

Universität für
Weiterbildung
Krems Österreichische
Nationalbibliothek TECHNISCHE
UNIVERSITÄT
WIEN universität
wien

di:'angewandte

 ÖAW
ÖSTERREICHISCHE
AKADEMIE DER
WISSENSCHAFTEN universität
innsbruck

Public Launch Event
12. Oktober 2026 – 10:00-11:30

Walter Scholger

Universität Graz – Institut für Digitale Geisteswissenschaften

DHinfra Public Launch Event

GPU-Cluster und große Modelle
für die Geistes- und
Kulturwissenschaften

Agenda

- 10:00 Begrüßung, Agenda
Was tun die (digitalen) **Geisteswissenschaften**?
- 10:10 Was ist DHinfra; was wurde 'gebaut'?
Was ist ein **Supercomputer**?
Was ist bei DHinfra anders?
- 10:30 Der **Einstieg**: die Konsole
Drei Wege auf den Cluster: Browser, Kommandozeile, API
- 10:50 **Use Cases**: eine Urkunde lesen, mit KI
Spielregeln am Cluster
- 11:15 DHinfra und die **Zukunft**; Q&A

Was tun die Geisteswissenschaften?

- (Digitale) Geisteswissenschaften == (Digital) Humanities
- Umfasst Sprachen, Geschichte, Literatur, Kunst, Musik, Philosophie etc.
- Gemeinsamer Nenner: **das kulturelle Erbe**

Was tun die digitalen Geisteswissenschaften?

Drei Aufgaben:

1. digitalisieren
2. bewahren
3. erforschen

DHinfra beschleunigt alle drei Aufgaben, am meisten die Forschung.

Was tun die digitalen Geisteswissenschaften?

- Digital Humanities seit drei Jahrzehnten in Graz
- Gewi-Lab → INIG → ZIM/ACDH → DDH

Siehe dh.uni-graz.at (Website)

Siehe gams.uni-graz.at (Repositorium)



CLARIAH-AT

- 11 österreichische Wissenschafts- und Kulturerbeinstitutionen
- CLARIAH verbindet CLARIN-ERIC und DARIAH-EU
- Österreichischer Repräsentant in europäischen ESFRI Forschungsinfrastrukturkonsortien (ERICs)
- Aufbau und Ausbau technischer und sozialer Infrastrukturen für die Geisteswissenschaften: Infrastrukturen - Daten - Methoden - Wissen



clariah.at



„SUPER-COMPUTER“

Millionen Zeitungsartikel in fünf Tagen erforschen

Mit der Hilfe von Künstlicher Intelligenz können Millionen Seiten historischer Zeitungen und Quellenangaben innerhalb kurzer Zeit durchforstet werden. Entwickelt wurde der "Super-Computer-Cluster" unter der Leitung der Universität Graz. Statt Jahrzehnte werden dafür nur fünf Tage benötigt.

7. Oktober 2026, 13.19 Uhr

Teilen



Siehe auch dinfra.at/impact

Beispiel: historische Stellenanzeigen



[historical-job-ads.uni-graz.at/en/about-the-project/
static.uni-graz.at/fileadmin/_files/_project_sites/_historical-job-ads/Bilder/CLARIN_job_ads.pdf](https://historical-job-ads.uni-graz.at/en/about-the-project/static.uni-graz.at/fileadmin/_files/_project_sites/_historical-job-ads/Bilder/CLARIN_job_ads.pdf)

Florian Atzenhofer-Baumgartner; David Fleischhacker
Universität Graz – Institut für Digitale Geisteswissenschaften

Was ist DHinfra; was wurde 'gebaut'?

Vergleiche DHinfra Abschlussevent

dhinfra.at/2026-06-08-dhinfra-final-event-results

Projektbedingungen

Förderprogramm: “(Digitale) Forschungsinfrastrukturen”

Fördergeber: BMFWF und EU Recovery & Resilience Facility (NextGen)

Förderziele: Beschaffung, Modernisierung, besserer Zugang

Infrastrukturprojekt – kein Forschungsprojekt



 Bundesministerium
Frauen, Wissenschaft
und Forschung

DHinfra in Zahlen

9 Institutionen

3,5 Jahre (2023 - 2026)

4 Arbeitsgruppen

8 Arbeitspakete

28 Meilensteine

50+ Mitglieder

100+ Assoziierte



di:'angewandte

DHinfra und CLARIAH-AT

- CLARIAH-AT als österreichischer Repräsentant in europäischen ESFRI Forschungsinfrastrukturkonsortien (ERICs)
- DHinfra **als Projekt und** nun **als Infrastruktur**
- Graz als CLARIN-B-Zentrum (GÉANT Hyper Connection POI)



Projektziele (1/2)

DHinfra schließt Lücken ...

- ... bei der Standard-Digitalisierung innerhalb von Bibliotheken, Archiven, Museen.
- ... von generischen Repositorien und Datenbanken.
- ... in Verfügbarkeit von Hochleistungsrechnern ('HPC') außerhalb der Natur-, Technik- und Lebenswissenschaften

DHinfra ermöglicht, ...

- ... **neue** und/oder **große Datenbestände** (an Kulturerbe) zu **speichern**
- ... diese Datenbestände mit **Maschinellern Lernen** (bzw. **KI**) zu verarbeiten.

Projektziele (2/2)

DHinfra bewerkstelligt das in 4 Bereichen:

- 1) Fortgeschrittene Digitalisierung
- 2) Fachspezifische Datenbanken und Repositorien
- 3) domänen-affine ML- und Recheninfrastruktur
- 4) Open-Source-Software

Wir eröffnen damit neue und **kompetitive Forschung**, ermöglichen **sichere Verarbeitung** von Kulturerbe-Daten, **stärken** die existierende DH-**Community** in Österreich, **fördern Nachwuchs** und **Interdisziplinarität** und sichern eine **nachhaltige Nutzung** von Infrastruktur.

Was hat DHinfra 'gebaut'?

Digitalisierung: Multispektralkameras, Scanroboter, Rechner zur Prozessierung

Siehe dhinfra.at/digitization

Datenbanken und Server: Cluster für Virtuelle Maschinen (für Projektdatenbanken)

Siehe dhinfra.at/saas

Offene Software: Entwicklungsbegleitung und Publikation

Siehe dhinfra.at/software

Policies: Service Policy, Privacy Policy, Technical Policy Appendix etc.

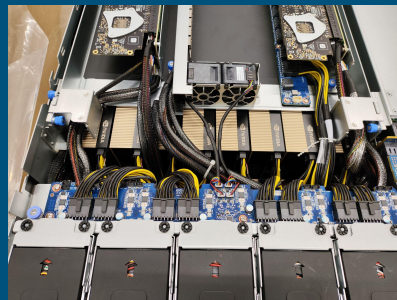
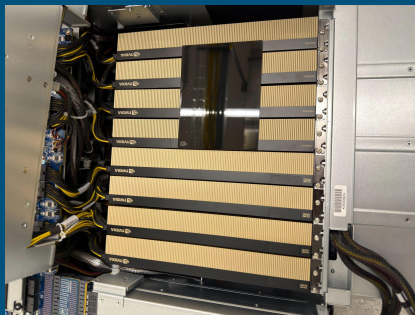
Siehe dhinfra.at/policy

Was hat DHinfra 'gebaut'?

Maschinelles Lernen (& KI):

GPU-Cluster inkl. Speicher (Graz+Krems) (und GPUs an einzelnen Standorten)

Siehe dhinfra.at/gpu-cluster







Was ist ein Supercomputer?





<https://commons.wikimedia.org/wiki/File:Server.svg>













Was läuft auf Supercomputern?

Simulation

Wetter, Klima, Material

Physik-Formeln; programmiert

CPU, immer öfter GPU

Physik, Chemie, Meteorologie etc.

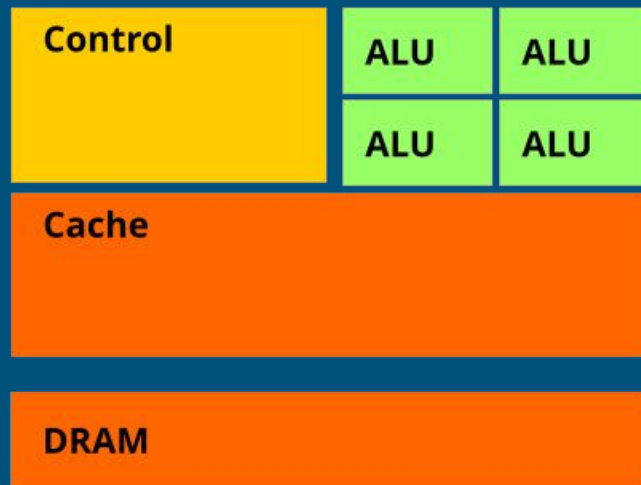
“KI” / Maschinelles Lernen

Texte, Sprache, Bilder

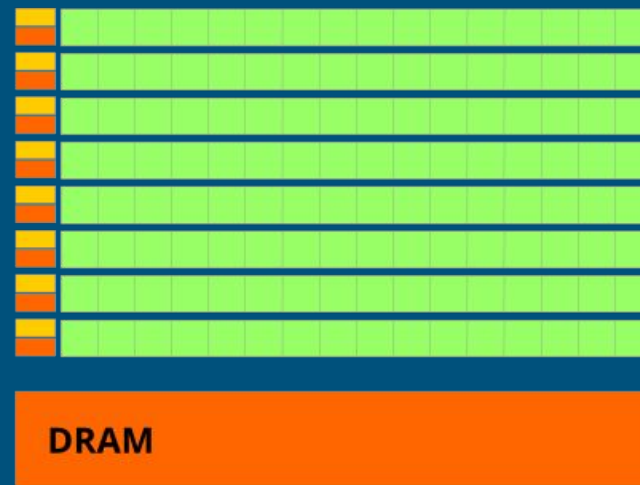
aus Beispieldaten gelernt

GPU

z.B. DH



CPU



GPU

Was ist bei DHinfra anders?

Leistungen (Hardware)

Aktuelle Nvidia GPUs und Knotenkonfiguration

- 18 x H200; 11 x RTXPro6K; 4 x L40s
- Knoten für unterschiedliche Zwecke optimiert (und somit **flexibel!**)
- ~4 % an vRAM des MUSICA-Clusters, aber mit mehr/besseren RAM + Cores

Infiniband Interconnect im Cluster

- (2x) 400Gb/s + RDMA als NFS: Daten sind überall gleich(schnell) erreichbar

ACOnet Uplinks (+ GÉANT) → **Single Sign-on** (SSO) + **Datentransport**

- Schnittstelle zu Krems: 10Gb/s, Schnittstelle von Graz: 25Gb/s (ggf. 100+!)

Leistungen (Software)

HPC-Elemente

- SSH (sicheres **Zugriffsprotokoll**)
- Batch Scheduling (SLURM; die **Warteschlange**)
- Interaktives Arbeiten (SLURM + Jupyterhub; die Programmier-**Umgebung**)

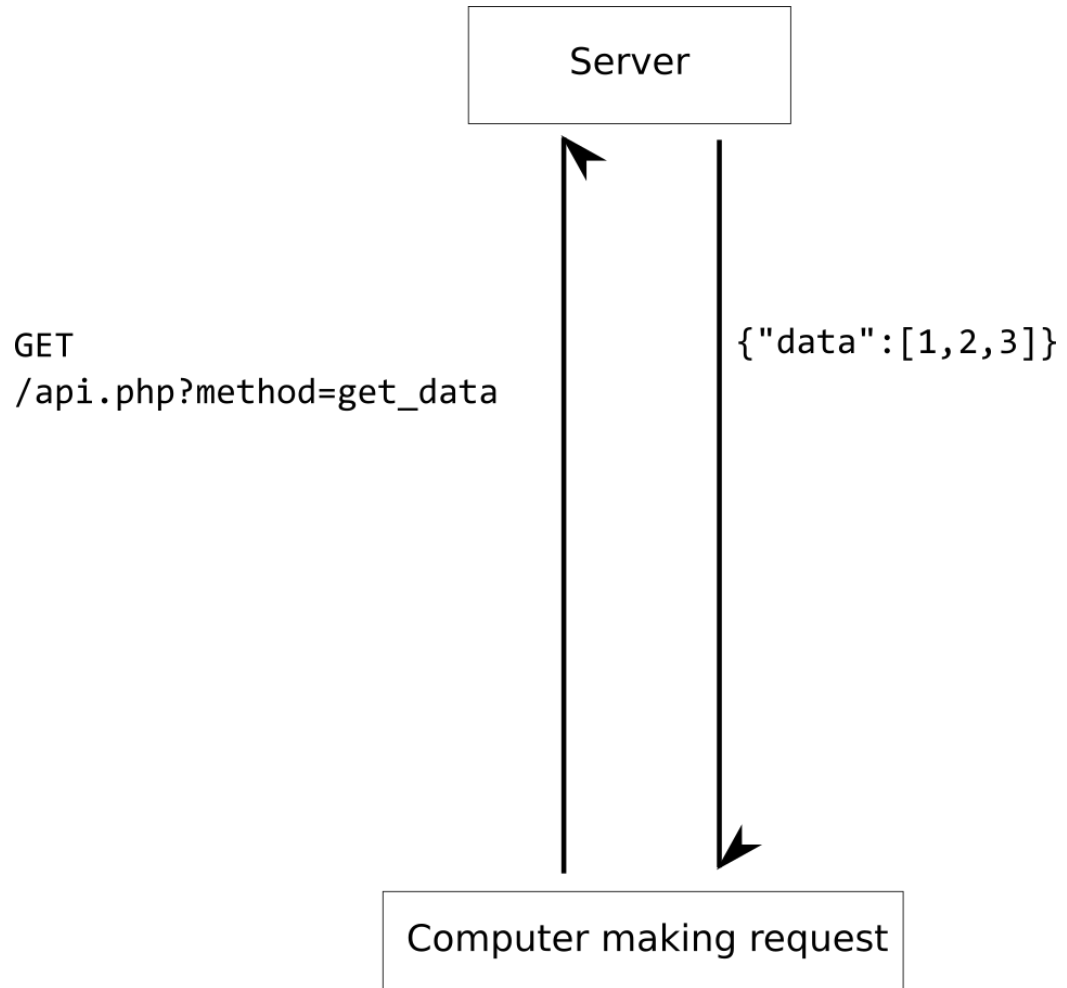
Cloud-Elemente → Alleinstellungsmerkmal

- Containerisierung (Docker, Apptainer, Pyxis, LXCs – **Isolierung**)
- Generische sowie maßgeschneiderte **APIs**
- **KI-Anwendungen** (via APIs; grafische Oberflächen möglich)

z.B. Große Modelle (z.B. 397B Parameter): über 2.000 Tokens/s (auf 4xH200)

API:
Application
Programming
Interface

Schnittstelle
für Programme



Gegenüber klassischem HPC

- **Nahbar**: Testprojekt in < 1 min; Standard-Projekt nach kurzer Prüfung
- **Flexibel**: 3 Wege hinein (Browser, Kommandozeile, anpassbare **APIs**)
- Hauseigenes, **wartungsfreundliches** Projektmanagement-Tool
- Entwickler:innen 'betreuen', die **DH-Arbeitsweise** und -daten kennen

Gegenüber kommerziellen KI-Anbietern

- Login mit dem **Konto der eigenen Institution*** (eduid.at); mehr in Arbeit
- **Open-Weight-Modelle** laufen auf dem Cluster; erreichbar über API
- Daten bleiben im akademischen Netz
- Prompts, Bilder und Antworten: **nichts** davon wird **gespeichert oder verkauft**
- **Demokratisch**: Nutzende stimmen u.a., über das nächste Modell ab

Liste an Identity Providers → gültige eduids: eduid.at/entities/idp

Was wir nicht sind

- kein Langzeitarchiv und keine Datenbank
- kein Nationales HPC-Center
- kein Chat-Produkt mit Support-Desk

Nachhaltigkeit

- Hardware bezahlt, läuft bis 2029/2030
- Software-Stack komplett **Open Source**
- Aufbau von **Expertise**; Pfad zur **Souveränität**
- Verstetigung geplant (mehr dazu später)

Der Einstieg: die Konsole

Der Einstieg

dhinfra.at/login leitet weiter an dhinfra.uni-graz.at

Link auch zu finden unter dhinfra.at/gpu-cluster



Welcome to DHinfra

 Login with your institution

Log in with a local account

Powered by authentik

Find Your Institution

Your university, organization or company

Examples: Science Institute, Lee@uni.edu, UCLA

☒ Remember this choice

[Learn More](#)

Find Your Institution

Your university, organization or company

graz



Examples: Science Institute, Lee@uni.edu, UCLA

☒ Remember this choice [Learn More](#)

[University of Graz](#)



Medical University of Graz

Graz University of Technology
tugraz.at

PPH Augustinum Graz
kphgraz.at

University of Music and Performing...
kug.ac.at

FH CAMPUS 02

[FAQ: Access to IT services with MFA](#)

uniLOGIN



**ACOnet Identity Federation eduID
/ eduGAIN**

Login with uniACCOUNT

[Password forgotten or expired?](#)

Login

Or sign in with

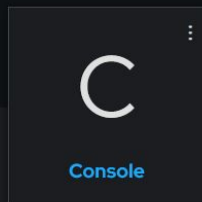
Sign in with Passkey

SATOSA IdP Proxy - zididp Shibboleth replacement



Application Dashboard

Search for an application by name...





Redirecting to Console



florian.atzenhofer-baumgartner@edu.uni-graz.at

[Not you?](#)

You're about to sign into Console.

Application requires following permissions:

- Email address
- General Profile Information

Continue

Your account

Profile

USERNAME

u2160

EMAIL

florian.atzenhofer-baumgartner@edu.uni-graz.at

NAME

Atzenhofer-Baumgartner, Florian

ROLE

MEMBER

SSH ACCESS

☒ none [Request access](#)

HOME QUOTA (DEFAULT)

not measured
of 50 GiB

Efficiency

No finished jobs in the last 30 days.[How this is computed](#)

Your projects

You're not in any project yet. Ask your project lead to add you, or create one above.[+ Create new project](#)

Available LLMs

Text

Vision

Tools

Reasoning

Embeddings

Sort

Popular

Past

Week

DeepSeek V4.1 Flash

deepseek-v4.1-flash

TEXT VISION TOOLS
REASONING

1,048,561 ctx

x1.00 / x1.00 / x0.25

Qwen3-VL-Embedding-8B

qwen3-vl-embedding-8b

TEXT VISION EMBEDDINGS

8,192 ctx · x0.10 / - / -

Qwen3-Embedding-8B

qwen3-embedding-8b

TEXT EMBEDDINGS

8,192 ctx · x0.10 / - / -

all-MiniLM-L6-v2

all-minilm-l6-v2

TEXT EMBEDDINGS

512 ctx · x0.10 / - / -

BGE-M3

bge-m3

TEXT EMBEDDINGS

8,192 ctx · x0.10 / - / -

BGE-small-en-v1.5

bge-small-en-v1.5

TEXT EMBEDDINGS

512 ctx · x0.10 / - / -

EmbeddingGemma-300M

embeddinggemma-300m

TEXT EMBEDDINGS

2,048 ctx · x0.10 / - / -

Jina Embeddings v3 (passage)

jina-embeddings-v3-passage

TEXT EMBEDDINGS

8,192 ctx · x0.10 / - / -

Jina Embeddings v3 (query)

jina-embeddings-v3-query

TEXT EMBEDDINGS

8,192 ctx · x0.10 / - / -

Jina Embeddings v4 (text retrieval)

jina-embeddings-v4-text-retrieval

TEXT EMBEDDINGS

8,192 ctx · x0.10 / - / -

Nomic Embed v1.5

nomic-embed-text-v1.5

TEXT EMBEDDINGS

2,048 ctx · x0.10 / - / -

x input token factor

x output token factor

x cached token factor

Send a request click a model above to use it; replace `$DHINFRA_API_KEY` with your project key · [full guide](#)

```
curl https://api.dhinfra.uni-graz.at/v1/chat/completions \
-H "Authorization: Bearer $DHINFRA_API_KEY" \
-H "Content-Type: application/json" \
-d '{"model": "deepseek-v4.1-flash", "messages": [{"role": "user", "content": "Hello!"}]}'
```



ACCOUNT

Dashboard

Quotas

Settings

Docs

Quotas

[What each limit means →](#)

Home directory

CURRENT QUOTA

50 GiB

cluster default

USED

< 1 MiB

excludes snapshots

REQUEST A DIFFERENT HOME QUOTA



API rate limit

CURRENT TIER

standard (120 req/min)

REQUEST A DIFFERENT RATE TIER



Settings

SSH keys

Paste your **public** key (e.g. the contents of `~/.ssh/id_ed25519.pub`). These are published to the cluster so you can SSH in once your SSH-access request is approved. Never paste a private key.

No SSH keys yet. Add one below.

ADD SSH KEY



Cluster SSH access

[How SSH access and the second factor work →](#)

none

Add an SSH key above first, then you can request access.

Two-factor authentication

SSH to the cluster needs a six-digit code from an authenticator app as well as your key. Set it up here, or re-pair if you replaced your phone.

not set up

Required before SSH access can be granted.

Set up authenticator

Danger zone

Deleting your account is permanent. There is no grace period.

Delete my account

ACCOUNT

Dashboard

Quotas

Settings

Docs

Using the API

API keys

SSH access

Slurm

JupyterHub

Models

Quotas and
budgets

Efficiency

Projects and
membershipAcknowledging
DHinfra

Getting help

Changelog

FAQ

YOUR NEXT STEP

Get into a project

Nothing on the cluster works until you are in one: keys, budgets and storage all belong to a project.

[How to join or create one →](#)

Welcome to the DHinfra console

The console is where you get access to the cluster and see what you are using: hosted language and embedding models over an API, and GPUs for running your own code.

Start here

Everything is scoped to a **project**. API keys, budgets, storage and Slurm jobs all belong to one.

1. **Join or create a project.** [Join one a colleague already leads](#), [create your own](#), or [make a test project](#) to try the cluster first.
2. **Generate an API key.** From your dashboard, up to five per project. A key is shown only once, so store it when you create it.
3. **Send a request.** The API is OpenAI-compatible: change the base URL and the key, and most clients work as they are.
4. **Check your limits** *(optional)*. Tokens, compute, storage and request rate are capped separately and fail in different ways.

Running your own code *(optional)*

Training runs, batch jobs and notebooks need a shell on the head node and a Slurm job, which has a separate approval step. [Get SSH access](#).

Create a new project

1 Type

2 Funding

3 Details

4 Compute

5 Review

Choose a project type

Standard

RECOMMENDED

A full research project. Budgets start at defaults and grow via requests.

- Workspace 100 GiB (growable)
- Compute budget requestable
- LLM tokens requestable
- No time limit

Test

STRICT LIMITS

A short-lived trial to try the cluster under tight, fixed limits.

- Workspace 10 GiB (fixed)
- Compute budget 2 CT (fixed)
- LLM tokens 100,000 (fixed)
- Expires after 7 days
- Max 1 per user, ever (archiving does not free the allowance)

Cancel

Next →

Create a new project

- 1 Type
- 2 Funding
- 3 Details
- 4 Compute
- 5 Review

Enter a project name.

Funding

Funding (required)

Funded (grant awarded, externally reviewed)

Not funded (no grant, not reviewed)

Visibility

☒ The project may be listed publicly (e.g. on the DHinfra impact page)

When this project publishes, please acknowledge the infrastructure. Keep the string "Digital Humanities Infrastructure Austria (DHinfra.at)" intact in whatever sentence you write: [wordings and the paper to cite](#).

Supporting document (optional)

Browse... No file selected.

One PDF, max 10 MB, visible to admins only.

← Back

Next →

Create a new project

1 Type 2 Funding 3 Details 4 Compute 5 Review

Enter a project name.

Project details

Name

ada-vision

Final name

-

Description

What's the project? Who's it for? Is it a student project? Why its own group / quota?

Lowercase letters, digits, `_` / `-`; the `project-` prefix and a unique index are added automatically.

Project record (optional)

DOI

10.1234/abcd or https://doi.org/...

Project / PI page

https://

Expected members

e.g. 4

Anticipated storage (GiB)

e.g. 500

End date (optional)

dd/mm/yyyy

With an end date set, the project is archived automatically after it passes (extendable later via request).

[← Back](#)[Next →](#)

- ACCOUNT
- Dashboard
- Quotas
- Settings
- Docs

Create a new project

- 1 Type
- 2 Funding
- 3 Details
- 4 Compute
- 5 Review

Project details

Name

testproject-student

Final name

project-testproject-student-0044

Description

I am submitting a project as a student, I am doing experiments for recommender systems for digital cultural heritage, and need data-center-grade GPUs

Lowercase letters, digits, _ / - ; the project- prefix and a unique index are added automatically.

Project record (optional)

DOI

10.1234/abcd or https://doi.org/...

Project / PI page

https://

Expected members

e.g. 4

Anticipated storage (GiB)

e.g. 500

End date (optional)

dd/mm/yyyy

With an end date set, the project is archived automatically after it passes (extendable later via request).

← Back

Next →

DHinfra console · Digital Humanities Infrastructure dhinfra.at · policies · status · v0.10.2

60

Create a new project

- 1 Type
- 2 Funding
- 3 Details
- 4 Compute
- 5 Review

Services & compute

What will the project use? (optional)

☒ Slurm (SSH)

☐ H200 GPU

☐ JupyterHub

☒ RTX-6000 Pro GPU

☐ LLM endpoint

☐ L40S GPU

☐ Other

Expected compute (for planning)

GPU	GPUs	Hours / week	For (weeks)
RTX-6000 Pro GPU	11	10	12
+			

≈ 2.640 CT over 12 weeks (220 CT/week, 1% of the cluster's weekly capacity).

← Back

Next →

ACCOUNT

Dashboard

Quotas

Settings

Docs

Create a new project

1 Type 2 Funding 3 Details 4 Compute 5 Review

Review & submit

An admin approves before the group is created. On approval you become its first member and leader, and an API key is auto-issued. Your request gets a reference number, listed below, for any follow-up.

Type	standard
Funding	not funded
Name	testproject-student
Description	I am submitting a project as a student, I am doing experiments for recommender systems for digital cultural heritage, and need data-center-grade GPUs
Services	Slurm (SSH), RTX-6000 Pro GPU
End date	none
Expected compute	≈ 2.640 CT over 12 weeks (220 CT/week, 1% of the cluster's weekly capacity).
Estimate based on	11x RTX-6000 Pro GPU, 10 h/week, 12 weeks

☐ I accept the [Service and Acceptable Use Policy](#) and the other [policies](#) that govern the cluster

[← Back](#)[Submit request](#)

Your account

Profile

USERNAME

u2160

EMAIL

florian.atzenhofer-baumgartner@edu.uni-graz.at

NAME

Atzenhofer-Baumgartner, Florian

ROLE

MEMBER

SSH ACCESS

☒ none [Request access](#)

HOME QUOTA (DEFAULT)

not measured
of 50 GiB

Efficiency

No finished jobs in the last 30 days.[How this is computed](#)

Your projects

You're not in any project yet. Ask your project lead to add you, or create one above.[+ Create new project](#)

ACCOUNT

Dashboard**Quotas****Settings****Docs**

MANAGE

Lead

ADMIN

Approvals**Users****Groups****Model catalogue****Model votes****Slurm****Settings****Ops****Maintenance****Audit log**

ACCOUNT

[Dashboard](#)[Quotas](#)[Settings](#)[Docs](#)

MANAGE

[Lead](#)

ADMIN

[Approvals](#)[Users](#)[Groups](#)[Model catalogue](#)[Model votes](#)[Slurm](#)[Settings](#)[Ops](#)[Maintenance](#)[Audit log](#)

Groups you lead or manage

Active

● project-admin-test-0005 MANAGER MEMBER	2 members	lead David Fleischhacker	● no key	50 GiB	>
● project-admin-test-0019 LEAD	1 member	lead Florian Atzenhofer-Baumgartner	● 2 keys	100 GiB	>
● project-rag-plus-0017 LEAD	12 members	lead Florian Atzenhofer-Baumgartner	● 3 keys	100 GiB	>

ACCOUNT

Dashboard

Quotas

Settings

Docs

MANAGE

Lead

ADMIN

Approvals

Users

Groups

Model catalogue

Model votes

Slurm

Settings

Ops

Maintenance

Audit log

project-rag-plus-0017 LEAD

12 members

lead Florian Atzenhofer-Baumgartner

3 keys

100 GiB



COMPUTE

0.0%

0.0 / 50 CT

STORAGE

0%

< 1 MiB / 100 GiB

TOKENS IN (TOTAL)

25.0M

since day one

TOKENS OUT (TOTAL)

1.1M

since day one

TODAY IN

0%

0 / 20.0M

TODAY OUT

0%

0 / 30.0M

WEEK IN

0%

0 / 100.0M

WEEK OUT

0%

0 / 150.0M

WHO USED THIS PROJECT'S BUDGET



YOUR API KEYS (3 OF 5)

Only you can see or use these. Every member of this project has their own 5. [How keys work](#) →

dhinfra_6368a82e9ce8cede.

120 req/min · expires 2026-11-17

Rotate

Revoke

Rate limit

dhinfra_5ff538c73af4f2f6.

workshop-shared

120 req/min · expires 2026-12-14

Rotate

Revoke

Rate limit

dhinfra_856687f1172568d8.

mine

120 req/min · expires 2026-12-14

Rotate

Revoke

Rate limit

label (optional)

New key

Manage project →

COMPUTE

0.0%

0.0 / 50 CT

STORAGE

0%

< 1 MiB / 100 GiB

TOKENS IN (TOTAL)

25.0M

since day one

TOKENS OUT (TOTAL)

1.1M

since day one

TODAY IN

0%

0 / 20.0M

TODAY OUT

0%

0 / 30.0M

WEEK IN

0%

0 / 100.0M

WEEK OUT

0%

0 / 150.0M

WHO USED THIS PROJECT'S BUDGET

Today in

0 / 20.0M tokens

Today out

0 / 30.0M tokens

Week in

0 / 100.0M tokens

Week out

0 / 150.0M tokens

Total in

25.0M tokens

Total out

1.1M tokens

MEMBER

TODAY IN

TODAY OUT

WEEK IN

WEEK OUT

TOTAL IN

TOTAL OUT

project-rag-plus-0017

← Groups you lead

Usage & limits

[What each limit means →](#)

WORKSPACE STORAGE

100 GiB (default)

COMPUTE BUDGET

50 CT

lifetime total

Aggregate token usage

DAILY IN

0%

0 / 20.0M

DAILY OUT

0%

0 / 30.0M

WEEKLY IN

0%

0 / 100.0M

WEEKLY OUT

0%

0 / 150.0M

WHO USED THIS PROJECT'S BUDGET >

Workspace storage

STORAGE

0%

< 1 MiB / 100 GiB

Project record

Correct these whenever they change. Every edit is logged below. Type, funding status, end date and accepted terms are fixed at approval.

DOI

10.1234/abcd or https://doi.org/...

Project page

https://

Submit a new request

ADD MEMBER



CHANGE LLM TOKEN BUDGET



REQUEST WORKSPACE STORAGE QUOTA



REQUEST COMPUTE BUDGET



REQUEST MAX-RUNTIME TIER



Danger zone

Reason

why is this project being archived? grant ended, replaced by another project, ...

Request archive

Submitted for admin approval. During the 30-day grace an admin can restore; after 30 days the workspace is permanently destroyed.

Recent admin actions on this group

No actions logged yet.

Submit a new request

ADD MEMBER

Member email(s)

anna.huber@uni-graz.at
b.mayr@uni-graz.at

Add member

One per line to add a whole course at once. They must have signed in to the console at least once.

CHANGE LLM TOKEN BUDGET

Daily tokens-in now 20M

new cap, M

reason (optional)

Request

Daily tokens-out now 30M

new cap, M

reason (optional)

Request

Weekly tokens-in now 100M

new cap, M

reason (optional)

Request

Weekly tokens-out now 150M

new cap, M

reason (optional)

Request

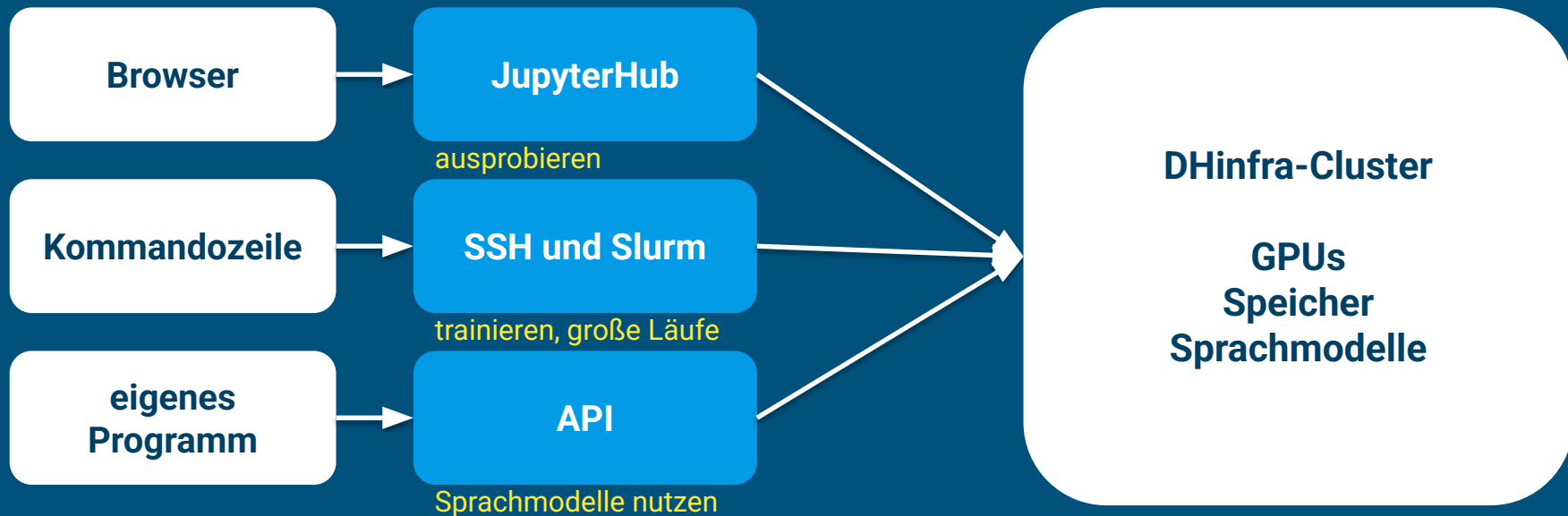
REQUEST WORKSPACE STORAGE QUOTA

REQUEST COMPUTE BUDGET

REQUEST MAX-RUNTIME TIER

Drei Wege auf den Cluster

Drei Wege auf den Cluster

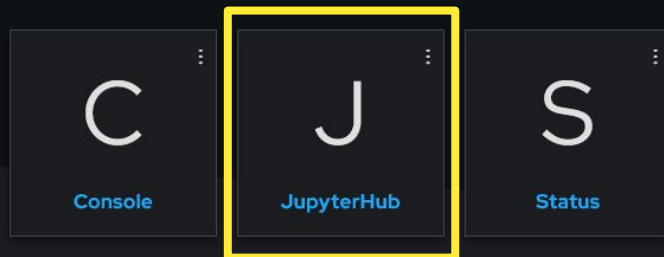


Jupyterhub

Scripting + Experimente



Application Dashboard



[Start My Server](#)

Server Options

Project (Slurm account -- Compute-Tokens are billed here)

admin-test-0005

Environment

PyTorch + CUDA 13.2 (GPU work)

GPUs

0 (CPU only)

CPU cores

8

RAM (GB)

16

Estimated cost

0.16 CT / hour (up to 1.3 CT for a full 8 h session)

At current rates an L40S-hour costs 1.00 CT. Billed on what you reserve; the notebook runs until the project's budget is used up (max 8 h).

Start

Server Options

Project (Slurm account -- Compute-Tokens are billed here)

admin-test-0005

Environment

PyTorch + CUDA 13.2 (GPU work)

GPUs

1

GPU type

RTX-6000 Pro (compute)

CPU cores

16

RAM (GB)

48

Estimated cost

2.35 CT / hour (up to 18.8 CT for a full 8 h session)

At current rates an L40S-hour costs 1.00 CT. Billed on what you reserve; the notebook runs until the project's budget is used up (max 8 h).

Start

Your server is starting up.

You will be redirected automatically when it's ready for you.

Cluster job running, waiting for the notebook to answer.

▼ Event log

Server requested

Cluster job running, waiting for the notebook to answer.

dhinfracomp03

MIXED

0/7 active 1 alloc 0.29 kw gpu 0.97 kw wall

RTX 6000 Pro

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

status.dhinfra.uni-graz.at

Slurm



Training

Docs

Using the API

API keys

SSH access

Slurm

JupyterHub

Models

Quotas and budgets

Efficiency

Projects and membership

Acknowledging DHinfra

Getting help

Changelog

FAQ

Your first job

`srunch` runs a command straight away and waits for it to finish, which is what you want while you are still working out whether the command does the right thing at all.

```
srunch -p compute --ntasks=1 --time=00:10:00 hostname
```

For real work, write the job into a script and submit it with `sbatch`. The job then survives you closing your terminal, and Slurm holds it in the queue until the resources it asked for are free.

```
#!/bin/bash
#SBATCH --job-name=my-training-run
#SBATCH --partition=compute
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --gres=gpu:h200:1
#SBATCH --cpus-per-task=8
#SBATCH --mem=64G
#SBATCH --time=04:00:00
#SBATCH --output=runs/slurm-%j.out
#SBATCH --error=runs/slurm-%j.err

set -euo pipefail
cd "${SLURM_SUBMIT_DIR}"
mkdir -p runs
```

```
atzenhof@dhlnfrahead01:~$ whoami
atzenhof
```

API

Systematisches prozessieren

Using the API

The API is OpenAI-compatible: point an existing client at our base URL with your key and it will usually work unchanged.

Base URL: `https://api.dhinfra.uni-graz.at/v1`

What you need

- **Be part of a project.** Requests are billed to the project, not to you personally. See [Projects and membership](#).
- **An API key.** Keys belong to one person and one project; you can hold several per project (e.g. production and testing). See [API keys](#).
- **A model slug.** The identifier that goes in the `model` field. Your dashboard lists every model you can reach.

Your first request

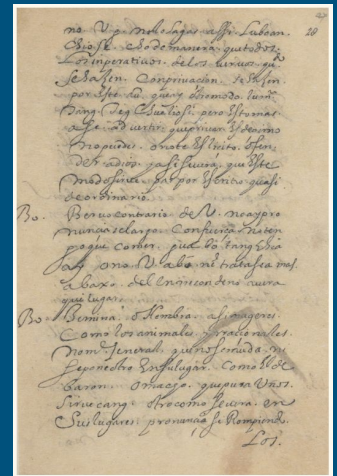
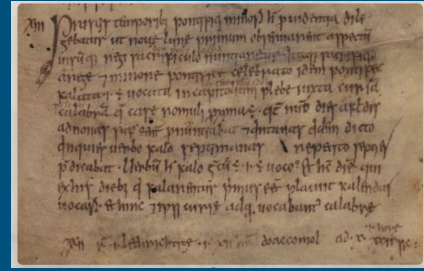
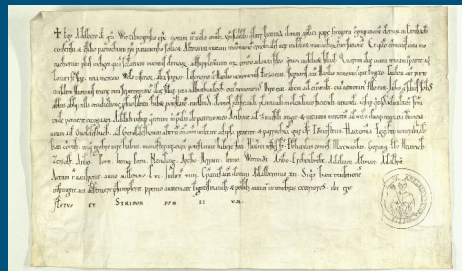
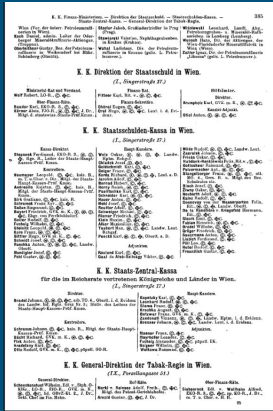
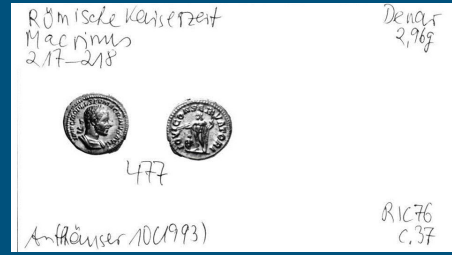
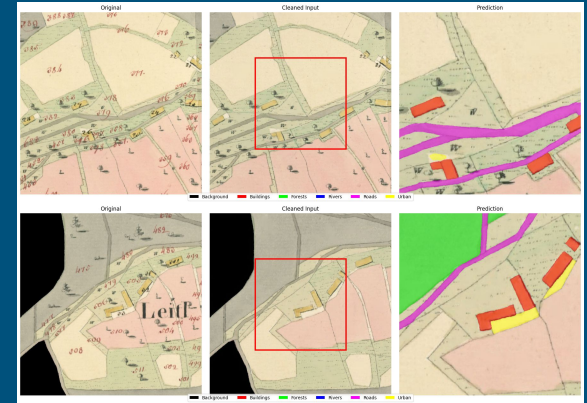
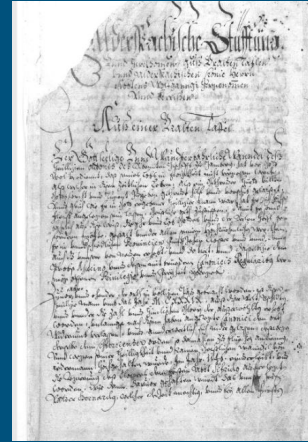
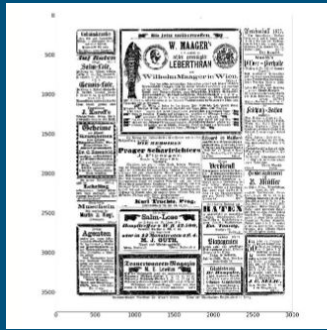
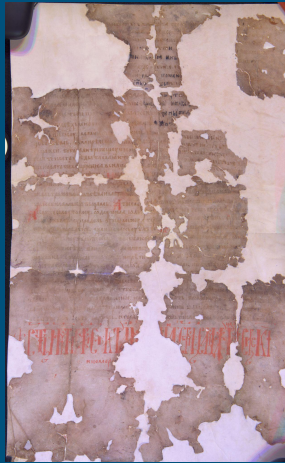
```
curl https://api.dhinfra.uni-graz.at/v1/chat/completions \  
-H "Authorization: Bearer $DHINFRA_API_KEY" \  
-H "Content-Type: application/json" \  
-d '{"model": "<model-slug>", "messages": [{"role": "user", "content": "Hello!"}]}'
```

The dashboard fills this in for you: click a model and the example updates with its slug.

What comes back

The response is the OpenAI chat-completion shape:

Use Cases



Bildnachweise

<https://gams.uni-graz.at/context:emho>

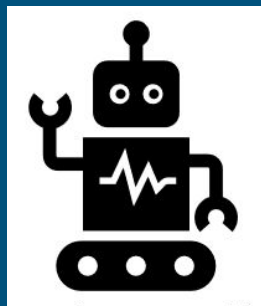
<https://gams.uni-graz.at/context:glossvibe>

<https://gams.uni-graz.at/o:cord.a470b8ce.8a90.301f.9e94.8e36c80c12a1/sdef:TEI/get?locale=de&mode=datelist&context=context:cord.at.stial-lambachosb>

<https://gams.uni-graz.at/o:aled.1#F1>

© außer nach Angaben in Folien der Abschlusspräsentation:

<https://www.dhinfra.at/2026-06-08-dhinfra-final-event-results/>



use-casesPublic

Edit PinsWatch0Fork3Star0

main1 Branch0 Tags

Go to file

Add file

Code

atzenhofer Merge pull request #10 from suzanaSagadin/ri14-1-reconciliation1f6ad5e · 53 minutes ago29 Commits

.github	Sort use cases into candidates and curated, add submission d...	last month
candidates	Merge pull request #10 from suzanaSagadin/ri14-1-reconcili...	53 minutes ago
curated	Sort use cases into candidates and curated, add submission d...	last month
.gitignore	Allow .env.example files past the credentials ignore rule	1 hour ago
LICENSE	Spell the name DHinfra in the README and the copyright line	last month
README.md	Replace em-dashes in the README and the use case template	1 hour ago
USE-CASE-TEMPLATE.md	Replace em-dashes in the README and the use case template	1 hour ago

READMEApache-2.0 license

DHinfra cluster use cases

Research pipelines that run on the DHinfra cluster: notebooks, scripts, and small sample datasets you can open and run.

Start with the console documentation at console.dhinfra.uni-graz.at/console/docs. It already carries a lot of example code: getting a session, requesting a GPU, loading data, running a job. This repository holds the larger examples that build on it.

This repository covers the integrated cluster environment: JupyterHub, GPUs, batch jobs, shared storage, the model

About

No description, website, or topics provided.

- Readme
- Apache-2.0 license
- Activity
- Custom properties
- 0 stars
- 0 watching
- 3 forks
- Audit log
- Report repository

Releases

No releases published
[Create a new release](#)

Packages

No packages published
[Publish your first package](#)

Contributors7



Name



..



001-job-ads



002-anno-download



003-anno-newspaper-processing



004-kraken-training-inscription



005-Ilmprocessing-medieval-accountingbooks



006-letter-tei-encoding

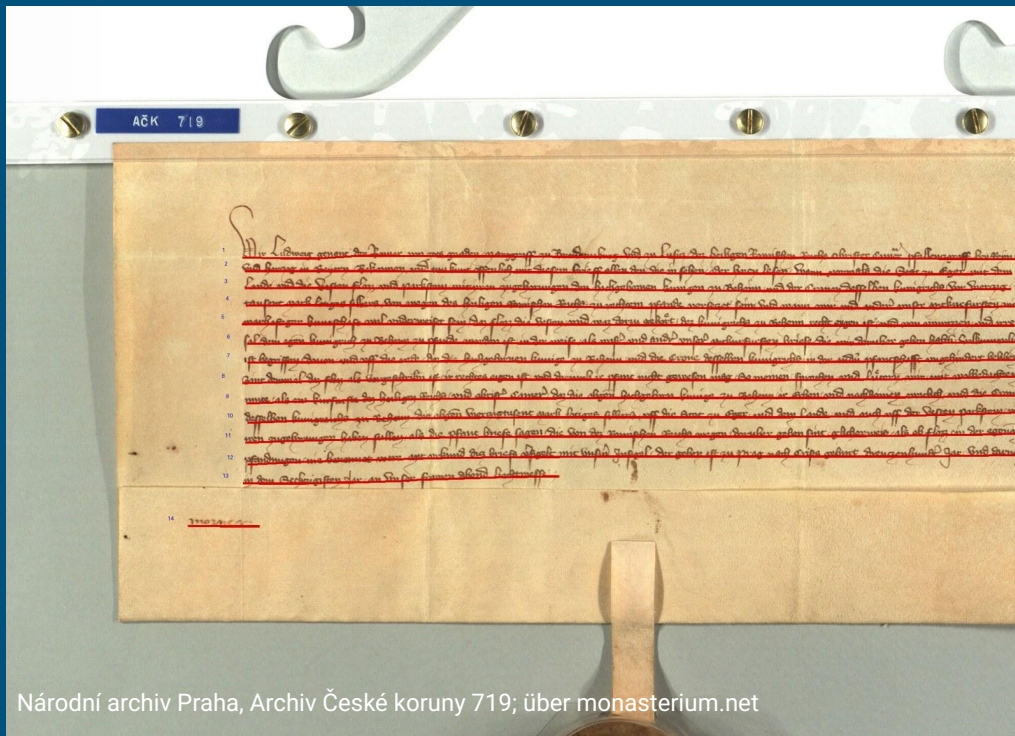


007-ri14-1-reconciliation

Use Cases: eine Urkunde lesen, mit KI



Das Beispiel: eine Urkunde lesen

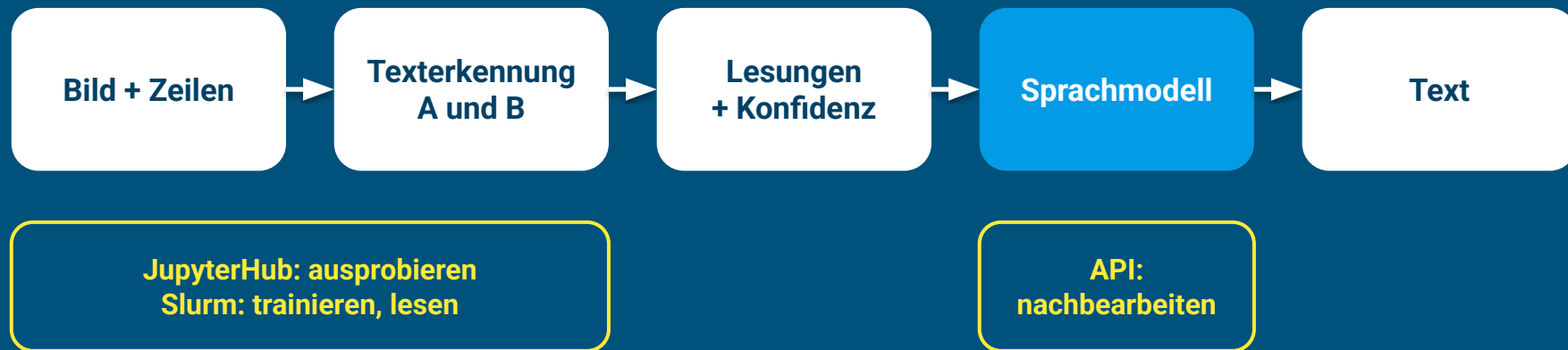


Ludwig der Römer, Prag 1360
Archiv der böhmischen Krone,
Nationalarchiv Prag

14 Zeilen, jede transkribiert
(Ground Truth aus dem Projekt HOME)

Gemessen wird die Zeichenfehlerrate
(CER): wie viele Zeichen falsch sind

Vom Bild zum Text



Texterkennung mit kraken:

A = Manicule 2026 (löst Kürzungen auf), B = CATMuS Medieval (behält sie); Sprachmodell: DeepSeek V4.1 Flash

Line-Modell: Renet, Nicolas, Anguelos Nicolaou, Georg Vogeler. 2026. Consistent Line Extraction on Medieval Charters: Mask R-CNN and U-Net Architectures. Accepted (after peer review) in the ICDAR 2026 workshop "HIP'26: 8th International Workshop on Historical Document Imaging and Processing".

HTR-Modelle: doi.org/10.5281/zenodo.20676471 bzw. doi.org/10.5281/zenodo.21488839

1 · Ausprobieren: JupyterHub

```
[2]: from kraken import rpred
      from kraken.containers import BaselineLine, Segmentation
      from kraken.lib import models

      n = 2
      line_id, polygon, baseline = lines[n - 1]
      segmentation = Segmentation(type="baselines", imagename="gt/ack719.jpg", text_direction="horizontal-lr",
                                script_detection=False, lines=[BaselineLine(id=line_id, baseline=baseline, boundary=polygon)])
      model = models.load_any("models/catmus-medieval-1.6.0.mlmodel", device="cuda:0" if torch.cuda.is_available() else "cpu")
      record = next(rpred.rpred(model, image, segmentation))

      print("Modell:      ", record.prediction)
      print("Ground Truth:", truth[n - 1])
      print(f"Konfidenz {sum(record.confidences) / len(record.confidences):.2f}, CER {score.cer([record.prediction], [truth[n - 1]]):.1%}")
```

Modell: Önd hercog in Reyern Bekennen üñd tuu kunt offeulich mīt diesem brieff allen den die in sehen oder horen lesen Wami uormals di
e Stat zu Eger mit dem
Ground Truth: und herezog in Beyern, bekennen und tun kunt offenlich mit disem brieff allen den, die in sehen oder horen lesen, wann vormal
die stat zu Eger, mit dem
Konfidenz 0.96, CER 6.1%

Zeile 2, CATMuS:
Önd hercog in Reyern
Bekennen üñd tuu kunt
offeulich mīt diesem bri...

Transkription:
und herezog in Beyern,
bekennen und tun kunt
offenlich mit disem brief...

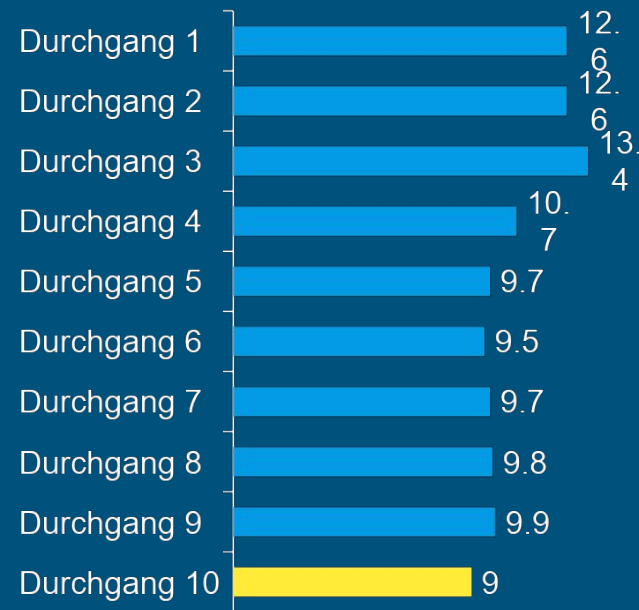
CER 6,1 %

2 · Trainieren: Slurm

```
#SBATCH --partition=compute
#SBATCH --gres=gpu:h200:1
#SBATCH --time=00:45:00

ketos train -i catmus-medieval.mlmodel \
  -N 10 train/pages/*.xml
```

16 Urkunden, 513 transkribierte Zeilen
10 Durchgänge, **7 Minuten auf einer H200**



Fehlerrate auf zurückgehaltenen Zeilen je Durchgang

3 · Lesen: Slurm

```
sbatch read.slurm  
squeue --me
```

A: Manicule: in Dem sechzigstensi jar au unser fraroen dbend Bechenessi.

B: CATMuS: in dem Sechaigisten lar an un ser fraien abend liechtmesse.

B nachtrainiert: in dem sechzigsten jat an unsee frawen abend liechtmesse.

Transkription: in dem sechzigsten iar, an unser frawen abend liechtmesse.

4 · Nachbearbeiten: die API



Hinter jedem Chatfenster steckt eine Anfrage an ein Modell.
Die API gibt Ihnen diese Anfrage selbst in die Hand.

'Zehn' Zeilen

```
from openai import OpenAI

client = OpenAI(base_url="https://api.dhinfra.uni-graz.at/v1",
                 api_key=os.environ["DHINFRA_API_KEY"])

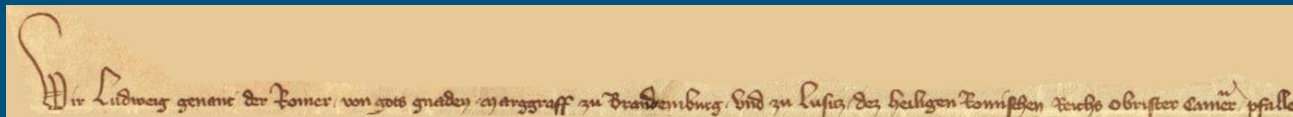
antwort = client.chat.completions.create(
    model="deepseek-v4.1-flash",
    messages=[{"role": "user",
               "content": anleitung + lesungen + zeilenbilder}])

print(antwort.choices[0].message.content)
```

Pro Zeile:
Lesung A, Lesung B,
je mit Konfidenz,
und das Bild der Zeile

Die Anleitung: Kürzungen
auflösen, Schreibung
behalten, nichts erfinden

Was das Modell bekommt, was es zurückgibt



Wir Ludwig genant der Romer, von gots gnaden marggraff zu Brandenburg, und zu Lusia, des heiligen Römischen Reichs obrister camer pfille



obristen camer pfallenczgraff bey Reyn

A: lo de b a a e d e co b ae e lis a cet ps

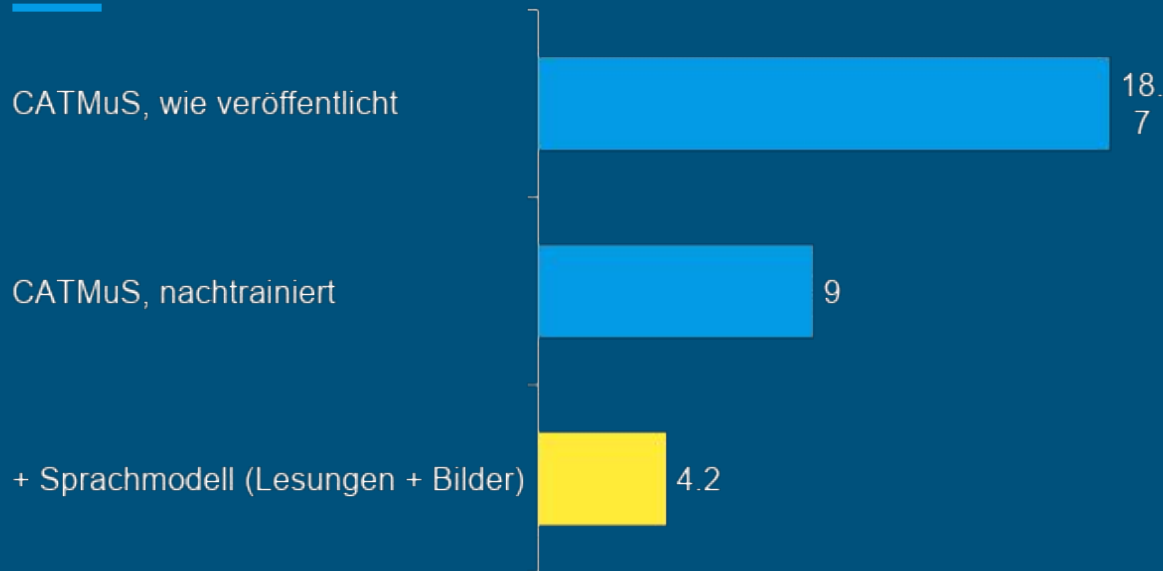
B nachtrainiert: de den eten er heanen den en peden nes ens en ben ben dan s bg en be en bernsten dch dersens
dend a en e s den de

(Anmerkung: In diesem konkreten Beispiel haben beide HTR-Modelle schlechte Resultate erzielt).

‘Sprach’modell: Wir Ludwig genant der Romer, von gots gnaden marggraff zu Brandenburg, und zu Lusia, des heiligen Römischen reichs obrister camer pfille

Transkription: Wir Ludweig genant der Romer, von Gots gnaden marggraff zu Brandemburg und zu Lusicz, dez heiligen römischen reichs obrister camerer, pfallenczgraff bey Reyn

Das Ergebnis



Zeichenfehlerrate, AČK 719

18,7 % → 4,2 %

Zeile 1 mit der Initiale: beide
Texterkennungen scheitern, das
Sprachmodell liest sie aus dem
Bild

Lesungen und Bild gehören zusammen

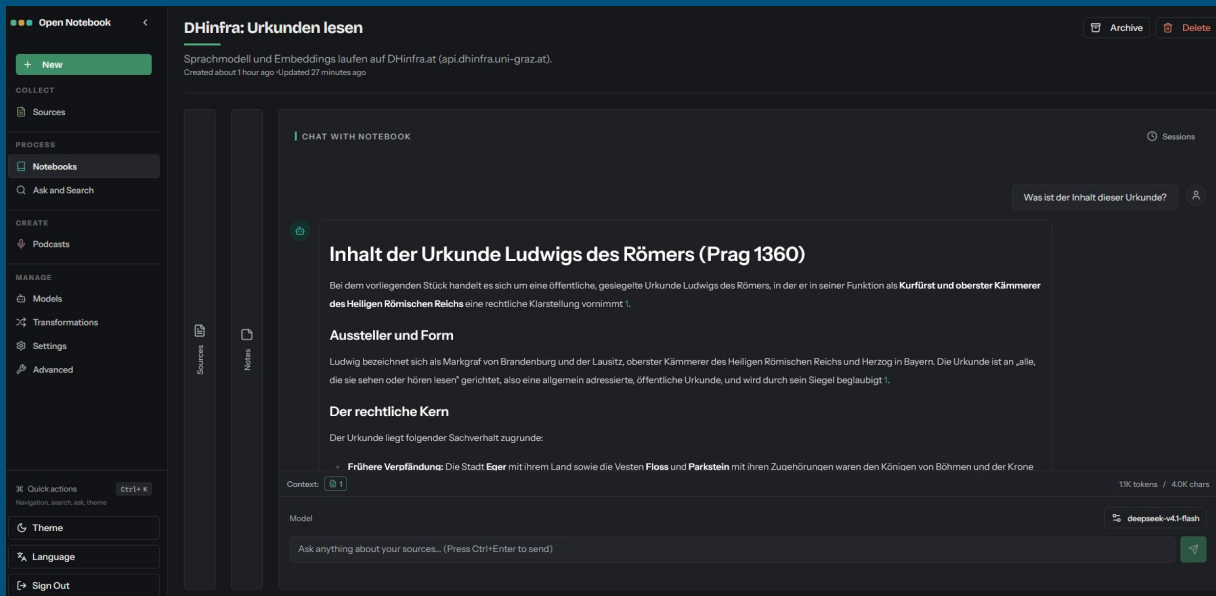
	Prag 1360 (Transkription)	Wien 1424 (Edition)
B: CATMuS	18,7 %	27,2 %
Sprachmodell, nur Lesungen	18,7 %	26,8 %
Sprachmodell, nur Bilder	18,1 %	65,4 %
Sprachmodell, Lesungen + Bilder	7,3 %	13,8 %

Wien gegen die Edition gemessen, die auch Text enthält, den keine Texterkennung gefunden hat.
Test zuvor mit Qwen 3.5 397B – mit Lesungen und Bildern: 9,7 %

Beispiel: GUI basierend auf API

github.com/lfnovo/open-notebook

Im Hintergrund: DHinfra



Open Notebook:
eine offene Alternative zu
NotebookLM,
am eigenen Laptop

Sprachmodell und Suche:
DHinfra

Nur Adresse und
Schlüssel eingetragen

Open Notebook

New

COLLECT

Sources

PROCESS

Notebooks

Ask and Search

CREATE

Podcasts

MANAGE

Models

Transformations

Settings

Advanced

Quick actions

Navigation, search, ask, theme

Theme

Language

Sign Out

DHinfra: Urkunden lesen

Archive

Delete

Sprachmodell und Embeddings laufen auf DHinfra.at (api.dhinfra.uni-graz.at).

Created about 1 hour ago · Updated 27 minutes ago

CHAT WITH NOTEBOOK

Sessions

Was ist der Inhalt dieser Urkunde?

Inhalt der Urkunde Ludwigs des Römers (Prag 1360)

Bei dem vorliegenden Stück handelt es sich um eine öffentliche, gesiegelte Urkunde Ludwigs des Römers, in der er in seiner Funktion als **Kurfürst und oberster Kämmerer des Heiligen Römischen Reichs** eine rechtliche Klarstellung vornimmt 1.

Aussteller und Form

Ludwig bezeichnet sich als Markgraf von Brandenburg und der Lausitz, oberster Kämmerer des Heiligen Römischen Reichs und Herzog in Bayern. Die Urkunde ist an „alle, die sie sehen oder hören lesen“ gerichtet, also eine allgemein adressierte, öffentliche Urkunde, und wird durch sein Siegel beglaubigt 1.

Der rechtliche Kern

Der Urkunde liegt folgender Sachverhalt zugrunde:

• **Frühere Verpfändung:** Die Stadt **Eger** mit ihrem Land sowie die Vesten **Floss** und **Parkstein** mit ihren Zugehörungen waren den Königen von Böhmen und der Krone

Context: 1

Model

deepseek-v4.1-flash

Ask anything about your sources... (Press Ctrl+Enter to send)

105

Spielregeln am Cluster

Nutzungsbedingungen

dhinfra.at/policy

dhinfra.at/policy/service

GPU POWER
1.59
kW

WALL POWER (IPMI)
4.16
kW

GPUS ACTIVE / ALLOC / TOTAL
0 / 7 / 27

JOBS RUNNING / QUEUED
2 / 0

FABRIC I/O (INFINIBAND)
0
GB/S

CLUSTER POWER (kW) - 300 SAMPLES

FABRIC I/O (GB/S) - 300 SAMPLES

dhinfracomp01

IDLE

0/8 active

0.41 kW gpu

0.97 kW wall

H200

0%

H200

0%

H200

0%

H200

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

CPU

0%

RAM

1%
32 / 2267 G

dhinfracomp02

MIXED

0/8 active

0 alloc

0.73 kW gpu

1.43 kW wall

H200

0%

H200

0%

H200

0%

H200

0%

H200

0%

H200

0%

H200

0%

CPU

0%

RAM

14%
309 / 2267 G

dhinfracomp03

IDLE

0/7 active

0.24 kW gpu

0.88 kW wall

RTX 6000 Pro

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

RTX 6000 Pro

0%

CPU

0%

RAM

3%
24 / 755 G

dhinfracomp04

MIXED

0/4 active

1 alloc

0.22 kW gpu

0.53 kW wall

L40S

0%

L40S

0%

L40S

0%

L40S

0%

CPU

0%

RAM

2%
16 / 755 G

dhinfracomp05

CONTROLLER

0.35 kW wall

CPU

RAM

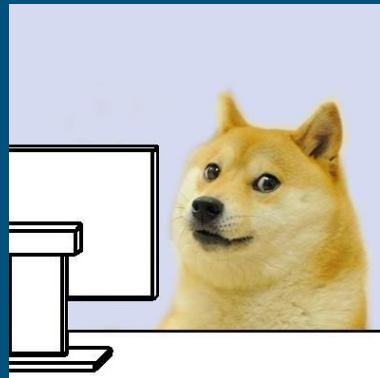
Best Practices

Wie nutze ich das am effizientesten?

Wie nutze ich das am einfachsten?

Wie nutze ich das am sichersten?

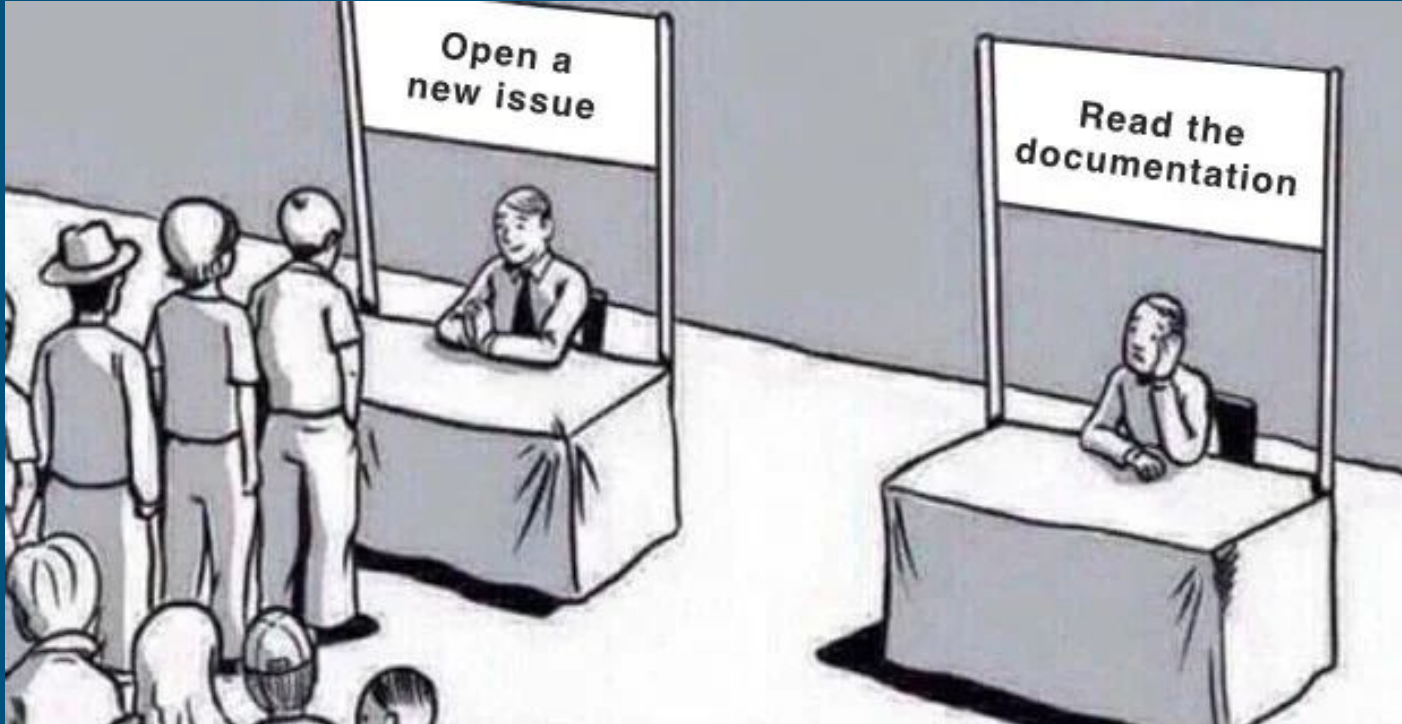
(oder: ohne etwas zu zerstören?)



Prinzip: Es ist die Aufgabe der Administrator:innen, nicht der User, das System robust zu gestalten.

Best Practices

Man lese die Dokumentation.



https://www.reddit.com/r/ProgrammerHumor/comments/grhntz/so_true_and_painful/

Best Practices

Man nutze bestehendes (gratis) Trainings-Material und Trainings-Events.

Home

Welcome to ASC – Austrian Scientific Computing

Austrian Scientific Computing (ASC) – formerly known as the Vienna Scientific Cluster (VSC) – is a collaboration of several Austrian universities that provides supercomputer resources and technical support to their users.

The current flagship systems, [VSC-4](#), [VSC-5](#) and [MUSICA](#), are the fastest supercomputers in Austria. They power advanced research and innovation in fields such as physics, chemistry, climate science, life sciences, engineering, and beyond.

For information on accessing ASC resources and a list of member universities visit [Access](#).

To see upcoming training events in high-performance computing (HPC), visit [Training](#).

LATEST NEWS

>> [Supercomputer MUSICA goes into operation](#)

>> [Join our online courses and webinars](#)



use-cases / candidates / 

Add file  

 **strubrina** minor updates e8a4545 · 3 days ago  History

Name	Last commit message	Last commit date
 ..		
 001-job-ads	Sort use cases into candidates and curated, add submission docs	last month
 002-anno-download	Sort use cases into candidates and curated, add submission docs	last month
 003-anno-newspaper-processing	Sort use cases into candidates and curated, add submission docs	last month
 004-kraken-training-inscription	Sort use cases into candidates and curated, add submission docs	last month
 005-Ilmprocessing-medieval-accountingbooks	added use case 005: accounting books	last month
 006-letter-tei-encoding	minor updates	3 days ago

Best Practices

Man beginne mit so wenig Ressourcen wie nötig.

Server Options

Project (Slurm account -- Compute-Tokens are billed here)

admin

Environment

PyTorch + CUDA 13.2 (GPU work)

GPUs

0 (CPU only)

CPU cores

8

RAM (GB)

16

Estimated cost

0.16 CT / hour (up to 1.3 CT for a full 8 h session)

At current rates an L40S-hour costs 1.00 CT. Billed on what you reserve; the notebook runs until the project's budget is used up (max 8 h).

Start

Best Practices

Man reflektiere seinen Verbrauch.

2726	2026-10-02 01:05	4 h 23 m	6 x RTX 6000, 128 cores, 500 GB	65.6 CT	55%	▼
GPU	6 x RTX 6000	52.6 CT charged	36.3 CT useful	GPU 0 busy 78%, peak 73 GB GPU 1 busy 68%, peak 66 GB GPU 2 busy 66%, peak 66 GB GPU 3 busy 74%, peak 66 GB GPU 4 busy 76%, peak 49 GB GPU 5 busy 52%, peak 63 GB		
2668	2026-09-27 17:32	28 m	1 x H200, 18 cores, 34 GB	2.1 CT	7%	▼
GPU	1 x H200	1.9 CT charged	0.1 CT useful	busy 19%, peak 11 GB		
CPU	18 cores	0.1 CT charged	0.0 CT useful	28% of 18 cores used		
Memory	34 GB	0.0 CT charged	0.0 CT useful	peaked at 25 GB		

So startet man

1. dhinfra.at/login oder login.dhinfra.uni-graz.at
2. (eduID) Login nutzen
3. Testprojekt anlegen und ausprobieren
4. Standardprojekt beantragen, wenn es 'ernst' wird
5. Dokumentation nach dem Login lesen; Fragen im Chatraum stellen

Walter Scholger

Universität Graz – Institut für Digitale Geisteswissenschaften

DHinfra und die Zukunft

Statistiken – derzeit

Ø fast 40 % Auslastung

41/87 (echte) User

26 echte Projekte

5+ Drittmittelanträge mit DHinfra-Nutzung in Arbeit/Begutachtung

Projektanfragen von Archiven, Bibliotheken, IT, Sozialwissenschaften uvm.

—

Tokens in: 4.700.035.544

Tokens out: 775.970.571

—

Ø pro Tag:

324 an unbekannten Login-Versuchen

36 Sperrungen entsprechender IPs

70 kWh Stromverbrauch (ohne Kühlung)

Verstetigung

- Einführungsphase bis 31. Dezember 2026 ohne Gebühren
- Policies durch CLARIAH-AT mitgetragen
- Absichtserklärung zur Verstetigung aller Projektpartner:innen

Verstetigung

- CLARIAH-AT Service Desk
- Zentrale Anlaufstelle für DH, die informiert weiterleiten kann
- Föderierte Expertise
- Bündelt Infrastrukturen und Expertise in Wien, Salzburg, Krems und Graz

Nächste Termine

09.11. CLARIAH-AT Roadshow
Wien (ACDH-ÖAW)
siehe clariah.at

24.11. DHd-AGKI (Design und Betrieb von DHinfra)
remote
siehe agki-dh.github.io

30.11. - 01.12. CRASH - Cultural Heritage Advanced Search Hackathon
Graz (Unicorn) – REGISTRIERUNG BIS 01.11.
siehe crash-this.github.io

Danksagungen

ASC/VSC sowie ACA/EuroCC für Expertise

ASHPC24, ASHPC26, DHd26: Community Feedback & Stipendien

IT Services aller Institutionen, v.a., Universität Graz und Krems

TU Graz, Institute of Machine Learning and Neural Computation

Uni Graz, Institut für Mathematik und Wissenschaftliches Rechnen

Universität Wien zur (finanziellen) Förderung des FOSS-Whitepapers

CLARIAH-AT für Finanzierung von DHinfra-Gov (Follow-up)

BMFWF und EU-RRF via “(Digitale) Forschungsinfrastrukturen”

Die Inhalte dieser Folien sind unter CC-BY-SA nachnutzbar, außer anders gekennzeichnet.



Zitate zu wissenschaftlichem Gebrauch unterliegen den Lizenzen der jeweils angegebenen Quellen.



Universität für
Weiterbildung
Krems



Österreichische
Nationalbibliothek



ÖAW
ÖSTERREICHISCHE
AKADEMIE DER
WISSENSCHAFTEN

universität
innsbruck

di:angewandte

Q&A

Fragen? Gerne JETZT, ansonsten dhinfra@uni-graz.at