CONTINUOUS INTEGRATION FOR OCR-D IN HPC

DevOps Strategies in HPC

By: Abdallah Abdelnaby
Supervisor: Giorgi Mamulashvili

Agenda

- Introduction
- Literature
- Project Objectives
- OCR-D
- Project Key Components and Structure
- Results
- Conclusion and Future Work

Inroduction

What is DevOps?

Collaborative culture and set of practices to improve software development and operations

- **Key Features:**
  - Continuous Integration/Continuous Delivery (CI/CD)
  - Infrastructure as Code (IaC)
  - Automation and Monitoring
- **Benefits**:
  - Faster development cycles
  - Improved collaboration
  - Increased reliability

Introduction

What is HPC?

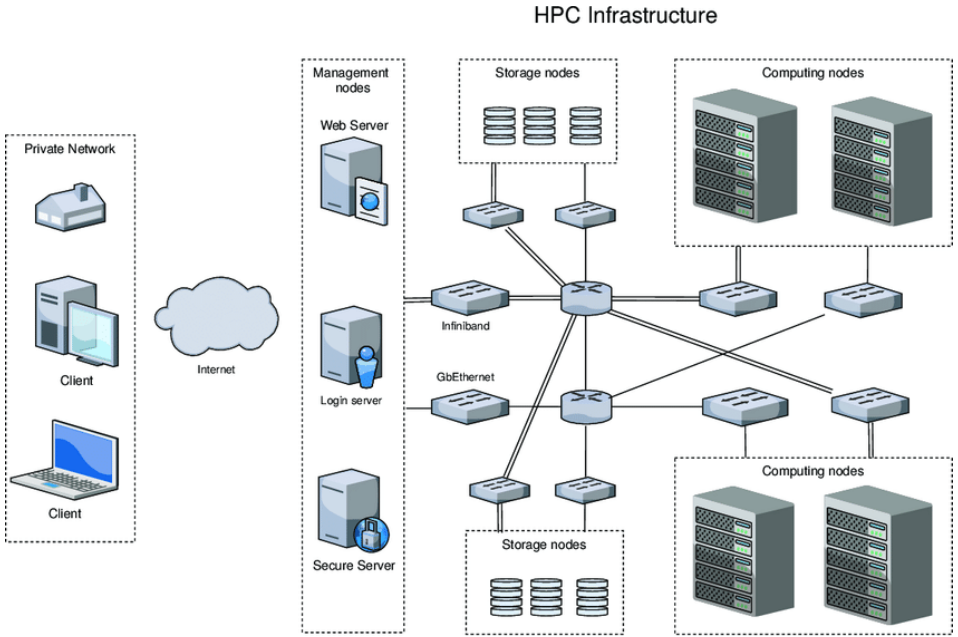


Figure 1: (2020) Timing Predictability in High-Performance Computing With Probabilistic Real-Time

Literature

- Virtualization and Containers
- Git Repositories
- CI/CD Engines
- Integration Examples:
  - Centralized Pipelines
  - Decentralized Pipelines

(2023) Leveraging DevOps for Scientific Computing

Virtualization and Containers

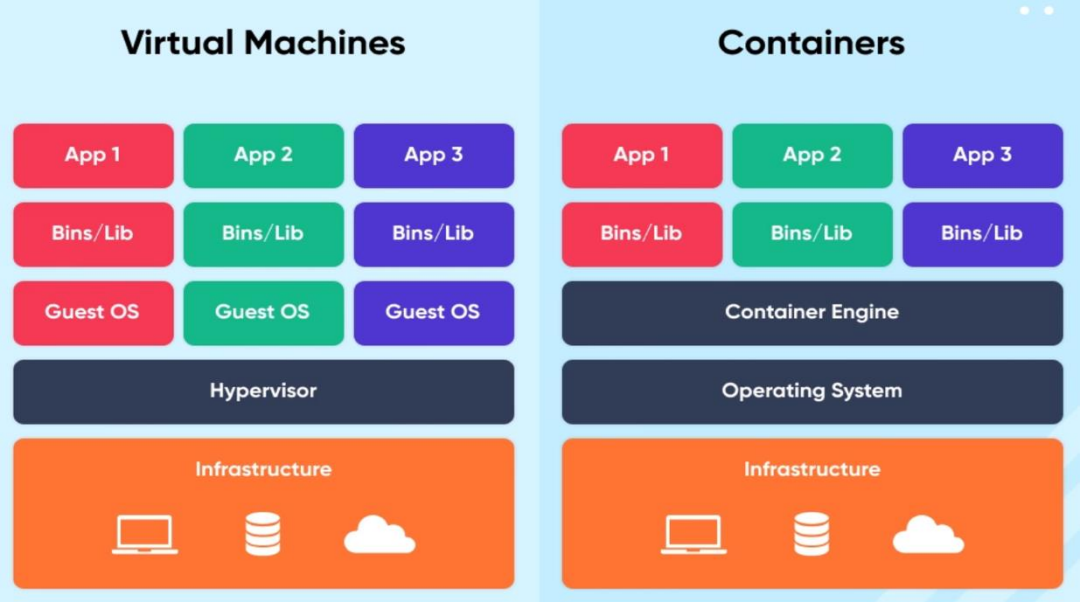


Figure 2: (2023) Net Solutions

Git Repositories

- Version control for code and documentation
- History tracking and collaboration
- Simplifies code synchronization
- Enables cryptographically secure attestations (RFC3161/RFC5816)

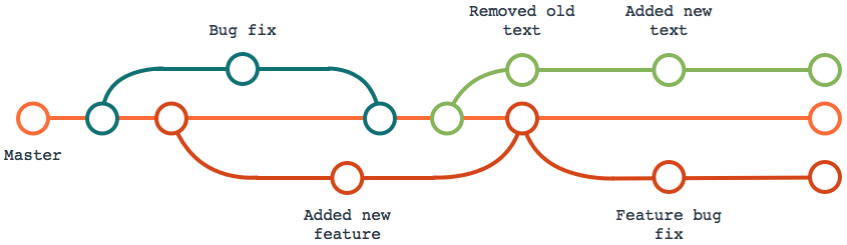


Figure 3: (2018) CPanel

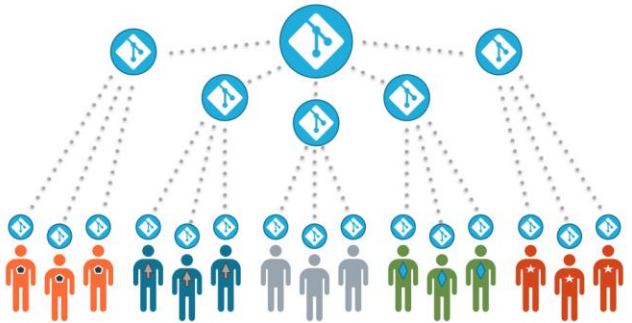


Figure 4: (2018) CPanel

CI/CD Engines

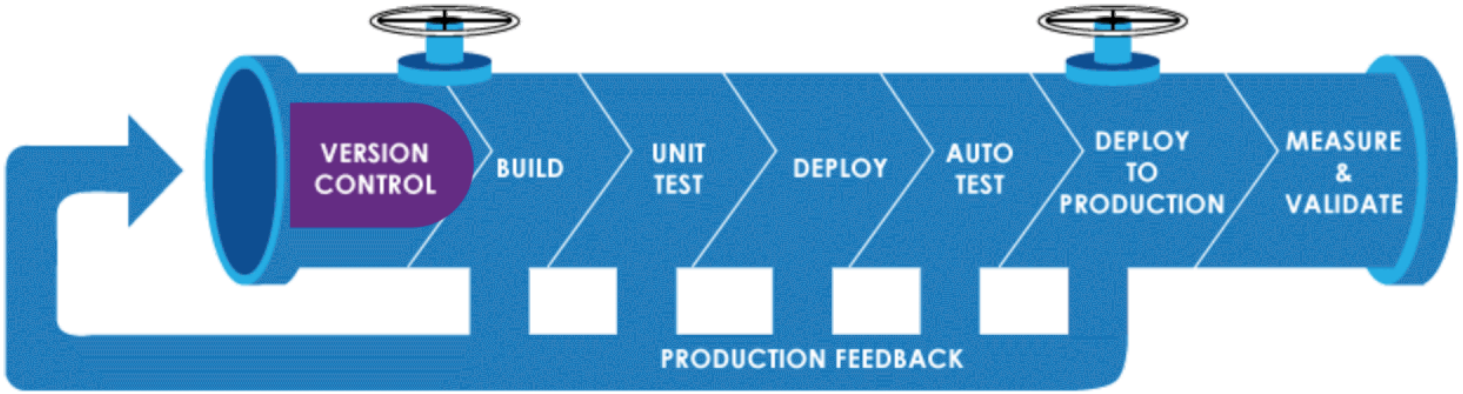


Figure 5: (2023) LinkedIn

Integration Examples

Centralized Pipeline

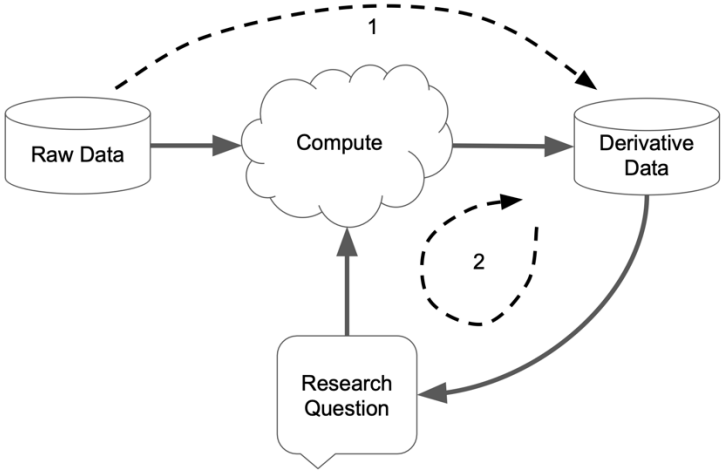


Figure 6: (2023) Leveraging DevOps for Scientific Computing

Decentralized Pipeline

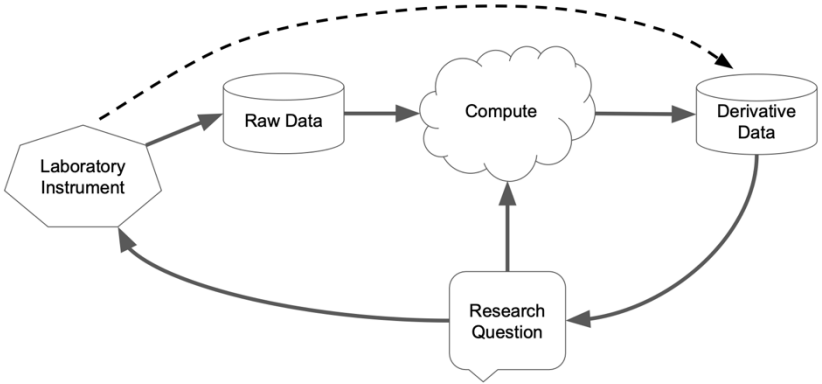


Figure 7: (2023) Leveraging DevOps for Scientific Computing

Project Objectives

- Provide a continuous integration for OCR-D in HPC
- Be able to run OCR-D workflows on HPC through the CI/CD Pipeline

But,
What is OCR-D? and what are the OCR-D workflows?!

OCR-D

- DFG-Funded Initiative for Optical Character Recognition Development
- https://ocr-d.de/

- Mainly used to make digital copies for old documents and manuscripts

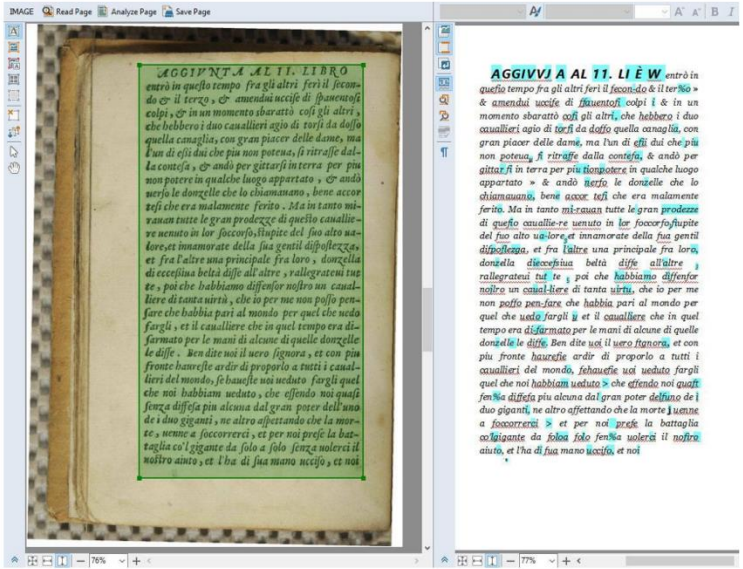


Figure 8: (2016). Early printed edition and OCR techniques

OCR-D Processors and Workflows

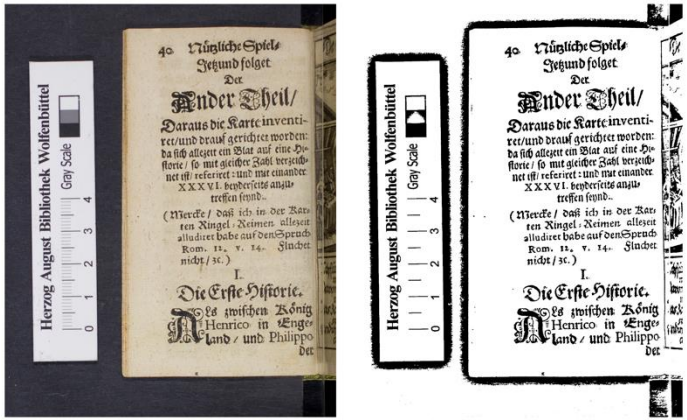


Figure 9: OCR-D Binarization

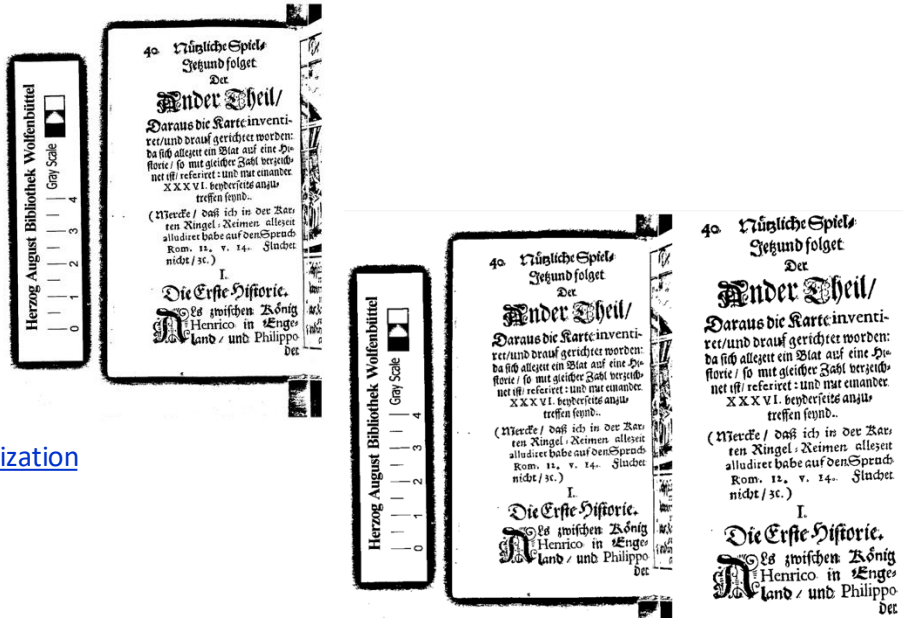


Figure 10: OCR-D Cropping

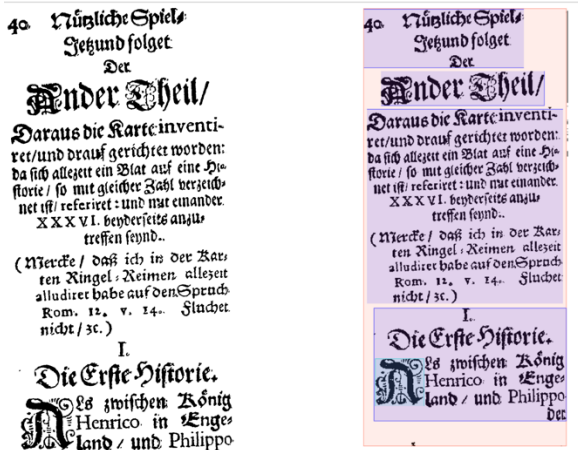


Figure 11: OCR-D Region Segmentation

Key Components

- GWDG HPC
- Containerization System
  - Singularity
- Git Repositories
  - Gitlab
- CI/CD Pipeline
- GitLab Runner

Project Structure

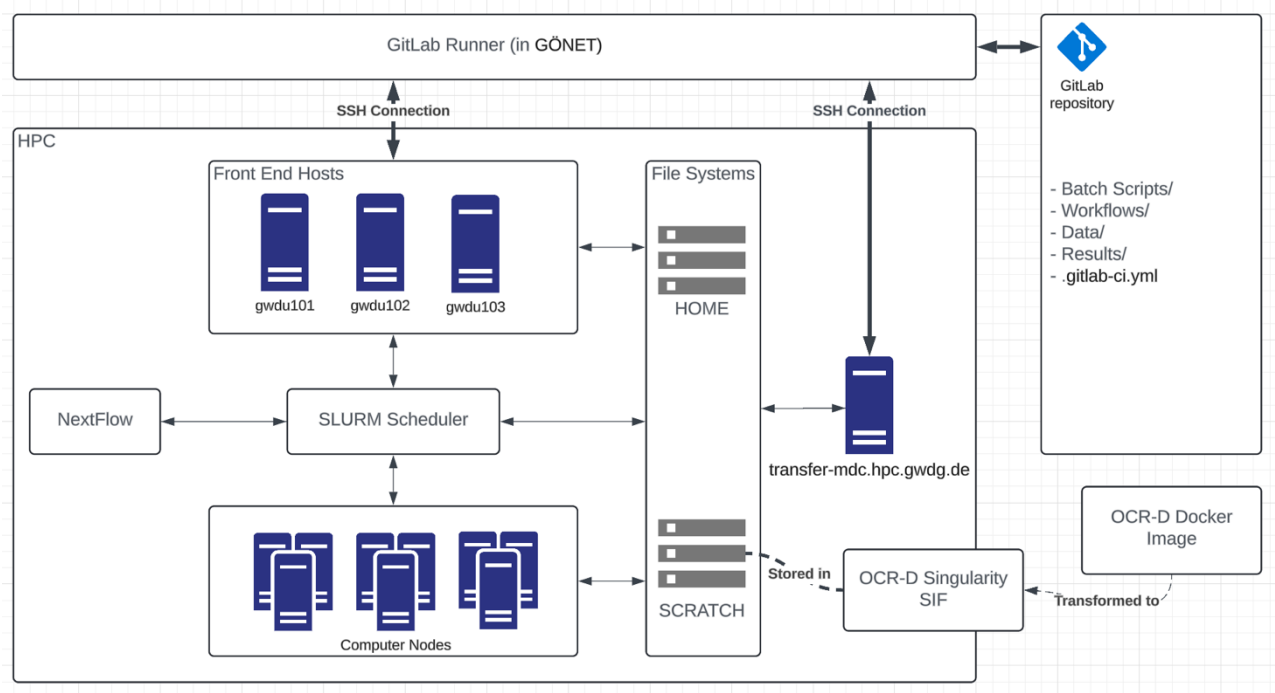


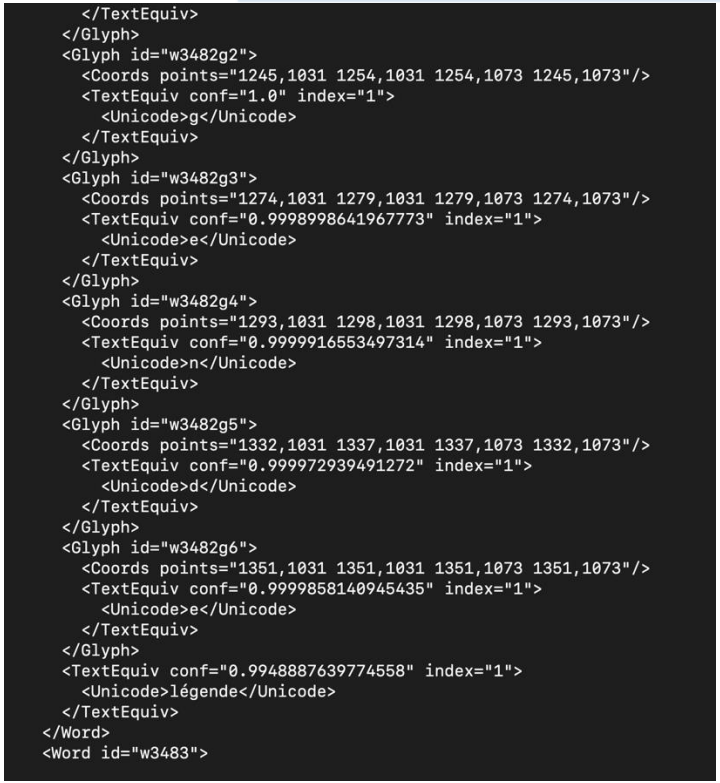
Figure 12

CI/CD Pipeline Stages

1. Connect to HPC (on any tag)
2. Upload data, workflow, the batch scripts to HPC (on any tag)
3. Pull OCR-D Docker image and Create a Singularity SIF (on build tag)
4. Download OCR-D Models (needs ”3”)
5. Submit the OCR Job (on {workflow_name} tag)
6. Retrieve the results to repo/ when available (needs “3”)
7. Cleanup ➔delete the work from SCRATCH or move it to HOME and disconnect (needs “4”)

Results

- Page XML (Harldy Readable)
- Solution: LAREX

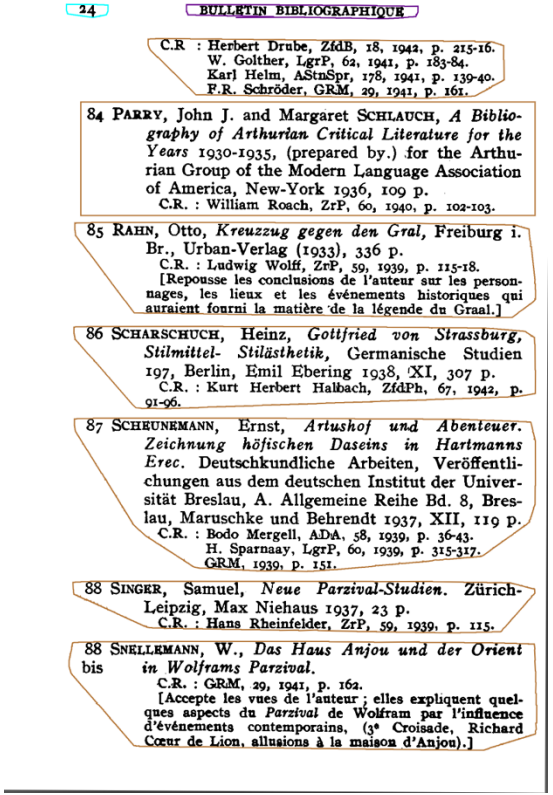


```
      </TextEquiv>
    </Glyph>
    <Glyph id="w3482g2">
      <Coords points="1245,1031 1254,1031 1254,1073 1245,1073"/>
      <TextEquiv conf="1.0" index="1">
        <Unicode>g</Unicode>
      </TextEquiv>
    </Glyph>
    <Glyph id="w3482g3">
      <Coords points="1274,1031 1279,1031 1279,1073 1274,1073"/>
      <TextEquiv conf="0.9998998641967773" index="1">
        <Unicode>e</Unicode>
      </TextEquiv>
    </Glyph>
    <Glyph id="w3482g4">
      <Coords points="1293,1031 1298,1031 1298,1073 1293,1073"/>
      <TextEquiv conf="0.9999916553497314" index="1">
        <Unicode>n</Unicode>
      </TextEquiv>
    </Glyph>
    <Glyph id="w3482g5">
      <Coords points="1332,1031 1337,1031 1337,1073 1332,1073"/>
      <TextEquiv conf="0.999972939491272" index="1">
        <Unicode>d</Unicode>
      </TextEquiv>
    </Glyph>
    <Glyph id="w3482g6">
      <Coords points="1351,1031 1351,1031 1351,1073 1351,1073"/>
      <TextEquiv conf="0.9999858140945435" index="1">
        <Unicode>e</Unicode>
      </TextEquiv>
    </Glyph>
    <TextEquiv conf="0.9948887639774558" index="1">
      <Unicode>légende</Unicode>
    </TextEquiv>
  </Word>
  <Word id="w3483">
```

Figure 13

Results

Segmentation Results



24 BULLETIN BIBLIOGRAPHIQUE

C.R : Herbert Drube, ZfdB, 18, 1942, p. 215-16.
W. Golther, LgrP, 62, 1941, p. 183-84.
Karl Helm, AStnSpr, 178, 1941, p. 139-40.
F.R. Schröder, GRM, 29, 1941, p. 161.

84 PARRY, John J. and Margaret SCHLAUCH, *A Bibliography of Arthurian Critical Literature for the Years* 1930-1935, (prepared by.) for the Arthurian Group of the Modern Language Association of America, New-York 1936, 109 p.
C.R. : William Roach, ZrP, 60, 1940, p. 102-103.

85 RAHN, Otto, *Kreuzzug gegen den Gral*, Freiburg i. Br., Urban-Verlag (1933), 336 p.
C.R. : Ludwig Wolff, ZrP, 59, 1939, p. 115-18.
[Repousse les conclusions de l'auteur sur les personnages, les lieux et les événements historiques qui auraient fourni la matière de la légende du Graal.]

86 SCHARSCHUCH, Heinz, *Gottfried von Strassburg, Stilmittel- Stilästhetik*, Germanische Studien 197, Berlin, Emil Ebering 1938, XI, 307 p.
C.R. : Kurt Herbert Halbach, ZfdPh, 67, 1942, p. 91-96.

87 SCHEUNEMANN, Ernst, *Artushof und Abenteuer. Zeichnung höfischen Daseins in Hartmanns Erec*. Deutschkundliche Arbeiten, Veröffentlichungen aus dem deutschen Institut der Universität Breslau, A. Allgemeine Reihe Bd. 8, Breslau, Maruschke und Behrendt 1937, XII, 119 p.
C.R. : Bodo Mergell, ADA, 58, 1939, p. 36-43.
H. Sparnaay, LgrP, 60, 1939, p. 315-317.
GRM, 1939, p. 151.

88 SINGER, Samuel, *Neue Parzival-Studien*. Zürich-Leipzig, Max Niehaus 1937, 23 p.
C.R. : Hans Rheinfelder, ZrP, 59, 1939, p. 115.

88 bis SNELLEMANN, W., *Das Haus Anjou und der Orient in Wolframs Parzival*.
C.R. : GRM, 29, 1941, p. 162.
[Accepte les vues de l'auteur ; elles expliquent quelques aspects du *Parzival* de Wolfram par l'influence d'événements contemporains, (3e Croisade, Richard Cœur de Lion, allusions à la maison d'Anjou).]

Figure 14

Results

Line Detection Results

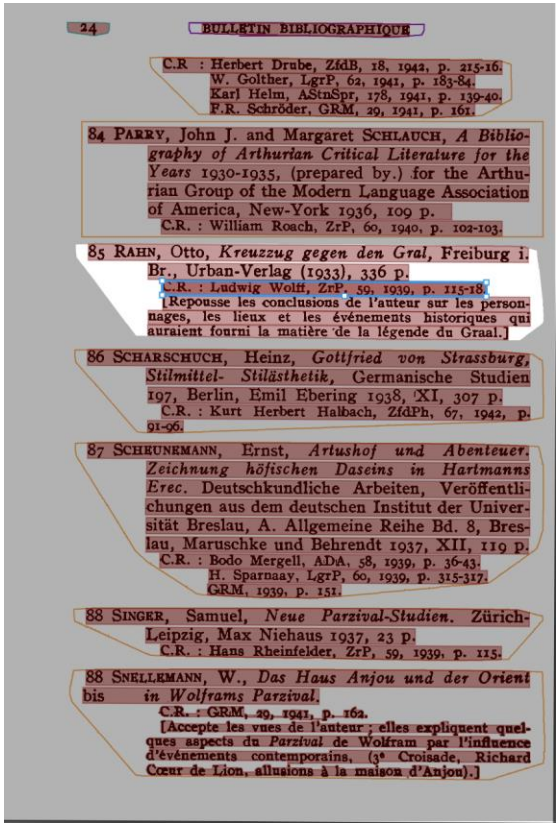


Figure 15

Results

Text Recognition Results

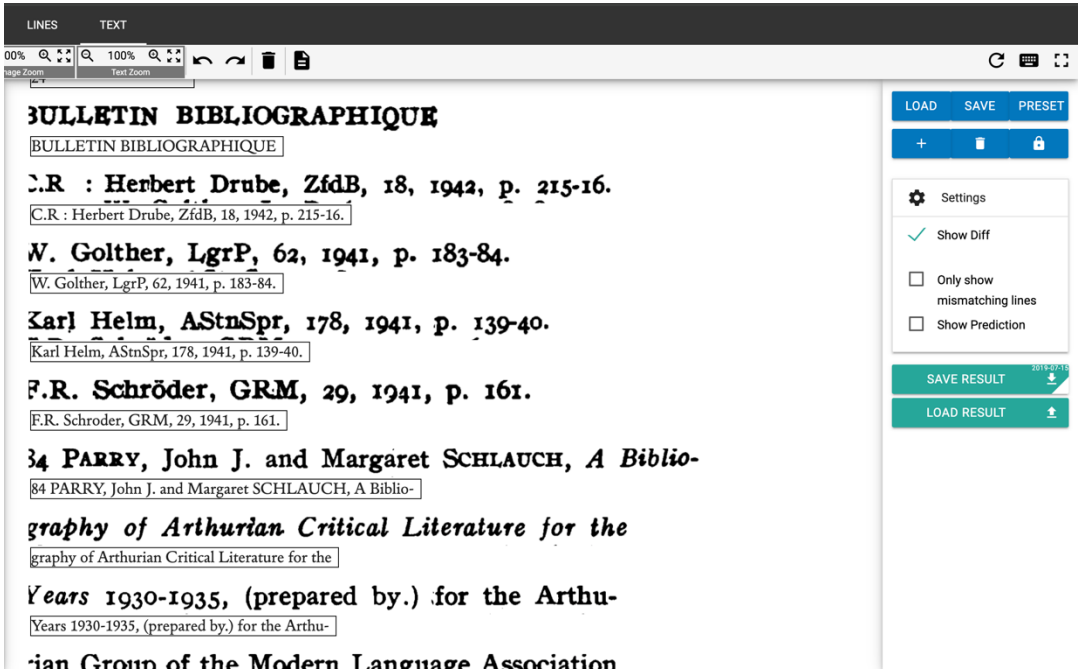


Figure 16

Results

Processing speed

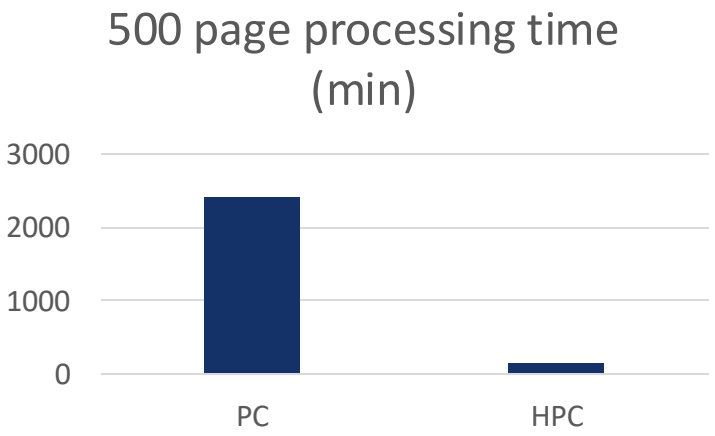


Figure 17

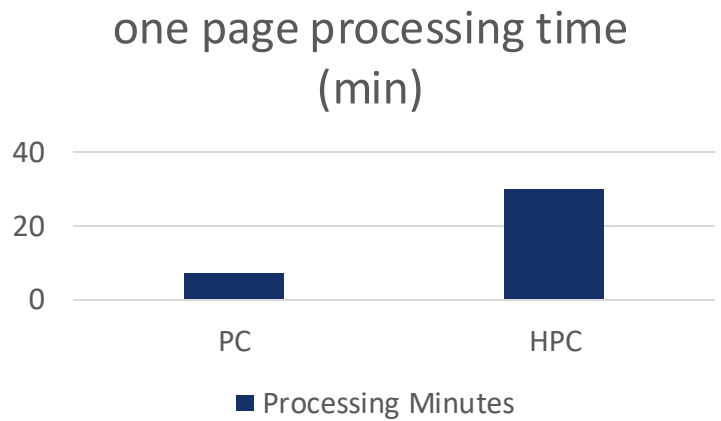


Figure 18

Conclusion and Future Work

- Use the GWDG proxy server
- Optimize the project to run parallel pipelines
- Deploy LAREX on a server and upload the results directly to LAREX

References

- Reghenzani, Federico & Massari, Giuseppe & Fornaciari, William. (2020). Timing Predictability in High-Performance Computing With Probabilistic Real-Time. IEEE Access. 8. 10.1109/ACCESS.2020.3038559.
- Mancinelli, T. (2016). Early printed edition and OCR techniques: what is the state-of-art? Strategies to be developed from the working-progress Mambrino project work. Historias Fingidas, No. 4 (2016). https://doi.org/10.13136/2284-2667/65
- Sampedro, Zebula & Holt, Aaron & Hauser, Thomas. (2018). Continuous Integration and Delivery for HPC: Using Singularity and Jenkins. 1-6. 10.1145/3219104.3219147.
- https://ocr-d.de/en/workflows
- https://github.com/subugoe/operandi/tree/main
- https://github.com/OCR4all/LAREX
- https://hpc.guix.info/blog/2023/03/contiguous-integration-and-continuous-delivery-for-hpc/
- https://brelje.net/blog/devops-scientific-computing/
- https://docs.sylabs.io/guides/3.0/user-guide/installation.html
- https://hps.vi4io.org/_media/teaching/summer_term_2024/pchpc/clusterintro.pdf