Marefa
Implementing an Epistemological Model

Fifth ICUDL
Pittsburgh, PA
November 7, 2009

Nayel Shafei

Table of Contents

Goal of Marefa

Marefa is an ambitious project; to preserve Arabic culture

Preserving Knowledge

- Disintegrating governments
- Other Projects are funded by/for governors, or to commemorate dead people.

Encourage Progress

- To bridge the digital divide, fostering a knowledgeable and technologically literate society, at the most individual levels
- To enable all segments of society to readily access a 21st century information age

Global Dialog

- The Initiative seeks to promote the preservation of Arab identity and heritage
- To drastically expand the amount of digital Arab content publically available
  - Statistics indicate that, at most, 0.3% of all digital content is available in Arabic

Marefa seeks to focus on these three underpinnings

Marefa: Advancing Arabic Content

Marefa seeks to provide, share and proliferate digital Arabic content in a rapidly changing Information Age

Overview

- Founded by Nayel Shafei in Feb. 17, 2007. Owned by Marefa Foundation, a New Jersey-based not-for-profit Corporation. With 62,500 genuine articles, 200,000 images and 2.5 million pages of manuscripts, Marefa is the largest encyclopedia in Arabic in any form. It is free and Open. One million unique visitors a month, downloading 250,000 books a month
- Internationally acclaimed project:
  - Selected by Library of Congress as one of five online Arabic resources
  - Chosen by University of North Carolina as one of top seven sites for Arabic manuscripts

Marefa: Advancing Arabic Content

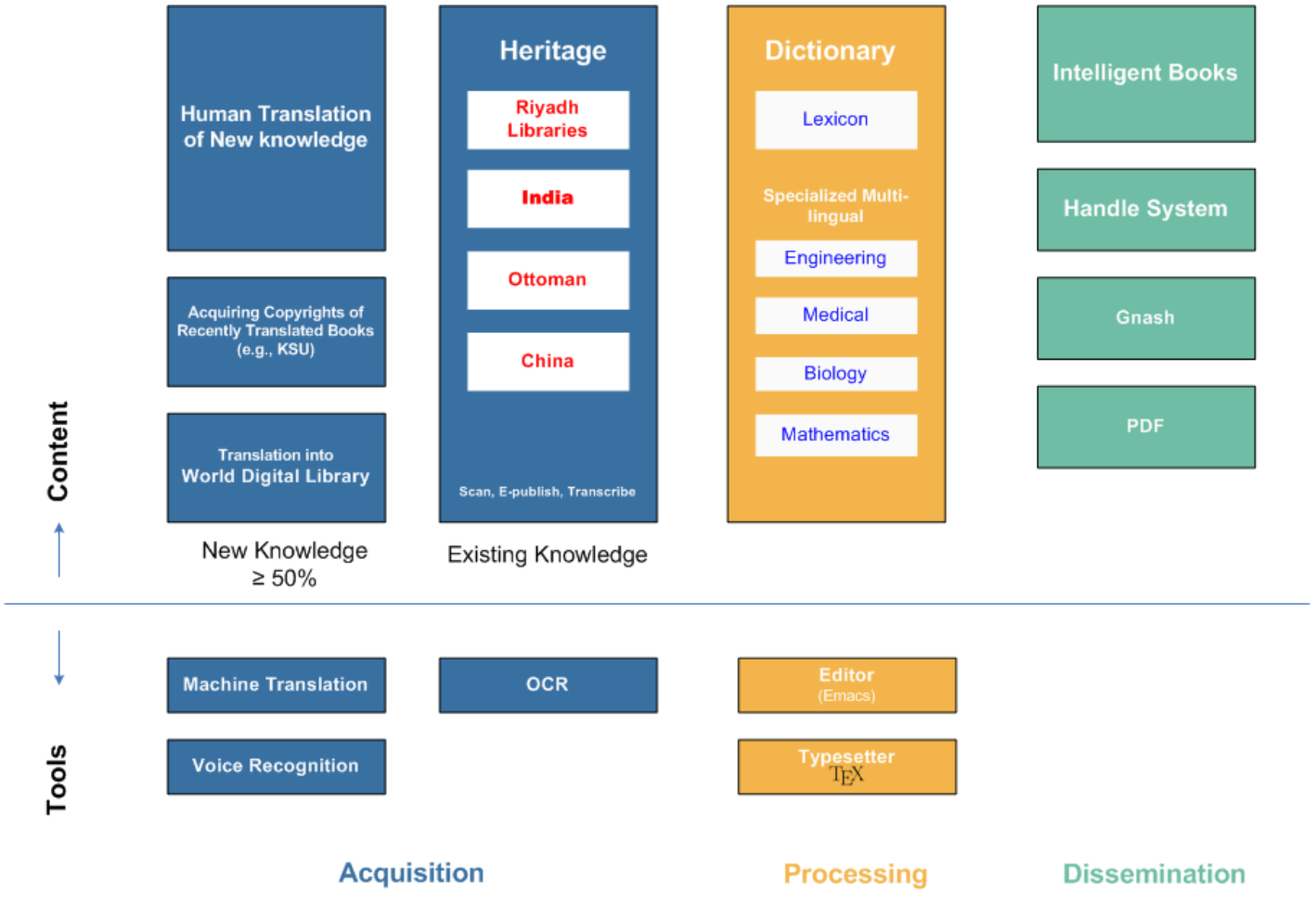
Sister Projects

- In addition the encyclopedia, Marefa project provides:
  - Mail, Video,
  - Webinar, 2,000 attendeess, every week. Largest live webinar , probably in the world.
  - Blogs, Forums, Sources, Manuscripts, Collaborative books

Guest Sites

- Marefa hosts, for free, other Arabic cultural websites, including:
  - *Egyptian Society for Historical Studies*
  - *Sharq Nameh*, the only publication in Arabic about Iran & Turkey
  - *Arab Institute for Translation*, Algiers
  - *Al-Oloom*, Arabic translation of Scientific American, Kuwait.
  - *Alwan for the Arts*, New York
  - New York Mid-Eastern & South Asian Film Festival

Marefa Model



Massachusetts Institute of Technology

Components Of The Model

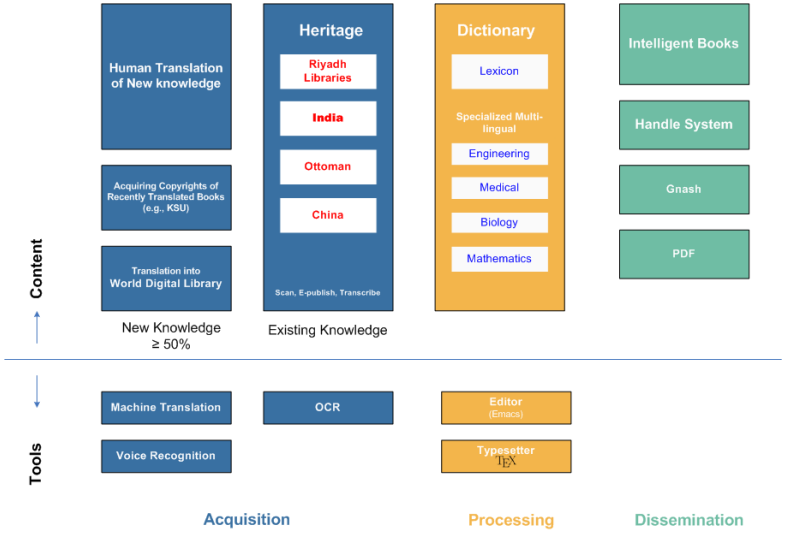
Three Phases of Content

Components of the Model

In order to encourage content-building, three phases are required:

- Acquisition
- Processing
- Dissemination

To build content in each of the three phases, the funding of software tools is also necessary

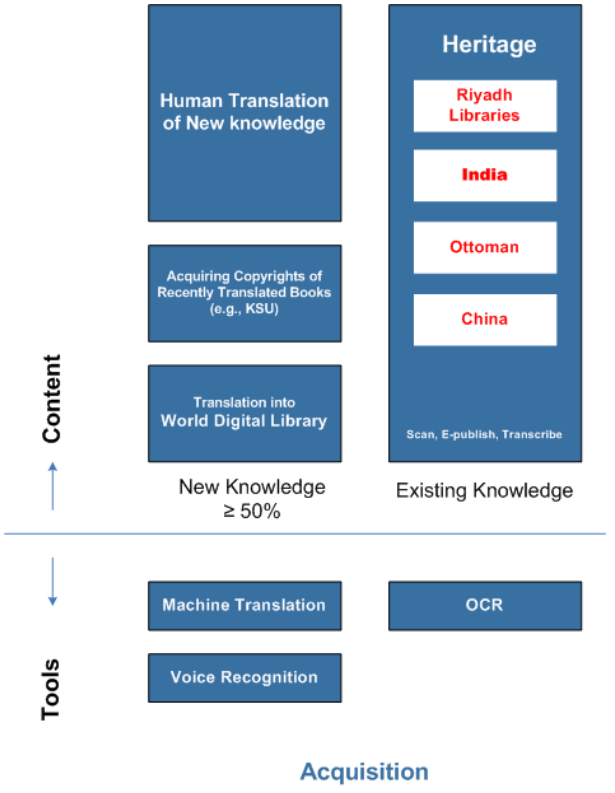


Phase (I): Acquisition

Components of the Model

In order for Arab society to move forward, we need to ensure that more than 50% of knowledge generated is new

New knowledge pertains to modern sciences and strategic technologies, as opposed to heritage: “old” knowledge currently preserved



Phase I: Acquiring New Knowledge

Components of the Model

Human Translation

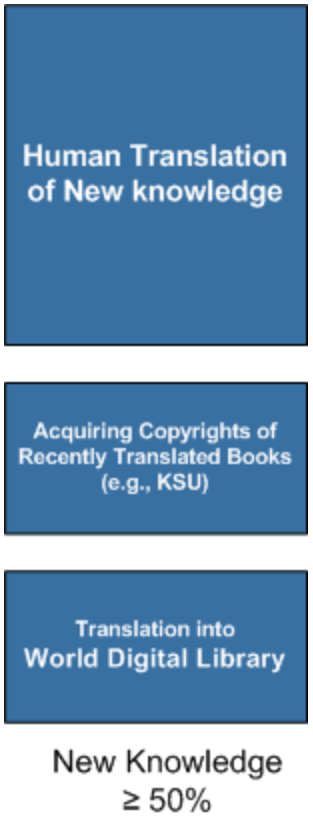
- Translation of 10,000 articles from English Wikipedia into free, open Marefa.

Acquiring Copyrights

- Convince publishers of recently translated books to allow electronic publishing of the books, in total, or in pieces
- Example: King Saud University Publisher

World Digital Library

- Selecting ten manuscripts that represent the contribution of the Arabic/Islamic Civilization, and translating it to the 6 official languages of UNESCO
- In collaboration with Library of Congress and World Digital Library



Massachusetts Institute of Technology

Phase I: Tools for Acquiring Knowledge

Components of the Model

Machine Translation

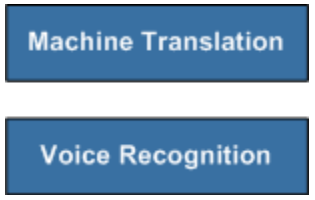
- Statistical machine translation is crucial in bridging the widening technological disparity between Arabs and the rest of the world. The adaptive translation engine should improve its performance as we feed it more translations, which is synergistic with all efforts for human translations
- The effort will be built in collaboration with MIT and the United Nations

Voice Recognition

- Sentence about voice recognition
- To be performed in collaboration with Prof. Victor Zue, MIT

Arabic OCR

- The finessing of existing OCR programs and algorithms for use in Arabic language is also crucial
- To be performed in collaboration with MIT and Carnegie Mellon University



Massachusetts Institute of Technology

Phase I: Acquiring Heritage

India

- Phase (I): Electronically publishing 25,000 Arabic manuscripts/books from Hyderabad, after selecting which ones to do
- Phase (II): Scanning the unrepeated manuscripts out of 100,000 Arabic manuscripts in India outside of Hyderabad

Ottoman Archive

- Publishing 40,000 manuscripts, mostly in Arabic, in collaboration with IRCICA and Islamic Countries Organization

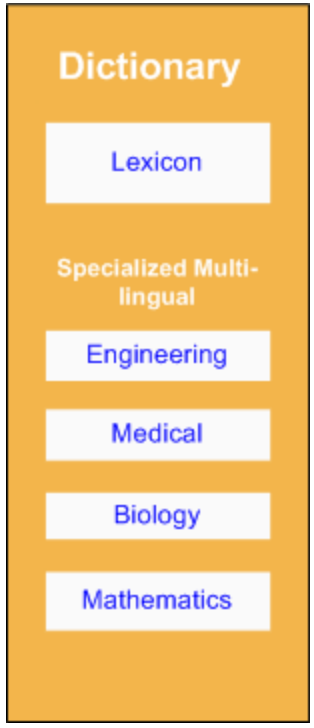
China

- Publishing 8,000 manuscripts in collaboration with Chinese Academy of Engineering and University of Tokyo

Components of the Model

Phase II: Processing Content

Components of the Model



Lexicon

- Ongoing project
  المعجم

Specialized Multilingual

Engineering

- Encoding The Unified Arabic Engineering Glossary, published in 1990, composed of 13 volumes (10,400 pages), and consisting of11,000 pages developed by the Federation of Arab Engineering Syndicates. The glossary is out of print; the original documents were destroyed in Baghdad.

Medical

- Encoding The Unified Arabic Medical Glossary, published in 1993 by the World Health Organization, making the glossary interactive and connecting it to encyclopedic articles.

Biology

- Compile the Binomial Classification of all species, which is the basis for any serious work in biology.

Phase II: Tools for Processing Content

Components of the Model

Editor Emacs

- Text and program Editor.
  in collaboration with GNU

Typesetter TeX

- Extending the world standard program for typesetting into Arabic.
- Project developed by Prof. Idrees Samawi Hamed, Colorado University.

Massachusetts Institute of Technology

مدينة الملك عبدالعزيز للعلوم والتقنية KACST

Phase III: Disseminating Content

Components of the Model

Intelligent Books / OCW

- The current development of MIT OpenCourseWare.
- Developed by Hal Abelson, the father of OCW,MIT-CSAIL.

Handle System

- The development of a mechanism for identifying all documents under Bob Kahn, the father of the Internet.
- Endorsed by all world authorities
- Collaboration for extending it to Arabic language, and to have a mirror site in the Arab World.

Gnash

- Creating the free version of Adobe Flash, from GNU
- In collaboration with Richard Stallman, founder of GNU, and world's chief hacker.

PDF

- Creating the free version of PDF.
- In collaboration with Richard Stallman

Massachusetts Institute of Technology