CLARIAH-AT

Bundesministerium
Frauen, Wissenschaft
und Forschung

Funded by
the European Union

DH INFRA

AGKI-DH Webinar
24. November 2026, online

Universität für
Weiterbildung
Krems

Donau Universität Krems

PARIS
LODRON
UNIVERSITÄT
SALZBURG

ÖAW Österreichische Akademie der Wissenschaften

universität
innsbruck

Österreichische
Nationalbibliothek

TU WIEN TECHNISCHE UNIVERSITÄT WIEN

UNI GRAZ

universität
wien

di:'ʌngewʌndtə

DHinfra.at

Design und Betrieb eines föderierten GPU-Clusters für die Digitalen Geisteswissenschaften

Florian Atzenhofer-Baumgartner
Universität Graz, Institut für Digitale Geisteswissenschaften

Worum es heute geht

1. **Was gibt es?** Cluster, Zugänge, Modelle, Nutzung
2. **Was heißt das für die AG?** Themen der AG auf dem Cluster, gemeinsame Projekte
3. **Wie baut man so etwas auf?** Der Weg, Entscheidungen, Architektur
4. **Wie betreibt man so etwas?** Menschen, Alltag, Grundsätze, Kosten
5. **Wie entwickelt man es weiter?** Was gerade entsteht, ab 2027, offene Fragen

Teil 1

Was gibt es?

DHinfra in Zahlen

9 Institutionen
3,5 Jahre (2023 bis 2026)
4 Arbeitsgruppen
8 Arbeitspakete
28 Meilensteine
50+ Mitglieder
100+ assoziierte Personen

Projektrahmen

Förderprogramm: „(Digitale) Forschungsinfrastrukturen“
Fördergeber: BMFWF und EU-Aufbau- und Resilienzfazilität
Ziele: Beschaffung, Modernisierung, besserer Zugang

Ein Infrastrukturprojekt, kein Forschungsprojekt
Fortführung gefördert von **CLARIAH-AT**: Governance und Verstetigung

Funded by the European Union

Bundesministerium Frauen, Wissenschaft und Forschung

Was hat DHinfra gebaut?

Digitalisierung: Multispektralkameras, Scanroboter, Arbeitsplätze für die Verarbeitung

- Siehe **dhinfra.at/digitization**

Datenbanken und Server: ein Cluster virtueller Maschinen für Projektdatenbanken

- Siehe **dhinfra.at/saas**

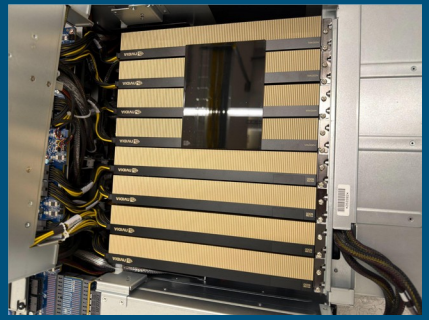
Offene Software: Förderung von Entwicklung und Veröffentlichung (QLever, liiive.now, Pletka)

- Siehe **dhinfra.at/software**

Maschinelles Lernen: GPU-Cluster mit Speicher in Graz und Krems

- Siehe **dhinfra.at/gpu-cluster**

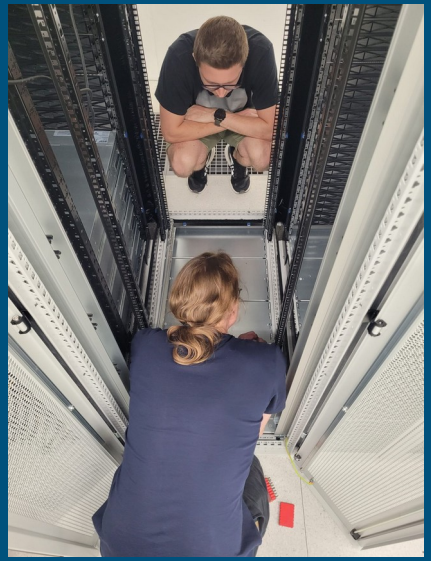


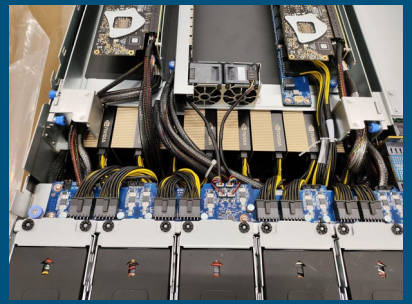












©

Hardware

Aktuelle NVIDIA-GPUs und Knotenkonfiguration

- **18 × H200, 11 × RTX PRO 6000, 4 × L40S**
- Knoten für unterschiedliche Zwecke konfiguriert (und damit **flexibel**)
- Rund 4 % des GPU-Speichers von MUSICA, dafür mehr und schnellerer RAM und Kerne

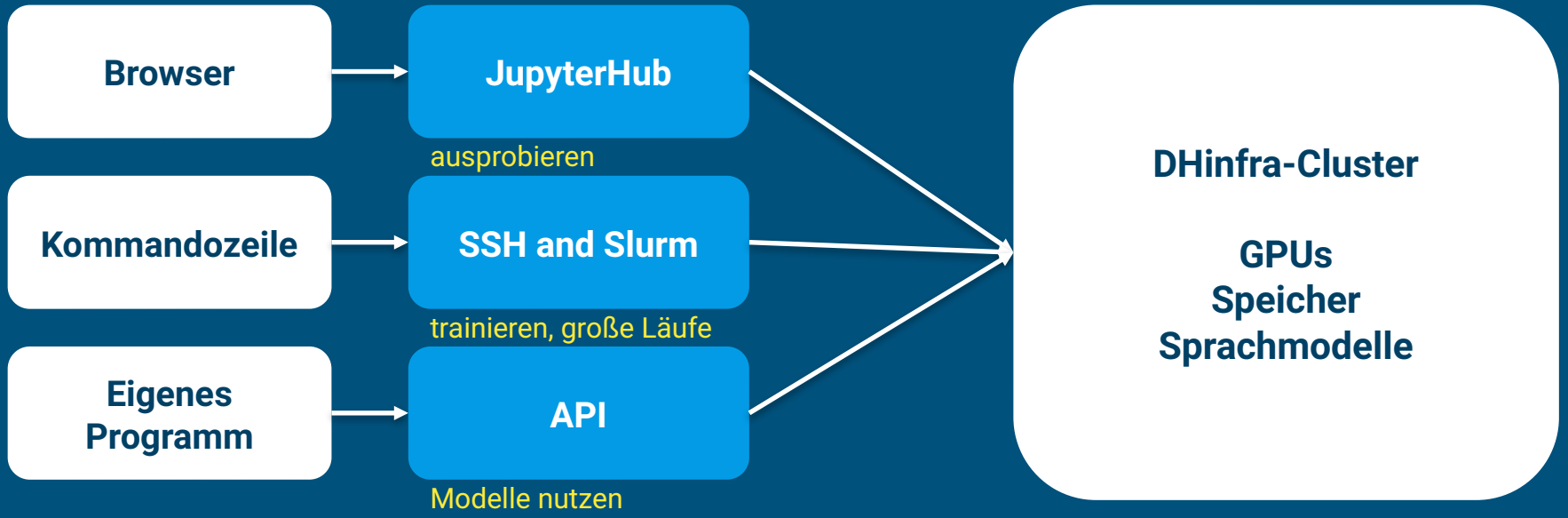
InfiniBand im Cluster

- 2 × 400 Gbit/s, NFS über RDMA: Daten sind von jedem Knoten gleich schnell erreichbar

ACOnet-Anbindung (und GÉANT): **Single Sign-on** und **Datentransfer**

- Verbindung nach Krems 10 Gbit/s, Uplink Graz 25 Gbit/s (100+ Gbit/s möglich)

Drei Wege auf den Cluster



API: Vertiefung

Schnittstelle: OpenAI-kompatibel (/v1/chat/completions, /v1/embeddings, /v1/models), dazu Anthropic Messages, /v1/rerank und /v1/usage
Sprachmodelle: Qwen3.5-397B,, DeepSeek-V4.1-Flash, GLM-5.3-Flash uvm.
Embeddings und Reranking: EmbeddingGemma 2, Qwen3-Embedding-8B, Qwen3-VL-Embedding-8B, Nemotron-3-Embed, Qwen3-Reranker-4B
Funktionen: Bilder als Eingabe, Streaming, Tool Calling, Reasoning (on/off), Prompt Caching
Werkzeuge: openai-Bibliothek, LangChain, LlamaIndex, Coding-Agents etc;
nur Adresse und Schlüssel ändern
Grenzen: Budget pro Projekt, täglich und wöchentlich; 120 Anfragen pro Minute und Schlüssel, mehr auf Antrag

Ein Aufruf aus Python

```
import os
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="qwen3.5-397b",
    messages=[{"role": "user",
               "content": "Welche Orte nennt die Urkunde?"}],
)
print(antwort.choices[0].message.content)
```

Schlüssel in der Konsole erzeugen; gilt 90 Tage

Gleiche Adresse für LangChain, LlamaIndex, Open Notebook oder Open WebUI

Batches: dieselbe Schleife, parallel

Verbrauch abfragen: /v1/usage

Was ist bei DHinfra anders?

Gegenüber klassischem HPC

- **Zugänglich**: Testprojekt in unter einer Minute, Standardprojekt nach kurzer Prüfung
- **Flexibel**: Browser, Kommandozeile, API; Container
- Betreuung durch Entwickler, die **DH-Methoden** und -Daten kennen

Gegenüber kommerziellen KI-Anbietern

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



Siehe auch **dhinfra.at/impact**

Nutzung bisher

Stand Mitte November 2026

- **43 %** durchschnittliche Auslastung
- **60+** aktive Nutzende bei 100+ Konten; **30+** Projekte
- 5+ Drittmittelanträge, die DHinfra einplanen, in Vorbereitung oder Begutachtung
- Projektanfragen von Archiven, Bibliotheken, IT-Diensten, aus den Sozialwissenschaften und mehr

Pro Tag im Schnitt: ~300 unbekannte Loginversuche, 35 gesperrte IP-Adressen, 100 kWh Strom (ohne Kühlung)

Teil 2

Was heißt das für die AG?

Themen der AG, auf dem Cluster

Offene Modelle
Große Open-Weight-Modelle ohne eigene Hardware; über das nächste Modell stimmt die Community ab

Benchmarking
Viele Modelle mit denselben Daten vergleichen: im Batch über die API oder eigene Modelle unter Slurm

RAG und LangChain
Embeddings, Reranker und Sprachmodell hinter einer Adresse und einem Schlüssel

Fine-Tuning
Bis zu 8 GPUs pro Job, auch über zwei Knoten; Container mit PyTorch

Modelle verschiedener Größen
Im Browser auf RTX PRO 6000 oder L40S vergleichen, bevor sie woanders genutzt/skaliet werden

Agents und Tool Calling
Tool Calling und Reasoning über die API; der Agent läuft am Laptop gegen die API

“Carbon Reporting”
Der Strom des Clusters wird gemessen; der Efficiency Score zeigt die Auslastung der eigenen Jobs

Zugang zu HPC
Testprojekt eine Minute nach dem ersten Login; für sehr große Läufe weiter zu ASC/VSC oder EuroHPC

Gemeinsame Projekte mit der AG

Für wen: Forschung und Lehre in den Geistes- und Kulturwissenschaften, in Kulturerbe und den Künsten
Der einfachste Weg: ein gemeinsames Projekt, das eine österreichische Einrichtung leitet

- Mit eduID.at: Login über die Heimatinstitution
- Ohne eduID.at: Einladung direkt in das Projekt oder per Endorsement
- SSH: öffentlicher Schlüssel, zweiter Faktor, Freigabe

Was ein Projekt teilt: Budget, Speicher und Daten
Sensible Daten: vorher melden, wir beraten
Andere Vorhaben: Anfrage an dhinfra@uni-graz.at

Wo die User Base mitwirken kann

- **Modellwahl**: in der Konsole über die nächsten Modelle abstimmen
- **Use Cases**: Pipelines als Pull Request an **github.com/dhinfra-at/use-cases**, mit Vorlage
- **Interessengruppen**: die erste zur automatischen Texterkennung
- **Benchmarks**: Datensätze der AG auf denselben offenen Modellen auswerten
- **Workshops und Lehre**: ein Kurs ist ein Projekt; Modelle für viele Teilnehmende gleichzeitig
- **Anträge**: DHinfra als Recheninfrastruktur in Drittmittelanträgen einplanen
- **Rückmeldung**: welche Modelle, welche Schnittstellen, was fehlt

Was DHinfra nicht ist

- **Kein Langzeitarchiv** und keine Forschungsdatenbank: Projekte haben ein Enddatum
- **Kein Chat-Produkt**: Die API ist die Schnittstelle; offene Oberflächen wie Open WebUI oder Open Notebook lassen sich anbinden
- **Kein nationales Rechenzentrum**: Für sehr große Läufe sind ASC/VSC (MUSICA) und EuroHPC die richtige Adresse; den Pilotlauf kann man bei uns machen
- **Kein kommerzieller Dienst**: kommerzielle Partner nur in einer akademischen Kooperation

Teil 3

Wie baut man so etwas auf?

Der Weg und das Ziel

2023	Projektstart mit 9 Partnern
Juli 2024	Umfrage: Bedarf an Rechenleistung und Speicher
2024–2025	Spezifikation und EU-weite Ausschreibung
August 2025	Installation in Graz
2025–2026	Prototypen: Inferenz, Login, Projektverwaltung; Partner testen
Februar 2026	Workshop auf der DHd 2026
April 2026	Architektur und Plattform-Paket
Juni 2026	Projektende; Partner testen 2.0
Juli 2026	Service Provider in ACOnet/eduID.at
Oktober 2026	Offen für das Fach; Launch am 12. Oktober
November 2026	Konten ohne eduID: Einladung und Endorsements
2027	Betrieb mit CLARIAH-AT

Entscheidungen beim Aufbau

Hardware: H200 wie in den AI Factories, über eine EU-weite Ausschreibung; drei GPU-Typen für Training, Inferenz und kleine Jobs
Netz und Speicher: InfiniBand; ZFS für Home, Projekte, Modelle und Scratch, über NFS mit RDMA an jedem Knoten
Scheduling: Slurm wie an den HPC-Zentren; Container über Pyxis/Enroot und Apptainer
Identität: Authentik mit LDAP und SSSD für die Linux-Konten; Keycloak und FreeIPA geprüft und verworfen
Modelle: llama-swap startet die Modell-Server als Slurm-Jobs; ein eigenes Gateway prüft Schlüssel, Rechte und Budget
Software: durchgehend Open Source; die eigene Plattform steht unter Apache 2.0

Architektur

Anmeldung	eduID.at über SAML-Proxy (SimpleSAMLphp)	Lokale Konten per Einladung	Authentik: OIDC, LDAP, zweiter Faktor		
Dienste	Traefik	Konsole und API-Gateway	JupyterHub	SSH-Einstieg	Statusseite
Rechnen	Slurm	Knoten: H200, RTX PRO 6000, L40S	llama-swap und Modell-Server		
Speicher	ZFS über NFS mit RDMA (InfiniBand)	Home, Projekte, Modelle, Scratch			

Überwachung: Prometheus, Grafana, GPU-Metriken (DCGM); Alarme; Statusseite

Föderiert heißt zweierlei

Föderierte Identität

- Login über die Heimatinstitution: DHinfra ist Service Provider in ACOnet/eduID.at
- GÉANT und REFEDS Data Protection Code of Conduct; Research and Scholarship, Personalized Access
- Logins aus anderen europäischen Föderationen (eduGAIN) sind technisch vorbereitet
- Der Login zeigt, wer jemand ist; Zugang gibt erst ein genehmigtes Projekt

Föderierte Standorte

- Graz und Krems, verbunden über ACOnet; jeder Partner trägt seine laufenden Kosten
- Krems meldet sich über das Grazer Authentik an; dort eigene VMs und Speicher, mit NixOS konfiguriert
- Plattform und Regelwerk sind so gebaut, dass weitere Standorte aufbauen können

Übernommen und selbst gebaut

Übernommen (Open Source)
Slurm, Pyxis/Enroot, Apptainer
Authentik, SimpleSAMLphp, Traefik
JupyterHub
llama-swap, vLLM
Prometheus, Grafana, DCGM-Exporter
ZFS, NFS mit RDMA, Ubuntu

Selbst gebaut
Konsole: Projekte, Anträge, Mitglieder, Budgets, Abstimmungen
API-Gateway: Schlüssel, Rechte, Abrechnung in Compute-Tokens und LLM-Tokens
Statusseite und Dokumentation
Testumgebung: der ganze Stack auf einem Laptop
Regelwerk und Kostenrechnung
Lizenz: Apache 2.0, bald öffentlich

Was wir beim Aufbau gelernt haben

- Spezifikationen und Governance: **früh und parallel einplanen**
- Die **EU-weite Ausschreibung** hat sich gelohnt: dieselbe Hardware wie in den AI Factories, zu ‘vernünftigen’ Kosten
- Ein Infrastrukturprojekt mit wenig Personalmitteln: viel **Weiterbildung** im Team
- Strom, Notstrom und Netzanbindung haben den Zeitplan mitbestimmt
- **Prototypen ersetzen**, sobald klar ist, was gebraucht wird
- **Details zu Lizenzfragen** bei der Spezifikation klären, nicht erst kurz vor Abnahme
- Kleine, fachspezifische Cluster **ergänzen** die großen Zentren

Teil 4

Wie betreibt man so etwas?

Die Menschen hinter dem Cluster







Fünf Personen haben die wesentlichen Teile des DHinfra-Clusters entwickelt:
Florian Atzenhofer-Baumgartner und **David Fleischhacker**: leitende Entwickler des GPU-Clusters, Administration
Max Resch: Beratung und Aufbau sowie Administration des Standorts Krems
Lukas Waldhofer: SAML-Service-Provider; **Michael Otto**: erste Versionen des Inferenz-Controllers und der Projektverwaltung

©

Was das Governance-Projekt geliefert hat

- **Governance-Modell**: Steuerungsgruppe, Technical Board, zentrales Operations-Team
- **Regelwerk**: Service-, Datenschutz- und technische Policy (**dhinfra.at/policy**)
- **Auftragsverarbeitungsverträge** zwischen den Standorten
- **Mitglied der Föderation** ACOnet/eduID.at (REFEDS und GÉANT Codes of Conduct)
- **Rechtliche Einordnung** des Dienstes nach dem EU AI Act
- **Kostenrechnung** als Grundlage für Finanzierungsszenarien

Verteilter Betrieb: Jeder Partner trägt seine laufenden Kosten selbst.
Ohne die Unterstützung von **CLARIAH-AT** wäre der Betrieb nach dem Projekt nicht möglich.

Betrieb im Alltag

Sicherheit
Advisories lesen, Updates für Kernel, Bibliotheken, Slurm und Login-Dienste; nur ein Rechner von außen erreichbar

Hardware
Defekte GPUs erkennen und tauschen, Knoten für Wartung aus der Warteschlange nehmen, Sensoren und Lüfter überwachen

Modelle
Neue Modelle testen und aufnehmen, laden und entladen; GPU-Reset auf Knoten, die ein Modell bedienen

Nutzende
Projektanträge prüfen, Fragen im Chatraum, Budgets und Speicher anpassen

Beobachten
Prometheus, Grafana und GPU-Metriken; Alarme; Statusseite für alle

Abwehr
Pro Tag 300+ unbekannte Loginversuche und 35+ gesperrte IP-Adressen

Wie wir arbeiten

- **Zwei Repositorien**: platform (das Paket für jeden Standort) und cluster-config (Standort, Runbooks)
- **Testumgebung am Laptop**: Login, Konsole und API laufen lokal; Systemtests über die ganze Konsole und die API
- **Issues beschreiben Verhalten**: am Live-System gefunden, im Code behoben, am Live-System geprüft
- **Runbooks**: was am Cluster getan wurde, steht im Konfigurationsrepositorium
- **Erst lesen, dann ändern**: Prüfungen am Cluster sind lesend; Änderungen werden geplant
- **Dokumentation** in der Konsole, mit Suche und Changelog

Grundsätze für den Betrieb

- **Zugang zuerst**: Kontingente werden nach der gemessenen Nutzung kalibriert
- Eine niedrigere Spitzenauslastung nehmen wir in Kauf, damit der **Einstieg leicht** bleibt
- **Beobachten und ansprechen**: Der Efficiency Score zeigt, welcher Anteil der gebuchten Rechenzeit gerechnet hat; er hat “keine Folgen”
- **Eigene Netze**, unabhängig vom zentralen HPC der Universität und ergänzend dazu
- **Institutionelle Ressourcen nutzen**: Netz, Föderation, Räume, IT-Dienste
- **Einfachheit, Sicherheit, Dokumentation**

Was der Betrieb (derzeit) kostet

> 700.000 EUR

Hardware Graz, EU-weit ausgeschrieben

100 kWh

Strom pro Tag (ohne Kühlung)

4 : 2 : 1

Compute-Tokens pro Stunde auf H200, RTX PRO 6000, L40S

- **Beispiel HabsburgLLM**: 18 Tage auf 8 H200, rund 1.000 EUR Strom; zu Cloud-Preisen 20.000+
- Laufend zählen vor allem **Personal** und die Erneuerung der Hardware ab 2029
- Bis 31. Dezember 2026 **kostenfrei**; ein Preismodell für 2027 ist in Arbeit

Teil 5

Wie entwickelt man es weiter?

Woran wir arbeiten

Ein **Preismodell** für jede Art von Anfrage

- GPU-Stunden, gemeinsame Modelle über die API, eigene Modelle, Batch-Jobs

Neue API-Endpunkte für weitere Arten von Anfragen

- Modelle jenseits der Sprachmodelle, etwa Layouterkennung und Zeilensegmentierung
- Gleicher Zugang wie heute: API-Schlüssel, Rechte, Budget

Logins jenseits von eduID

- Seit November im Betrieb: Einladung direkt in ein Projekt oder per Voucher, mit eigenem Konto und zweitem Faktor

Nächster Schritt: weitere Modelltypen

- Heute leitet das Gateway fünf Arten von Anfragen weiter: Chat, Completions, Messages, Embeddings, Rerank
- **Neu**: ein Endpunkt für Prediction-Modelle, etwa Layouterkennung und Zeilensegmentierung
- Gleicher Weg wie bei den Sprachmodellen: llama-swap startet das Modell unter Slurm, das Gateway prüft **Schlüssel, Rechte und Budget**
- **Private Modelle**: nur für freigegebene Nutzende sichtbar
- Danach: eine **Batch-Schnittstelle** für große Mengen und dynamische Nutzungsmuster

Zugang ohne eduID

- Mit eduID **bürgt die Heimatinstitution** für die Person
- Manche Berechtigte haben keine eduID.at: Bibliotheken, Museen, Partner im Ausland
- Breiterer Zugang muss das Vertrauen von Partnern und Fördergebern halten

Seit November im Betrieb: Bürgschaft wie beim arXiv-Endorsement

- **Einladung direkt in ein Projekt** oder **per Endorsement**
- Eine bestehende Person, etwa eine Projektleitung, bürgt für die neue
- Eigenes Konto mit Passwort und zweitem Faktor; SSH nach Freigabe
- **Vertrauen der Partner**: hinter jedem neuen Konto steht eine bekannte Person
- **Schlank**: wenig Verwaltungsaufwand, keine eigene Prüfstelle

Ab 2027

- Einführungsphase bis 31. Dezember 2026, **kostenfrei**
- Regelwerk von **CLARIAH-AT** unterstützt; alle Projektpartner haben eine Absichtserklärung für den Weiterbetrieb unterzeichnet
- Betrieb und Koordination durch CLARIAH-AT und die beteiligten DH-Zentren
- **CLARIAH-AT Service Desk** als zentrale Anlaufstelle
- **Föderierte Expertise**
- Weitere DH-Rechenstandorte können auf der Plattform (Apache 2.0, bald öffentlich auf **github.com/dhinfra-at**) und dem Regelwerk aufbauen
- **Interessengruppen**, zuerst zur automatischen Texterkennung

CRASH: ein Crashtest für die Infrastruktur

- **Cultural Heritage Advanced Search Hackathon**
- **30. November bis 1. Dezember 2026**, Graz (UNICORN)
- Suchwerkzeuge für Kulturerbe-Daten bauen und offen testen: „Vibes and Checks“
- Für Archive, Bibliotheken, Museen, Forschung, Studierende, Entwicklung und Design
- DHinfra stellt die Cluster mit sehr großen Open-Weight-Modellen bereit
- Daten aus Europeana und Kulturpool, auf DHinfra aufbereitet
- Gefördert von CLARIAH-AT

Nächste Woche: **crash-this.github.io**

Offene Fragen, die wir mitbringen

- Wie koppelt man föderierte Identität an **Kontingente pro Institution**?
- Wie hält man große Modelle auf **unterschiedlichen GPUs** bereit, ohne Speicher zu blockieren?
- Wie teilt man Karten zwischen **Dauerbetrieb** (API) und **Batch-Jobs**?
- Wie verrechnet man, ohne den **leichten Einstieg** zu verlieren?
- Wie bleiben Ergebnisse **reproduzierbar**, wenn die Community neue Modelle wählt?
- Wer entscheidet was, wenn **mehrere Standorte** Dienste hosten?

So beginnt man

- **dhinfra.at/login** oder **login.dhinfra.uni-graz.at** öffnen
- Mit eduID anmelden, oder mit einer Einladung ein eigenes Konto anlegen
- Ein Testprojekt anlegen und ausprobieren
- Ein Standardprojekt beantragen, wenn es ernst wird
- Die Dokumentation nach dem Login lesen; Fragen in den Chatraum

Kontakt: dhinfra@uni-graz.at

Danksagung

ASC/VSC und ACA/EuroCC für ihre Expertise
ASHPC24, ASHPC26, DHd 2026: Rückmeldungen aus der Community und Reisestipendien
Die IT-Dienste aller Institutionen, vor allem der Universität Graz und der Universität für Weiterbildung Krems
TU Graz, Institute of Machine Learning and Neural Computation
Universität Graz, Institut für Mathematik und Wissenschaftliches Rechnen
Universität Wien für die Finanzierung des FOSS-Whitepapers
CLARIAH-AT für die Förderung von DHinfra-Gov (Folgeprojekt)
BMFWF und EU-Aufbau- und Resilienzfazilität über „(Digitale) Forschungsinfrastrukturen“

Bildnachweise

- **Fotos mit ©**: nach den Angaben in den Folien der Abschlussveranstaltung, **dhinfra.at/2026-06-08-dhinfra-final-event-results**
- **Folie 15**: ORF Steiermark, „Millionen Zeitungsartikel in fünf Tagen erforschen“, **steiermark.orf.at/stories/3374244**
- **Logos**: die Rechte liegen bei den jeweiligen Partnerinstitutionen und Fördergebern

Lizenz

Die Inhalte dieser Folien sind unter **CC BY-SA 4.0** nachnutzbar, außer anders gekennzeichnet.



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

Weitere Folien zu DHinfra.at, ebenfalls unter CC BY-SA:

- Abschlussveranstaltung, 8. Juni 2026: **dhinfra.at/2026-06-08-dhinfra-final-event-results**
- Öffentlicher Launch, 12. Oktober 2026: **dhinfra.at/2026-10-12-public-launch-report**

CLARIAH-AT

Bundesministerium
Frauen, Wissenschaft
und Forschung

Funded by
the European Union

DH INFRA

Vielen Dank!

Universität für
Weiterbildung
Krems

DONAU UNIVERSITÄT KREMS

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

PARIS
LODRON
UNIVERSITÄT
SALZBURG

universität
innsbruck

Österreichische
Nationalbibliothek

TU WIEN TECHNISCHE UNIVERSITÄT WIEN

UNI GRAZ

universität
wien

di:'ʌngewʌndtə

Anhang: Daten und Sicherheit

- Login- und Zugriffsprotokolle: 90 Tage
- Nutzungsdaten und Verwaltungshandlungen: bis zu 3 Jahre
- Projektunterlagen: bis zu 7 Jahre nach Projektende
- Projekte haben ein Enddatum: danach schreibgeschützt, Schlüssel widerrufen, 30 Tage wiederherstellbar, dann gelöscht
- Testprojekte enden nach 7 Tagen
- Nur ein Rechner ist von außen erreichbar, mit Schlüssel und zweitem Faktor
- **Nutzende und Projekte sehen nur ihre eigenen Daten**

Anhang: Regelwerk

Drei Dokumente, öffentlich unter **dhinfra.at/policy**:

- Service and Acceptable Use Policy
- Privacy Policy
- Technical Policy Appendix

Dazu ein Governance-Konzept und ein Memorandum of Understanding

Anhang: Position zum EU AI Act

- Zweck: Forschung und Lehre
- Keine eigenen Produkte; wir hosten **Open-Weight-Modelle**
- Keine verbotenen und keine Hochrisiko-Anwendungen
- Anwendungen mit besonderen Anforderungen: **Klärung beim Antrag**
- Die Administratoren haben Expertise mit KI-Diensten
- KI-Kompetenz der Nutzenden: noch nicht geprüft
- API-Ausgaben sind nicht gekennzeichnet, weil das technisch kaum möglich ist

Anhang: Projekte und Kontingente

Testprojekt: 7 Tage, 2 Compute-Tokens, 100.000 LLM-Tokens, 10 GB; eines pro Person
Standardprojekt: Antrag mit Beschreibung, Förderung, erwarteten Ressourcen; jeder Antrag hat eine Nummer
Speicher: Home 50 GB, Projektverzeichnis 100 GB, beides auf Antrag mehr; gemeinsame Modelle und Datensätze schreibgeschützt
Slurm: Partitionen compute (H200, RTX PRO 6000) und inference (L40S); bis zu 8 GPUs pro Job, auch über zwei Knoten
Laufzeiten: bis 2, 4, 7 oder 14 Tage
JupyterHub: 0 bis 2 GPUs (RTX PRO 6000 oder L40S); eigene Umgebung mit uv als Kernel
API: Schlüssel einmal angezeigt, 90 Tage gültig; Budget täglich und wöchentlich