Supporting languages... all of them

Amir Aharoni
Gerard Meijssen

Language technology

- Wikimedia Foundation aims to share the sum of all knowledge
    - All knowledge with everyone
    - All knowledge is distributed over all languages and cultures
- To do this we need to support information in over 7000 languages
    - Information can be translated but we need to support sources in any language to provide provenance
- All languages have their requirements

Language technology II

- To write a language you need a script
- A script and its characters are in Unicode
  - the Unicode is a standard and can support any script
- However,
  - several scripts are not yet in Unicode
  - even a language like Polish or Slovenian is not supported for all its requirements in the Latin script
    - (old Polish or Slovenian for instance)

Language technology III

- To support languages in MediaWiki,
  - we need support for its scripts
  - we need to know how numbers and dates are written
- But also
  - we support plural in our software
  - we support gender in our software
  - we support formality in our software

Language technology IV

- We support a language on our server but...
- is it supported on the device of the reader
  - can the end user actually see the characters?
  - can the end user actually write the characters?

We support webfonts and input methods...
- for a small selection of languages
- using only freely licensed Unicode fonts

Language standards

We use language tools that are standards
We use language data from standards

- ISO-639-3 to know the languages to support
- Unicode for scripts
- CLDR for information about languages
- the US keyboard layout - it's just practical

Language standards II

The language standards support only a few languages

- MediaWiki supports more languages than the CLDR
- Not all scripts are defined in Unicode yet
- Not all characters are defined in Unicode yet
- Many people do not use the US keyboard

Language standards III

To "fix" the standards we need people to amend and append the information in the standards

● language teams are organized at translatewiki.net
● we communicate with standard organisations
● we communicate about our needs

as always, it is about the sum of all knowledge; this exists only in all languages

As always, it is about the sum of all knowledge; this exists only in all languages