

# An innovative platform to allow full translation of Internet sites

**Théo Hoffenberg**

CEO - Founder, Softissimo

theo@softissimo.com

**Christophe Brun-Franc**

Project Manager, Softissimo

cbf@softissimo.com

## Abstract

This paper introduces FLAVIUS, an innovative platform to allow full translation and indexing of Internet sites.

The EC-funded FLAVIUS project aims at providing an end-to-end solution for Internet content publishers of small to very large sites that would allow them to get their content translated and indexed in other languages, with a variety of options.

Most sites cannot be found easily by users who browse in other languages. A German user is unlikely to find a French provider if he looks for a collectible, a piece of furniture ... even if several sites provide instant translation.

This presentation will show how FLAVIUS can ease the process of translating and indexing a website; and the benefits that Internet content publishers can expect for the audience of their website. We will demonstrate through a concrete case the novelty and the advantages of this platform.

## 1 Introduction

Most Internet contents are available in a single language. Users may not find the information they are looking for, simply because it is not available in the language they are searching in.

Let's take a real example. A French person who wants some information about the red mountains Birmingham Alabama will probably search for "Montagnes rouges Birmingham Alabama" on Google. He will not get any relevant results. However, the Red Mountains Park's

website provides interesting information in English that could be useful to him.

The webmaster of the Red Mountains Park's website could have translated his website in French. But it would have implied significant costs and time that he might not want to spend. He could also have integrated an instant translation plug-in to his website so that the user could translate it, page by page. Anyway, his website would not have been found through a search in French.

FLAVIUS, which stands for Foreign Language Version of Internet and User-generated Site, is an EC-funded project that aims at bridging the communication gap between the Internet content publishers who write in their own language and the end-users that search for content in another language. Its objective is to provide Internet content publishers with a cost efficient, quick and easy solution to offer localized contents, through an online platform that translates a website as a whole in as many languages as needed and allows the translated versions to be indexed.

Reverso-Softissimo, which is a pioneer in language technologies, is the coordinator of the project that encompasses 6 other partners.<sup>1</sup>

## 2 A novel solution

Currently, several sites provide instant translation but this technology does not allow getting multilingual version of a website that can be indexed by search engines like any other handwritten page. Search engines, for example Google, offer cross-lingual search. However, the feature is not very visible to the user, for whom this way of search is not intuitive at all.

---

<sup>1</sup> 3 technological partners, namely Language Weaver, Across and Daedalus and 3 user partners, namely Qype, Overblog and TV Trip are part of the FLAVIUS project.

With FLAVIUS, Internet content publishers will just have to provide the content of their site in one language - their native language - and multilingual versions will be automatically generated through the FLAVIUS platform. They will have the option, either to give the URL of their website or - for those who are computer-savvy - to upload XML files. In this latter case, the FLAVIUS platform will keep their structure and return ready-to-edit translated XML files. In both cases, they will then publish all languages at the same time, which allows the entire content to be indexed in a multilingual way by search engines and thus, be easily accessible to all internet users, including in languages that the publisher does not understand.

Besides, everyone will contribute to enhancing the translation quality: if a website publishing team has foreign language skills, it will have the possibility to post-edit some of the translations, which will fill a translation memory that will be used for further translations. Moreover, visitors of the translated pages will have the option to suggest new translation for any piece of text, also contributing to increase the number of useful texts in the translation memory.

### **3 Target audience**

Through FLAVIUS, all internet content publishers - from webmasters of small websites, who are looking for a large number of visitors and cannot afford to pay for human translation, to large companies, popular websites (including websites with user-generated content), blog platforms, newspapers... which intend to reach new foreign customers at a very low cost - will be able to solve these issues.

In parallel, all internet users will benefit from this by better accessing the online information, even when it has not been written originally in a language they can read.

### **4 Technology and workflow**

The FLAVIUS platform is a web portal that interfaces with several technological components. Each of them is dedicated to a specific task and developed by a FLAVIUS partner.

The FLAVIUS approach encompasses several steps. Some of them are optional, and aim at improving the translation quality. They constitute the theoretical frame of the FLAVIUS project. Please note that some of them are not implemented yet on the platform. You will find further

in this paper the status of the project through a use case that presents what can be performed on the FLAVIUS platform as of today.

#### **4.1 Content retrieval**

Publishers can directly enter the URL of their site on the FLAVIUS platform or upload source files. XML and RSS formats are supported.

The FLAVIUS platform retrieves and extracts the data to be processed (any type of data can be processed, except images and Flash). When a URL is entered, the crawl depth can be chosen. Some control is done with the size of the data.

#### **4.2 Spell checking (optional)**

A larger part of internet content is now provided by users directly, who care less about spelling or grammar mistakes than the websites owners. The presence of these errors leads most of the time to wrong translations, in addition to making the reading experience less comfortable. Correcting or harmonizing the source text is a key step to enable quality translations.

That is the reason why the FLAVIUS platform contains an optional module to spell check the source text.

This component - developed by Daedalus - provides a spelling and grammar analysis of the submitted texts if they are written in Spanish, Italian, English or French. Moreover, it automatically corrects non-ambiguous errors and gives suggestions, when there is ambiguity.

#### **4.3 Terminology extraction (optional)**

As the final quality of the translated texts depends to a large extent on the translation accuracy of the key terms and expressions, publishers have the option to identify and extract the key expressions in their site content, using an analysis tool developed by Softissimo.

With this component, they can customize their translation and be sure of their accuracy. The revised segments (words or expressions) are stored in a dictionary that is then used to customize the automatic translation.

#### 4.4 Translation

Publishers select their choice of languages for the translation. For the moment, 8 languages are supported by the FLAVIUS platform, namely English, French, Spanish, Italian, German, Romanian, Swedish and Polish. But the project plans to cover the Europe's 23 official languages.

The translation process is made of two consecutive steps. First, the FLAVIUS platform looks up in a translation memory – powered by Across. It is generally assumed that the results given by the memory are better than those of the machine translation, since they originate from a human work. The translation memory can be shared by every platform's user. Then, if no result is found in the translation memory, the segment is translated by the translation engine. For the moment, a statistical Machine Translation (MT) engine - developed by Language Weaver – is used. But several MT engines can be plugged-in.

#### 4.5 Post-editing (optional)

Publishers can review the translated versions and modify them. The modified segments fill the translation memory that is used at the first step of translating.

#### 4.6 Translated content retrieval

Publishers retrieve the translated content. If they uploaded files, they get back as many translated files as the number of languages they had selected. If they submitted a URL, they get back as many new URLs as the number of languages they had selected. In one click, they can publish the new versions of their site, without worrying about technical problems. The translated versions are stored on the FLAVIUS servers.

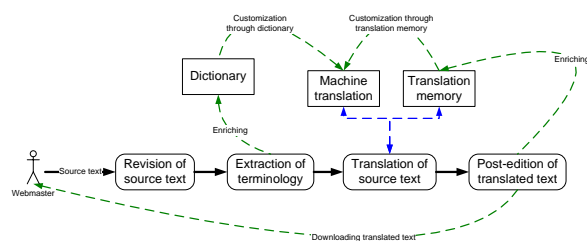


Figure 1. FLAVIUS workflow

### 5 Translation enhancement

The FLAVIUS workflow is a virtuous circle that allows a continuous enhancement of the translation quality. As mentioned above, spell checking is a key factor to enable quality translations.

Terminology extraction is a way to customize and enhance the automatic translation, as post-editing actions fuel the translation memory that in turn improves the translation quality.

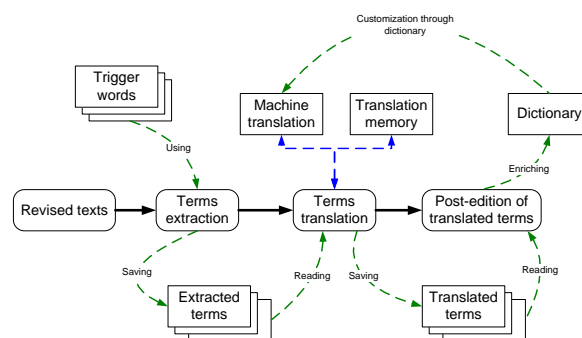


Figure 2. Translation enhancement

### 6 SEO management

Besides, one of the key issues for the success of a website is to be easily reached by potential customers and visitors through search engine. The FLAVIUS platform enhances the Search Engine Optimization (SEO) in several ways.

First of all, it enables the indexation of website in a multilingual environment.


Secondly, terminology extraction allows to identify the key expressions and thus to work on optimizing the keywords and Meta tags.

Finally, Internet content publishers whose translated site versions are hosted on FLAVIUS servers can benefit from the good indexation of the FLAVIUS platform itself. Moreover, as the number of in-coming links is important for indexation, links redirecting to the translated websites will be available on the Flavius platform.

Thanks to the combination of linguistic and SEO techniques, the FLAVIUS platform constitutes a real progress in the way to optimize the indexing of foreign versions of websites.

### 7 Use case

A webmaster has a website of 40 pages developed in a CMS and translated in French and in English. He intends to reach the Italian and Romanian markets but does not want to manage translation in those languages manually. He uses the FLAVIUS platform to generate versions of his website in Romanian and Italian.

 He first needs to create an account on the platform. Once authenticated, he can see on a dashboard the jobs already created (in progress or completed) and launch new ones.

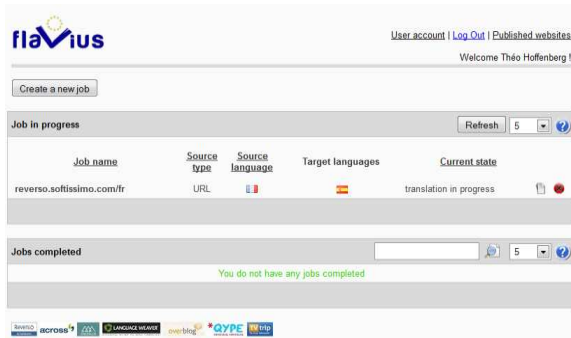


Figure 3. The dashboard

To start the translation process, he creates a new job by clicking on the “Create new job” button above the dashboard.

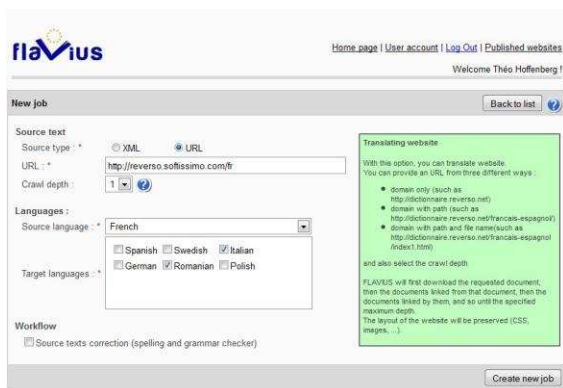


Figure 4. The “Create a new job” form

On the “Create new job” form, he types out the URL of his website and selects the source language and the target languages (e.g., Romanian and Italian).

He can also set the depth of crawling that defines the portion of his website that he wants to be translated.

He has then the option to apply a spelling and grammar correction to the source text. As previously mentioned in this paper, this operation aims at enhancing the quality of the translation.

To start the translation process, he clicks on the button “Create new job”. He is automatically redirected to the dashboard where he can follow the processing.

At any time, he can access to the details and status of the translation processing.

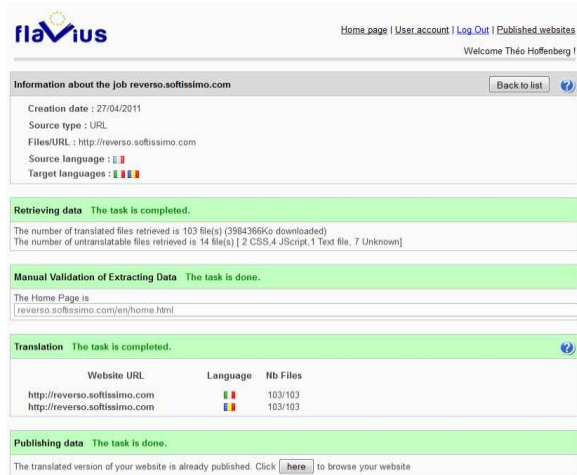


Figure 5. Details about the translation processing

The pages of his website are downloaded on the FLAVIUS server and translated according to the parameters he defined.

Once the job is completed, it disappears from the list of jobs in progress and appears on the list of completed jobs.

From this list, he can publish in real-time the whole translated website on the FLAVIUS server, by simply clicking on the “Publish” button.


Once the website is published, it is available on the internet and can be indexed by search engines.

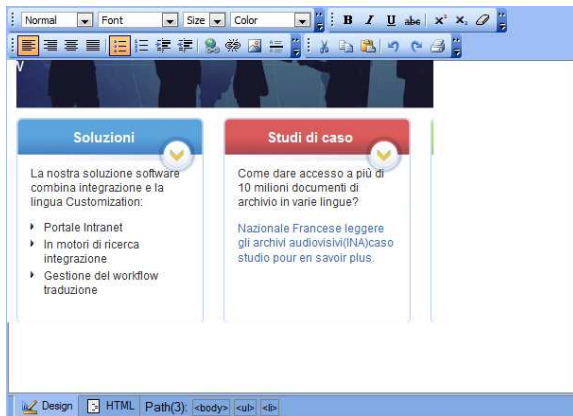


Figure 6. The published translated website

On the top of the translated website, a non-intrusive top banner allows visitors to select the languages of the website and switch among them.

Visitors of the translated website can submit comments and feedback on the translation by clicking on the “Feedback” button from the top banner.

 At any time, the webmaster can post-edit the translated content and modify it.



**Figure 7.** Post-editing my translated website

This use case has presented the actions that can be done on the platform as of today. As it is getting continuously upgraded, the best way to keep informed of the evolutions of the platform is to visit the project website: <http://www.project-flavius.eu/>.

We want to thank particularly the European Commission for its support and also all the partners who allowed us to reach this phase and achieve a working prototype.