NRC·CNRC

Institute for
Information
Technology

Université du Québec en Outaouais

# Making WikiMedia Resources More Useful for Translators

Alain Désilets
alain.desilets@nrc-cnrc.gc.ca

Caroline Barrière
caroline.barriere@nrc-cnrc.gc.ca

Jean Quirion
jean.quirion@uqo.ca

National Research Council of Canada

Université du Québec en Outaouais

# The WikiTerm Project

Evaluate, improve and create **open, wiki-based multilingual resources for translators**.

Part of **wider project called OPLT** (French acronym for Observing the Use of Translation Technology).

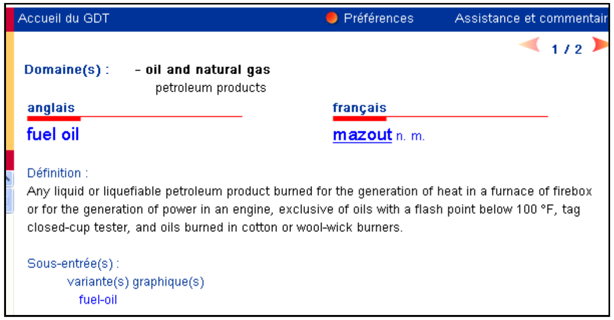
- Collect data by **observing translators** at work.
- Use this data to guide development of Computer Assisted Translation technologies
  - Improve **existing tools**
  - Inspire ideas for **new and innovative tools**

# The WikiTerm Project (2)

Translators are very heavy users of online multilingual resources (ex: bilingual dictionaries, terminology databases).

The resources they employ are:

- **Proprietary**
- Highly **controled** in terms of edit rights
- Sometimes **expensive**



Domaine(s) : - oil and natural gas
petroleum products

anglais: fuel oil

français: mazout n. m.

Définition :
Any liquid or liquefiable petroleum product burned for the generation of heat in a furnace of firebox or for the generation of power in an engine, exclusive of oils with a flash point below 100 °F, tag closed-cup tester, and oils burned in cotton or wool-wick burners.

Sous-entrée(s) :
variante(s) graphique(s)
fuel-oil

Inspired by the WikiPedia phenomenon, many of the large terminology databases (Termium, GDT, IATE), are **slowly but carefully** opening themselves to contributors from selected and trusted outside contributors.

But what  if we completely opened the gate and created a **wiki-based multilingual resource that anyone in the world can edit**?

We call such a resource a **WikiTerm**.

# The WikiTerm Project (3)

The work we present today is our first investigation of the **WikiTerm** idea, and it looks at the following questions:

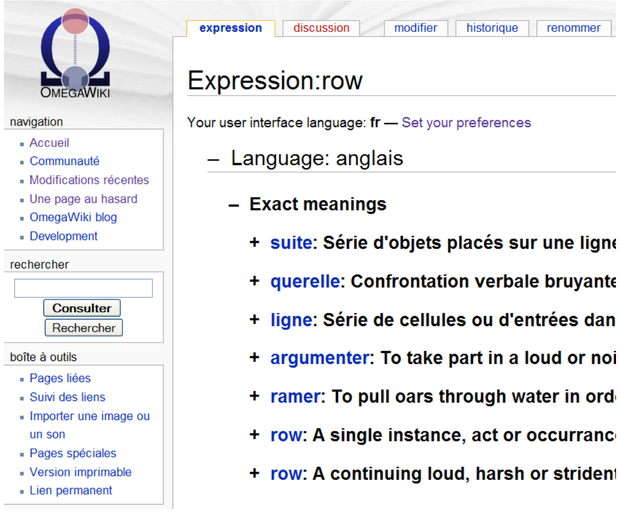
- *Do existing wiki resources like WikiPedia, Wiktionary or OmegaWiki already form a reasonable WikiTerm?*
- *If not, what requirements of a WikiTerm are not met by those existing resources?*
- *Could existing resources be modified or expanded to meet those requirements, or should WikiTerm be created as a new separate resource?*

# OmegaWiki: The New Kid on the Block

http://www.omegawiki.org/

This is a wiki-based **multilingual dictionary**.

At the moment, mostly an empty shell with little content.



Expression:row

Your user interface language: fr — Set your preferences

– Language: anglais

– **Exact meanings**

+ **suite: Série d'objets placés sur une ligne**
+ **querelle: Confrontation verbale bruyante**
+ **ligne: Série de cellules ou d'entrées dan**
+ **argumenter: To take part in a loud or noi**
+ **ramer: To pull oars through water in ord**
+ **row: A single instance, act or occurranc**
+ **row: A continuing loud, harsh or strident**

# Research Methodology

**Minimize speculation through observation** of translators involved in actual translation work (**Contextual Inquiry**)

- Translator translates in front of one or more **observers**.
- Translator **verbalises** what he is doing and why.
- Observers interrupt frequently with **probing questions** to clarify what was said, or bring to the foreground behavior that was not verbalised explicitely.
- Audio is **recorded**, screen is captured.

# Research Methodology (2)

**Collected data**

- 5 subjects so far (still collecting more)
- Different types of working environments (home-based freelance, medium size translation corporation, academic)
- 250 mins of video

# Analyzing the data

**From this raw data, we extracted 59 cases of translation difficulties that our subjects experienced**

- English term or expression in source language.
- French equivalent chosen by the translator.
- List of resources consulted by the translator in trying to resolve the translation difficulty.

# Analyzing the data (2)

| Type of difficulty | Examples |
|---|---|
| Terminology | subsidiary, fuel-oil |
| Phraseology | on short notice, for more than a decade |
| General language | grave, fiery, step |
| Cultural or Country-Specific Realities | Go Huskies!, liberal Indian Affairs critic |
| World Knowlege | Sun (name of a computer company), former Rep. Joseph Kennedy |

# Analyzing the data (3)

Original 59 cases filtered down to only retain **42 cases for which one might expect to find answers in a dictionary or terminological database, i.e either:**

- When  trying to resolve the difficulty, **subject actually searched** in a dictionary or terminological database.
- The authors could find the answer to the difficulty **in at least one of the following resources**: Grand Dictionaire Terminologique, Termium, Robert-Collins (these are the three resources that were most commonly used by our subjects).

Note: In two case (*"education gap"* and *"head salesman"*) the terminology expert in the research team (Jean Quirion) deemed that it would be **useful and realistic** for a WikiTerm to have an entry for this particular difficulty, even though it did not meet the above criteria.

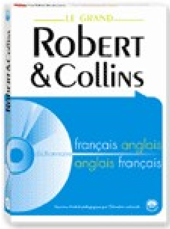
# How much do translators use online linguistic resources?

**Observation 1: Translators make heavy use of online linguistic resources**

- Our subjects used a dictionary or terminological resources to resolve a difficulty, **approximately at every 5 sentence** (assuming 10 words per sentence).
- **Only once** did a subject use a **paper resource.**

**Corollary: Translators are very heavy users of online linguistic resources, and are therefore an interesting constituency for wiki resources a la WikiPedia and Wiktionary.**

Le grand dictionnaire terminologique



# Do translators use wiki resources?

**Observation 2: None of our subjects seemed to use existing wiki resources**

- In 250 minutes of observation with 5 subjects, **we never once** saw a translator **going explicitly to a wiki resources** to resolve a translation difficulty.
- On **one occasion**, we saw a translator doing a Google search for a term (*“burough”*), then follow the WikiPedia link in the hits list.

**Corollary: Either translators do not know about these resources, OR, they are not useful to them as they currently stand.**

**Note:** Further analysis seems to indicate that it is the later.

# COULD translators use wiki resources to resolve translation difficulties?

Summary of relevant information found in various sources, for the 42 cases.

| | WikiPedia | Wiktionary | OmegaWiki (Early June 2007) | OmegaWiki (Aug 2, 2007) | Termium |
|---|---|---|---|---|---|
| **Has English entry** | 71.4% | 47.6% | 0% | 28.6% | 80.1% |
| **Has English entry in right sense** | 57.1% | 45.2% | 0% | 23.8% | 76.2% |
| **Has French equivalent** | 33.3 % | 35.7% | 0% | 14.3% | 76.2% |
| **Has French equivalent in correct sense** | **26.1 %** | **33.3%** | **0%** | **11.9%** | **76.2%** |

**Observation 3: For the needs of translators, WPedia, WNary and OWiki have poor content coverage when compared to traditional terminology databases like Termium (but OmegaWiki catching up quickly).**

# COULD translators use wiki resources to resolve translation difficulties? (2)

Solution might be to just encourage translators to contribute more to WikiPedia and Wiktionary…

But presentation of the information in those resources is really not well suited to their needs.

Given the frequency at which translators consult dictionaries and term banks (once every 5 sentence), information must be presented in a way that enables fast decision-making.

If it isn’t, translators will not be motivated to contribute.

NRC·CNRC
Institute for Information Technology

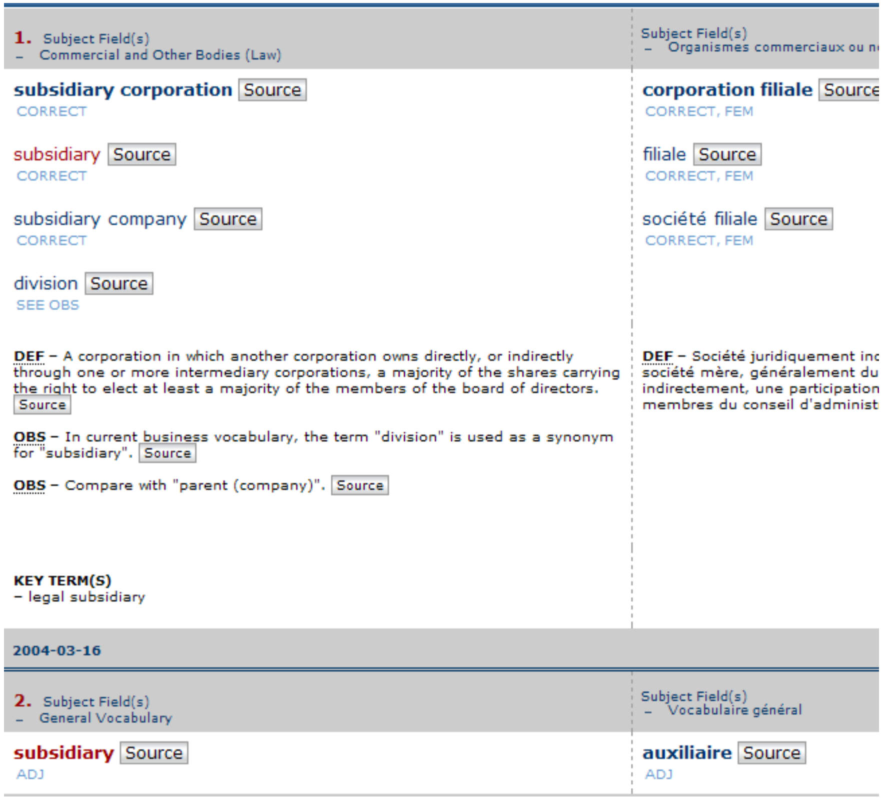
Université du Québec en Outaouais

# Presentation of the information

## Termium

**Easy and fast to scan:**

- Only information that is **relevant for translation** purposes.
- See both English and French **at same time**.
- **Multiple translations** on a single screen.
- **Consistent presentation** for all entries.



**1.** Subject Field(s)
– Commercial and Other Bodies (Law)

**subsidiary corporation** Source
CORRECT

subsidiary Source
CORRECT

subsidiary company Source
CORRECT

division Source
SEE OBS

**DEF** – A corporation in which another corporation owns directly, or indirectly through one or more intermediary corporations, a majority of the shares carrying the right to elect at least a majority of the members of the board of directors. Source

**OBS** – In current business vocabulary, the term "division" is used as a synonym for "subsidiary". Source

**OBS** – Compare with "parent (company)". Source

**KEY TERM(S)**
– legal subsidiary

Subject Field(s)
– Organismes commerciaux ou n

**corporation filiale** Source
CORRECT, FEM

filiale Source
CORRECT, FEM

société filiale Source
CORRECT, FEM

**DEF** – Société juridiquement inc
société mère, généralement du
indirectement, une participation
membres du conseil d'administ

**2004-03-16**

**2.** Subject Field(s)
– General Vocabulary

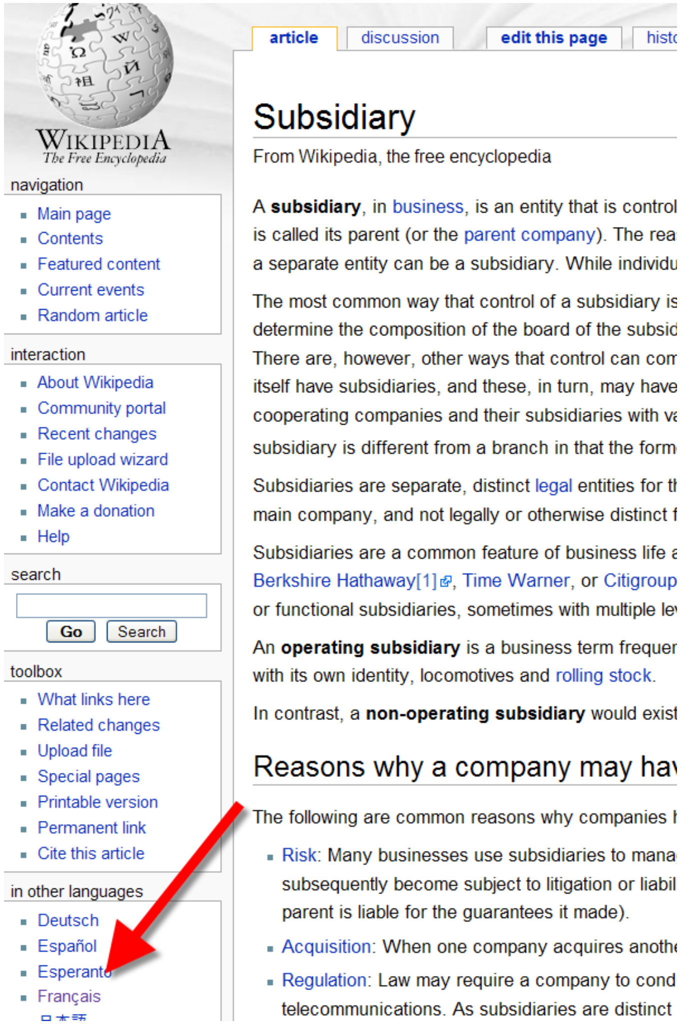
**subsidiary** Source
ADJ

Subject Field(s)
– Vocabulaire général

**auxiliaire** Source
ADJ

# Presentation of the information (2)

- Too much information that is **irrelevant for translators**.
- English entry and its translations are on **separate pages**.
- *is-translation-of* relation not garanteed to be symetrical
- If we follow the *Français* link, we get…



## Subsidiary

From Wikipedia, the free encyclopedia

A **subsidiary**, in business, is an entity that is control
is called its parent (or the parent company). The rea
a separate entity can be a subsidiary. While individu

The most common way that control of a subsidiary is
determine the composition of the board of the subsid
There are, however, other ways that control can com
itself have subsidiaries, and these, in turn, may have
cooperating companies and their subsidiaries with va
subsidiary is different from a branch in that the form

Subsidiaries are separate, distinct legal entities for th
main company, and not legally or otherwise distinct f

Subsidiaries are a common feature of business life a
Berkshire Hathaway[1], Time Warner, or Citigroup
or functional subsidiaries, sometimes with multiple le

An **operating subsidiary** is a business term frequer
with its own identity, locomotives and rolling stock.

In contrast, a **non-operating subsidiary** would exist

## Reasons why a company may hav

The following are common reasons why companies h

- Risk: Many businesses use subsidiaries to mana
subsequently become subject to litigation or liabil
parent is liable for the guarantees it made).
- Acquisition: When one company acquires anothe
- Regulation: Law may require a company to cond
telecommunications. As subsidiaries are distinct

navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

interaction

- About Wikipedia
- Community portal
- Recent changes
- File upload wizard
- Contact Wikipedia
- Make a donation
- Help

search

Go Search

toolbox

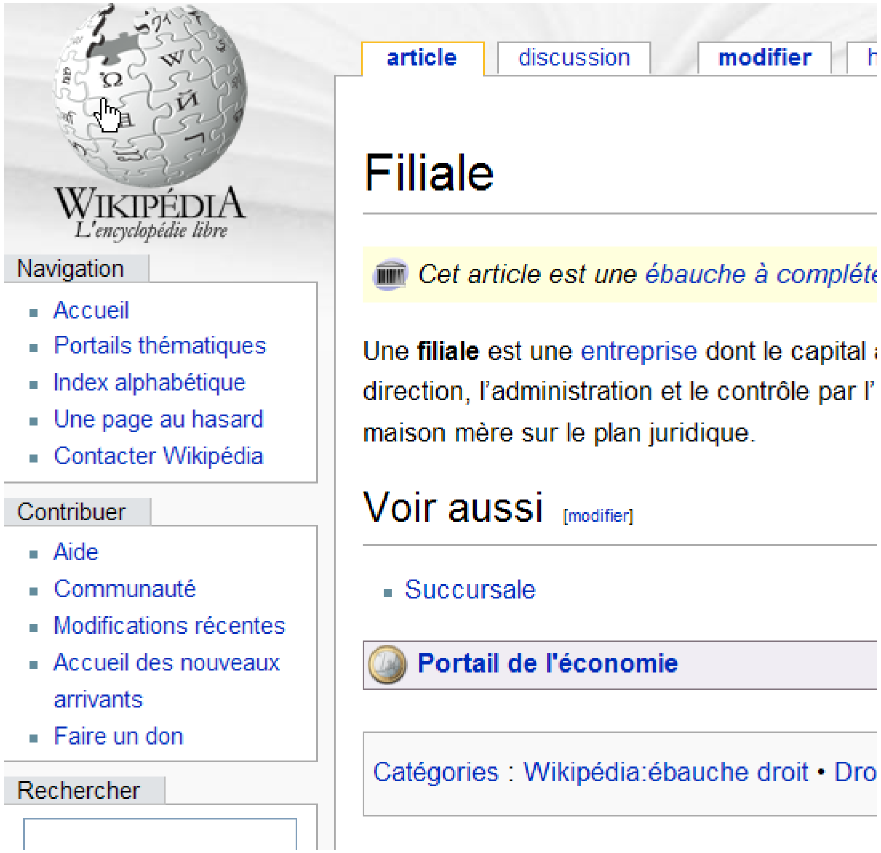
- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this article

in other languages

- Deutsch
- Español
- Esperanto
- Français

# Presentation of the information (3)

- This is one and only one possible translation of “subsidiary”.
- Other translations accessible from the *“Voir aussi”*.
- But a consolidated presentation of all the French translations of *“subsidiary”* would be better.
- Also, we have lost the definition of the original English term.



WIKIPÉDIA
L’encyclopédie libre

article | discussion | modifier | h

Navigation
- Accueil
- Portails thématiques
- Index alphabétique
- Une page au hasard
- Contacter Wikipédia

Contribuer
- Aide
- Communauté
- Modifications récentes
- Accueil des nouveaux arrivants
- Faire un don

Rechercher

## Filiale

*Cet article est une ébauche à complét*

Une **filiale** est une entreprise dont le capital direction, l’administration et le contrôle par l’ maison mère sur le plan juridique.

### Voir aussi [modifier]

- Succursale

**Portail de l'économie**

Catégories : Wikipédia:ébauche droit • Dro

NRC·CNRC
Institute for
Information
Technology

Université du Québec en Outaouais

# Presentation of the information (4)

Wiktionary
['wɪkʃənrɪ] *n.*,
a wiki-based Open
Content dictionary

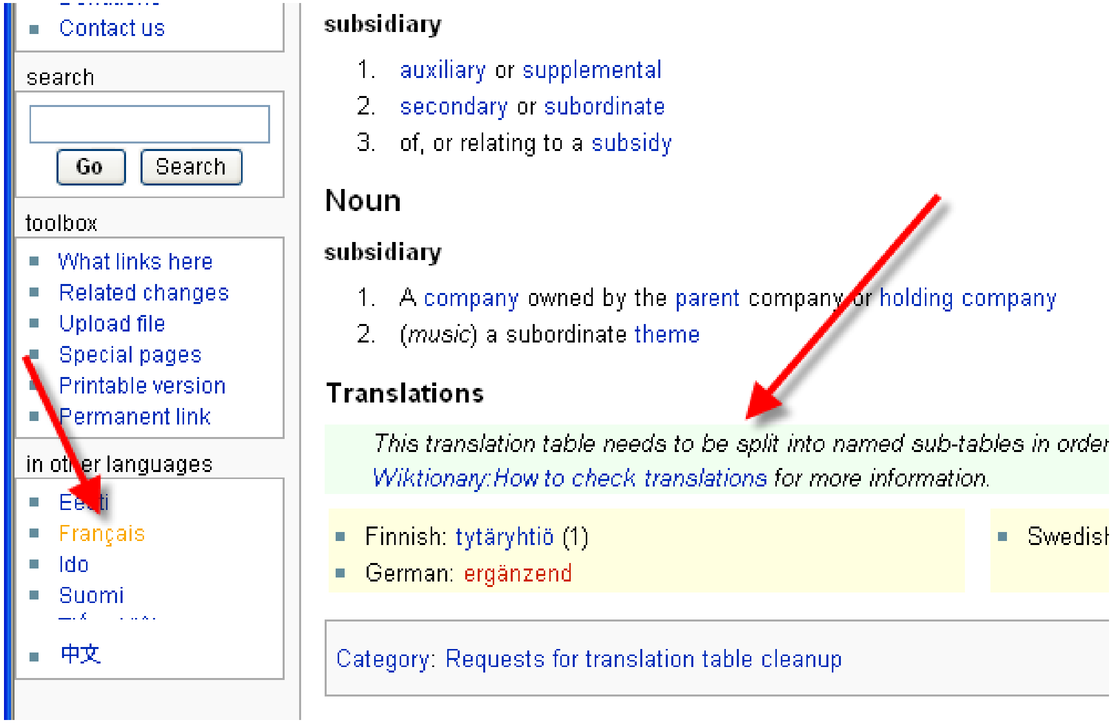
### Inconsistencies

- Translation for some languages is in *“other languages”* box.
- For some languages, it is in the *“Translations”* section.
- *is-translation-of* relation not garanteed to be symetrical

### Translations section has too much irrelevant information

- Shows translations in all languages, even though a given translator usually translates between two languages only.
- Situation gets worse as coverage of languages increases, and if different senses of the English translate differently.

If we follow the *“Français”* link…



- Contact us

search

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link

in other languages

- Eesti
- Français
- Ido
- Suomi
- 中文

**subsidiary**

1. auxiliary or supplemental
2. secondary or subordinate
3. of, or relating to a subsidy

Noun

**subsidiary**

1. A company owned by the parent company or holding company
2. (*music*) a subordinate theme

**Translations**

*This translation table needs to be split into named sub-tables in order*
*Wiktionary:How to check translations for more information.*

- Finnish: tytäryhtiö (1)
- German: ergänzend

- Swedish

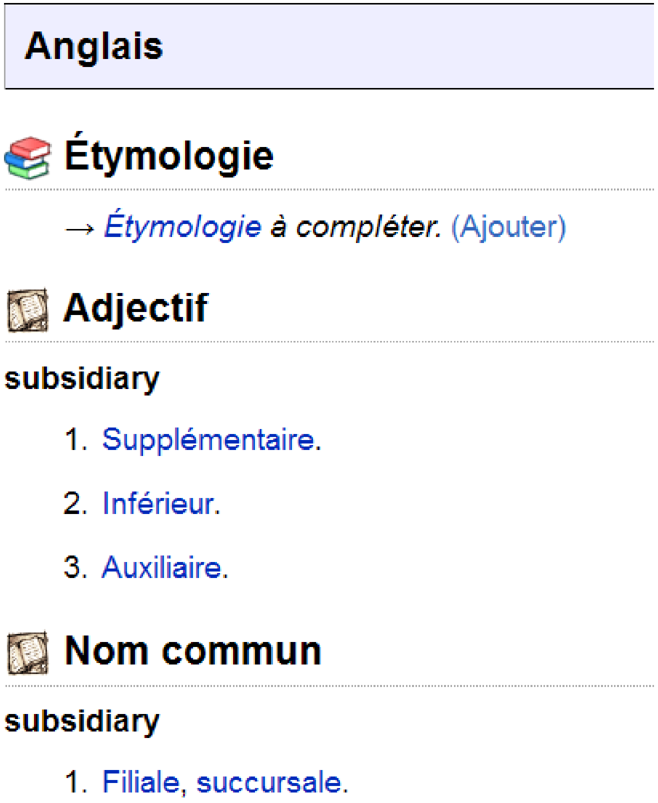
Category: Requests for translation table cleanup

# Presentation of the information (5)

**Presentation on the *"Français"* page is better.**

- Just the two relevant languages.
- Multiple choices of equivalents on a single page.
- But need to click on each equivalent to see its definition.
- It would be better to see a consolidated list of short definitions for each of those equivalents.

Wiktionary
['wɪkʃənrɪ] *n.*,
a wiki-based Open
Content dictionary



**Anglais**

**Étymologie**

→ *Étymologie à compléter.* (Ajouter)

**Adjectif**

**subsidiary**

1. Supplémentaire.
2. Inférieur.
3. Auxiliaire.

**Nom commun**

**subsidiary**

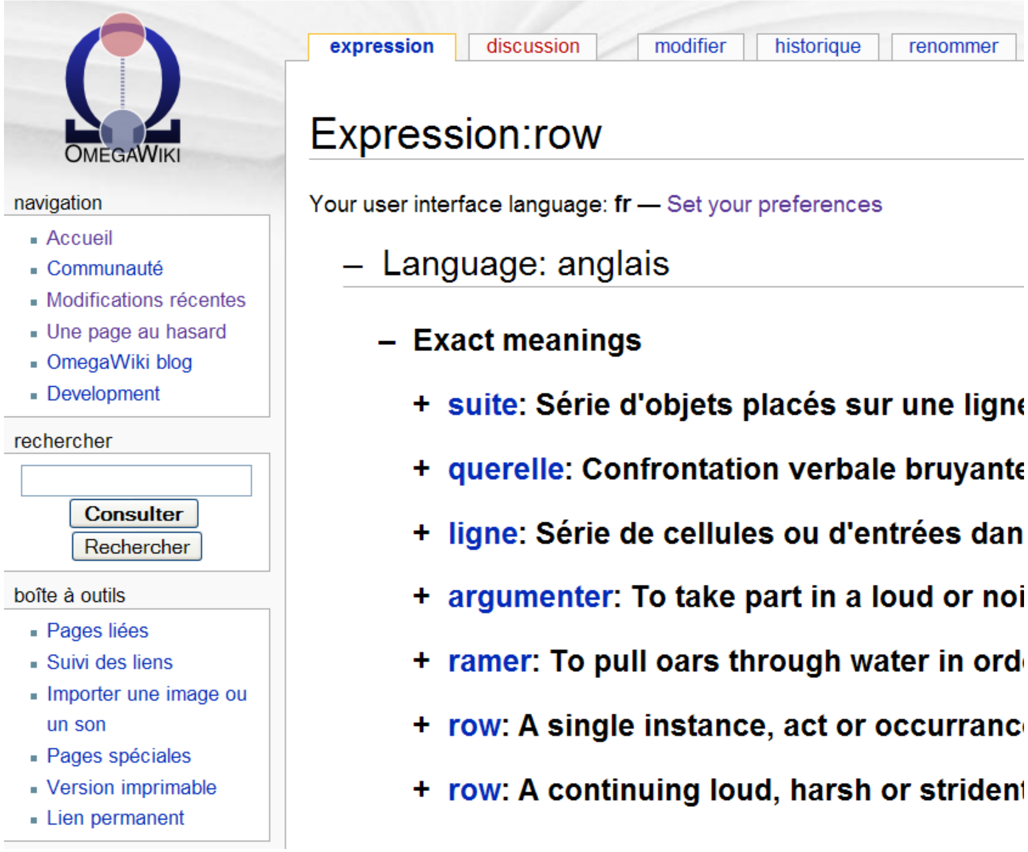
1. Filiale, succursale.

# Presentation of the information (6)

Much, Much, better!

- Single language pair.
- Includes short definitions of the French translations.
- *is-translation-of* relation IS garanteed to be symetrical

But term coverage so sparse that translators may not feel motivated to user nor contribute



Expression:row

Your user interface language: fr — Set your preferences

– Language: anglais

– Exact meanings

+ suite: Série d'objets placés sur une ligne
+ querelle: Confrontation verbale bruyante
+ ligne: Série de cellules ou d'entrées dan
+ argumenter: To take part in a loud or noi
+ ramer: To pull oars through water in ord
+ row: A single instance, act or occurranc
+ row: A continuing loud, harsh or strident

# Are existing wiki resources well suited to translator needs?

Short answer: **NO**.

None of them has the necessary content.

In addition, WikiPedia and Wiktionary do not present information in the way that translators need it.

The one that does a good job at presentation (OmegaWiki) has virtually no coverage of the terms and expressions translators search for.

# Building a WikiTerm

**Approach 1: Modifying and expanding WikiPedia and Wiktionary**

- Create a **specialized interface** in WPedia and WNary, which will extract and present only that information that translators care about (short definition, translation in the language user is translating to)
- Start a publicity campaign targeted at translators, to encourage them to contribute to them.

**Advantages:**

- These resources already cover a good portion of cases (25% for WPedia, 33% for WNary).
- They are well known.
- So translators may be motivated to contribute to its expansion.

**Disadvantages:**

- May be hard to automatically extract relevant information from the pages (especially WPedia).
- When users modify the content through the specialized interface, it may be hard to insert that content back into the original page.
- Some expressions (ex: *“value for the money”*) which translators would like to find in a WikiTerm, may not belong in WPedia or WNary.

# Building a WikiTerm (2)

**Approach 2: Expand content of OmegaWiki**

- Start a publicity campaign targeted at translators, to encourage them to contribute to it.
- Automatically seed the resource with content mined from WPedia, WNary.

**Advantages:**

- Presentation already very well targeted to the needs of translators.
- Relational structure of *is-translation-of* relationship allows automatic export of translators own terminology bases to OWiki.

**Disadvantages:**

- Not as well known at WPedia and WNary.
- May be hard to automatically extract relevant information from the pages (especially WPedia).
- Some expressions (ex: *"value for the money"*) which translators would like to find in a WikiTerm, may not belong in OWiki either.

# A Common Technical Challenge

Both of those approaches will face a common technical challenge, namely:

*How to automatically extract translation-relevant information from WPedia and WNary*.

Extracting terms and their equivalents

- Not trivial, because translations tagged in many different ways in WPedia and WNary.

Extracting short definitions

- Even harder, because no markup element for short definition.
  - Grab first paragraph? First sentence?

# Summary

Translators are heavy users of online multilingual resources.

They are currently dependant on proprietary closed resources.

The existing wiki resources lack the content and presentation that translators need.

The two approaches to create an open, wiki-like multilingual resources for translators is not clear…

But both require the ability to automatically extract translation-relevant information from WPedia and WNary.

… we are looking forward to conversations and feedback with the wiki community on that issue.

# Contact Info

Alain Désilets
alain.desilets@nrc-cnrc.gc.ca

Caroline Barrière
caroline.barriere@nrc-cnrc.gc.ca

Jean Quirion
jean.quirion@uqo.ca

National Research Council of Canada

Université du Québec en Outaouais

NRC CNRC

Institute for Information Technology

Science at work for Canada

National Research Council Canada

Conseil national de recherches Canada

Canada