Thai
Optical Character
Recognition

Open-Source Initiative Project

Last Updated: 11 Jun 2023

เป้าหมายหลัก

1. ต้องการให้เห็นถึงความสำคัญของ open-source software ในประเทศไทย
2. ต้องการขับเคลื่อนให้ทุกปัจเจกในประเทศไทยทำ digitization และสามารถทำ data analytics ขั้นต่ำได้โดยไม่มีค่าใช้จ่าย
3. ความสามารถในการเข้าถึงความรู้ของประชาชนไทยต้องเท่าเทียมเสมอกัน
4. ทำให้ภาครัฐสามารถทำ data transformation โดยไม่มีความจำเป็นต้องจ้างจากเอกชน ซึ่งสามารถป้องกันเรื่องข้อมูลความมั่นคงและการขัดกันทางผลประโยชน์

ความฝันอันสูงสุด

Data-driven nation

Digitization for everyone

Data analytics for everyone

Open data-information-knowledge-wisdom for everyone

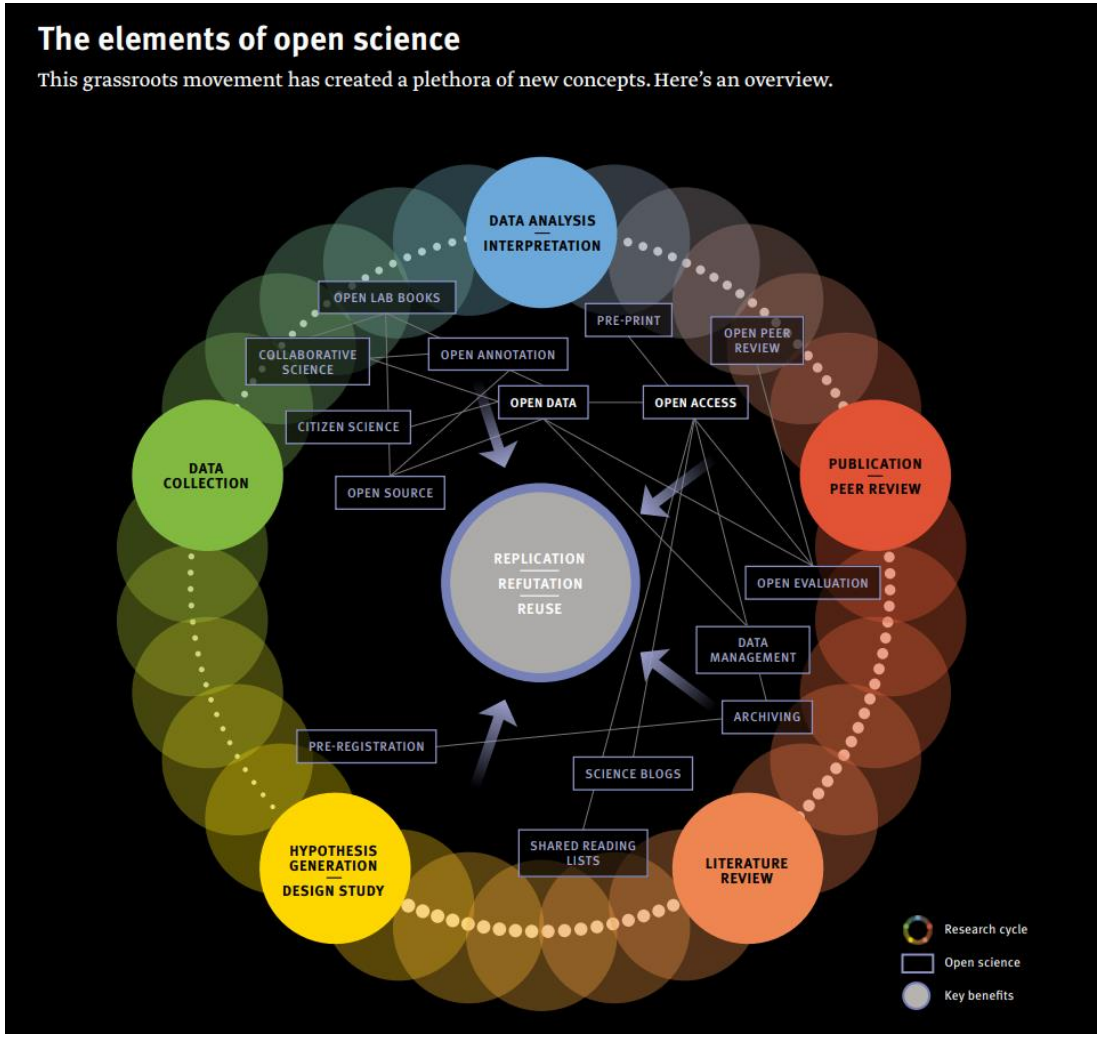
ความพยายามในเป้าหมายนั้น

My approach

1. Thai OCR engine open-source initiative
2. Thai open repository for advanced studies (more detailed than Wikipedia Thai; beyond the language barrier)

Governmental approach

1. Open National Data
2. Big Data and AI-related governmental organization
3. Smart city



Market Analysis

Current Progress

Future Direction

Business Model

Key partners
Existing open-source models
(Tesseract-OCR, PyThaiNLP, OpenCV)

Key activities
- Friendly OCR tools
- Knowledge graph construction over documents

Customer segments
(Prioritizing from most to least)

- Governmental organizations
- SMEs
- Individuals who'd like to have personal archives

Cost structure
- Man-day costs
- Data analytics tools and pricing for renting cloud services (maybe)
- Specialist consults/lectures

Value propositions
General
- Easier digitization
- Easier knowledge discovery process
- **"Always" open-source tools for everyone**

- Reducing processing time for data archiving
- Open-data initiation

- Importing data for data analytics process
- Precision accounting
- Data-driven organization in no-price

- Easy-to-archive

Revenue Streams
This project is for charity.
NO ANY DIRECT INCOME.

Indirect income sources:
- Open Science
- Open National Data Initiation

Customer Relationships
- Support Community in GitHub
- Bug report and feature proposal

Channels
- Developer society
- GitHub
- Governmental/academic organization (for free tools)
- Release notes

More concerns

การการันตีกลุ่มลูกค้า

องค์กรของรัฐ

- เอกสารมีความหลากหลายและเก่าจำนวนมาก
- ความจำเป็นในการวิเคราะห์ข้อมูลและนโยบายของรัฐที่พยายามผลักดันให้เกิด data-driven organization
- ประวัติขององค์กรและการตามเอกสาร

ตัวอย่างองค์กร

- สนง.คกก. กฤษฎีกา
- หจช. และ หสช.
- หน่วยงานภาครัฐอื่น ๆ ที่ต้องการทำ record management systems

- **SMEs**
- **บุคคลผู้สนใจงานจดหมายเหตุและงานทางโบราณคดี**

ความเข้าใจที่ต้องมี

องค์กรของรัฐ

- หน่วยงานต้องการทราบว่าตัวเองถือข้อมูลอะไรอยู่
- หน่วยงานต้องการทำ data analytics กับเอกสารที่มี
- หน่วยงานมีภารกิจที่ต้องปล่อยข้อมูลลง GD Catalogue, GDX และระบบสารบรรณอิเล็กทรอนิกส์อื่น ๆ

คุณภาพของข้อมูล

- ต้องสามารถวิเคราะห์ได้โดยง่ายและไม่คลุมเครือ (5-star open data https://5stardata.info/en/)
- ต้อง digitize ทุกรายละเอียดของเอกสาร การมี metadata ที่ดีเป็นสิ่งสำคัญ
- Versioning เป็นสิ่งจำเป็นในการเก็บเอกสาร

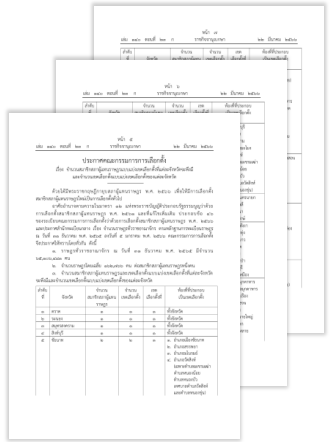
Market Analysis

Current Progress

Future Direction

ภาพรวมของ OCR

ไฟล์ PDF



ไฟล์ภาพ lossless
(PNG)

ตรวจจับเส้นตาราง
(Straight Line Detection)

ตรวจจับกลุ่มตัวอักษร
(Region of interest detection)

สร้างโครงสร้างเป็น
Markdown

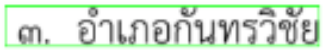
แปลงตารางเป็น CSV

๓. อำเภอกันทรวิชัย

รู้จำตัวอักษร
(Text recognition)

Tesseract OCR

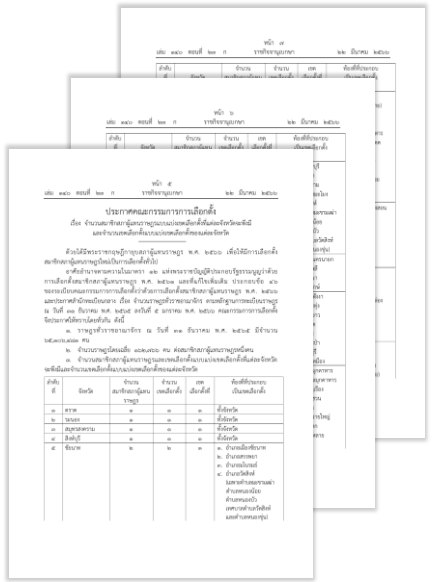
Customized
Tesseract
Engine



Text region

Thai Optical Character Recognition
Open-Source
Initiative Project

Straight line detection



ไฟล์ภาพ lossless (PNG)

Convert Color from BGR to grayscale

Canny Edge Detector

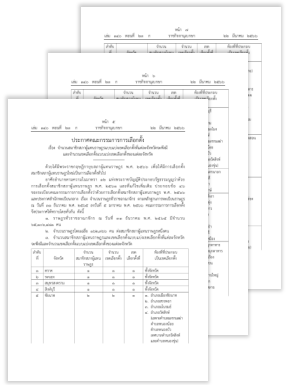
Hough Lines Detection

Lines detected

Thai Optical Character Recognition

Open-Source Initiative Project

Region of interest detection



ไฟล์ภาพ lossless (PNG)

Convert Color from BGR to grayscale

Gaussian blur (with kernel size = 3)

Canny Edge Detector

Contours detection (Topological Structural Analysis)

Hyperparameter auto-tuning: using Grid search for optimization

DBSCAN only in y-axis and grouping
- Epsilon
- Min samples

Rectangle collision detection and grouping
- Max gap of x-axis and y-axis
- y-scale

DBSCAN only in y-axis and grouping
- Epsilon
- Min samples

Rectangle collision detection and grouping
- Max gap of x-axis and y-axis
- y-scale

ตรวจจับกลุ่มตัวอักษร (Region of interest detection)

Objective $\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$

$$L(\boldsymbol{\theta}) = \sum_{i} \text{Area}(\text{Rect}(\boldsymbol{\theta}, i))$$

สิ่งที่ต้องการใน optimization คือ ทำให้พื้นที่กล่องที่ใช้คลุมตัวอักษรจะต้องน้อยที่สุด เพื่อป้องกันการคลุมกล่องในหลายบรรทัด

Text recognition

Tesseract OCR

Customized Tesseract Engine

Trained Model จากเอกสารราชการไทย

ใช้ในทุกกรณี ยกเว้นเลขไทย

Extended version of Thai OCR engine

ใช้เมื่ออ่านเลขไทย

รวมเส้นที่ detect ได้จาก Hough transform

๓. อำเภอกันทรวิชัย

Text region

๓.

อำเภอกันทรวิชัย

๓.
อำเภอกันทรวิชัย

แยก token ที่เกี่ยวข้องกับ space " " หรือ \t

ตัดแบ่งออกตาม space ที่ปรากฏ

||| ๓. อำเภอกันทรวิชัย |

รวมตาม space ที่ถูกตัดแบ่งออกไป

Thai Optical Character Recognition

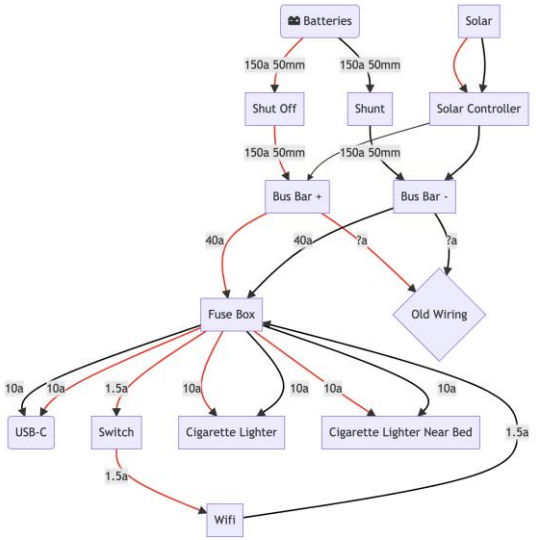
Open-Source Initiative Project

Market Analysis
Current Progress

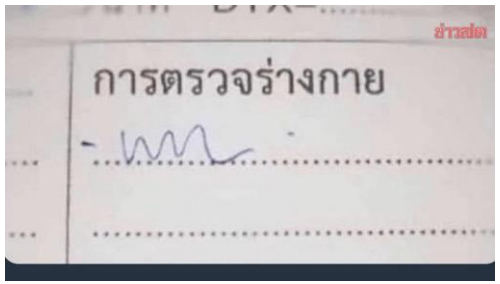
Future Direction

Pain points

Diagram export from OCR engine



We will format it as HTML/Markdown format.



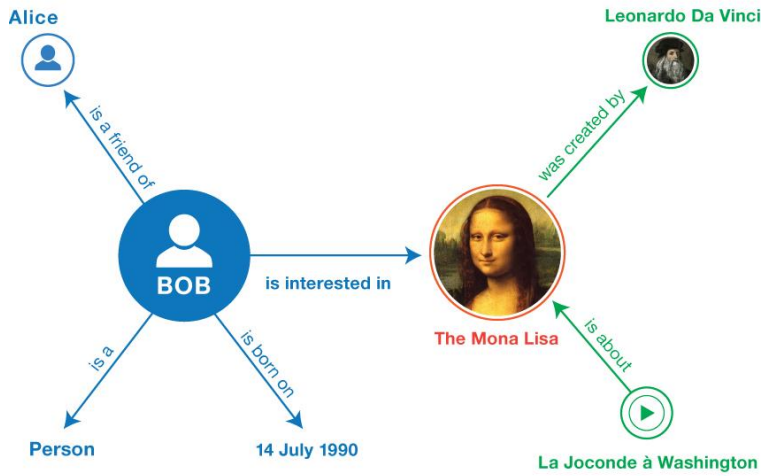
Handwriting/Typing classification

More fine-tuning of OCR model

Tesseract OCR

Variety of textual communi-cation

Knowledge graph construction from metadata which is defined in the document

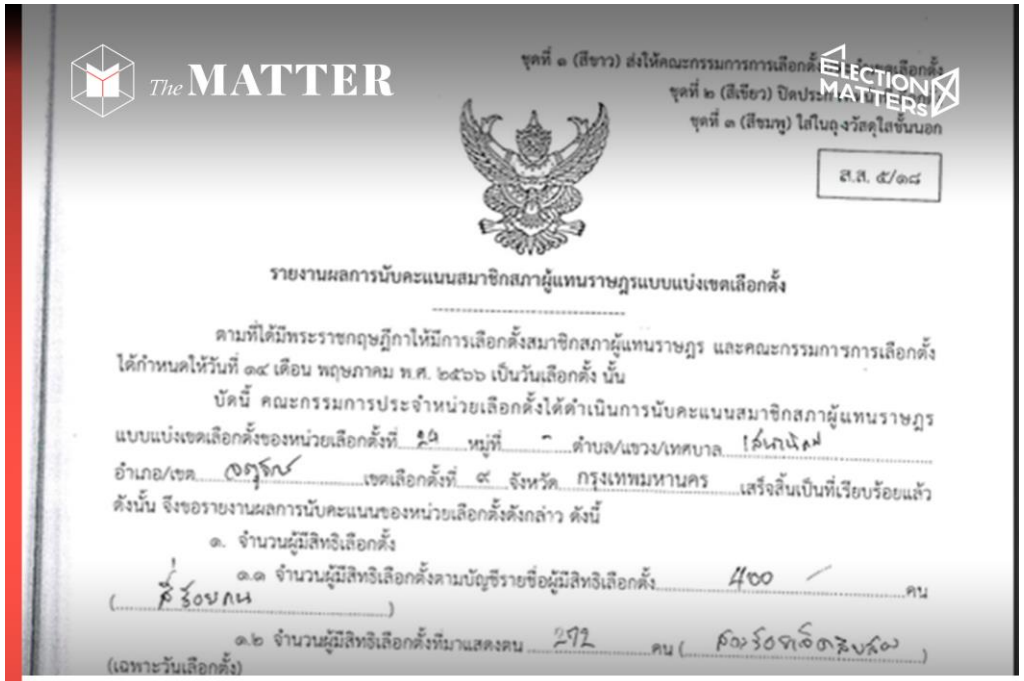


- Quick-win project
- Metadata generation

Thai Optical Character Recognition

Open-Source Initiative Project

The quick-win project



21 พ.ค. 2565 cr.กกต.

เข้าไปตรวจสอบกัน กกต.เปิดคะแนนเลือกตั้งรายหน่วย ให้ดูในเว็บไซต์แล้ว

แบบ ส.ส. 5/18

- มีจำนวน ~95,000 หน่วยเลือกตั้ง
- ไฟล์อยู่ในรูปแบบ scanned PDF
- ขนาดไฟล์รวม ~320 GB

Tasks

- Metadata generation
- Directory tree rearrangement
- Digitization process

Opportunities

- New datasets for Thai OCR
- Large-scale digitization prototyping

Thai Optical Character Recognition
Open-Source Initiative Project

Assignments

The quick-win project

Project manager: นิว

Dev: (ตามความเหมาะสม)

Tasks ที่ต้องทำก่อนเริ่มงาน

1. Epic planning
2. Sprint backlogging
3. Man-day estimation
4. The protocol for work management
5. Repository management
6. Wiki management
7. Sprint review scheduling