Practical Open Source AI "stubs" for archives – PoC solutions

Ph.D Anssi Jääskeläinen

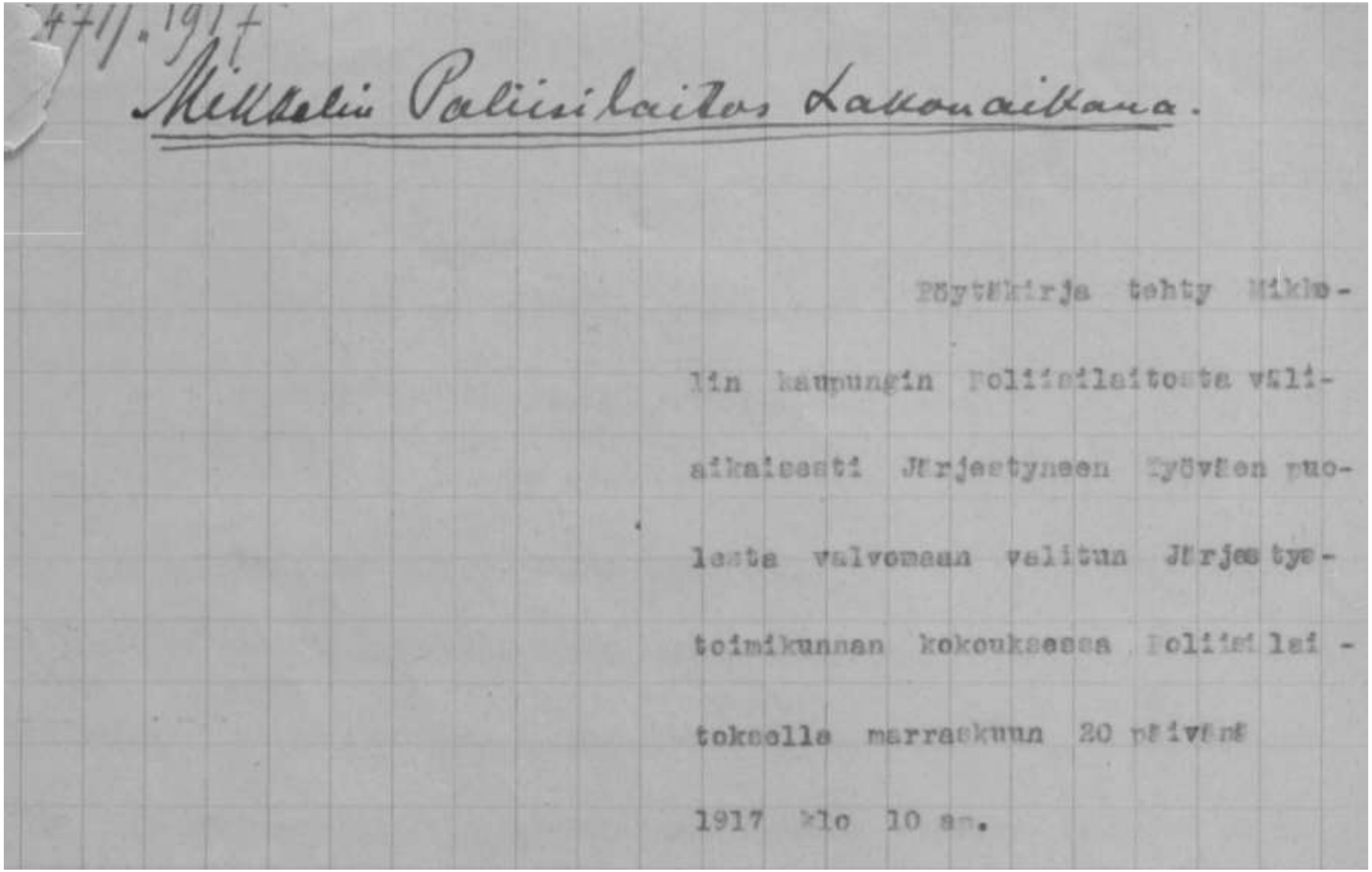
Xamk / Digitalia reseach center

Mikkeli, Finland

Agenda: From files to archive with AI
Set of AI tools

Image enhancements	Image to text	Enhanced OCR & LLM Fixes	Segmentation	Keywords from text
Similarity analysis	Image recognition	LLM translations	Speech to Text	Summaries
Text to Image	Classification / tagging	Q & A	RAG	Extending information

Sample image used for the presentation



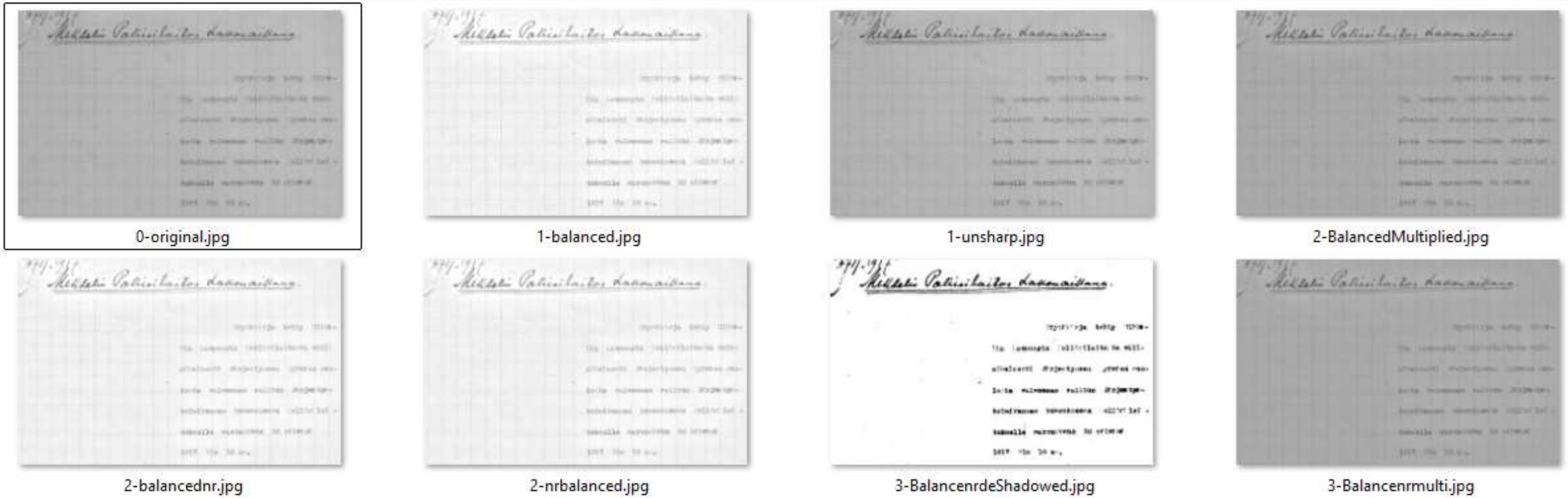
471/.1917

Mikkelin Poliisilaitos Lakkoaikana.

Pöytäkirja tehty Mikkelin kaupungin Poliisilaitosta väliaikaisesti Järjestyneen Työväen puolesta valvomaan valitun Järjestystoimikunnan kokouksessa Poliisilaitoksella marraskuun 20 päivänä 1917 klo 10 a.p.

https://yksa.disec.fi/Yksa4/download/141813411048700/file/e2983bac-70a9-4cc4-8135-8b552d87230a

Image enhancements



0-original.jpg

1-balanced.jpg

1-unsharp.jpg

2-BalancedMultiplied.jpg

2-balancednr.jpg

2-nrbalanced.jpg

3-BalancenrdeShadowed.jpg

3-Balancenrmulti.jpg

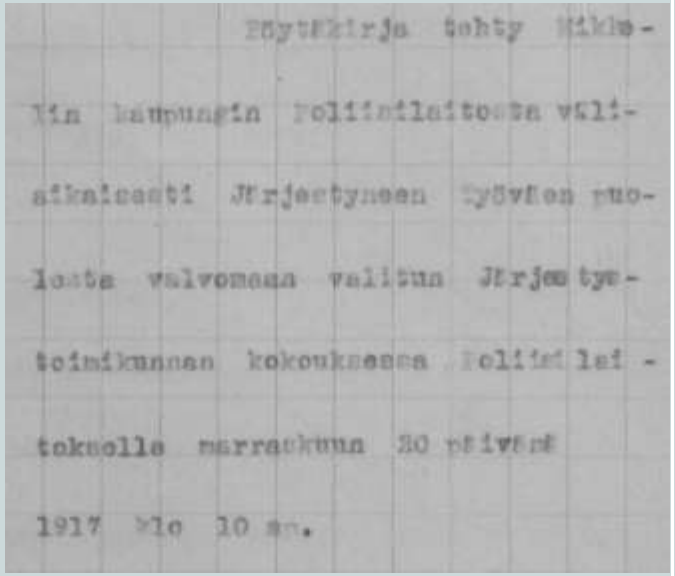
OCR quality: Key to enhanced Data Mining, **DOI:** 10.1109/ICECCME57830.2023.10252214

A. Jääskeläinen, M. Lipsanen, A. Föhr and T. Räisänen, "OCR quality: Key to enhanced Data Mining," 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Tenerife, Canary Islands, Spain, 2023, pp. 1-6, doi: 10.1109/ICECCME57830.2023.10252214.

Image to text

- You are all familiar with OCR/HTR technologies
  - Tesseract, PeroOCR, TrOCR, Abbyy fine reader, etc.
  - Based on trained AI models
  - We utilize finetuned PaddleOCR engine
- Steals parts from DLM presentation

Enhanced OCR & LLM fixes

Image	Original OCR	Tesseract OCR (best of enhanced)	Enhanced PaddleOCR
 Pöytäkirja tehty Mikke- lin kaupungin Poliisilaitosta väli- aikaisesti Järjestyneen Työväen puo- lesta valvomaan valitun Järjes tye- toimikunnan kokouksessa Poliisi lai - toksella marraskuun 20 päivänä 1917 klo 10 ap.	PKytrk tohty TM .& ituunkin rollir ilaito. ta vsii- aikai eeelti Jfrjäätyneen lyövf.ei^ ruo- Ij.-.uö iblvomeiqA yaii tua J* r Jeu tyr * toimikunnan kflikookcacje.a . olilr4 lei tokcolle rr ptkunn0 pr in n* 'lo	PAY atl . } Metitet: Cp leiat Le cYoq Habeas alefn-tal | Péythyirja tehty Wikig- lin | katpungin | potttetrat tos te we1t- | aikatcestt Itnjeetyneen lydvfien puo- | leste valvomaga yalitun Jirjes tye- toimikunnan kokouksessa Polite! lei -| | toksella marraskunn 20 ptivint 1917 lo 10 a.	Pöytäkirja tehty Mikle. lin kaupungin Foliisilaitosta Vali aikaisesti Jrjestyneen Työvlen puo- lesta valvomaan Valitun JLrjes tye- toimikunnan kokouksessa Foliisi lai toksella marraskuun 20 ptivanf 1917 klo10 an,

OCR enhacements → CER

Model	ELKA test (4273 row pictures)	NAF1 test (3475 row pictures)	NAF2 test (3247 row pictures)	ELKA test handwritten (714 row pictures)
Tesseract	4,6%	2,7%	4,4%	79,3%
Original PaddleOCR	6,7%	3,9%	6,8%	50,2%
Enhanced PaddleOCR	2%	1,2%	2,3%	20,7%

LLMs in OCR corrections

- **Best OCR result**

Pöytäkirja tehty Mikle.
lin kaupungin Foliisilaitosta Vali
aikaisesti
Jrjestyneen
Työvlen puo-
lesta valvomaan Valitun
JLrjes tye-
toimikunnan
kokouksessa Foliisi lai
toksella
marraskuun 20 ptivanf
1917 klo10 an,

- **Could LLM(s) be used to fix the rest..?**
  - Copilot: Pöytäkirja tehty Mikkelin kaupungin Poliisilaitosta vali- aikaisesti järjestyneen Työväen puolesta valvomaan valitun järjestyneen työtoimikunnan kokouksessa Poliisilaitoksella marraskuun 20. päivänä 1917 klo 10 aam.
  - Gemini: Pöytäkirja tehty Mikkelin kaupungin Poliisilaitoksella aikaisesti Järjestyksen Ylläpitäjän puolesta valvomaan Valitun Järjestys-toimikunnan kokouksessa Poliisilaitoksella marraskuun 20 päivänä 1917 klo 10 a.m.
- There are great open LLMs also for Finnish & over 100 other languages

Translations

- Huggingface: Helsinki-NLP/opus-mt-fi-en
    - Minutes made at the meeting of the elected organised working committee at the Police Department on 20 November 1917 at 10 a.m.
- CoPilot:
    - Minutes made at the Police Station of the city of Mikkeli on behalf of the temporarily organized Workers' Committee meeting at the Police Station on November 20, 1917, at 10 a.m.
- Gemini:
    - Minutes of the meeting of the organized labor committee elected to oversee the interests of the working class, held at the Mikkeli City Police Station on November 20, 1917, at 10 AM.
- Google translate:
    - The minutes were made on behalf of the labor committee organized by election from the Police Department of the city of Mikkeli to oversee the meeting of the labor committee at the Police Department on November 20, 1917 at 10 a.m.

Keywords from text

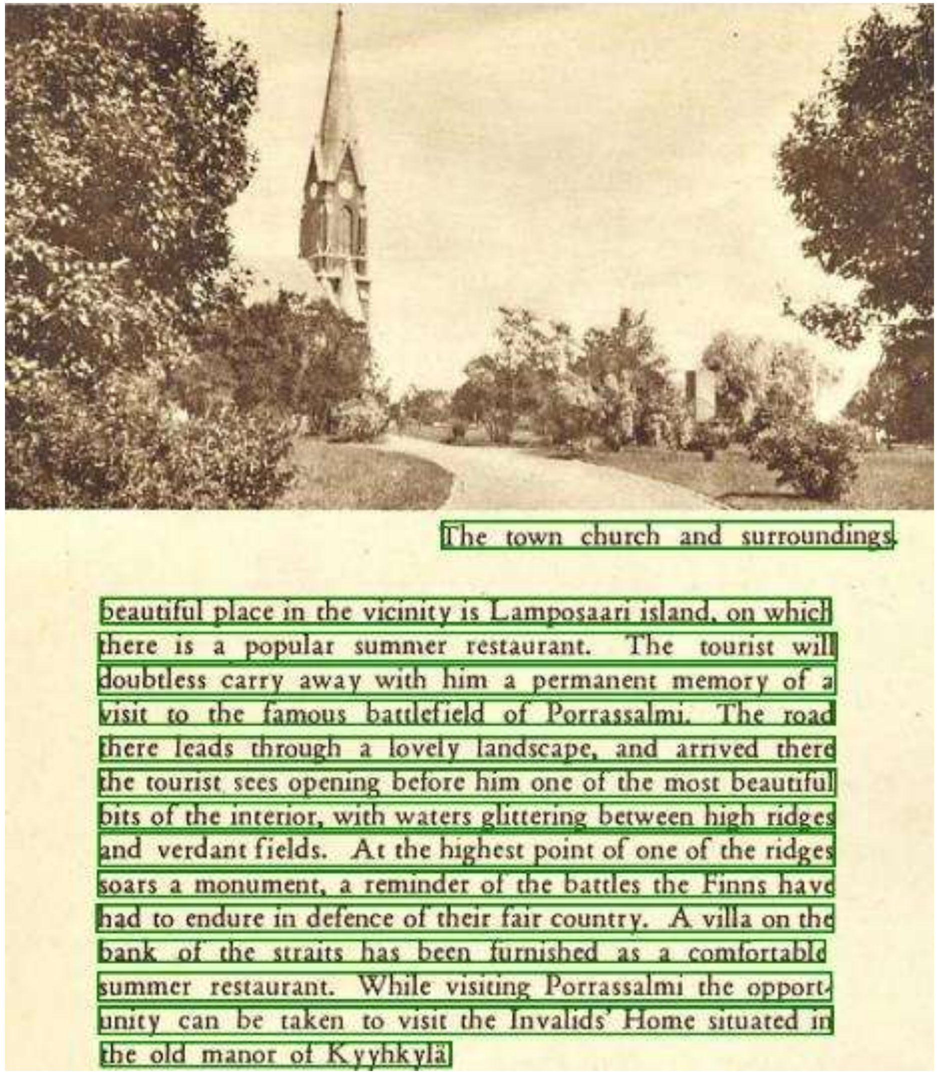
- Finnish National Library (Annif tool, https://annif.org/) embedded into https://arkkiivi.memorylab.fi/

	Original OCR	PaddleOCR	LLM fixed	Eng translated
Index terms	tyrni, idut, suomen kieli, toimikunnat, sähkökitara, laboratoriotekniikka, terveysvaikutukset, paneelit, viljely, siirtolapuutarhat	pöytäkirjat, kokoukset, historia, järjestöt, toimikunnat, hallinto, lehtialaindeksi, poliisi (organisaatiot), kaupungit, historiikit	historia, pöytäkirjat, poliisilaitokset, poliisi (organisaatiot), historiikit, poliisihallinto, kaupungit, kokoukset, lainsäädäntö, rikokset	police (organisations), towns and cities, history, police stations, commissions (organs), police (occupations), local histories (literary works), workers, meetings (arranged events), guides (literary works)
Lang	fi	fi	fi	en
Date		marraskuun 20 ptivanf 1917	marraskuun 20. päivänä	November 20, 1917

Summaries & classification / tagging

- In addition to keywords, LLMs support a lot more tasks
- Public demonstration: https://memorylab.fi/AIKO/cluetail-demo/
- Tested with the translated text
  - Similarity of two translations: 70.96%
  - Tone: neutral : 95.91%, positive : 2.81%, negative : 1.29%
  - Summary: Minutes made at the Police Station of the city of Mikkeli on behalf of the temporarily organized Workers' Committee.
  - Tags: Mikkeli, : 99.36%, November 20 1917, : 99.02%, Workers' Committee meeting, : 98.96% 10 a.m. : 98.69% police station, : 94.81%

Segmentation / Image recognition & classification



The town church and surroundings.

beautiful place in the vicinity is Lamposaari island, on which there is a popular summer restaurant. The tourist will doubtless carry away with him a permanent memory of a visit to the famous battlefield of Porrassalmi. The road there leads through a lovely landscape, and arrived there the tourist sees opening before him one of the most beautiful bits of the interior, with waters glittering between high ridges and verdant fields. At the highest point of one of the ridges soars a monument, a reminder of the battles the Finns have had to endure in defence of their fair country. A villa on the bank of the straits has been furnished as a comfortable summer restaurant. While visiting Porrassalmi the opportunity can be taken to visit the Invalids' Home situated in the old manor of Kyyhkylä.

- Huggingface:
  - **microsoft/resnet-50:** church, church building 34,5%
  - **hustvl/yolos-small:** Clock 73,8%
  - **nvidia/segformer-b0-finetuned-ade-512-512:** wall, building, sky, tree, earth, plant, painting, signboard, clock

Extending information

The town church and surroundings

beautiful place in the vicinity is **Lamposaari** island, on which there is a popular sumnner restaurant. The tourist will doubtless carry away with him a permanent memory of a visit to the famous **battlefield of Porrassalmi**. The road there leads through a lovely landscape, and arrived there the tourist sees opening before him one of the most beautiful bits of the interior, with waters glittering between high ridges and verdant fields. At the highest point of one of the ridges soars a monument, a reminder of the Bateles the Finns have had to endure in defence of their fair country. A villa on the bank of the straits has been furnished as a comfortable summer restaurant. While visiting Porrassalmi the opport unity can be taken to visit the Invalids' Home situated in the old **manor of Kyyhkylä.**

Copilot: Prompt Read and understand this text and extend the historical knowledge of the text topic: <text>

Lamposaari is a picturesque island located near Mikkeli, known for its natural beauty and recreational opportunities. The island features a popular summer restaurant, making it a favorite spot for both locals and tourists. The serene environment and scenic views of Lake Saimaa add to its charm.

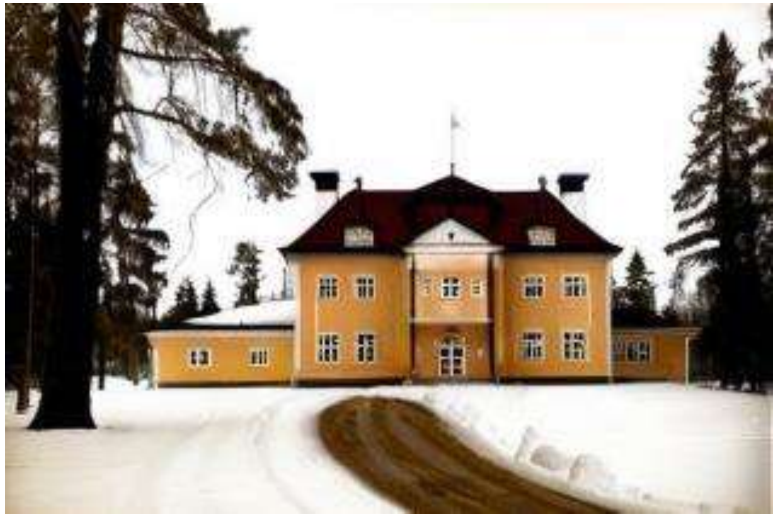
The Battle of Porrassalmi, fought on June 13, 1789, during the Russo-Swedish War, is a significant historical event for Mikkeli. A Swedish force of about 750 men successfully defended against a much larger Russian force of 5,000 men. The battlefield is now a historical site with a monument commemorating the bravery and sacrifices of the Finnish soldiers. The area offers stunning views of the surrounding landscape, with glittering waters and verdant fields, making it a memorable visit for history enthusiasts and nature lovers alike.

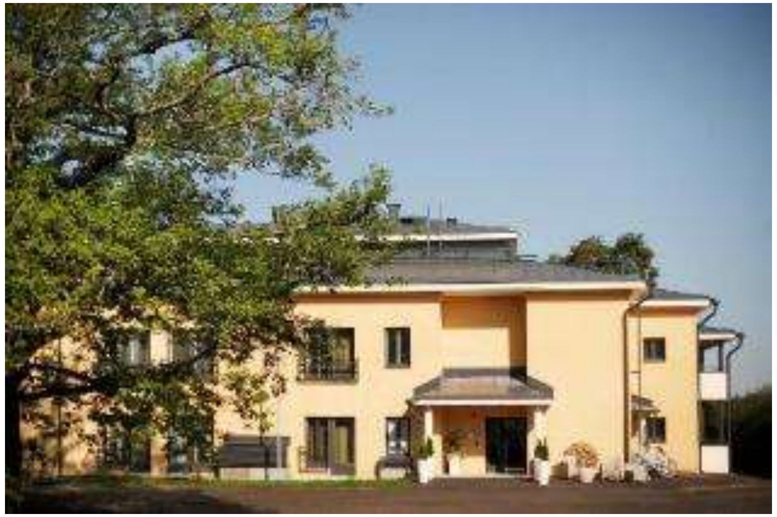
Kyyhkylä Manor, established in the 1630s, has a rich history and has served various purposes over the centuries. It was originally formed from two neighboring houses and has been owned by several notable figures, including Otto Carl von Fieandt. Today, it functions as a hotel and wellness resort, offering a blend of historical charm and modern amenities. The manor’s beautiful surroundings and historical significance make it a unique destination for visitors.

Visualizing information (Stable diffusion)

Which one is real Kyyhkylä Manor?

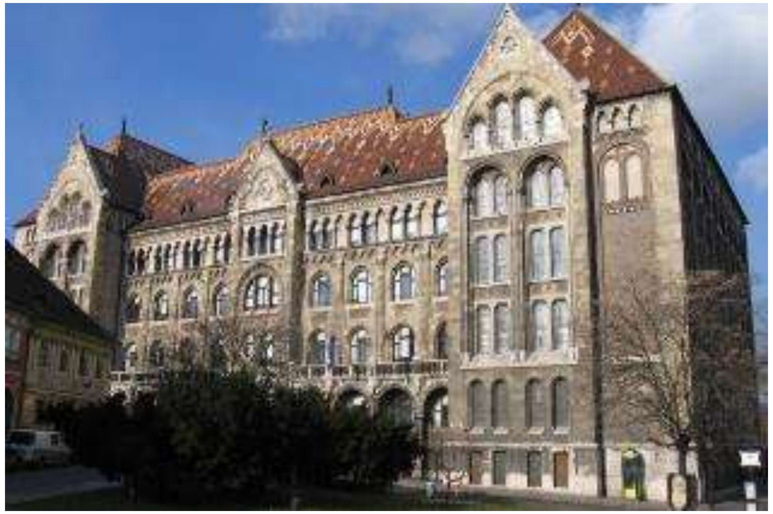
Is it this one

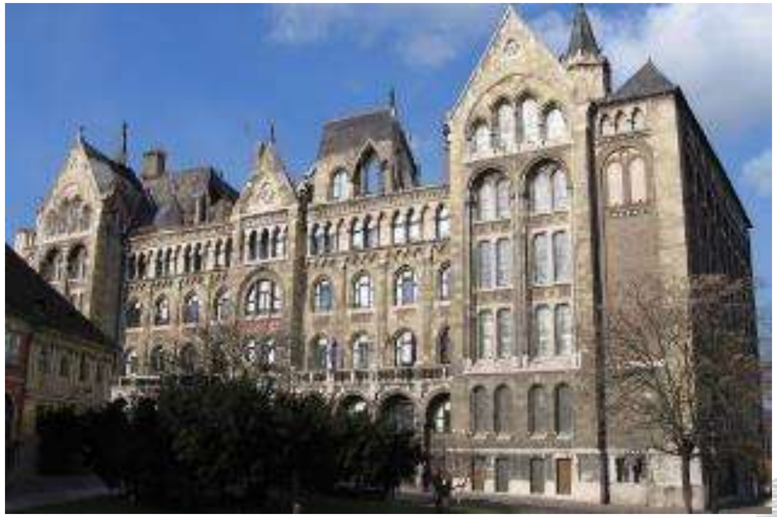




Or maybe this one

Original





More medieval..?

By József Rozsnyai - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=30768133

Thank you

Contact:
Anssi.jaaskelainen@xamk.fi
LinkedIn
Digitalia.fi / memorylab.fi

Image by: Stable Diffusion
Prompt: bump mapped metallic 3d ultra detailed outer shining question mark standing on a fractured black marble rock podium. Two invisible spotlights cause deep shadows:
Details: Steps: 150, Sampler: DPM++ 2M, Schedule type: Karras, CFG scale: 13, Seed: 3193114130, Size: 1024x1024, Model hash: 33e58e8668,
Model: juggernautXL_juggXIByRundiffusion, Version: v1.9.4

European Commission