Importing data in an Open Source Wiki

FOSDEM'21 Online

by Ludovic Dubost, XWiki SAS
Who am I?

Ludovic Dubost
Creator of XWiki & Founder of XWiki SAS

- XWiki & XWiki SAS
- CryptPad
- All Open Source
- 16 years of Open Source business
Why use a Wiki and XWiki?

Improved Knowledge sharing and organization of information

Information is easier to create and organize than in Office files (Microsoft Sharepoint)

XWiki is a prime Open Source alternative to Atlassian Confluence
Why importing?

Even when users want to go for Open Source, the transition process is a key difficulty.

A Wiki works much better with data in it, to help users understand the value of the Wiki concept.
Types of Imports

Sources of data are multiple

- Office files
- A database
- Another wiki (Confluence, MediaWiki, GitHub wiki)
- HTML Source
Available Methods

Manual / Coding

- Manual Wiki syntax conversion & Office file import
- Script using APIs of XWiki
- XWiki Batch Import including Office batch import
- Filter streams & Confluence, MediaWiki, DokuWiki, GitHub
- Wiki Filters

Tools
The Manual Way

Wiki syntax conversion
Office File Importing
Manual Tools

Wiki & HTML Syntax Importing

1. Install and/or activate additional syntaxes in XWiki
2. Copy paste HTML or Wiki Syntax from an outside source

3. Create XWiki document using the source syntax

4. Copy paste the content in it
5. Edit, change to XWiki syntax and accept the conversion
Manual Tools: Wiki & HTML Syntax Importing

CryptPad would not be alive without our financial contributors allowing to fund the development of the CryptPad software as an Open Source project.

2019 Funding

Thanks to the NLNet fundings and our subscribers we have been able to fund the March-December 2019 period which required around 98K funding. The source of this funding is:

- we have received a second 50,000 Euros grant from the European Community through the NLNet NGI Zero PETF fund to improve CryptPad for our Community. We have assigned half this second NLNet project on the 2019 budget.
- The first 50,000 Euros grant was to develop CryptPad Teams to build teams features inside CryptPad and improve the security of document sharing.
- we are receiving donations for CryptPad through the CryptPad Open-Collective or directly on Cryptpad.fr. We have received 4,700.00 Euros from July 2018 to December 2019.
- we are selling subscriptions to the Cryptpad.fr instance. We have currently sold 9,800.00 Euros of subscriptions in two years up until December 2019.

This funding allows to pay for our team of 2 employees with the help also of the French "Credt Impot Recherche" and the remaining funding from the BPI France project "OpenRAAS NG" which ended in March 2019.

2020 Funding

We are happy to have been able to secure our 2020 funding budget in April 2020. We needed 142K funding to be able to secure this budget.

- We have been starting the year with 25K from the second NLNet PETF Zero funding "CryptPad communities"
- We have been happy also to have our project "CryptPad SMC" (Secure Mobile Communication) being accepted by NGI Trust with a 75K funding for this project
- We have also received a funding of 9,1k from the MOSS program from the Mozilla Foundation after our pitch at FOSDEM
- Finally we are now expecting over 30K of revenue from subscriptions to Cryptpad.fr and from donation on OpenCollective, which is higher that what we initially expected.
- A third NLNet PETF Zero project "CryptPad Dialogue" has also been accepted for 50k funding, from which 3k will be assigned in 2020.

This brings our total funding for 2020 to 142k Euros

HTML can also be copy-pasted in the Wysiwyg editor
Uses LibreOffice server to convert to HTML then to XWiki syntax. Possibility to filter styles & split document in multiple pages. Office Import is also possible from the Wysiwyg editor.

See: https://extensions.xwiki.org/xwiki/bin/view/Extension/Office%20Importer%20Application
Manual Tools: Office Import

Uses LibreOffice server to convert to HTML then to XWiki syntax. Possibility to filter styles & split document in multiple pages.
Imports images, tables, footnotes and others elements.
The developer way

Script using APIs
The Developer way

• Many APIs availables to convert HTML or other syntaxes to XWiki

• Java available through Groovy

• Job Macro allows to run long tasks

• Allows to customize fully the import process

• Examples: HCL Connections import for Lenovo, HTML imports with custom cleanup
Batch Importing

Importing databases and automate file uploading and conversion in XWiki
Batch Import: General Principles

- Prepare a data structure in XWiki or use an existing one (Client list, Project list, Countries, Users)
- Prepare an Excel file of the data to import (from a database, by exporting data)
- List attached files in a column and upload the files to the server as a zip file
- Batch import to configure a mapping of the spreadsheet
- Preview and run the final import
Batch Import: create an App

AppWithinMinutes is a powerful tool to create XWiki data structure

Your data structure will receive the imported data

It's a great feature of XWiki allowing to have better navigation in information
Batch Import: prepare the spreadsheet/csv

Add a page name, space and title column to prepare where the data will go in XWiki.

A column will refer to the attached files (images, word file).

Clean up format of numbers and dates if necessary.
Batch Import: upload the file

Both CSV and Excel are supported. The Excel plugin needs a configuration in xwiki.cfg and a restart
Batch Import: configure the mapping

Each column in the mapping can be mapped to a column in the App Within Minutes application.
Some hidden settings are necessary to perform file attachment or file conversion (from word to XWiki)

The files to be uploaded need to be transferred on the server as a zip file. The XWiki servers needs to be able to access it
Batch Import: preview imports

Overview of data mapping and parsing

<table>
<thead>
<tr>
<th>Document full name</th>
<th>Population</th>
<th>Quarantine</th>
<th>Density</th>
<th>Age</th>
<th>median</th>
<th>urban</th>
<th>gdp2019</th>
<th>First Case</th>
<th>Total Cases</th>
<th>Death per million</th>
<th>Document title</th>
<th>Document space</th>
<th>Document attachments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Countries.AF</td>
<td>38928346</td>
<td>60.0</td>
<td>18.0</td>
<td>25.0</td>
<td>18734</td>
<td>02/11/2021</td>
<td>273</td>
<td>0.2</td>
<td>Afghanistan</td>
<td>Countries</td>
<td>af.png</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Countries.AL</td>
<td>2877797</td>
<td>105.0</td>
<td>36.0</td>
<td>63.0</td>
<td>15418</td>
<td>03/07/2020</td>
<td>304</td>
<td>6.0</td>
<td>Albania</td>
<td>Countries</td>
<td>al.png</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Countries.DZ</td>
<td>43851044</td>
<td>18.0</td>
<td>29.0</td>
<td>73.0</td>
<td>172781</td>
<td>02/12/2021</td>
<td>986</td>
<td>2.0</td>
<td>Algeria</td>
<td>Countries</td>
<td>dt.png</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Countries.AO</td>
<td>77265</td>
<td>164.0</td>
<td>45.0</td>
<td>88.0</td>
<td>91527</td>
<td>03/01/2020</td>
<td>439</td>
<td>207.0</td>
<td>Andorra</td>
<td>Countries</td>
<td>ad.png</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Countries.AG</td>
<td>97929</td>
<td>223.0</td>
<td>34.0</td>
<td>26.0</td>
<td>1688</td>
<td>03/12/2020</td>
<td>9</td>
<td>9</td>
<td>Antigua and Barbuda</td>
<td>Countries</td>
<td>ag.png</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Countries.AR</td>
<td>45195774</td>
<td>17.0</td>
<td>32.0</td>
<td>93.0</td>
<td>445469</td>
<td>03/02/2020</td>
<td>1265</td>
<td>0.9</td>
<td>Argentina</td>
<td>Countries</td>
<td>ar.png</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Countries.AM</td>
<td>2963243</td>
<td>104.0</td>
<td>35.0</td>
<td>63.0</td>
<td>13444</td>
<td>02/05/2022</td>
<td>736</td>
<td>2.0</td>
<td>Armenia</td>
<td>Countries</td>
<td>am.png</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Countries.AW</td>
<td>106766</td>
<td>593.0</td>
<td>41.0</td>
<td>44.0</td>
<td>2903</td>
<td>03/12/2020</td>
<td>60</td>
<td>1.0</td>
<td>Aruba</td>
<td>Countries</td>
<td>aw.png</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Countries.AU</td>
<td>25499884</td>
<td>3.0</td>
<td>38.0</td>
<td>86.0</td>
<td>1376255</td>
<td>01/12/2021</td>
<td>5350</td>
<td>1.0</td>
<td>Australia</td>
<td>Countries</td>
<td>au.png</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Previewing allows to check if the data is converted properly.

In particular check the document full name.

Date conversions are possible.

You can verify the attached file column is properly read.
Batch Import: deduplication settings

Countries
Last modified by Administrator on 2021/01/16 22:35

Data deduplication and overwrite

If by any chance, the column chosen for document name does not contain unique names, what should happen to the duplicate rows?

- Ignore
- Update: use the data as part of the other rows with the same document name (if the same column is set on multiple rows, the last one will win).
- Generate a new document name, by adding a number at the end of the document name

If then, for the document name built as indicated above, a document already exists in the wiki with that name, what should we do to the existing document in the wiki?

- Leave unchanged (ignore data from the file)
- Update with the data coming from the file (the fields mapped on the columns of the file will be updated, the rest will remain unchanged)
- Replace with the data coming from the file (the existing document will be deleted and a new one will be created instead, containing only the data in the file)
- Generate a new document name, by adding a number at the end of the document name built as indicated above. Note that 2 subsequent imports with this option checked will not yield the same results since the document names used depend on the documents existing on the wiki.

Back to mapping preview  Do an import simulation

Created by Administrator on 2021/01/16 19:48
**Batch Import: enjoy the results**

<table>
<thead>
<tr>
<th>Images</th>
<th>Page Title</th>
<th>Population</th>
<th>median Age</th>
<th>gdp2019</th>
<th>Total Cases</th>
<th>Death per million</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="San Marino" /> 33931</td>
<td>San Marino</td>
<td>47.0</td>
<td>1591</td>
<td>245</td>
<td>884.0</td>
<td>Edit Del</td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="Spain" /> 46754778</td>
<td>Spain</td>
<td>45.0</td>
<td>13897870</td>
<td>117710</td>
<td>234.0</td>
<td>Edit Del</td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="Italy" /> 60461826</td>
<td>Italy</td>
<td>47.0</td>
<td>1988636</td>
<td>115242</td>
<td>230.0</td>
<td>Edit Del</td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="Andorra" /> 77265</td>
<td>Andorra</td>
<td>45.0</td>
<td>91527</td>
<td>439</td>
<td>207.0</td>
<td>Edit Del</td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="Belgium" /> 11589623</td>
<td>Belgium</td>
<td>42.0</td>
<td>517909</td>
<td>16770</td>
<td>99.0</td>
<td>Edit Del</td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="Netherlands" /> 17134872</td>
<td>Netherlands</td>
<td>43.0</td>
<td>902355</td>
<td>15723</td>
<td>87.0</td>
<td>Edit Del</td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="France" /> 65273511</td>
<td>France</td>
<td>42.0</td>
<td>2707074</td>
<td>59105</td>
<td>83.0</td>
<td>Edit Del</td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="Switzerland" /> 8654622</td>
<td>Switzerland</td>
<td>43.0</td>
<td>715360</td>
<td>19303</td>
<td>66.0</td>
<td>Edit Del</td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="Luxembourg" /> 625678</td>
<td>Luxembourg</td>
<td>49.0</td>
<td>69453</td>
<td>2487</td>
<td>48.0</td>
<td>Edit Del</td>
<td></td>
</tr>
</tbody>
</table>
Batch Import: additional possibilities

• Convert files from Word to XWiki
• Convert files from Office formats to PDF
• Write custom processes at each stage of the import process (reading CVS, reading each line, after line import, etc..)
• Convert HTML or many Wiki syntax in fields to XWiki syntax
• Work on the spreadsheet to create meta-data prior to XWiki import (useful for Sharepoint or DMS)

See https://extensions.xwiki.org/xwiki/bin/view/Extension/Batch+Import+Application
XWiki Filter Streams

Confluence, MediaWiki, GitHub Wiki, DokuWiki
XWiki Filter Streams & Syntax Conversion

• A framework to import/export from to and from XWiki

• Uses XWiki's syntax conversion capabilities

• Also can import/export content in XWiki format (useful for migration from old XWiki instances)

See https://extensions.xwiki.org/xwiki/bin/view/Extension/Filter%20Module
Supported Wikis

- Confluence
- MediaWiki
- DokuWiki
- GitHub Wiki

Warning:
No import is perfect. There is always work to do. It can be because of custom features or macros, absolute links or changes in recent versions.
A word about Confluence

- Wiki used by many developers, including FOSS fans

- Major pricing changes by Atlassian over time
  - Regular pricing increase on Cloud and on-Premise
  - Major shift away from onPremise (only Data Center offer)

- XWiki receiving more calls for Confluence migration
Confluence Import process

- Export from Confluence as XML
- Run Confluence Filter
- Run Nested Migrator
- List & adapt macros or replace them

See https://extensions.xwiki.org/xwiki/bin/view/Extension/Confluence/XML/
Confluence Import: export Confluence XML

Exporting is available in Space Settings / Content Tools / Export

Choose XML output. It's possible to export all or only selected pages

Note that not being an admin might lead to a partial export
Confluence Import: configure filter export

Choose confluence+xml

The Confluence export zip file should be uploaded on the server in an area available
Confluence Import: configure filter import

Choose xwiki+instance

With or without versions

Run import

The result should show up in a space named with the Confluence space Key all pages under that space as XWiki "terminal pages"
Confluence Import: follow the progress page by page

You can see each page import. Errors should not stop the process

If there are massive errors, check disk space. Unfinished import runs can lead data behind in tmp directories
Confluence Import: Nested Pages Migrator

Choose space to convert to nested pages (usually the Confluence space key)

Uncheck all advanced settings

Run "compute plan"

See https://extensions.xwiki.org/xwiki/bin/view/Extension/Nested%20Pages%20Migrator%20Application
Confluence Import: check compute plan result

The tool lists how the pages will be organized

Check the result which should be similar to the Confluence hierarchy

If errors are listed of "too long page names", XWiki has a full page name limit of 255 characters. Page Name need to be changed but not the title. Find the intermediary page with a too long name.
Confluence Import: execute plan

The plan is being executed and it could take some time. Please wait...

LOGS:

1. Starting job of type [npmig.executor] with identifier [[npmig, executemigrationplan, xwiki]]
2. Start the execution of the plan.
Confluence Import: enjoy the results

This document will explain how to prepare and perform an upgrade of a client project. Note that I would really like to emphasize the fact that we're talking about a project upgrade and not an installation upgrade, meaning that the purpose of an upgrade should not be seen as upgrading the installation of the client (this wiki) but of the whole project: making sure that the build is up to date, all sources are in place, and most importantly, that the upgrade is reproducible.

Note that, although this topic has been addressed by How to migrate XWiki Enterprise and How to migrate XWiki Enterprise Manager, this document differentiates in the fact that it focuses on the upgrade as a project upgrade, not as an installation upgrade, and describes how to analyze an upgrade and perform it from the sources point of view rather than from a "do it and fix any visible bugs" perspective. Those 2 documents remain interesting lectures, for the "install and test" phase of the upgrade, which you will find below.

An upgrade must always be reproducible. This means that any point during the upgrade process you need to be able to do again what you've just done (be it on dev, staging or production), 100 times even, with the same result, and you should even be able to write instructions so that somebody else can do it following your instructions, and the result is the same.

In particular, this means that you need to keep track of everything that you've done:
- what did you upgrade (which version, which installation, was it custom, etc.)
- war that you installed. If they're not standard platform wars, they'd better be somewhere online (in the son of the project, as a build or as a sar) not on your own machine
- modifications that you've done in the configuration files of this war
- any jars / plugins / templates, etc that you install / replaced in the war. These also better be somewhere online, don't odd jars built on your own machine
- wars that you imported. If they're not standard platform wars, they'd better be somewhere online (in the son of the project, as a build or as a sar) not on your own machine
- pages that you've deleted from those wars
- modifications that you've made in the settings / preferences of the wars
- little code patches, modifications in wiki pages that you made because something was not working properly in order to hack it

This is not wasted time. It will transform later in upgrade instructions.

Procedure for using the Distribution Wizard is described in a separate procedure. Please refer to it in addition to this document.

Prepare the upgrade

Do your homework.
Confluence Import: what about Macros?

What if you see red boxes?

Confluence has a lot of macros, many are not supported

Some macros can be bridged to XWiki macros (need Wiki Macro development)

We are working on it. Some macros will be paying.
We are building a tool to list which macros are being used.

Contact us if you want to check which are the macros used in your instance, and help us decide which macros are more important.
Confluence: a work in progress

- Import more content
  - PR by contributor for comments / labels

- Support more macros of Confluence

- Maintain importer for future Confluence evolution (Cloud & On-premise). Confluence export format is not officially documented

If your company wants to migrate, sponsoring the work is possible
Conclusion
Conclusion

• Many tools available for importing content in XWiki

• Many of these tools have been build during service missions from XWiki SAS for clients. Thanks for the support.

• Thanks to the contributors who are reporting issues and also wrote code

• These tools will improve and in particular the Confluence one, with more users coming to XWiki
THANK YOU

Talk build with Open Source:

• Operating System: QubesOS / Fedora
• Presentation software: CryptPad OnlyOffice Slides (alpha)
• Screenshots: Shotwell
• Recording: Open Broadcasting Software
• XWiki and XWiki Extensions!