

Editing challenges on multi-script wikis

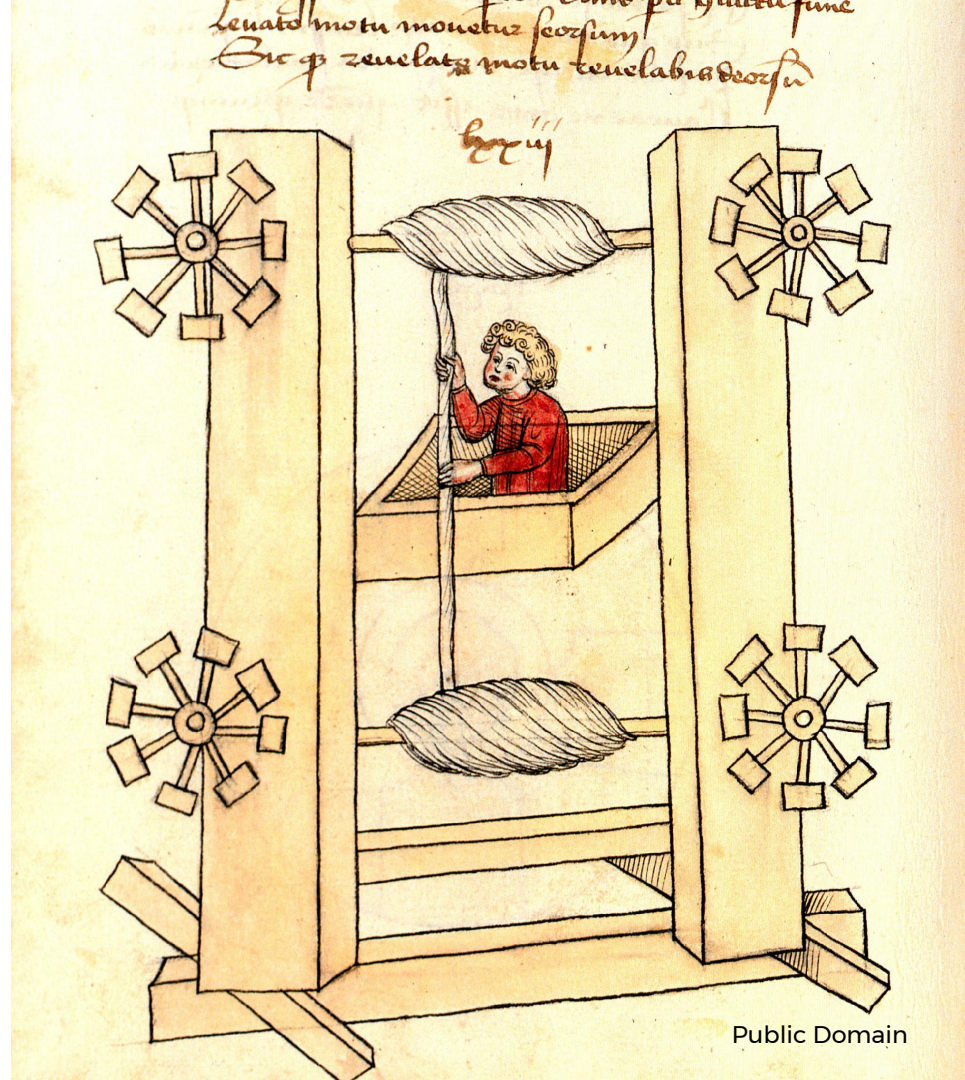
C. Scott Ananian
Wikimania 2017



August 12, 2017
Montréal, Québec, Canada

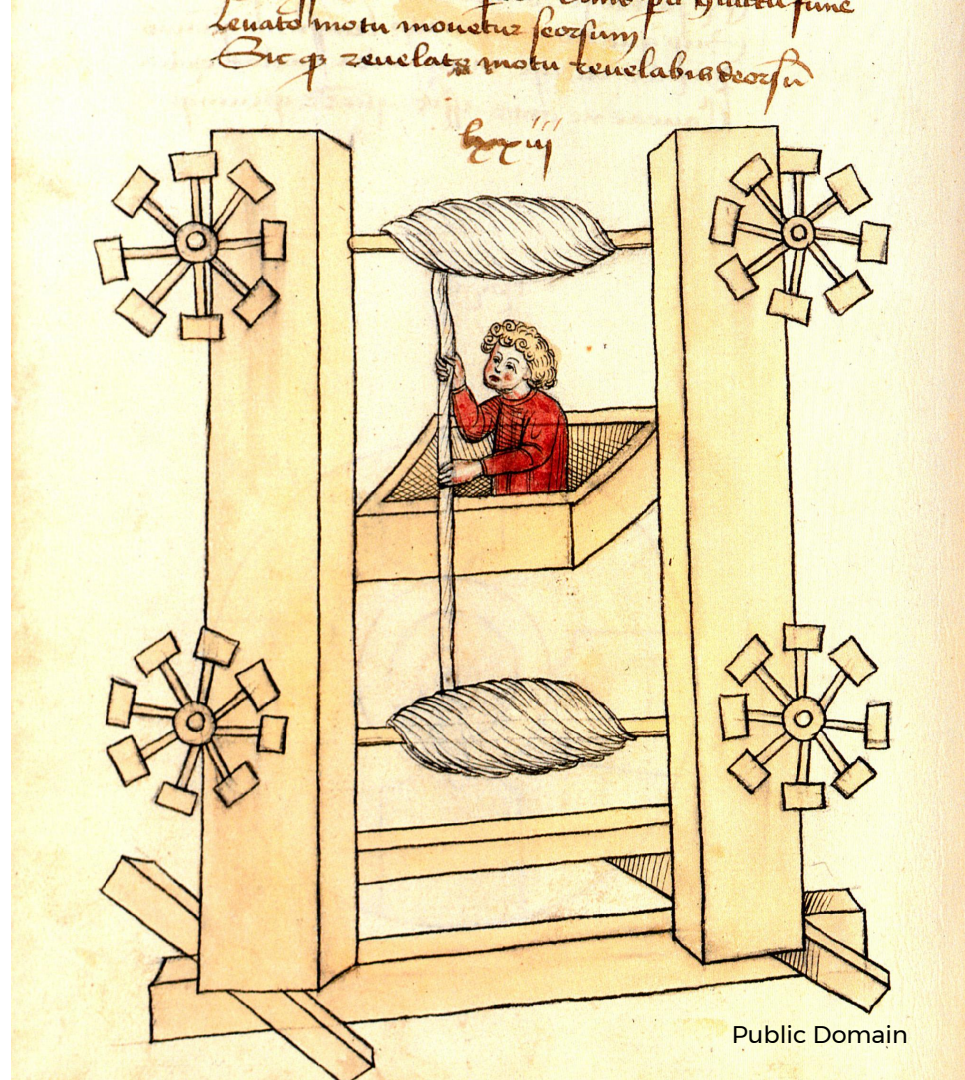
[[Elevator]]

Elevator design
by the German
engineer Konrad
Kyeser (1405)



[[Lift]]

Lift design by the German engineer Konrad Kyeser (1405)



[[Lift]]

Lift je uređaj za transport ljudi ili tereta među spratovima zgrada ili radnih platformi.



WIKIMEDIA
FOUNDATION



[[Лифт]]

Лифт је уређај за транспорт људи или терета међу спратовима зграда или радних платформи.



WIKIMEDIA
FOUNDATION



[[電梯]]

電梯，亦稱升降機、垂直
電梯。在馬來西亞、新加
坡和香港俗稱
「𨋖」(lift的譯音)，
是一種垂直運送行人或
貨物的運輸工具。



[[电梯]]

电梯，亦称升降机、垂直电梯。在马来西亚、新加坡和香港俗称“**𨋖**”(lift的译音)，是一种垂直运送行人或货物的运输工具。

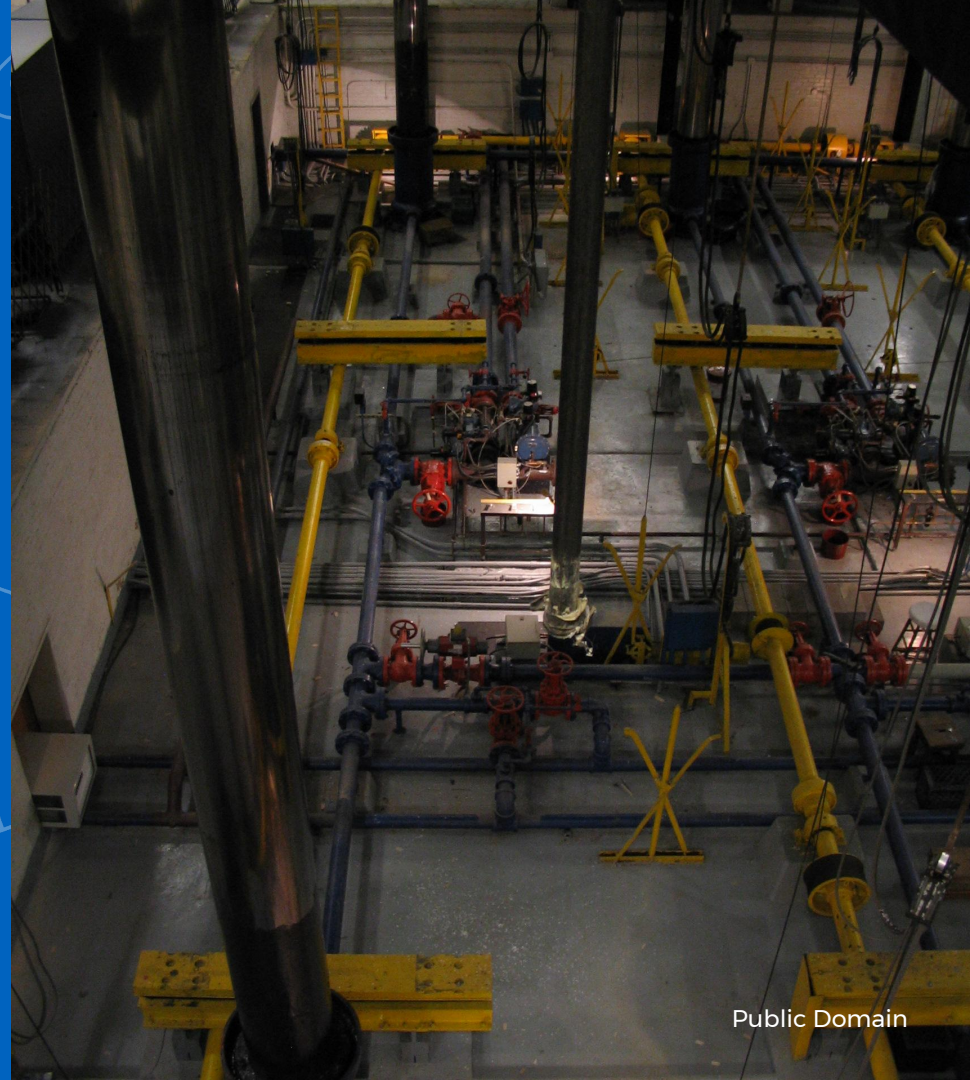


[[رافعہ]]

انتصابی نقل و حمل کی گل۔ جدید عمارتوں، جہازوں اور کانوں میں استعمال ہونے والی تمام کھلی اور بند ساختوں اور لگاتار چلنے والے ان پٹوں کو بھی کہا جاتا ہے جو (**Elevator**: رافعہ یا) انگریزی بھاری چیزوں کو ایک جگہ سے دوسری جگہ پہنچاتے ہیں۔ طاقت سے چلنے والے رافعہ جو عام طور پر بھاپ سے کام کرتے تھے انیسویں صدی عیسوی سے استعمال ہو رہے تھے جبکہ اس صدی کے اواخر میں برقی رافعہ عام ہو گیا۔



WIKIMEDIA
FOUNDATION

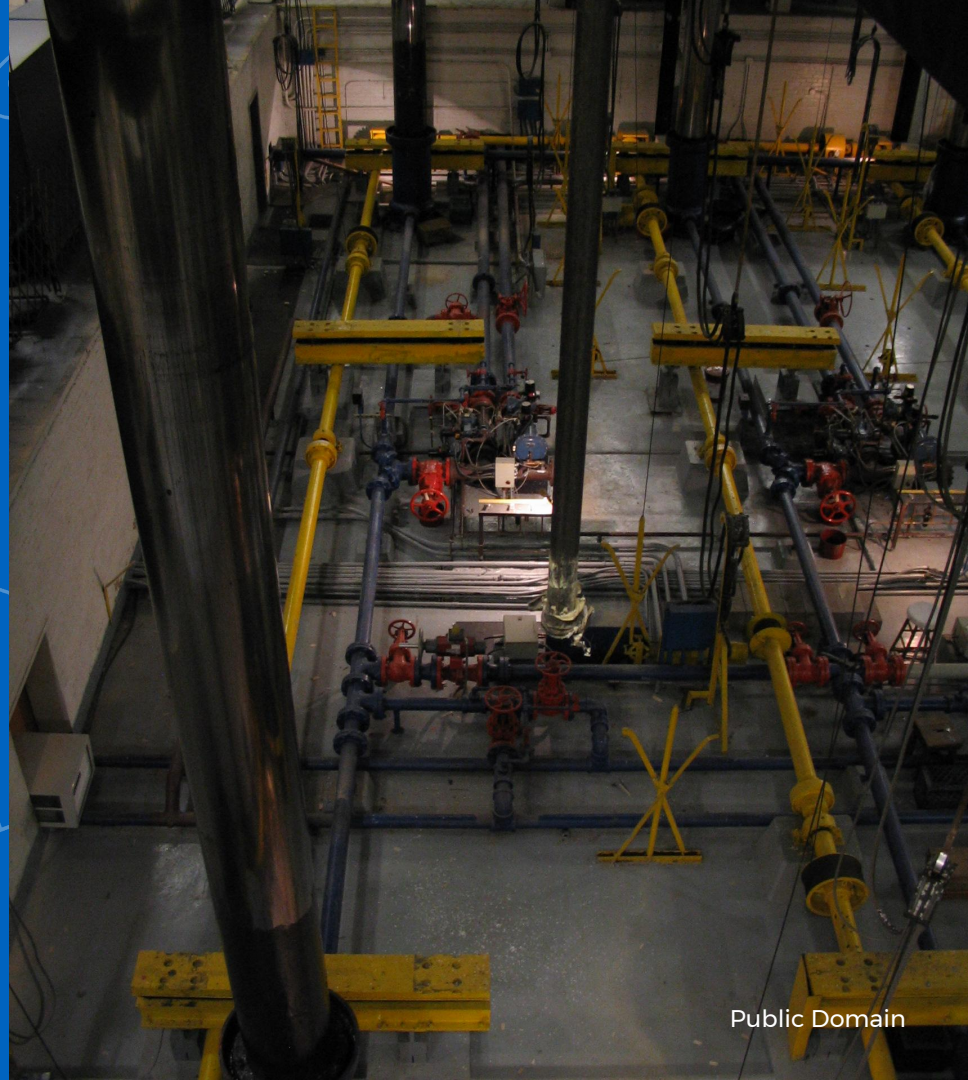


[[उत्थापक]]

उत्थापक, उच्चालित्र अथवा एलिवेटर (lift या elevator) एक युक्ति है वस्तुओं एवं व्यक्तिओं को उर्ध्व दिशा में चढ़ाने-उतारने के काम आती है। प्रायः किसी बहुमंजिला ऊँचे भवन, जलपोत एवं अन्य संरचनाओं में उत्थापक लगा होता है जो गोलों को या सामान आदि को एक मंजिल से दूसरी मंजिल या एक स्तर से दूसरे स्तर पर लाता और ले जाता है। उत्थापक प्रायः विद्युत मोटर द्वारा चलते हैं।



WIKIMEDIA
FOUNDATION



**LanguageConverter
converts words,
scripts, and (if you
stretch) languages**

Conversion in use on 11 wikis; wanted on ~35 more

[[[meta:Wikipedias in multiple writing systems](#)]]

- Chinese
- Serbian
- Kazakh
- Kurdish
- Inuktitut
- Anglo-Saxon
- Shilha
- Tajik (partial)
- Uzbek (partial)
- Gan (partial)
- Cantonese (client-side)



It brings our wikis together

(Although fighting over orthography is part of the fun?)



A photograph of two horses wading through a flooded field. In the foreground, a light-colored horse is partially submerged, its head and neck above water. In the background, a darker horse is also wading. The field is filled with water, and several wooden posts are visible, some standing upright and others partially submerged. The text "But it can make editing harder." is overlaid in large white letters.

**But it can make editing
harder.**

Mixed script editing

Списак 118 познатих хемијских елемената [\[уреди \]](#)

Следећа табела садржи 118 познатих хемијских елемената.

- **Атомски број, име, и симбол** служе независно као јединствени идентификатори.
- **Имена** су она која су прихваћена од стране [IUPAC](#); провизиона имена за недавно произведене елементе који нису формално именовани су дата у заградама.
- **Група, периода, и блок** се односе на позицију елемента у [периодном систему](#). Бројеви група су у тренутно званично прихваћеној нотацији; за старије алтернативне нотације погледајте [Група периодног система елемената](#).
- **Станје материје** (*Čvrsto, tečno, ili gasovito*) се односи на standardne uslove [temperature](#) i [pritiska](#) (STP).
- **Појављивање** прави разлику између елемената који се јављају у природи, категорисане као било *Praiskonski* или *Prolazni* (у смислу raspada), и *Sintetički* елементи који су произведени технолошким путем, и нису природно познати.
- **Опис** сумира својства елемента користећи опширне категорије које су присутне у периодном систему: [актинoid](#), [алкални метал](#), [земноалкални метал](#), [halogen](#), [лантанoid](#), [метал](#), [металoid](#), [пlemeniti gas](#), [nemetal](#), и [prelazni метал](#).

Status



Parsoid support

Visual Editor support



CC by SA 4.0. C. Scott Ananian

Reading variants

- Google indexing
- Kiwix offline reader
- Android app
- REST API
- PDF

Future



WIKIMEDIA
FOUNDATION

Source language annotation

Current implementation relies on character set to implicitly mark source language.

For reliable operation in some variants, we need explicit hints.

Ideally we can use an [annotation service](#), and not direct wikitext markup.

Initial work will focus on wikis where articles are stored in a consistent variant; for example, articles on yue.wiki stored in yue-hant variant.

Лифт => Lift

I think I have learned that
the best way to lift one's self
up is to help someone else.

— [Booker T. Washington](#)

Single-variant editing

Parsoid/Visual Editor use [Selective Serialization](#) in order to convert only the edited portion of a page back to wikitext.

Same technology can be used to edit in your native variant but avoid disrupting the variant used in unedited portions when saving changes.

Please lift those boxes and place them in the elevator.

Please `-{lift}`- those boxes and place them in the lift.

Glossaries

Word-based conversion dictionaries can get large and accumulate topic-specific entries.

Chinese wikipedia uses [templates](#) and a [custom gadget](#) to add new conversion rules.

Improve support by bringing this into core as article [glossaries](#).

[[Glossary:Books]]

-{en-uk:Sorcerer;
en-us:Philosopher}-
-{en-uk:Minister for Magic;
en-us:Minister of Magic}-
-{en-uk:sherbet lemon;
en-us:lemon drop}-

Content Translation

LanguageConverter can be viewed as a simple implementation of machine translation.

At the moment we only use machine translation to generate the initial revision of articles.

A future [Content Translation Tool](#) could be used to keep parallel texts in different languages (or variants) in sync.

TL;DR



Summary

- Not everyone uses Latin letters!
- Some texts which look different are really quite similar.
- LanguageConverter is a tool for sharing content when texts are closely related.
- **We can editing easier on wikis using LanguageConverter!**
- These wikis could be good incubators for more general cross-wiki machine translation tools.



THANK YOU



WIKIMEDIA
FOUNDATION

The graveyard of unused slides



WIKIMEDIA
FOUNDATION

Example markup

Marking up text which has variants:

- Bidirectional rules:
-`{zh-hans:computer; zh-hant:ELECTRONICBRAIN;}`-
- Unidirectional rules:
-`{HUGEBLOCK=>zh-cn:macro;}`-
- Disable conversion:
-`{R|SimpTrad}`-

Also a bunch of stateful options for adding/removing rules (more on this later), and for working around title limitations.

See https://www.mediawiki.org/wiki/Writing_systems/Syntax

Glossary Templates

From [\[\[zh:鋼鐵人3\]\]](#) ([\[\[en:Iron Man 3\]\]](#)):

{{noteTA

|G1=Movie

|G2=Show

|G3=美国漫画

|1=zh-hans:罗伯特; zh-tw:勞勃; zh-hk:羅拔;

|2=zh-hans:奧德利奇·齊連安; zh-tw:奧德奇·齊禮安; zh-hk:奧德奇·齊禮安;

|3=zh-hans:羅德斯; zh-tw:羅德; zh-hk:羅德;

}}



WIKIMEDIA
FOUNDATION

Transliteration code

```
class SrConverter extends LanguageConverter {  
    public $mToLatin = array(  
        'a' => 'a', 'б' => 'b', 'в' => 'v', 'г' => 'g', 'д' => 'd',  
        'ђ' => 'đ', 'е' => 'e', 'ж' => 'ž', 'з' => 'z', 'и' => 'i',  
        [...]  
        'X' => 'H', 'Ц' => 'C', 'Ч' => 'Č', 'Љ' => 'Dž', 'Ш' => 'Š',  
    );  
    public $mToCyrillics = array(  
        'a' => 'a', 'b' => 'б', 'c' => 'ц', 'č' => 'ч', 'ć' => 'ћ',  
        'd' => 'д', 'dž' => 'џ', 'đ' => 'ђ', 'e' => 'е', 'f' => 'ф',  
        [...]  
        'Nj' => 'Њ', 'n!j' => 'њ', 'N!j' => 'Hj', 'N!J' => 'HJ'  
    );  
};
```

Word conversion code

```
$zh2Hant = array(  
'佩' => '佩',  
'佻' => '佻',  
[...]  
';克制' => ';剋制',  
'? 克制' => '? 剋制',  
);  
$zh2Hans = array(  
'傾' => '倾',  
'佻' => '佻',  
[...]
```