CORE147

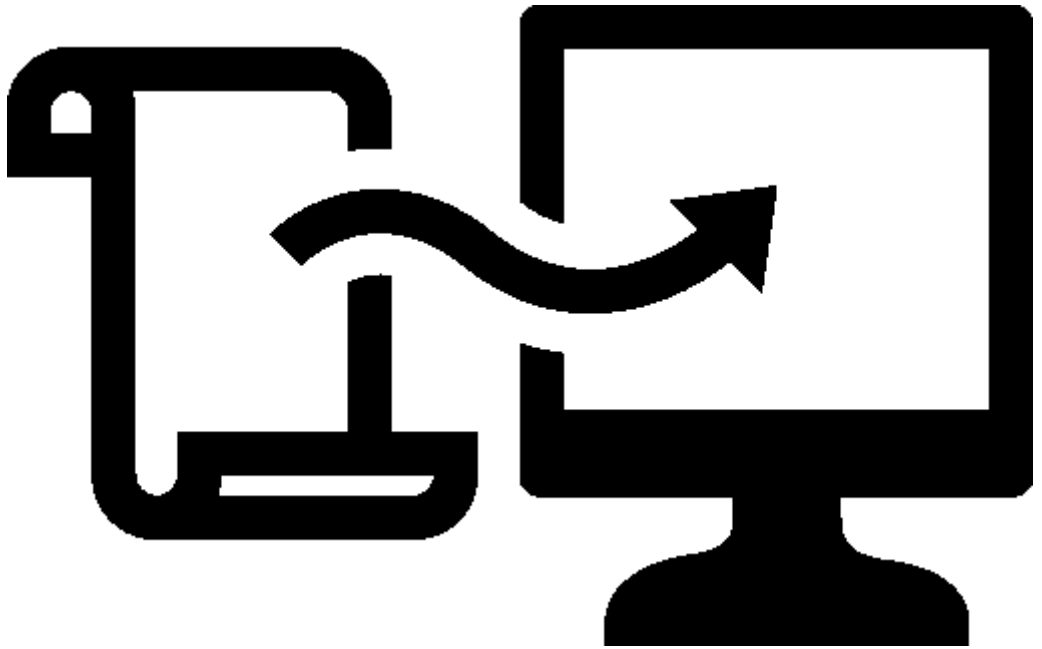
DIGITÁLNÍ DATA V HUMANITNÍCH A SOCIÁLNÍCH

Zuzana Nevěřilová
Hana Žižková

2024/25

DIGITÁLNÍ DOKUMENTY

- Můžeme je zobrazit na displeji
- Mohou mít různou kvalitu z hlediska počítačového zpracování
  - Můžeme dokument měnit?
  - Potřebujeme speciální program?
  - Používá program proprietární (firemní) standard?
  - Bude možné dokument použít v jiné verzi programu?
  - Má program alternativu?



https://thenounproject.com/icon/digitization-21031/

DIGITÁLNÍ DOKUMENTY A DATA

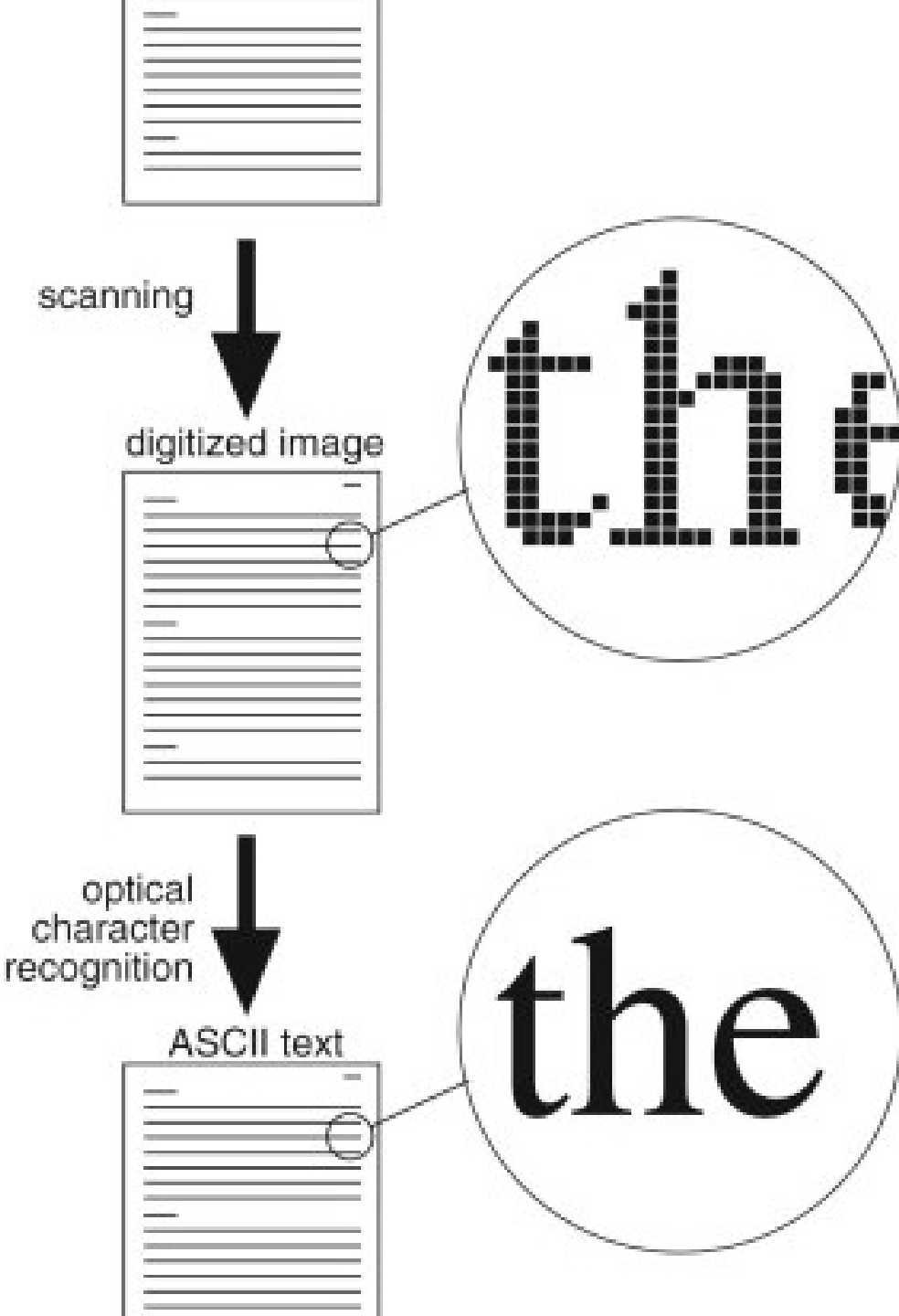
- Digitální dokument je binární soubor
  - Informace **v bitech a bajtech**
  - Obrázek
    - Rozlišení obrázku
  - Barevné spektrum obrázku
  - Popis obrázku
  - Přepis textu
  - Hypertext
- Digitální dokument obsahuje metadata
  - Kdo a kdy jej vytvořil
  - Co obsahuje (nalezitelnost)

https://www.sydle.com/blog/document-digitization-61b8e03c876cf6271dfbe88a

DIGITALIZACE

DIGITIZATION | DIGITALIZATION

- Digitální dokumenty
  - Digitalizované (z analogových)
  - Born-digital
- Technologie
  - Skener
  - Fotoaparát
  - …
- Surová („raw”) data
  - obohacená metadaty



https://www.sciencedirect.com/topics/computer-science/digitization-project

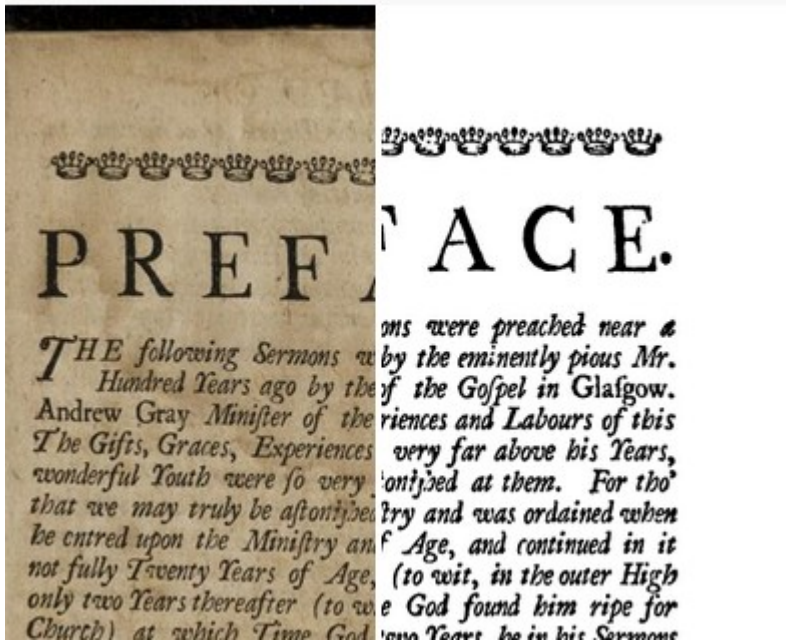
DIGITALIZACE
DIGITIZATION

Fyzický objekt

PREFACE

The following Sermons were preached near a Hundred Years ago by the eminently pious Mr. Andrew Gray Minister of the Gospel in Glasgow. The Gifts, Graces, Experiences and Labours of this wonderful Youth were …

PŘEDZPRACOVÁNÍ
PREPROCESSING



OPTICKÉ ROZPOZNÁVÁNÍ ZNAKŮ
OPTICAL CHARACTER RECOGNITION

1. Rozpoznání oblasti s textem
2. Rozpoznání jednotlivých tvarů znaků
3. Pravděpodobnost konkrétního znaku v kontextu ostatních znaků
4. Uspořádání textů do bloků

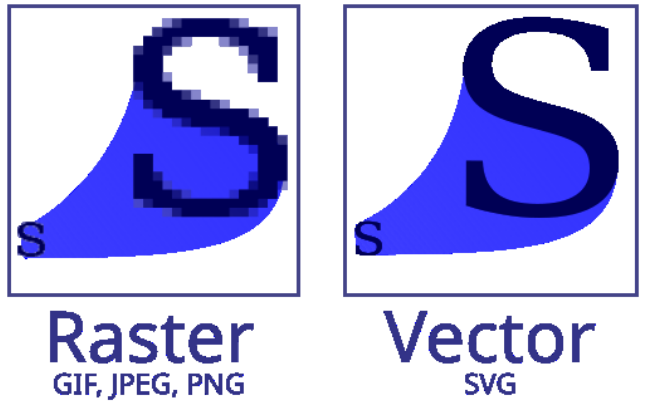
OPTICKÉ ROZPOZNÁVÁNÍ ZNAKŮ
OPTICAL CHARACTER RECOGNITION

1. Rozpoznání oblasti s textem
2. Rozpoznání jednotlivých tvarů znaků
3. Pravděpodobnost konkrétního znaku v kontextu ostatních znaků
4. Uspořádání textů do bloků

Problémy: šum, kvalita výchozího snímku, mezery, bloky, netextové prvky
Problémy s písmeny: různá písma, různé velikosti, různé jazyky a jejich specifické znaky, překlepy, ...

DIGITALIZACE: VSTUP A VÝSTUP

- Vstup: bitmapový obrázek (PNG, JPG, TIFF, …)
- Výstup: vícevrstvé (multilayered) PDF



ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM

By A. M. TURING.

[Received 28 May, 1936.—Read 12 November, 1936.]

The "computable" numbers may be described briefly as the real

https://commons.wikimedia.org/wiki/File:Bitmap_VS_SVG.svg

DIGITALIZAČNÍ SOFTWARE

- Předzpracování
  - Otočení, narovnání, odstranění šumu
- Obecné OCR
- Speciální OCR
- Proces digitalizace (digitalization)
  - Nastavení pro hromadné zpracování
  - Speciální skenery

- Předzpracování: BIQE, Book Restorer, grafický software (ImageMagick)
- Obecné OCR: ABBY, InftyReader, Tesseract, Google Vision (Cloud service)
- Speciální OCR: Kraken, Infty Project, Pero

ANALOGOVÉ ↔ DIGITÁLNÍ

- Analogové dokumenty: fyzické artefakty (nepřístupné)
- Digitální dokumenty
  - retro-digital documents: kopie analogových dokumentů (sken, OCR)
  - born-digital documents: digitální objekty vytvořené jako digitální objekty (z ničeho fyzického)
  - retro-born-digital documents: digitální dokumenty před érou digitálních knihoven (bez metadat, hypertextu atd.)

Digitalizační centrum Knihovny AV ČR v.v.i.

https://digit.lib.cas.cz/

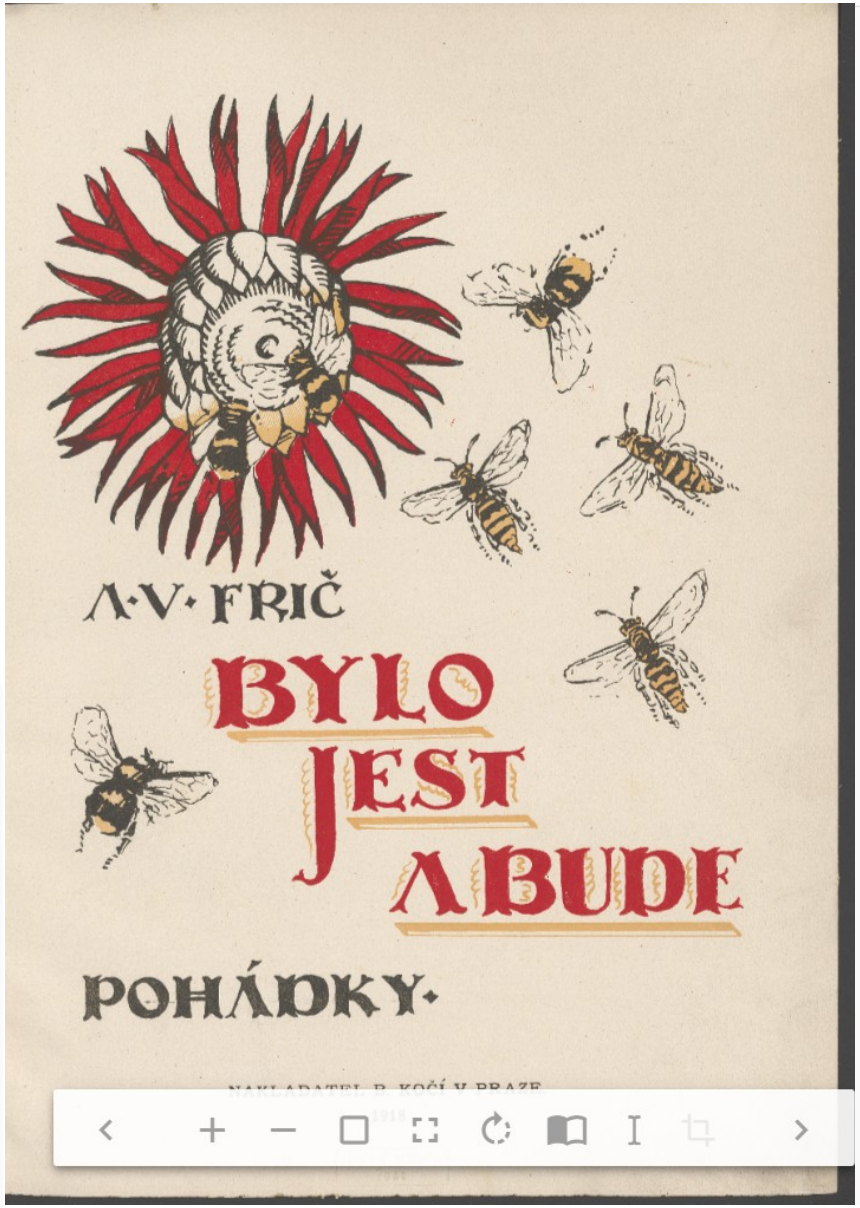
Projekt Kramerius

Jenštejnské kroniky

Navigace

Digitalizace

Metadata



Bylo jest a bude
pohádky

Author
Frič, Alberto Vojtěch

Publishing details
V Praze: B. Kočí, 1918

Document type
Book

Keywords
Česká poezie
Česká literatura

Genre
Poezie

Language
Czech

Location
Library of the ASCR
Shelf/Call number: UC 84

ČESKÉ DIGITALIZOVANÉ ZDROJE

- Vokabulář webový: https://vokabular.ujc.cas.cz/
- Samizdat a exilová literatura: https://scriptum.cz/
- Kramářské tisky: https://www.spalicek.net/
- Heraldická knihovna: http://www.historie.hranet.cz/heraldika/
- České archivy https://digi.ceskearchivy.cz/
- AHISTO https://nlp.fi.muni.cz/projekty/ahisto/portal/

DALŠÍ DIGITALIZOVANÉ ZDROJE

- Shakespeare Electronic Archive https://shea.mit.edu/shakespeare/
- Manuscriptorium https://www.manuscriptorium.com/
- ANNO Historische Zeitungen und Zeitschriften https://anno.onb.ac.at/
- ÖNB Digital https://onb.digital/
- DiFMOE https://www.difmoe.eu/
- DigiPress https://digipress.digitale-sammlungen.de/
- Śląska Biblioteka Cyfrowa https://www.sbc.org.pl/dlibra
- Europeana https://www.europeana.eu/en

- Netexty: https://compositor.bham.ac.uk/ - databáze ornamentů

INFRASTRUKTURY PRO DIGITALIZACI A ARCHIVACI

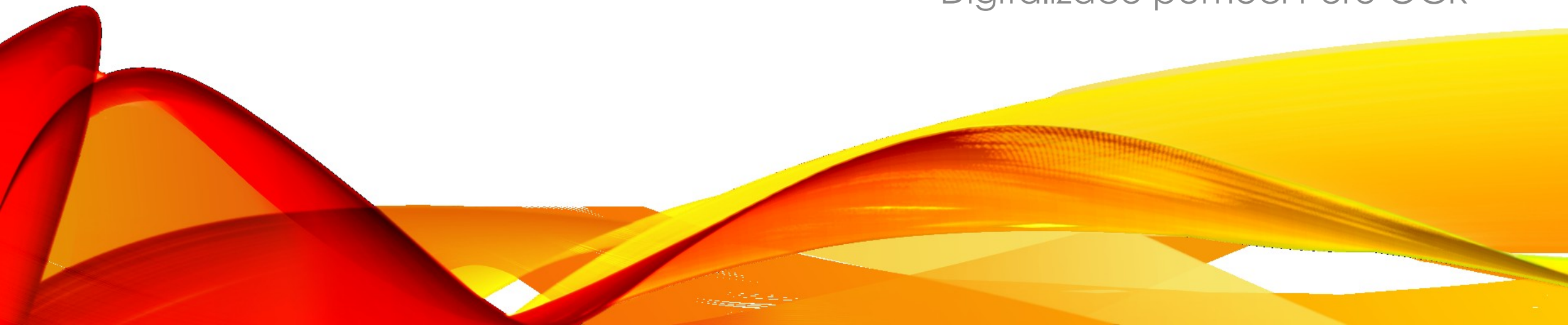
- Národní infrastruktura: http://www.digitalniknihovna.cz/
  - Národní digitální knihovna
  - Česká digitální knihovna
  - Registr digitalizace https://registrdigitalizace.cz/

- Europeana
- World Digital Library
- Internet Archive https://archive.org/search
- Speciální infrastruktury: EuDML, DML (digital mathematics library)

PRAKTICKÁ UKÁZKA

Digitalizace pomocí Google Cloud Vision

Digitalizace pomocí Pero OCR



DALŠÍ ČTENÍ

- Digitization Project. In subject area: Computer Science. Science Direct. <https://www.sciencedirect.com/topics/computer-science/digitization-project>
- SAP: Digitization vs digitalization. SAP. <https://www.sap.com/cz/products/erp/digitization-vs-digitalization.html>