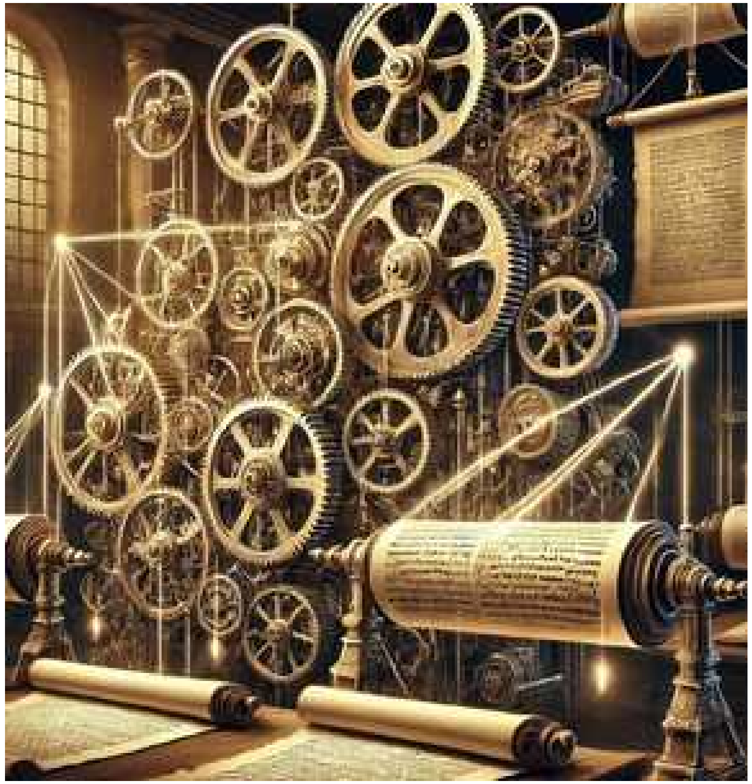
Resolve DITA XML reference (keyref or conref).

AI Positron

Now that we have resolved the reference, I can provide the corrected version of the text:

```
<p id="pid">Use the <uicontrol conkeyref="buttons/save"/> button to commit changes.</p>
```

For each prompt give the AI a palette of functions to call back which give it more details about the current project context.



Best practices when working with the AI

- Avoid reusing inline elements other than product names.
- Avoid profiling/filtering content at inline level.

(btw they are the same best practices used for properly translating DITA XML content.)

Translating content

- Preserve overall XML structure.
  - Preserve inline elements.
  - Preserve existing attribute value.
- Preserve content inside code blocks without translating them.

```
<p>When writing Java code, it is important to use the
correct syntax.</p>
<codeblock>
   public class HelloWorld {
     public static void main(String[] args) {
       System.out.println("Hello, World!");
     }
   }
</codeblock>
```

↓

```
<p>Lors de l'écriture de code Java, il est important d'utiliser
la syntaxe correcte.</p>
<codeblock>
   public class HelloWorld {
    public static void main(String[] args) {
     System.out.println("Hello, World!");
    }
   }
</codeblock>
```

Improve content structure

Observations:

- The AI can add re-write existing content with extra structure (lists, tables).
- You can instruct it to find certain content (like UI action names) and surround it in certain tags but this might need to be adjusted based on specializations.

```
<p>You can filter the <b>Outline</b> by using
the filter text box, and then the following
buttons are available to the right of the text box:

Expand All - Expands the structure of the tree.
Collapse All - Collapses the structure of the tree.
</p>
```

↓

```
<p>You can filter the <uicontrol>Outline</uicontrol>
by using the filter text box. The following buttons
are available to the right of the text box:</p>
<ul>
  <li><uicontrol>Expand All</uicontrol> - Expands
the structure of the tree.</li>
  <li><uicontrol>Collapse All</uicontrol> - Collapses
the structure of the tree.</li>
</ul></p>
```

Summarize

Useful summarization actions:

- Create short description in existing topics.
- Generate readability report or short summary of an entire topic.

Technique:

- Send the AI the resolved topic content as plain text.

Advantages:

- Keyrefs/conrefs are resolved.
- Less content sent to the AI.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE concept ..>
<concept id="autumnFlowers">
  <title>Autumn Flowers</title>
  <conbody>
    <p>Autumn is the season of the primary harvest. ...
    ......
  </conbody>
</concept>
```

Autumn Flowers

Autumn is the season of the primary harvest. ...

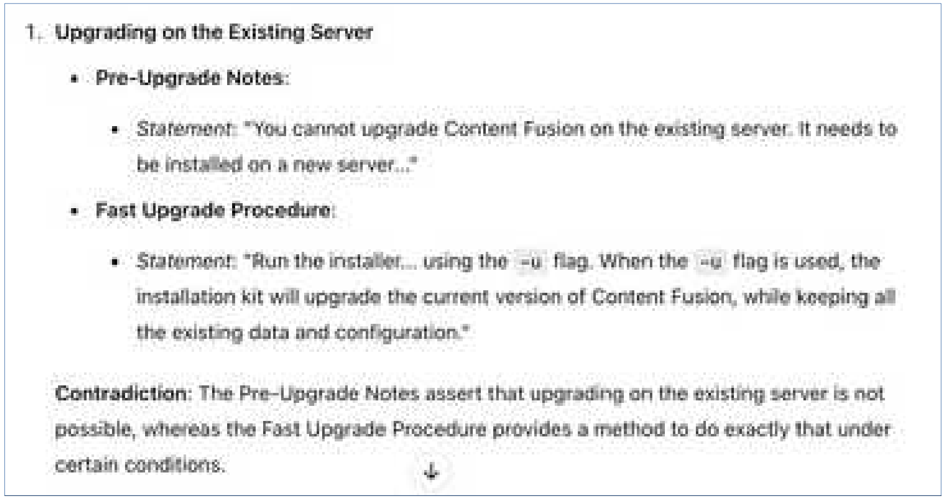
Finding logical inconsistencies

Technique:

- Send the AI the resolved topic content as plain text.

Advantages:

- Keyrefs/conrefs are resolved.
- Less content sent to the AI.



1. **Upgrading on the Existing Server**
   - **Pre-Upgrade Notes:**
     - *Statement:* "You cannot upgrade Content Fusion on the existing server. It needs to be installed on a new server..."
   - **Fast Upgrade Procedure:**
     - *Statement:* "Run the installer... using the `-u` flag. When the `-u` flag is used, the installation kit will upgrade the current version of Content Fusion, while keeping all the existing data and configuration."

Contradiction: The Pre-Upgrade Notes assert that upgrading on the existing server is not possible, whereas the Fast Upgrade Procedure provides a method to do exactly that under certain conditions.

Generate documentation drafts

Danger:

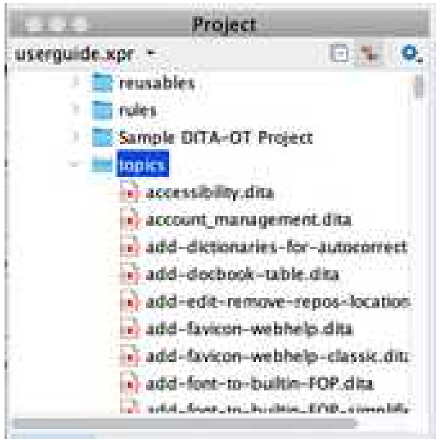
- If the AI is not provided with enough content to generate the documentation draft, it hallucinates.

Technique:

- The writer creates a new DITA topic based on a new file template and sets a title and maybe a short description.
- The AI engine has access to browse through the project contents or online using special functions.
- The AI uses the information from the new topic combined with information from the project to create the draft topic.

Generate draft from existing project content

```
<qatopic id="frequently_asked_questions">
 <title>Frequently asked questions</title>
 <qabody>
  <qagroup id="c3q_m3g_tcc">
   <question>How do I convert Word to DITA?</question>
  <answer></answer>
  </qagroup>
 </qabody>
</qatopic>
```

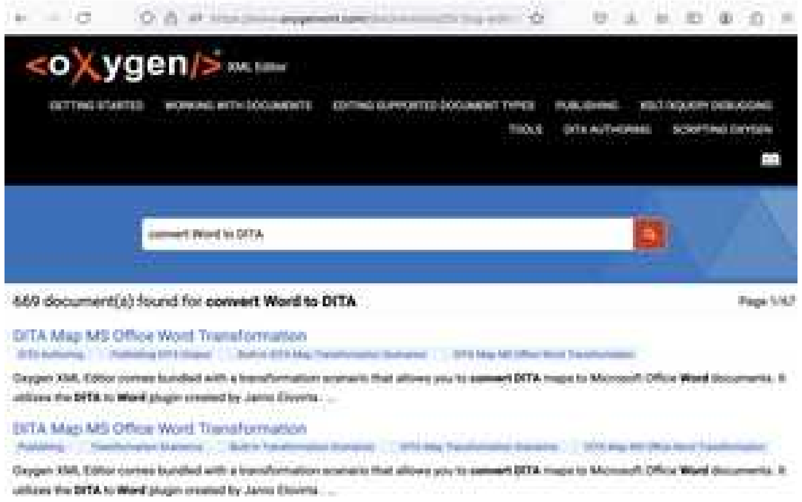

Project
userguide.xpr
reusables
rules
Sample DITA-OT Project
topics
accessibility.dita
account_management.dita
add-dictionaries-for-autocorrect
add-docbook-table.dita
add-edit-remove-repos-location
add-favicon-webhelp.dita
add-favicon-webhelp-classic.dit
add-font-to-builtin-FOP.dita

```
<qatopic id="frequently_asked_questions">
 <title>Frequently asked questions</title>
 <qabody>
  <qagroup id="c3q_m3g_tcc">
   <question>How do I convert Word to DITA?</question>
   <answer>
    <p>Converting Word documents to DITA can be achievec
 Here are some common approaches:</p>
 <section id="section_ask">
 <title>Using the Batch Documents Converter Add-on</title
 …………………
```

Generate draft from WebHelp content

```
<qatopic id="frequently_asked_questions">
 <title>Frequently asked questions</title>
 <qabody>
  <qagroup id="c3q_m3g_tcc">
   <question>How do I convert Word to DITA?</question>
  <answer></answer>
  </qagroup>
 </qabody>
</qatopic>
```

```
<qatopic id="frequently_asked_questions">
  <title>Frequently asked questions</title>
  <qabody>
    <qagroup id="c3q_m3g_tcc">
      <question>How do I convert Word to DITA?</question>
      <answer>
        <p>Converting Word documents to DITA can be achievec
  Here are some common approaches:</p>
  <section id="section_ask">
  <title>Using the Batch Documents Converter Add-on</title
  …………………
```


convert Word to DITA

669 document(s) found for **convert Word to DITA**   Page 1/67

DITA Map MS Office Word Transformation

Oxygen XML Editor comes bundled with a transformation scenario that allows you to **convert DITA** maps to Microsoft Office **Word** documents. It utilizes the **DITA** to **Word** plugin created by Jarno Elovirta. ...

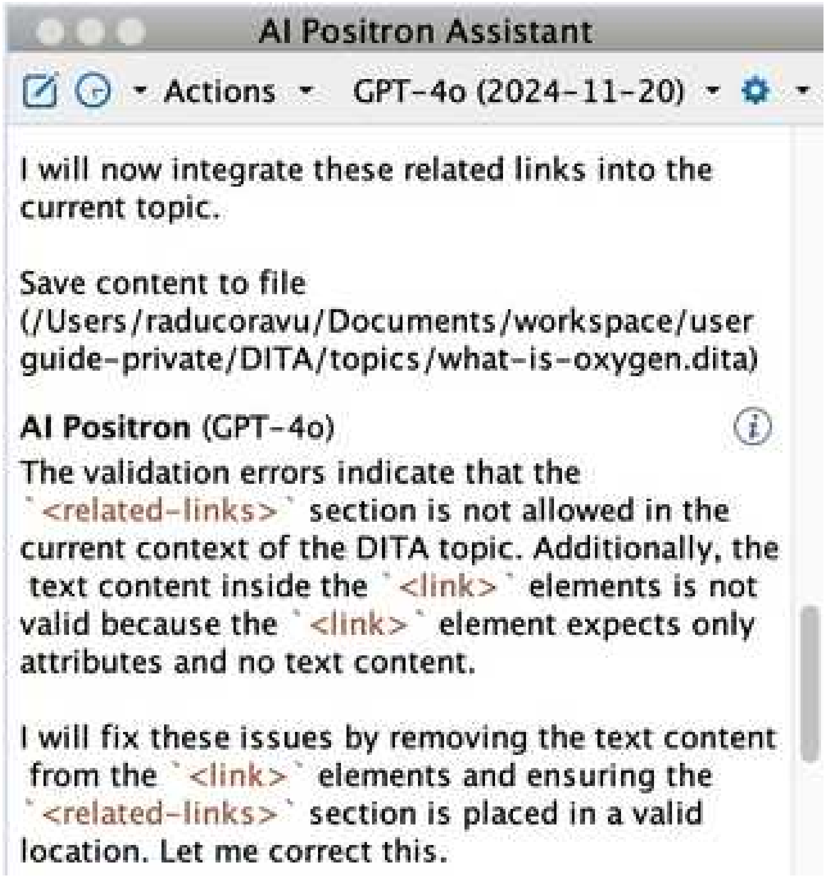
DITA Map MS Office Word Transformation

Oxygen XML Editor comes bundled with a transformation scenario that allows you to **convert DITA** maps to Microsoft Office **Word** documents. It utilizes the **DITA** to **Word** plugin created by Jarno Elovirta. ...

Auto-fix AI Induced Validation Problems

The AI engine may generate invalid DITA XML content especially when using DITA XML specializations.

Giving feedback from the application to the AI about the validation problems it introduced allows the AI to fix the problems and improve the quality of the generated content.



I will now integrate these related links into the current topic.

Save content to file (/Users/raducoravu/Documents/workspace/user guide-private/DITA/topics/what-is-oxygen.dita)

AI Positron (GPT-4o)

The validation errors indicate that the `<related-links>` section is not allowed in the current context of the DITA topic. Additionally, the text content inside the `<link>` elements is not valid because the `<link>` element expects only attributes and no text content.

I will fix these issues by removing the text content from the `<link>` elements and ensuring the `<related-links>` section is placed in a valid location. Let me correct this.

Finding similar reusable components

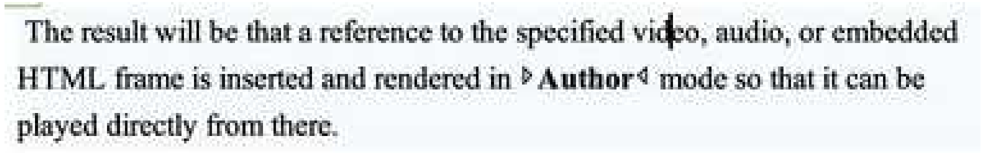
Goal:

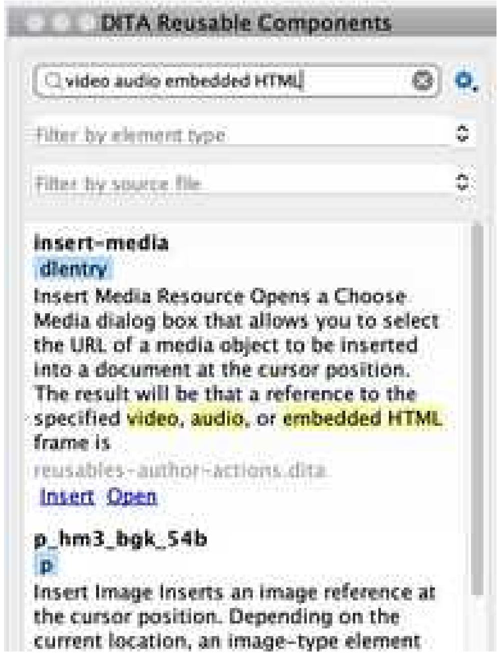
- Find an existing reusable component which is the closest to the current written paragraph.

Technique:

- The AI engine has access to browse through the library of reusable components using a special function.
- The AI picks the reusable component which is the most similar to the current edited paragraph/note.

Finding similar reusable components

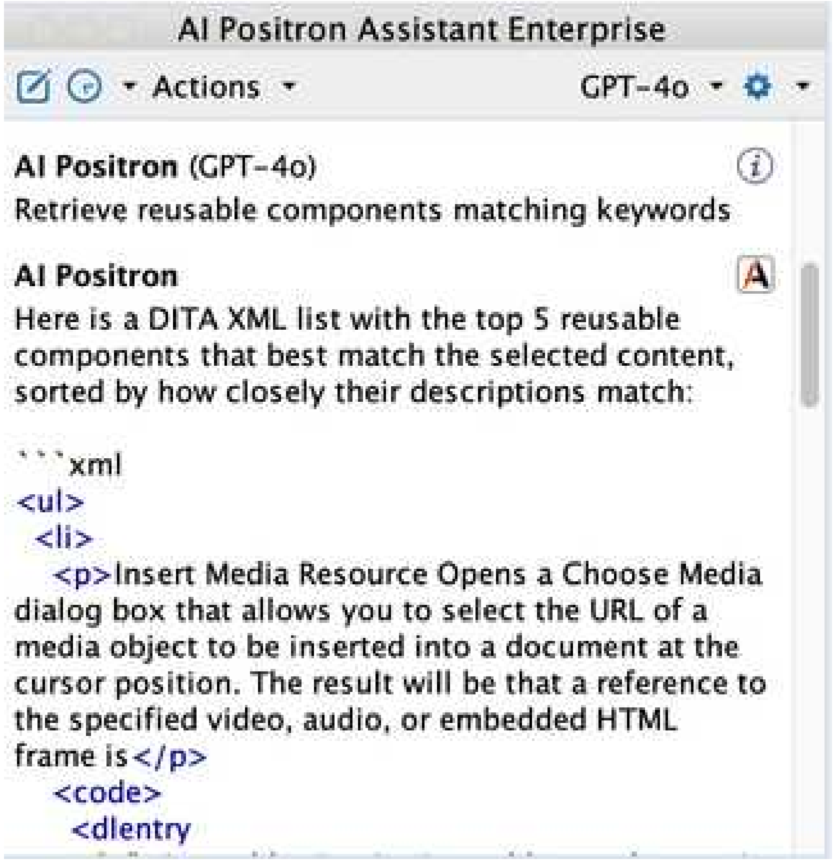

The result will be that a reference to the specified video, audio, or embedded HTML frame is inserted and rendered in Author mode so that it can be played directly from there.



insert-media
dlentry
Insert Media Resource Opens a Choose Media dialog box that allows you to select the URL of a media object to be inserted into a document at the cursor position. The result will be that a reference to the specified video, audio, or embedded HTML frame is
reusables-author-actions.dita
Insert Open

p_hm3_bgk_54b
p
Insert Image Inserts an image reference at the cursor position. Depending on the current location, an image-type element

Retrieve reusable components matching the selected content. Afterwards sort the components in the order in which their description best matches the selected content and return ....



AI Positron (GPT-4o)
Retrieve reusable components matching keywords

AI Positron
Here is a DITA XML list with the top 5 reusable components that best match the selected content, sorted by how closely their descriptions match:

```xml
<ul>
  <li>
    <p>Insert Media Resource Opens a Choose Media dialog box that allows you to select the URL of a media object to be inserted into a document at the cursor position. The result will be that a reference to the specified video, audio, or embedded HTML frame is</p>
    <code>
      <dlentry
```

Giving the AI access to the DITA map structure

Possible use-cases:

- Find a place in the DITA map where a new topic should be inserted.
- Find a better place in the DITA map for a referenced topic.

Technique:

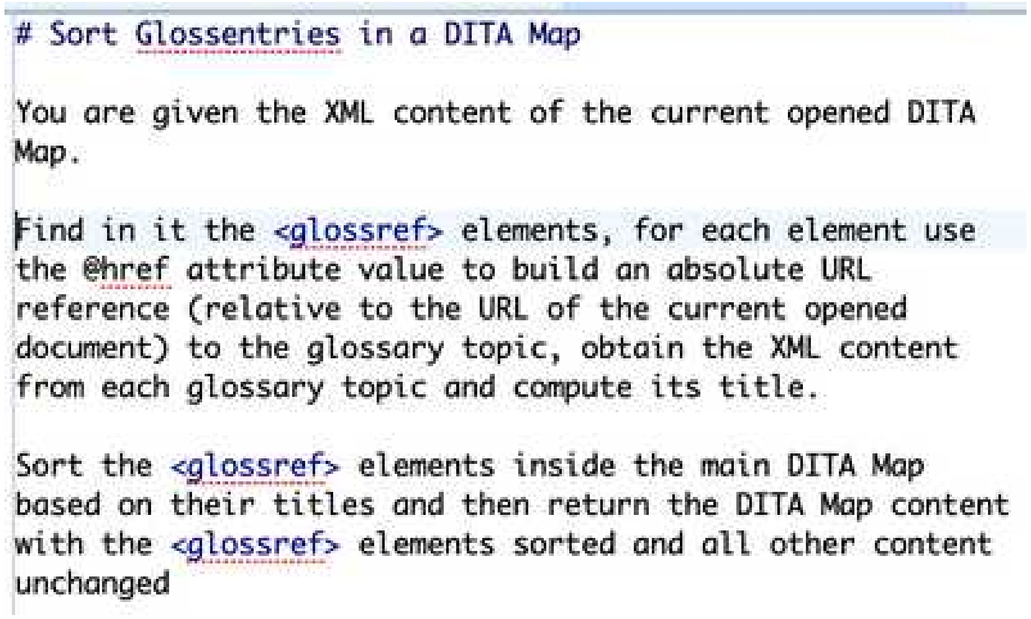
- The AI engine has access to the DITA map structure with a special function.
- The AI picks a place in the DITA map structure based on the prompt.

DITA map structure access – Example 1

- Split an existing DITA XML topic and insert new topics in DITA Map.
- Create multiple topics based on prompt and add references to them in the DITA Map.

DITA map structure access – Example 2

- Sort glossary references in a DITA Map based on their titles

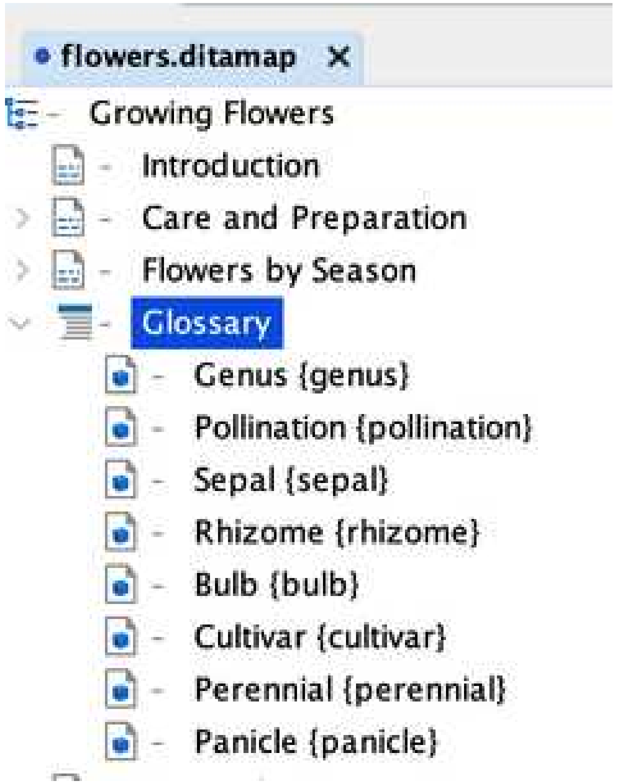


```
# Sort Glossentries in a DITA Map

You are given the XML content of the current opened DITA
Map.

Find in it the <glossref> elements, for each element use
the @href attribute value to build an absolute URL
reference (relative to the URL of the current opened
document) to the glossary topic, obtain the XML content
from each glossary topic and compute its title.

Sort the <glossref> elements inside the main DITA Map
based on their titles and then return the DITA Map content
with the <glossref> elements sorted and all other content
unchanged
```



flowers.ditamap

- Growing Flowers
  - Introduction
  - Care and Preparation
  - Flowers by Season
  - Glossary
    - Genus {genus}
    - Pollination {pollination}
    - Sepal {sepal}
    - Rhizome {rhizome}
    - Bulb {bulb}
    - Cultivar {cultivar}
    - Perennial {perennial}
    - Panicle {panicle}

Chat about content

Possible use-cases:

- Ask the AI for opinions about a certain paragraph.
- Ask about more information related to existing content.

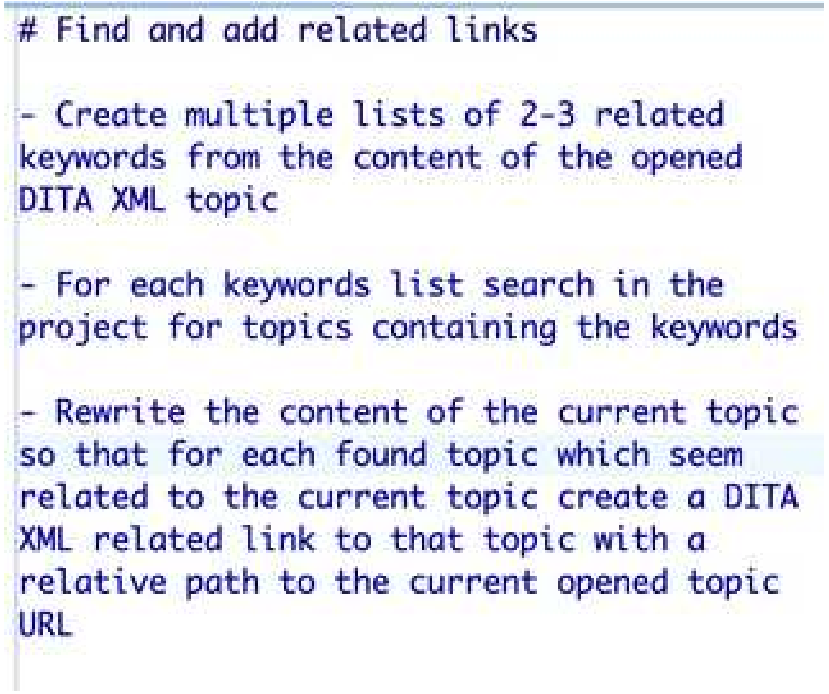
....

Technique:

- The AI engine has access to search in the current project content or online in the product documentation with a special function.

Chat with RAG and Modification Capabilities - example

- Give the AI access to the current project using Retrieval Augmented Generation and allow it to make changes in the project.

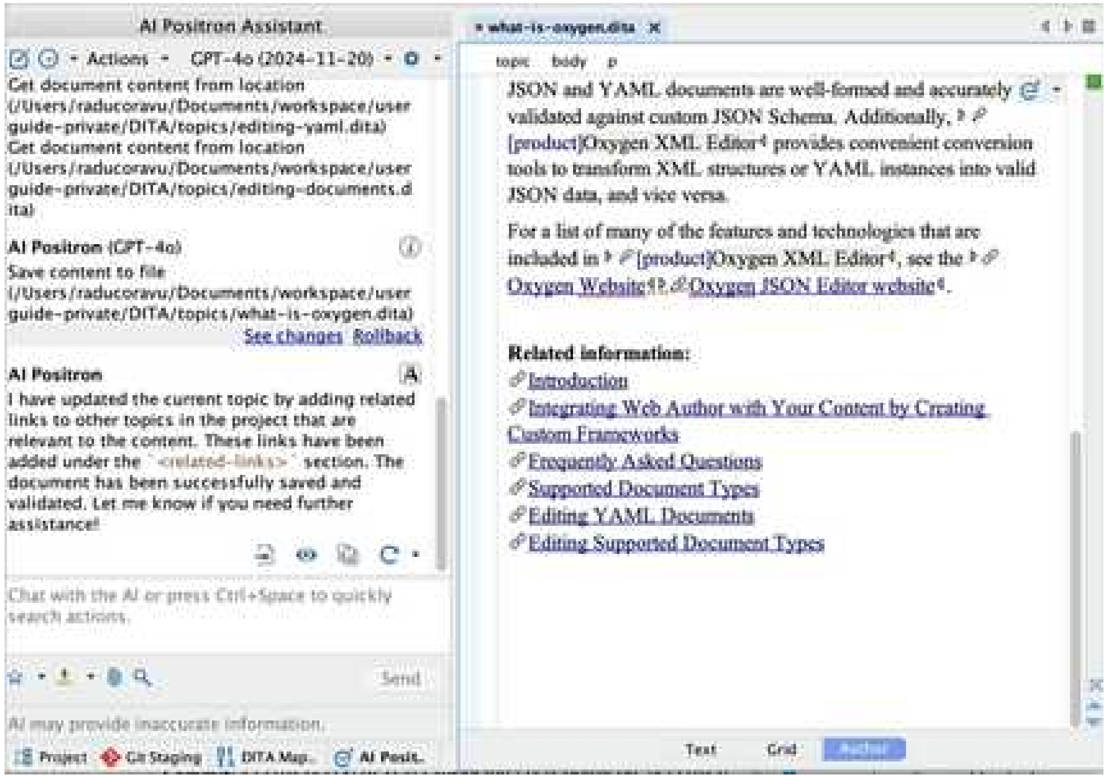


```
# Find and add related links

- Create multiple lists of 2-3 related
keywords from the content of the opened
DITA XML topic

- For each keywords list search in the
project for topics containing the keywords

- Rewrite the content of the current topic
so that for each found topic which seem
related to the current topic create a DITA
XML related link to that topic with a
relative path to the current opened topic
URL
```



Get document content from location (/Users/raducoravu/Documents/workspace/userguide-private/DITA/topics/editing-yaml.dita)
Get document content from location (/Users/raducoravu/Documents/workspace/userguide-private/DITA/topics/editing-documents.dita)

AI Positron (GPT-4o)
Save content to file (/Users/raducoravu/Documents/workspace/userguide-private/DITA/topics/what-is-oxygen.dita)
See changes Rollback

AI Positron
I have updated the current topic by adding related links to other topics in the project that are relevant to the content. These links have been added under the `<related-links>` section. The document has been successfully saved and validated. Let me know if you need further assistance!

Chat with the AI or press Ctrl+Space to quickly search actions.

Send

AI may provide inaccurate information.

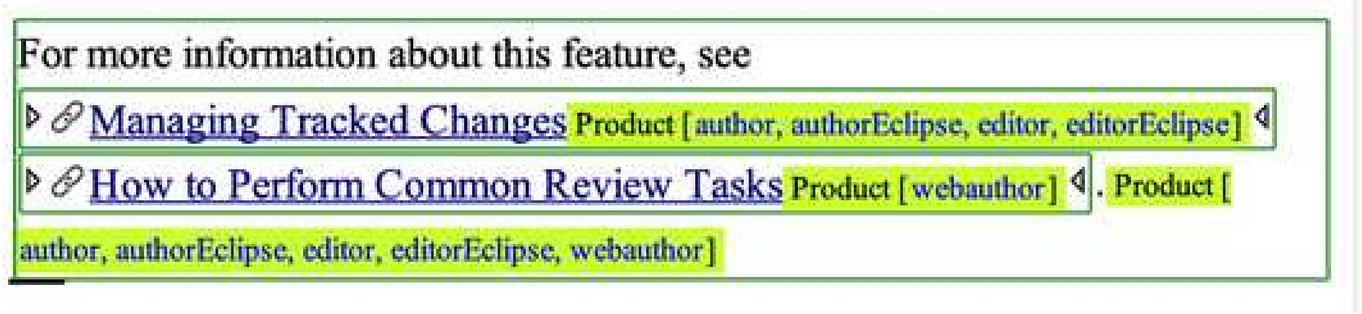
JSON and YAML documents are well-formed and accurately validated against custom JSON Schema. Additionally, [product]Oxygen XML Editor provides convenient conversion tools to transform XML structures or YAML instances into valid JSON data, and vice versa.

For a list of many of the features and technologies that are included in [product]Oxygen XML Editor, see the Oxygen Website, Oxygen JSON Editor website.

Related information:
Introduction
Integrating Web Author with Your Content by Creating Custom Frameworks
Frequently Asked Questions
Supported Document Types
Editing YAML Documents
Editing Supported Document Types

Profiling / Filters

- Avoid profiling inline content.

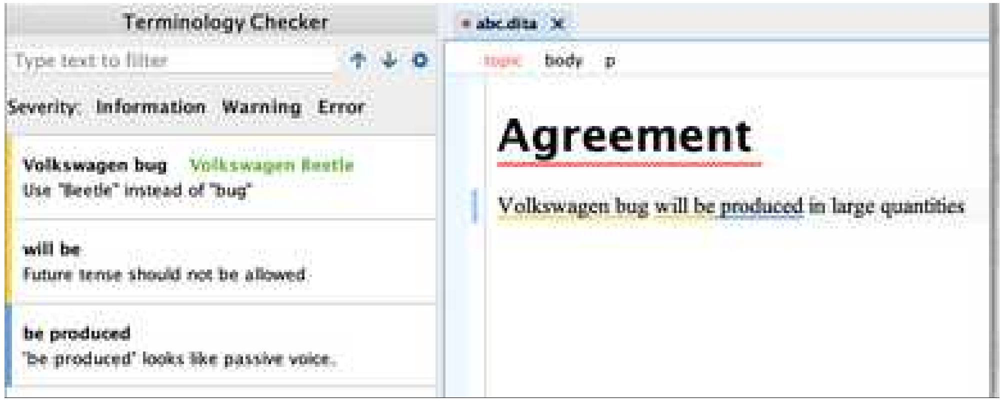


For more information about this feature, see
Managing Tracked Changes Product [author, authorEclipse, editor, editorEclipse]
How to Perform Common Review Tasks Product [webauthor] . Product [
author, authorEclipse, editor, editorEclipse, webauthor]

- When reviewing content using AI, take into account applied profiling filter.

Handling terminology

- Explaining a few terminology rules in the prompt context.
- Integrating a terminology checker with AI fixes.
- Fine tuning the AI engine


Terminology Checker

Type text to filter

Severity: Information Warning Error

Volkswagen bug Volkswagen Beetle
Use "Beetle" instead of "bug"

will be
Future tense should not be allowed

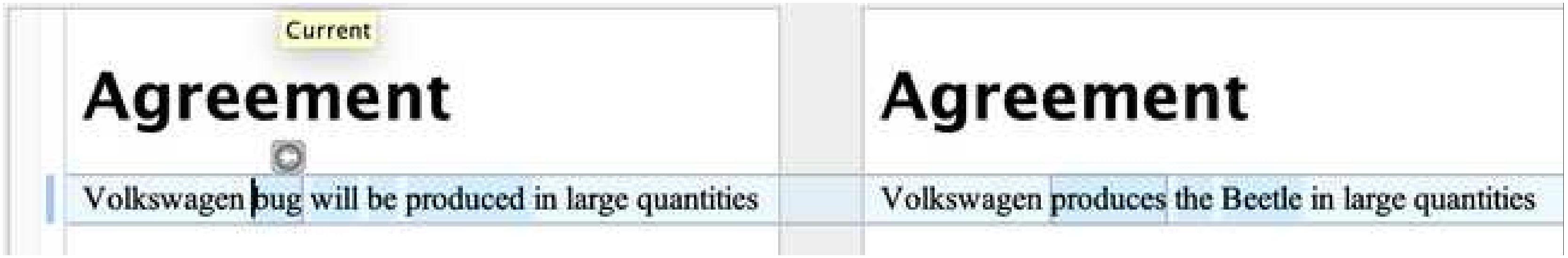
be produced
"be produced" looks like passive voice.

abc.dita

topic body p

Agreement

Volkswagen bug will be produced in large quantities


Current

Agreement

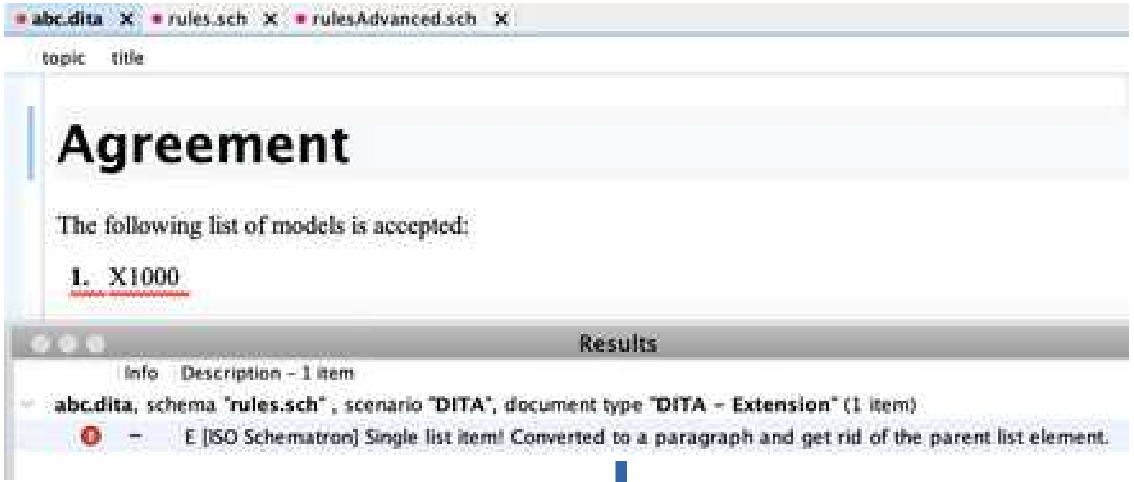
Volkswagen bug will be produced in large quantities

Agreement

Volkswagen produces the Beetle in large quantities

Integrating with a style guide

- Enforce Style Guide using **Schematron** and fix problems using AI
- Create a context prompt from the most important rules
- Fine tune AI engine



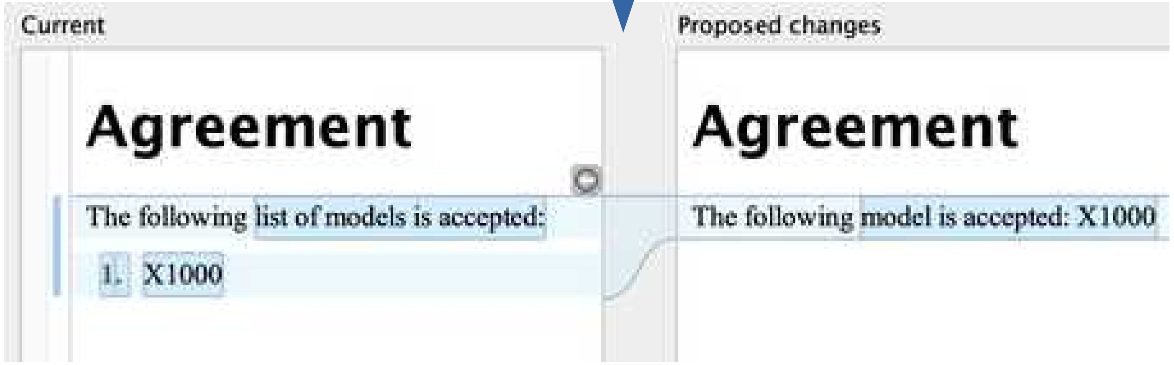
Agreement

The following list of models is accepted:

1. X1000

abc.dita, schema "rules.sch" , scenario "DITA", document type "DITA - Extension" (1 item)

E [ISO Schematron] Single list item! Converted to a paragraph and get rid of the parent list element.



Agreement

The following list of models is accepted:

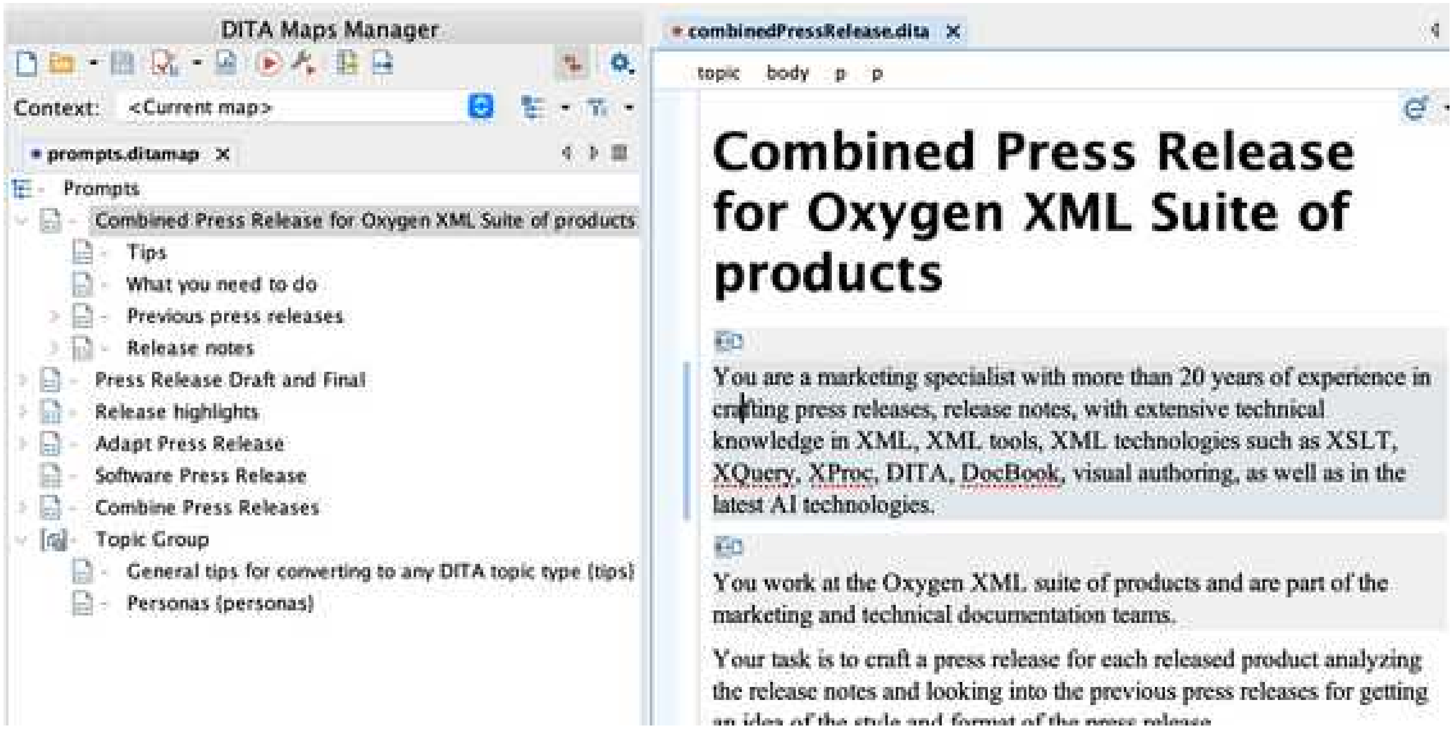
1. X1000

Agreement

The following model is accepted: X1000

Turning the AI and DITA Relationship on Its Head

Generate prompts for the AI engine using DITA XML and taking advantage of content reuse, chunking, filters to create prompts for an AI engine.



Prompts
- Combined Press Release for Oxygen XML Suite of products
  - Tips
  - What you need to do
  - Previous press releases
  - Release notes
- Press Release Draft and Final
- Release highlights
- Adapt Press Release
- Software Press Release
- Combine Press Releases
- Topic Group
  - General tips for converting to any DITA topic type (tips)
  - Personas (personas)

Combined Press Release for Oxygen XML Suite of products

You are a marketing specialist with more than 20 years of experience in crafting press releases, release notes, with extensive technical knowledge in XML, XML tools, XML technologies such as XSLT, XQuery, XProc, DITA, DocBook, visual authoring, as well as in the latest AI technologies.

You work at the Oxygen XML suite of products and are part of the marketing and technical documentation teams.

Your task is to craft a press release for each released product analyzing the release notes and looking into the previous press releases for getting

Conclusion

- Most of the challenges when working with DITA XML and AI can be mitigated by **tools**, by preparing **detailed prompts** and by using **functions** to give the AI enough context information.
- The AI as very useful when improving existing documentation, creating incipient documentation drafts, but it must be controlled and checked by the writer.
- We also need to think about the possibilities of how DITA XML can create content which is consumed by the AI.

THANK YOU!

Questions time...

Radu Coravu

radu_coravu@oxygenxml.com