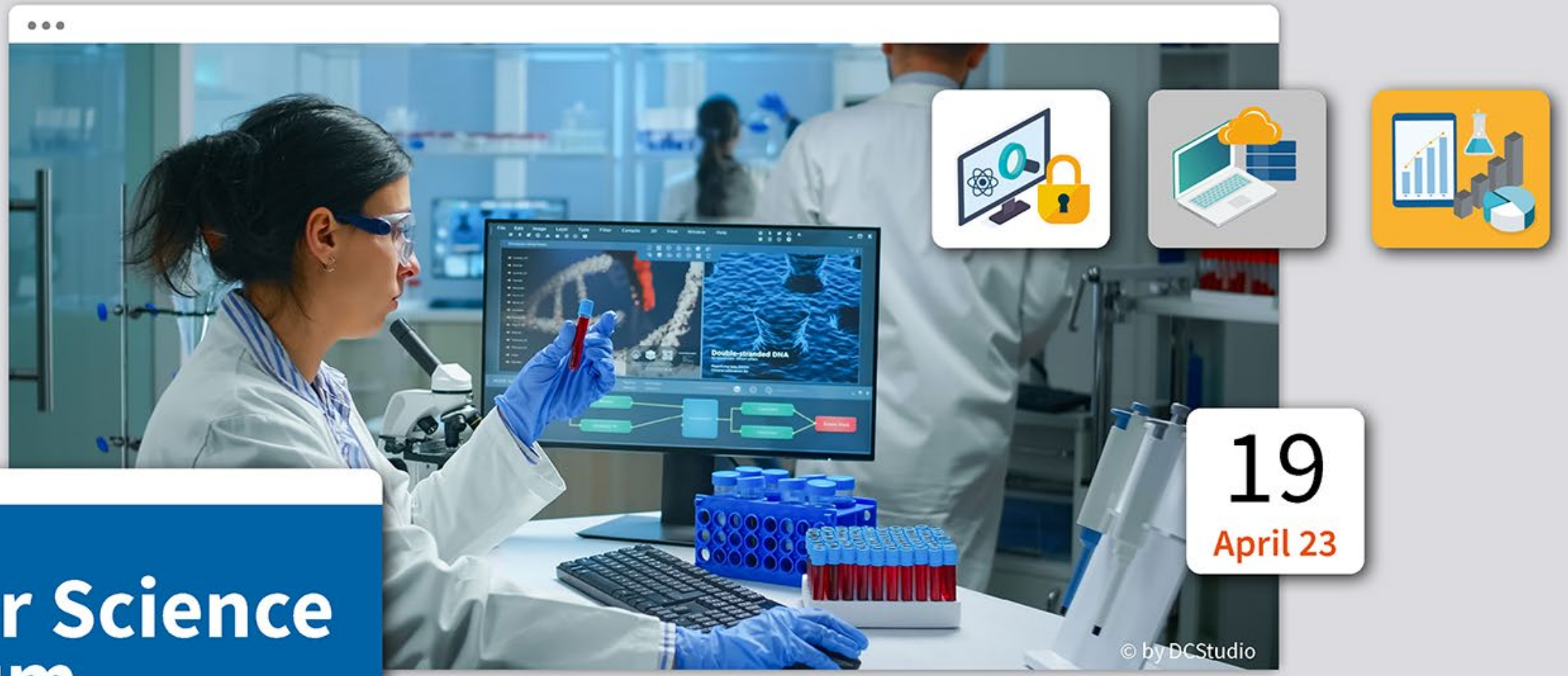
universität wien

IT4S IT for Science FORUM



19
April 23

IT for Science Forum

The speakers

Ronald Maier

Ronald Maier has been Vice-Rector for Digitalisation and Knowledge Transfer at the University of Vienna since 1 October 2019. He is Professor of Information Systems at the University of Innsbruck.

Ulf Busch

Ulf Busch has been CIO of the University of Vienna since 2010. He is the Head of Zentraler Informatikdienst.

Leopold Haimberger

Leopold Haimberger is Professor of Meteorology and Vice Dean for Infrastructure at the Faculty for Geosciences, Geography and Astronomy.

The speakers

Simon Engelberger

Simon Engelberger studied zoology at the University of Vienna. Since 2020 he is the Collections Manager in charge of the University of Vienna Zoological Collection.

János Békési

János Békési is at the ZID since 2017, supervising UNIDAM, a digital asset management system realized with Easydb, and several services for researchers mainly in the digital humanites.

Thomas Rattei

Thomas Rattei is Head of the Division of Computational Systems Biology and Vice-Head of the Department of Microbiology and Ecosystem Science.

The speakers

Ulrich Kiermayr
Ulrich Kiermayr is Head of Communication Networks & Infrastructure and Vice Head of Zentraler Informatikdienst.

Eva Karall
Eva Karall is Head of E-Learning at Zentraler Informatikdienst. The staff unit is responsible for the technical support and further development of the e-learning platform Moodle and tools for digital teaching (BigBlueButton, Zoom) as well as the u:stream services at the University of Vienna.

Michaela Bociurko, Moderator
Michaela Bociurko is Head of IT Communications and Marketing at Zentraler Informatikdienst.

Sara Curtis, Moderator
Sara Curtis is a member of the IT Communications and Marketing team at Zentraler Informatikdienst.

The topics

Opening

- Ronald Maier
- Ulf Busch

Innovative IT solutions in selected scientific projects

- IT for Meteorology and Climate Science, Leopold Haimberger
- Digitalisation of zoological specimens, […] held by the zoological collection of the University of Vienna, Simon Engelberger, János Békési
- DataLife – Data Infrastructure for Life Sciences, Thomas Rattei

IT services of the ZID for science and research

- Supporting projects in all stages, Ulrich Kiermayr
- Overleaf – Collaborative Online LaTeX Editor, Eva Karall

Summary and Outlook

- Ronald Maier

Opening

- Ronald Maier
- Ulf Busch

Innovative IT solutions in selected scientific projects

- IT for Meteorology and Climate Science
- Digitalisation of zoological specimens, historical photographs and other archival materials held by the zoological collection of the University of Vienna (UVZC)
- DataLife – Data Infrastructure for Life Sciences

universität wien

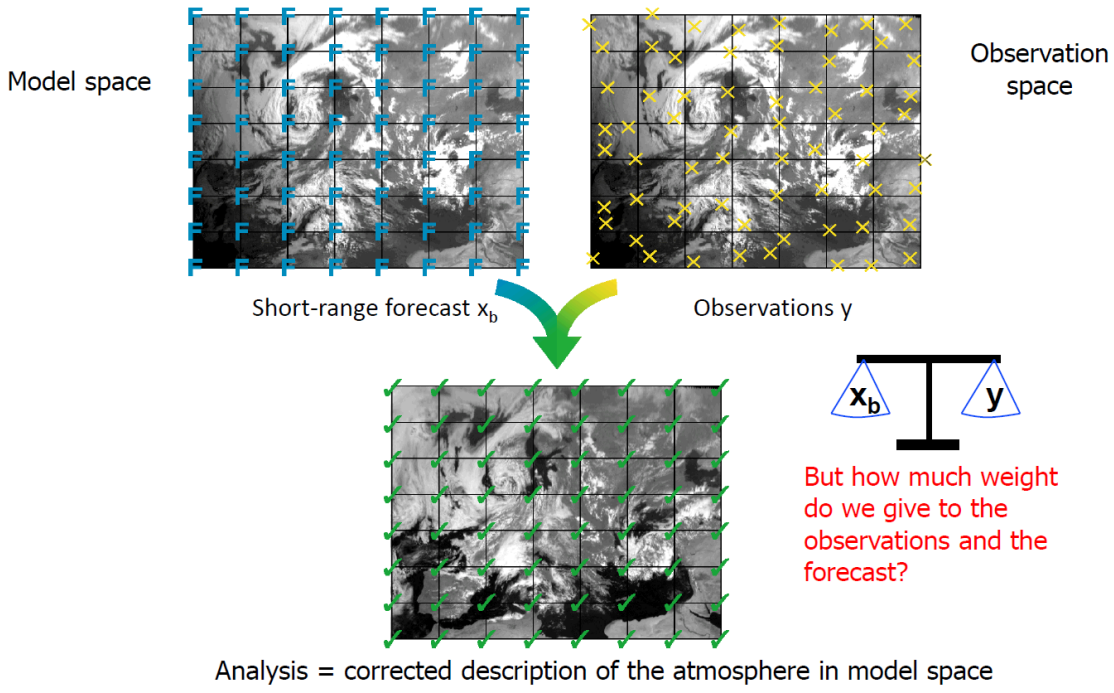
IT4S IT for Science FORUM

IT for Meteorology and Climate Science

Presentation by Leopold Haimberger

Outline

- **Our research topics**
  - Data assimilation
    - How to get satellite and weather observations onto a model grid
  - Climate Data Analysis and Climate Modeling
  - Ensemble forecasts
  - Monitoring trace substances
- **Our computing environment**
- **Contributing to FAIR data**
- **Towards cloud computing**
- **Challenges**

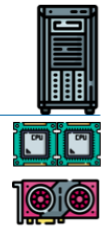


Our local resources

Bare metal @ Department – SRVX – JET – VSC Nodes

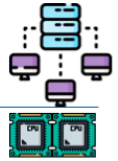
SRVX1

- Arsenal
- 4x CPU
- **Development Node**
- **Teaching Node**
- Storage Node
  - 400 TB (users)
  - 400 TB (scratch)



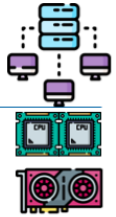
SRVX8

- Arsenal
- 2x CPU
- 1x GPU
- **Visual Node** (VNC)
- Storage Node
  - SSD RAID (11TB)



JET

- Arsenal
- 2x Login
- 7x Compute CPU
- **Computing/Post Processing Cluster**
- Storage 1.5 PB Global Storage
- Scheduler (**slurm**)
- Planed: Extend num. Nodes + Storage



VSC 4 / 5

- Arsenal
- **VSC4** - 5x CPU
- **VSC5** – 11x CPU
- VSC5 – 1x GPU
- **Computing Cluster**
- Shared HOME (200GB)
- Shared DATA (100TB)
- Scheduler (**slurm**)
- Projects can be requested (resources)

- HPC software stacks (FORTRAN, MPI, OpenMP, Python)
- Jupyter hubs on local servers and VSC
- Both for research and teaching

Main External computing facilities

ECMWF+Copernicus
- reliable partner since decades
- HPC and 100 PB data archive

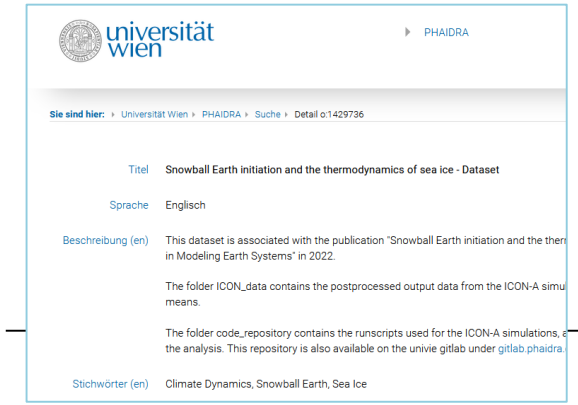
K-computer (Japan)
1000 member forecast ensemble

PHAIDRA, GitLab and Mattermost, Confluence – Don’t switch it off!

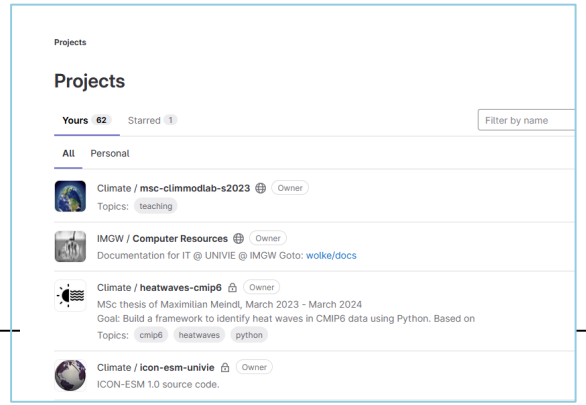
- Phaidra for long-term archiving of climate model data and analysis scripts
- GitLab for code development and project management (incl. external collaborators)
- Mattermost for fast internal communication
- Departmental Wiki pages for internal administration and communication

tricco
Project ID: 631

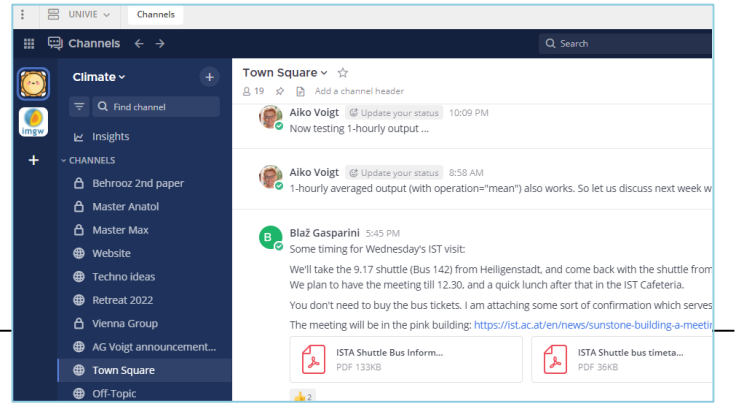
PHAIDRA



GitLab

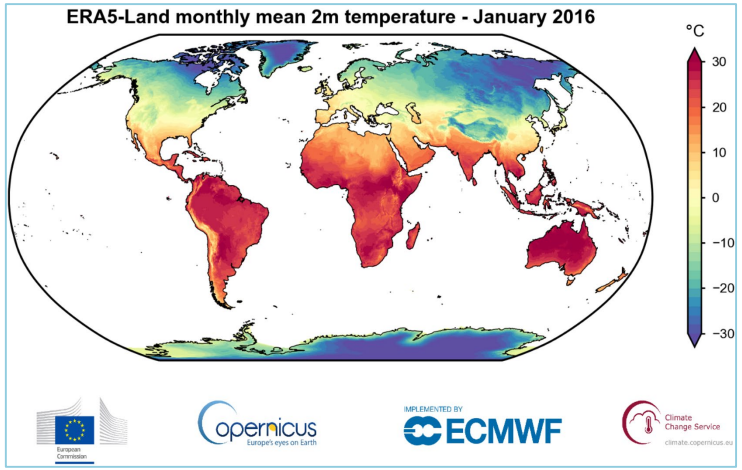


Mattermost



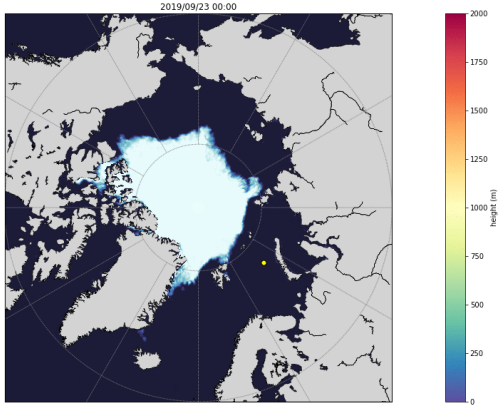
Copernicus Climate Reanalysis – ERA5

cds.climate.copernicus.eu/



- Gridded 3D atmospheric data set of the past 80 years (1940-)
- Ideal to learn about the state and evolution of recent climate
- Quality-certified, free download via Copernicus
- ~10 PB, 3000 cites/year, hugely popular

- We have copied small fraction, would like to have more, already uses 15% of our disk space.

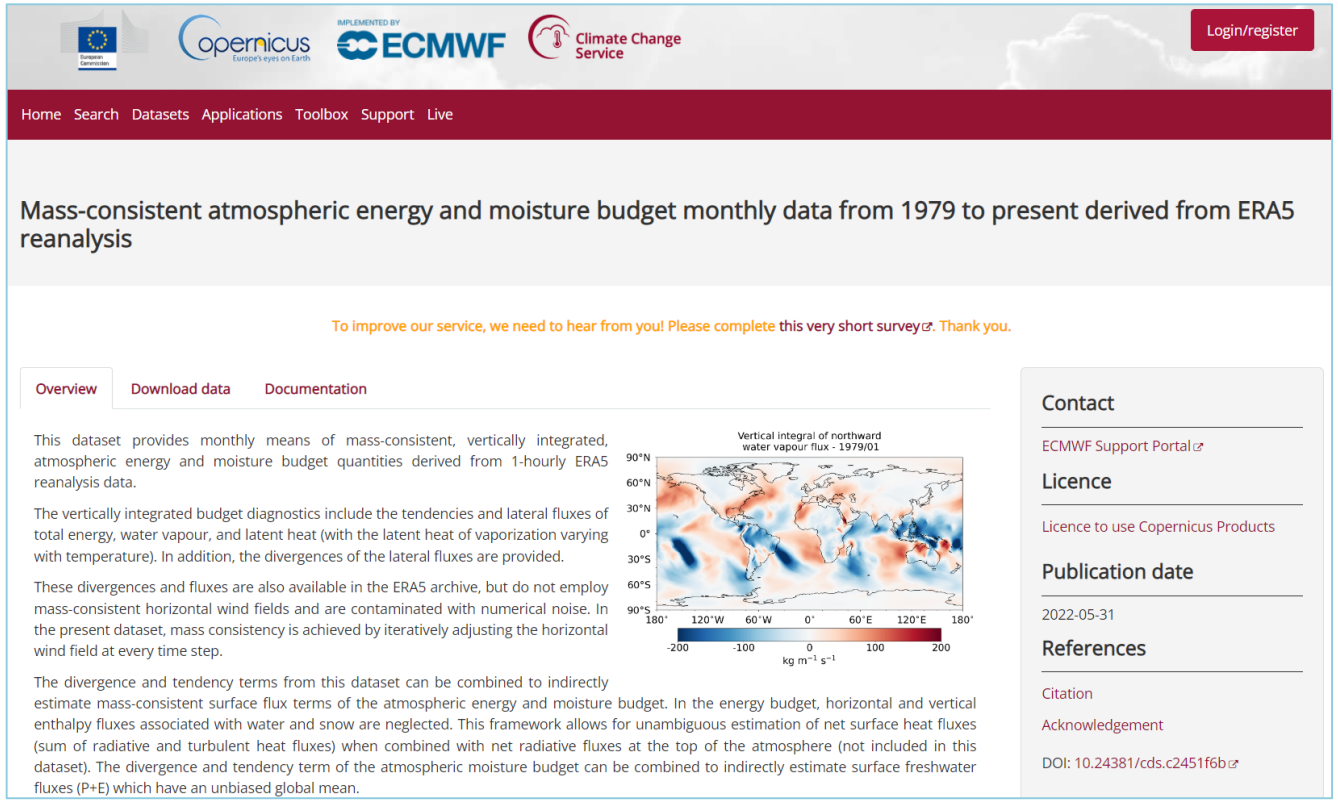


Wikimedia commons

https://flexteam.univie.ac.at/research/

Publishing@Copernicus

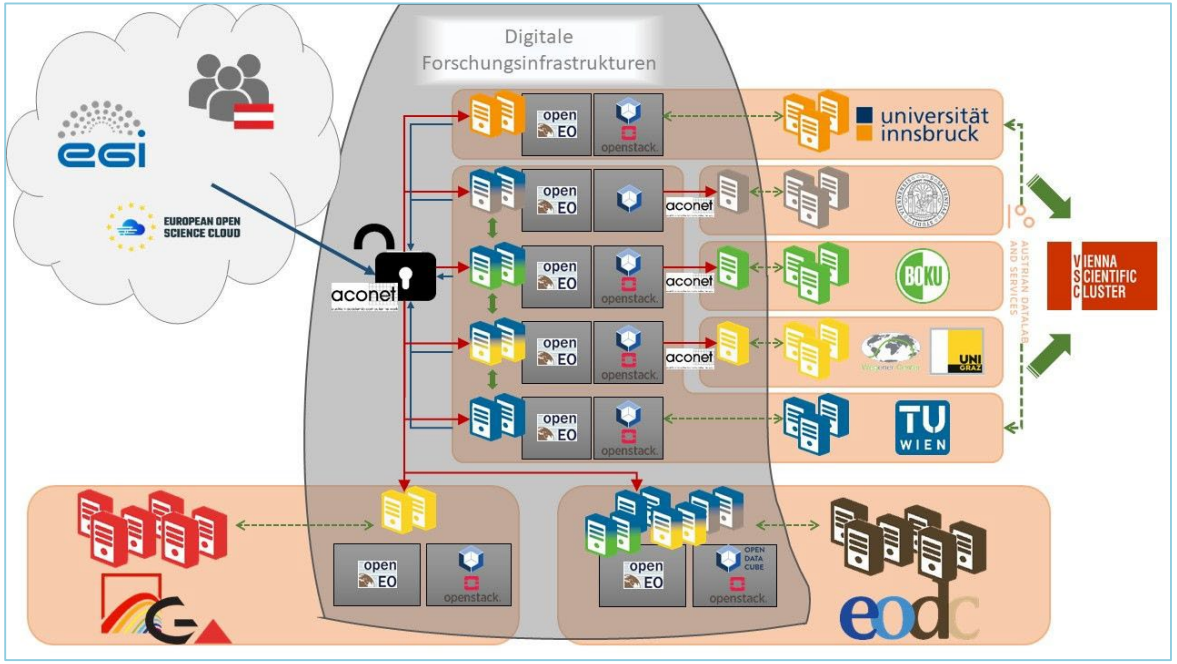
- Data store has all ingredients of a FAIR data repository
- Very liberal license
- Copernicus takes care of data lifecycle
- **But where is the logo?**
- Should we have published it on PHAIDRA?



Bundle Geoscience IT resources in Austria – Cloud4GEO

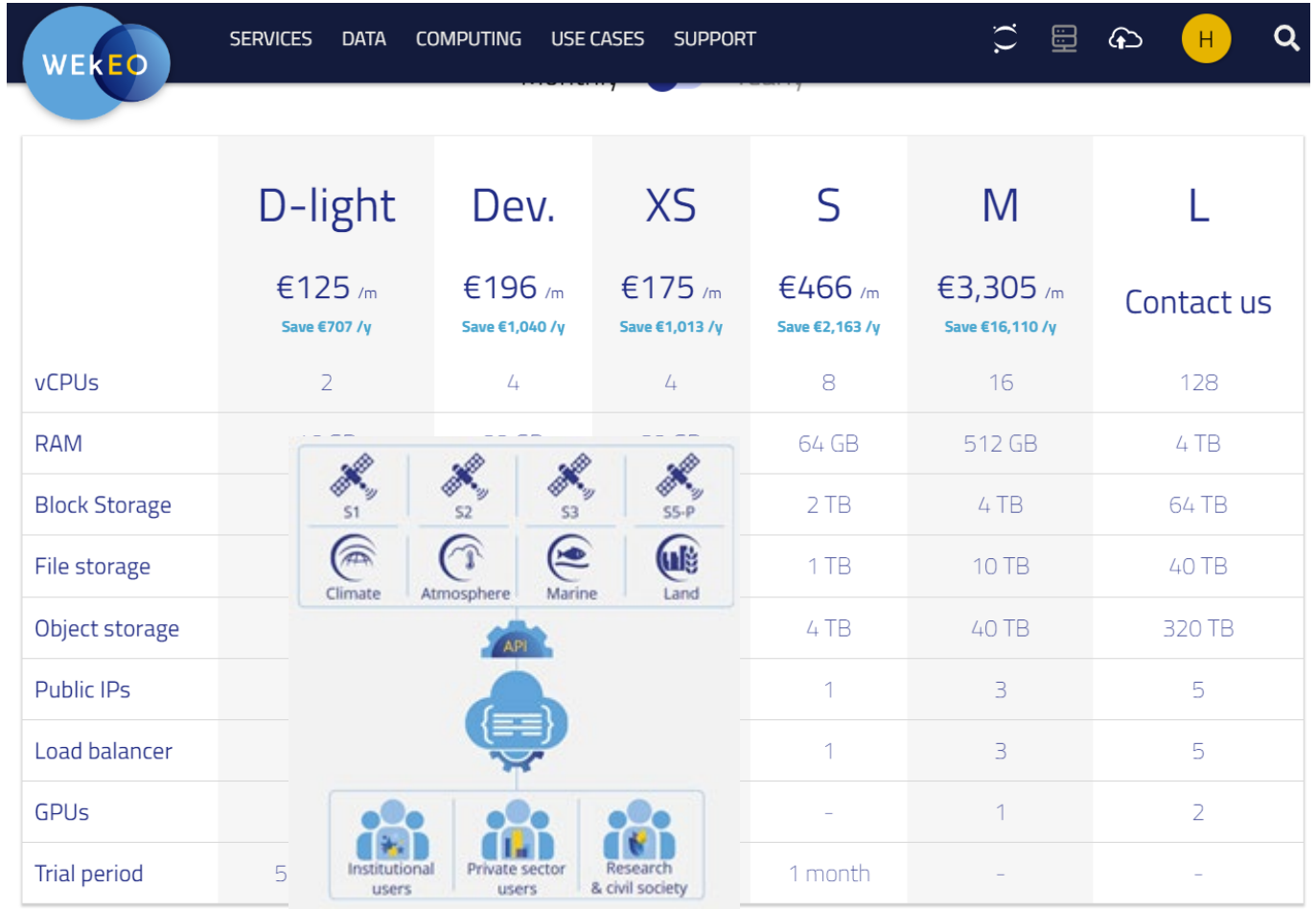
Bundesministerium
Bildung, Wissenschaft
und Forschung

- **Create** a powerful distributed cloud for geospatial data processing
- Single Sign-on
- Focus on interoperability: Open Standards
- Link to EOSC and EGI, Copernicus
  - not only as user, but also
  - share our data there (also compute)
- FAIR gives credit, attracts users!
- Data are „near“: high bandwidths, low latency
- Expensive to build and maintain



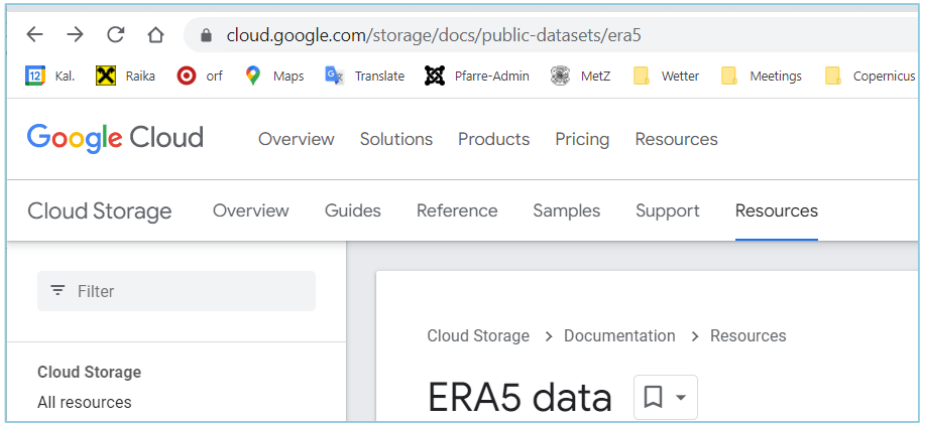
Are there other choices?

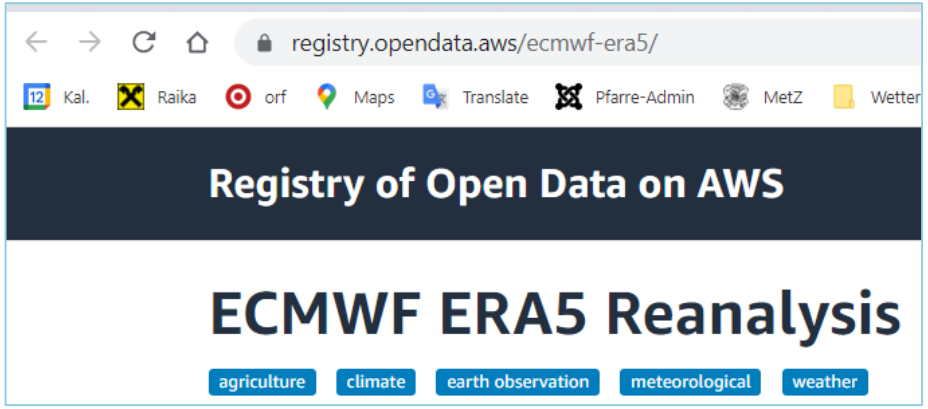
- Run our codes where the data are.
- Copernicus DIAS (e.g. WEkEO https://www.wekeo.eu/about)
- Free Trial
- Access via JupyterHub, VM
- Provides **harmonised, high bandwidth** data access to many PB of EO data
- Data free, only computing costs
- Some teething troubles, but improving.
- Next Generation: PHIDIAS Prototype of HPC/Data Infrastructure for On-demand Services



Business likes FAIR data!

- Copernicus' aim is to support companies with climate and monitoring data
- Well received by small companies in Europe …
- but also the big fishes …
- **Google and Amazon hold copies of ERA5**





Make money with climate data

- Well supported cloud environment
- 23 USD/TB/month
- 70,000 USD for 300 TB per year

We will not buy from them!

Google Cloud

Pricing tables

Important: Pricing updates for Cloud Storage took effect on Oct. 1, 2022 and on Apr. 1, 2023.

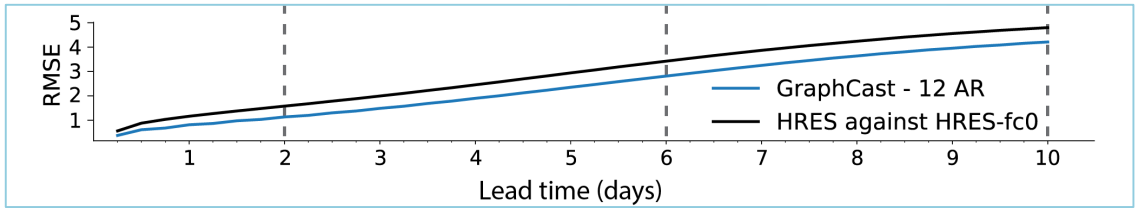
The pricing tables below show what charges apply when using Cloud Storage.

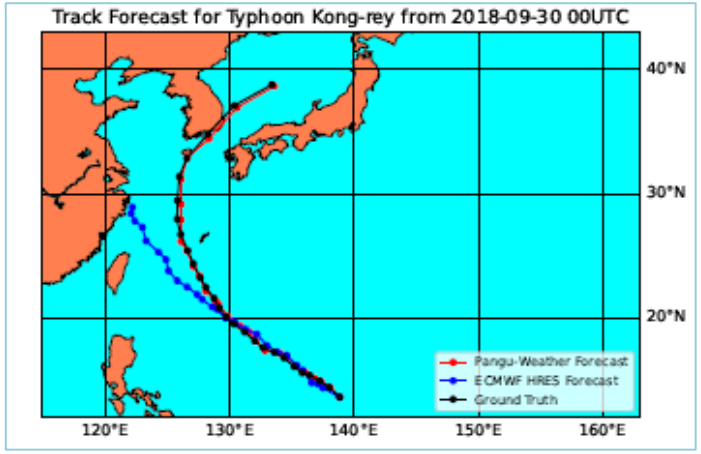
North America | South America | Europe | Middle East | Asia | Indonesia | Australia

Location	Standard storage (per GB per Month)	Nearline storage (per GB per Month)	Coldline storage (per GB per Month)	Archive storage (per GB per Month)
Warsaw (`europe-central2`)	$0.023	$0.013	$0.006	$0.0025

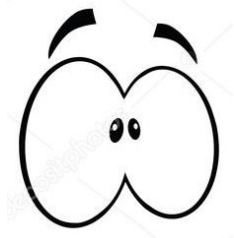
What Google does with ERA5

Train AI (GNN) for weather forecasting





Bi et al. 2022

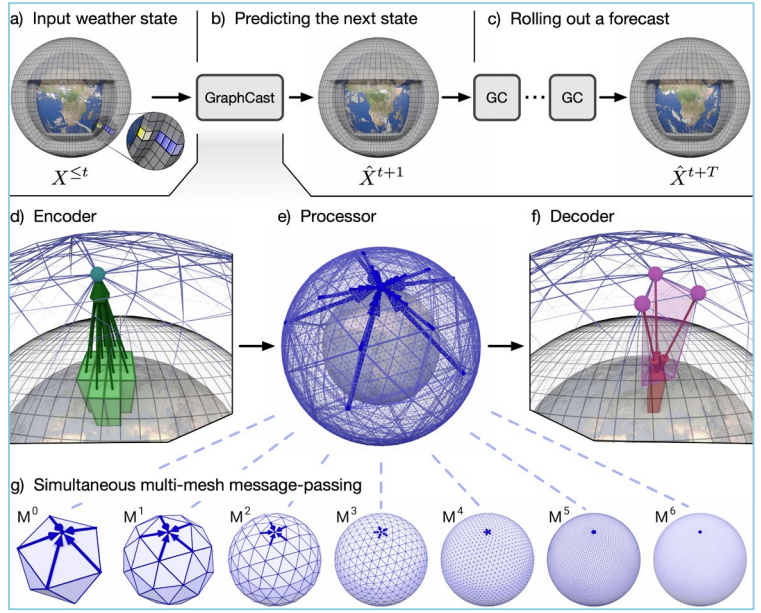


„We also thank ECMWF for providing invaluable datasets for the research community“

GraphCast: Learning skillful medium-range global weather forecasting

Remi Lam*,1, Alvaro Sanchez-Gonzalez*,1, Matthew Willson*,1, Peter Wirnsberger*,1, Meire Fortunato*,1, Alexander Pritzel*,1, Suman Ravuri[1], Timo Ewalds[1], Ferran Alet[1], Zach Eaton-Rosen[1], Weihua Hu[1], Alexander Merose[2], Stephan Hoyer[2], George Holland[1], Jacklynn Stott[1], Oriol Vinyals[1], Shakir Mohamed[1] and Peter Battaglia[1]

*equal contribution, [1]DeepMind, [2]Google



Conclusions

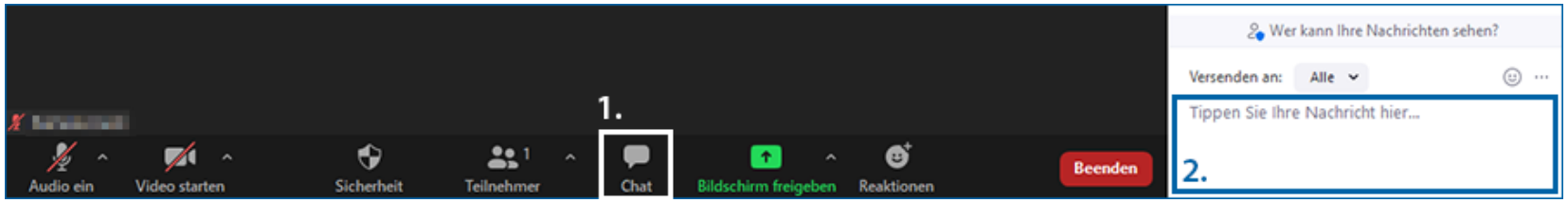
- We like the UNIVIE RDM tools a lot!
- Power of FAIR EO data quickly unfolding
  - Data assimilation plays an important role
  - We are small players, but let the big picture shine a little more
- Frequent quick scans through large amounts of data
- We need affordable „near“ cloud computing service
  - Build one for EO in Austria – ongoing
  - Tap into European resources – just starting
- Strong winds of change – AI challenge
  - Do we invest wisely, can we adapt quickly enough?

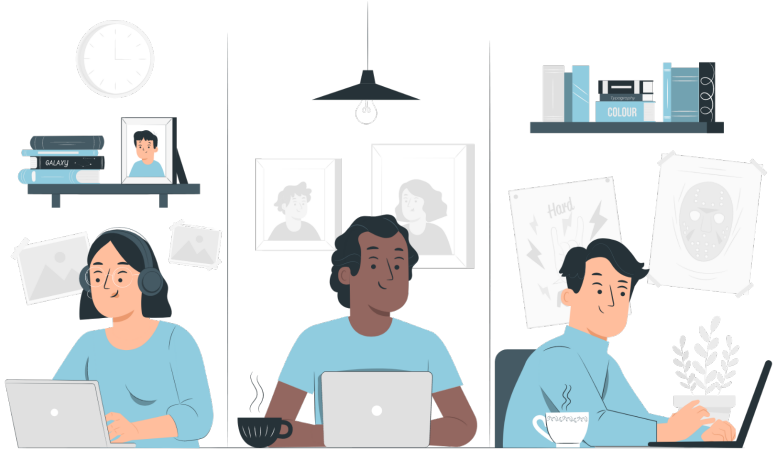
© Image by Suez Canal Authority / AP

© Image by https://www.flickr.com/photos/jnandreae/)

Questions about IT for Meteorology and Climate Science?

- Please ask your questions as a chat comment.

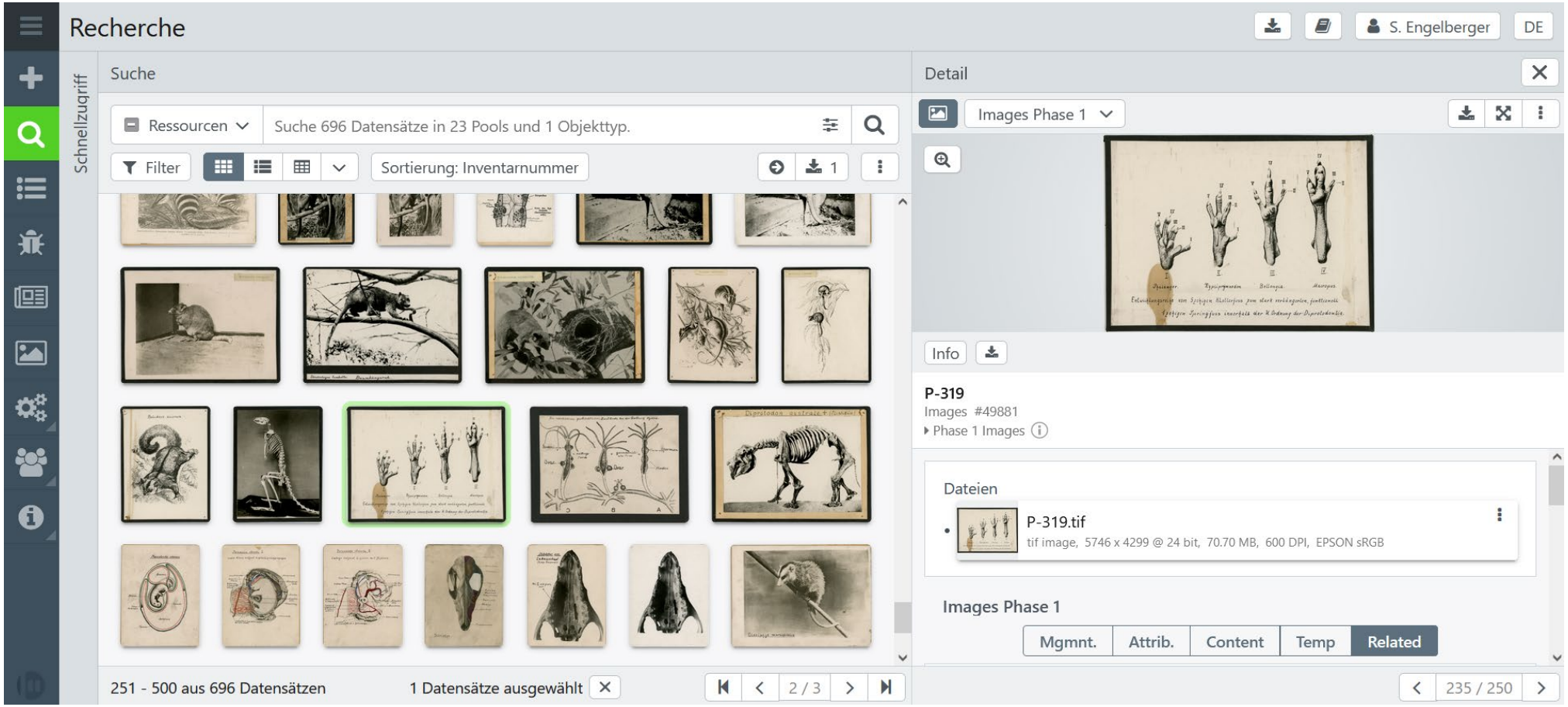




Digitalisation of zoological specimens, historical photographs and other archival materials held by the zoological collection of the University of Vienna (UVZC)

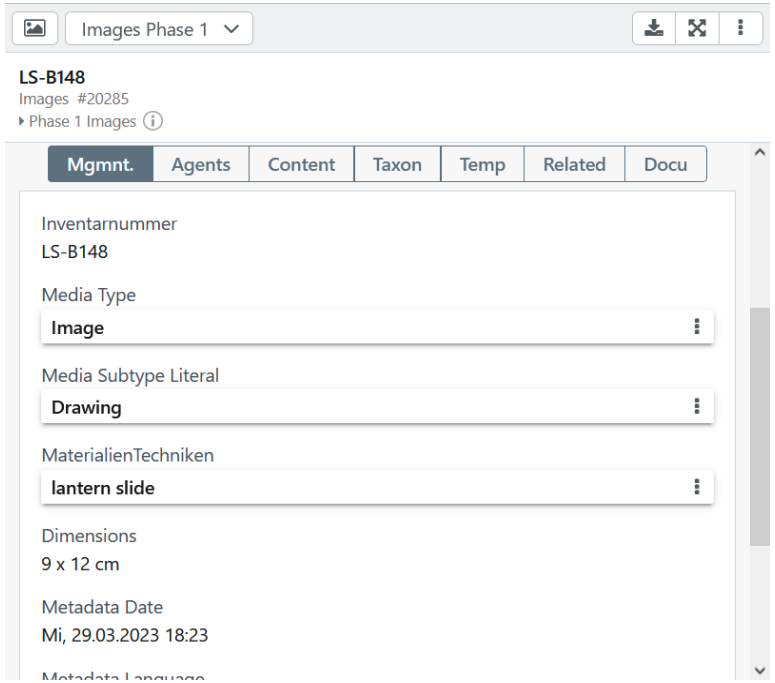
Presentation by Simon Engelberger and János Békési

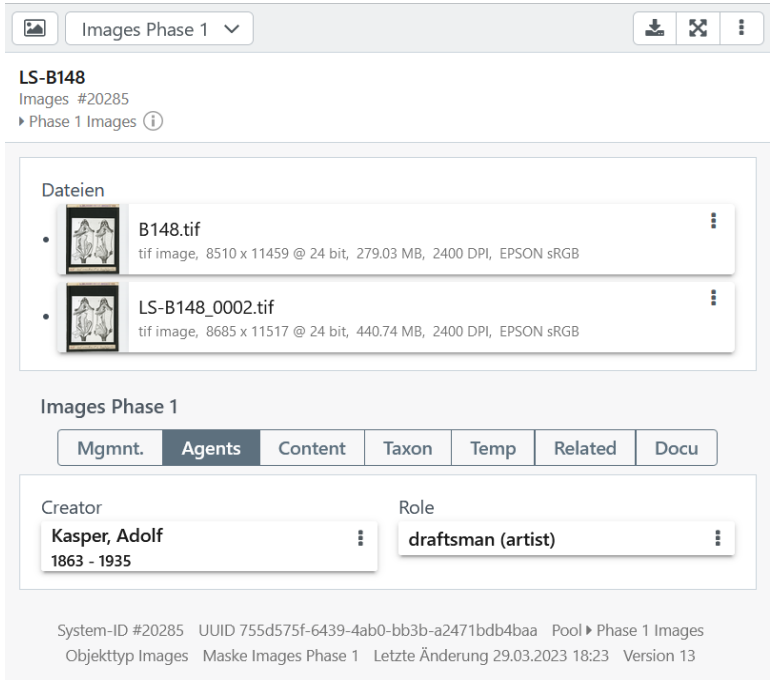
AV Media

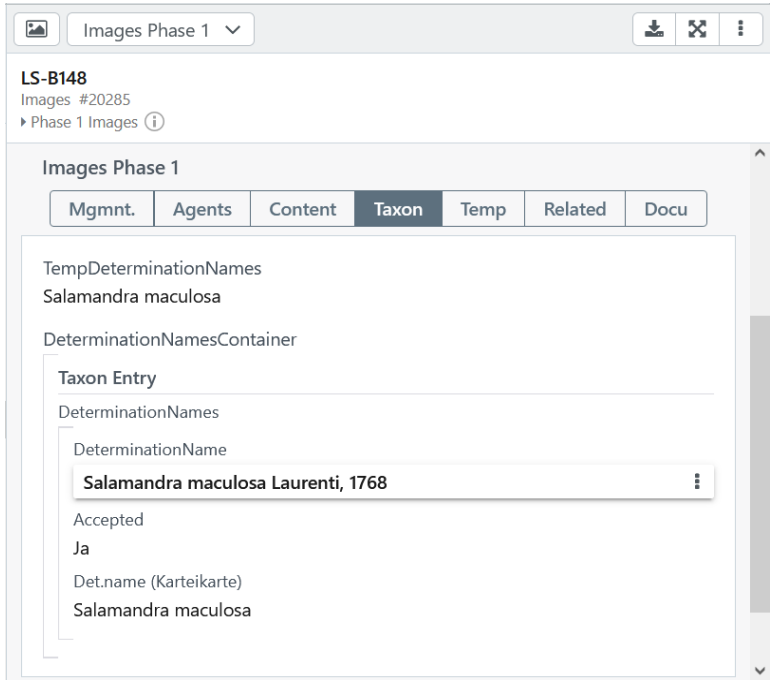

Recherche
S. Engelberger
DE
Schnellzugriff
Suche
Ressourcen
Suche 696 Datensätze in 23 Pools und 1 Objekttyp.
Filter
Sortierung: Inventarnummer
1
251 - 500 aus 696 Datensätzen
1 Datensätze ausgewählt
2 / 3
Detail
Images Phase 1
Info
P-319
Images #49881
Phase 1 Images
Dateien
P-319.tif
tif image, 5746 x 4299 @ 24 bit, 70.70 MB, 600 DPI, EPSON sRGB
Images Phase 1
Mgmnt.
Attrib.
Content
Temp
Related
235 / 250

AV Media: Standards

- Standard: Audiovisual Core (Audubon Core), AC
- Controlled Vocabularies
  - AC
  - Art & Architecture Thesaurus® (AAT)

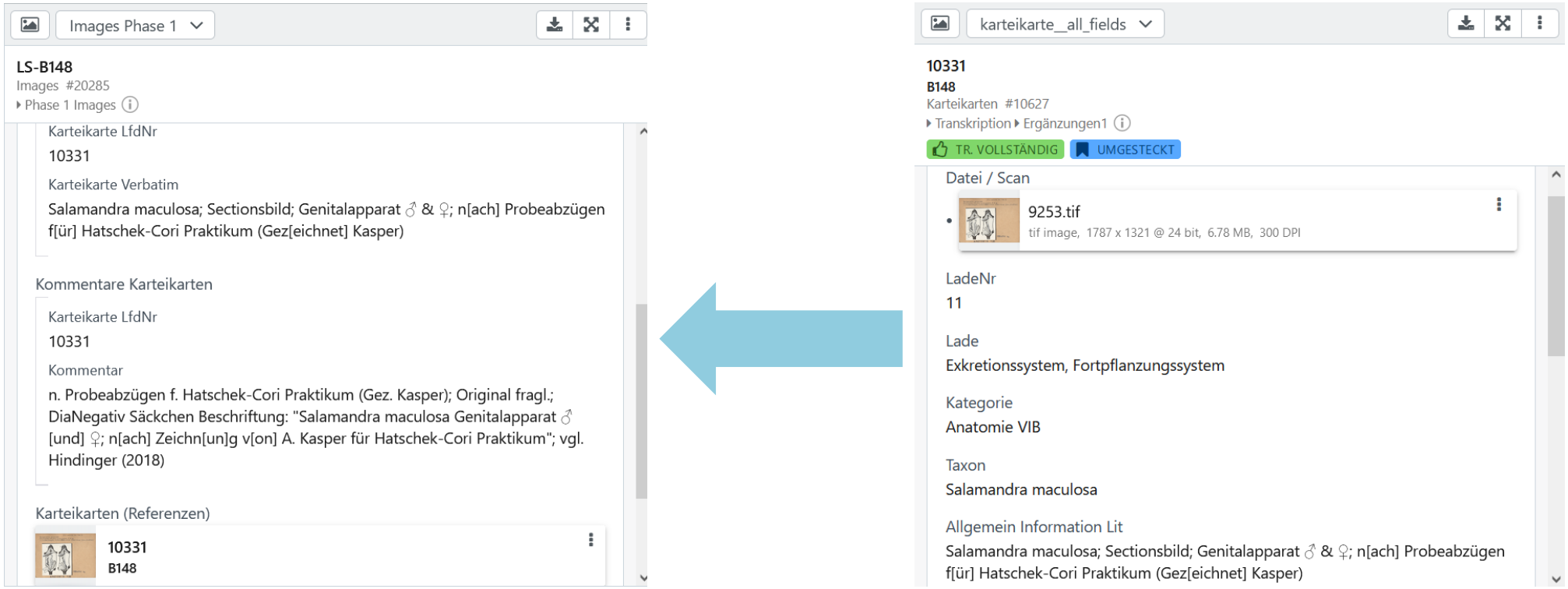






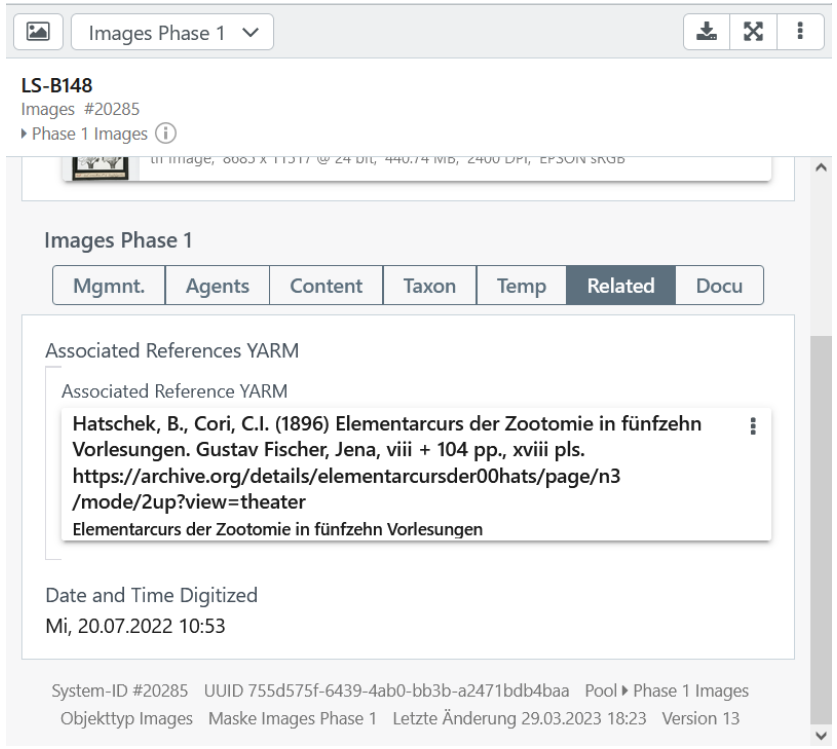
AV Media: Interlinks

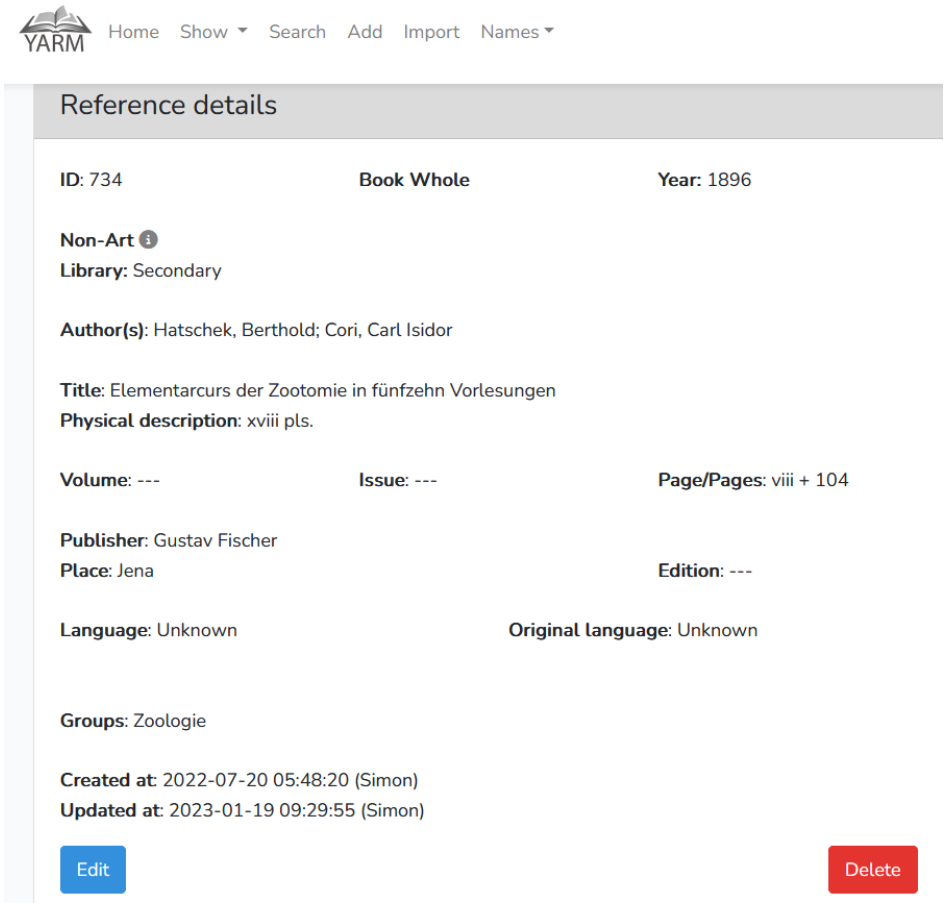
- Linked with digitised Card Files



AV Media: Linked bibliographic data

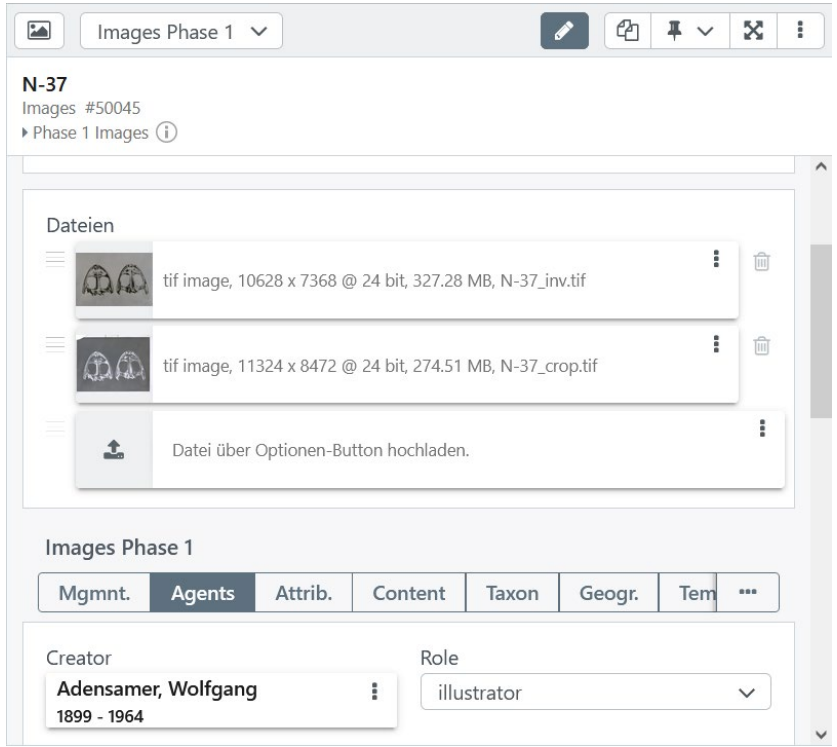
- Linked with YARM Reference Manager

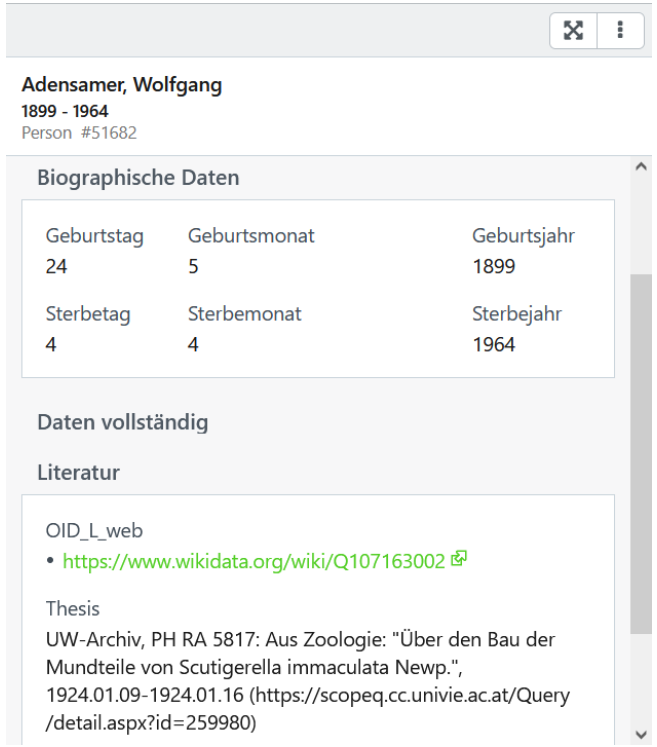




AV Media: Person IDs

- Controlled list of persons
- Including wikidataID for all persons



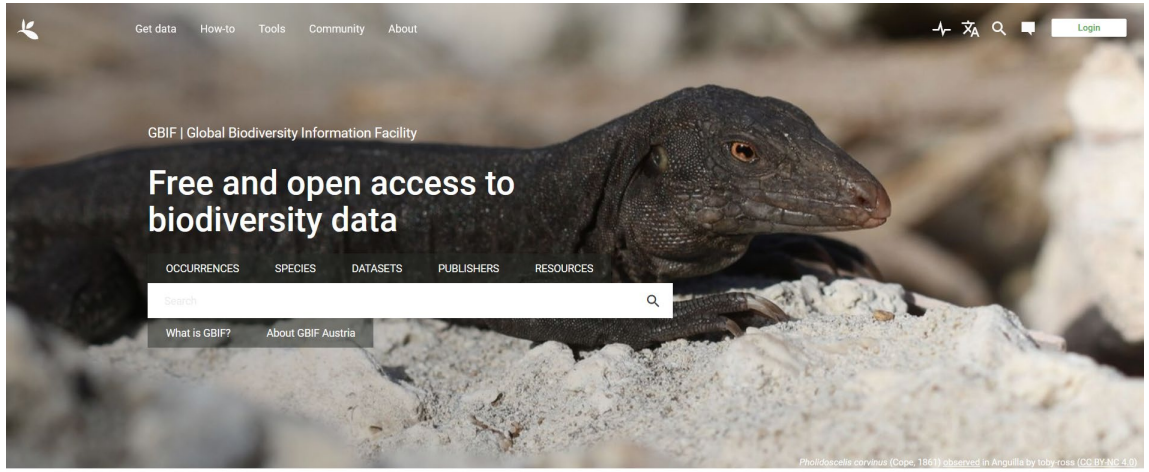


Future steps

- DWCA export (xml/csv)
- AC export (xml/csv)

- Link with GBIF Taxonomic Backbone (Global Biodiversity Facility, https://www.gbif.org/)

- Automated export of DWCA dataset to GBIF



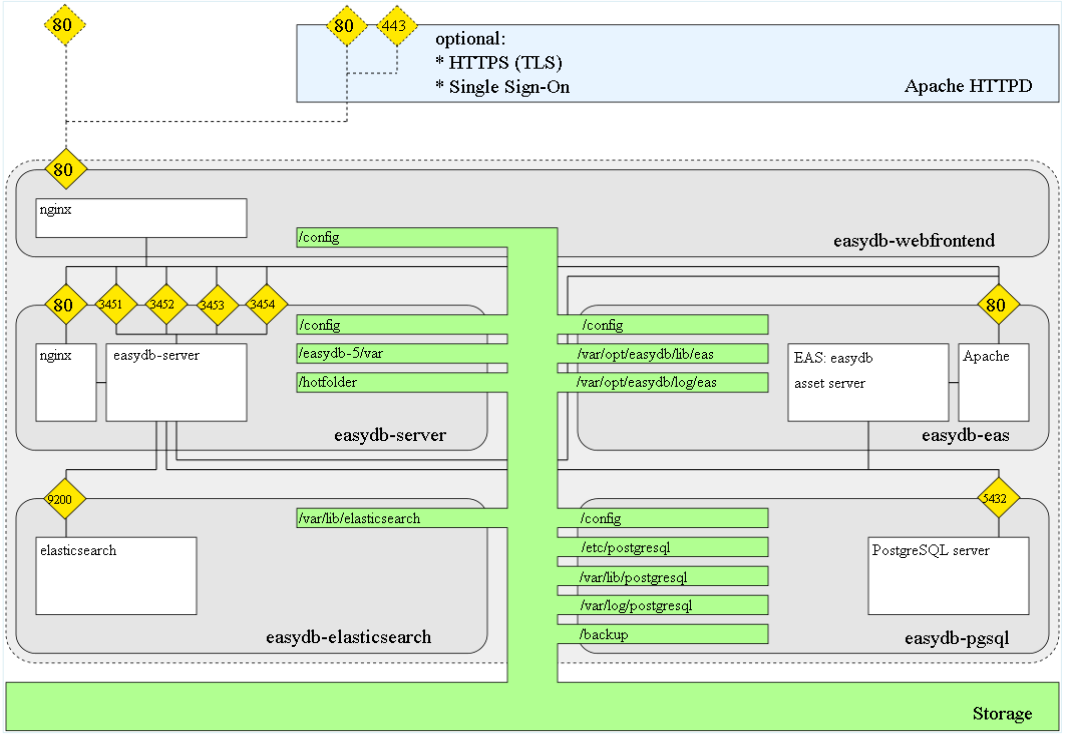
© Image by https://www.gbif.org/

ZID-related: Infrastructure

- Virtual machine in ZID virtual infrastructure (with backup and monitoring)
  - Presently 4 CPUs, 12 GB RAM and 1 TB storage
- Several Docker containers comprise easydb functionality
- File system storage
- Update by pulling new docker images (every 2–3 weeks)
- Mirrored development instance for tests and try-outs (plugins)

ZID-related: Components/containers

- File server („Eas“) in C
- Elasticsearch
- Postgresql
- Base server in Python
- Web frontend in Coffeescript/Javascript



Easydb Docker Configuration

ZID-related: Access and data

Access by other means than by browsers

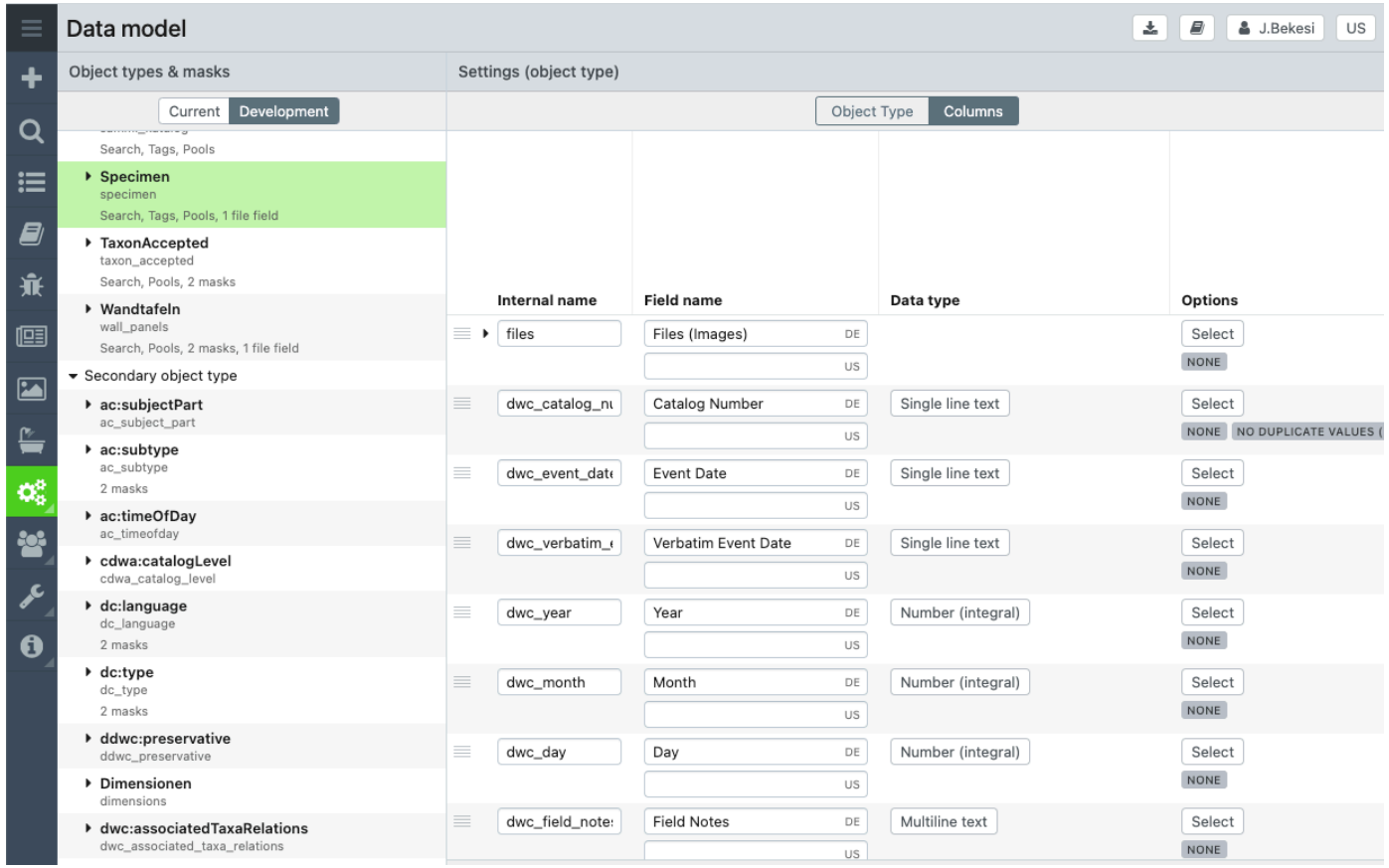
- Plugins for web frontend (Coffeescript)
- Plugins for server and/or API (Python)

© Image by storyset on Freepik

Current data

- 16,000 images
- 54 object types (mainly imported helper objects, 7 main objects)
- 56,000 objects (e.g. taxon or YARM entries) in total

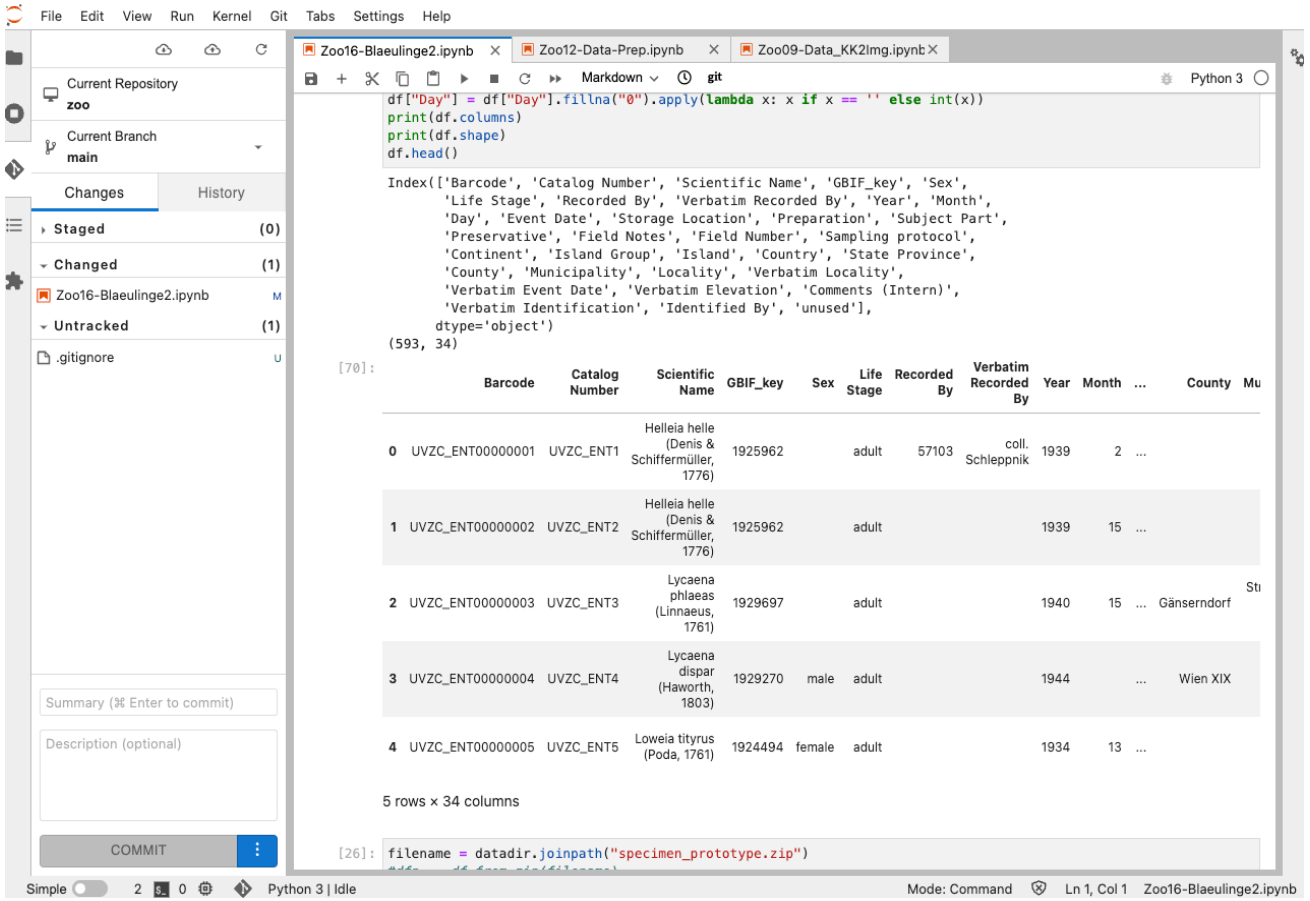
ZID-related: easydb



Data model

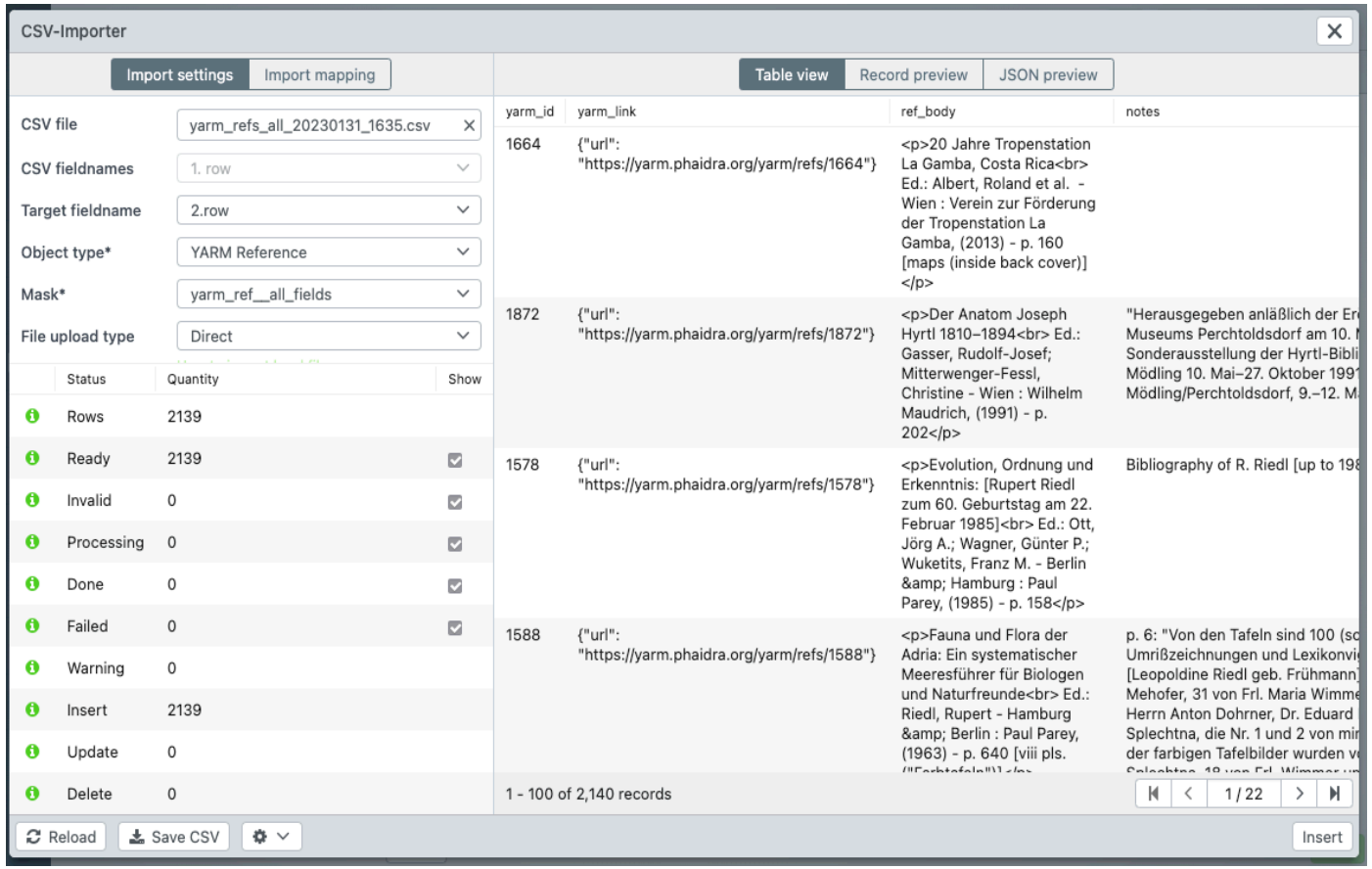
- Specimen object type has about 90 fields

ZID-related: Data refining



- Data is **refined/processed** in **Jupyter notebooks** and **exported** as **CSV files**

ZID-related: Import of metadata



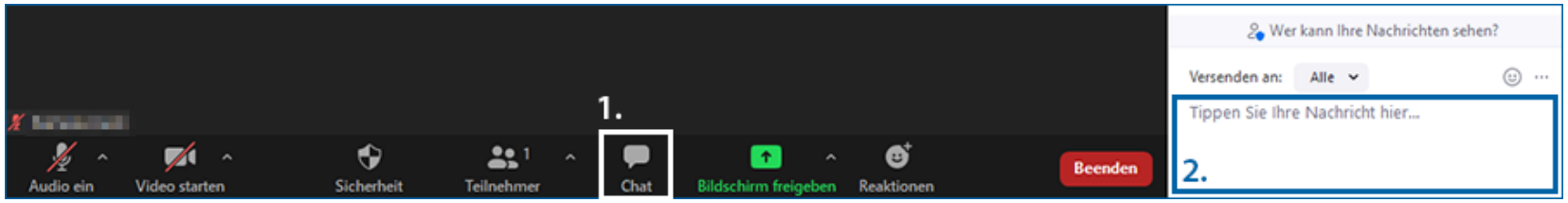
- Metadata are imported and/or bulk updated as **CSV files**

URLs and workflow

- URLs of interest:
  - Documentation of easydb: https://docs.easydb.de/en/
  - General info for easydb: https://www.programmfabrik.de/en/easydb/
- Our workflow: Nothing special → e-mail, telephone, feedback loops, try-outs
- Contact:
  - simon.engelberger@univie.ac.at
  - janos.bekesi@univie.ac.at

Questions about Digitalisation of archival materials?

- Please ask your questions as a chat comment.



© Image by storyset on Freepik

universität wien

IT4S IT for Science FORUM

DataLife – Data Infrastructure for Life Sciences

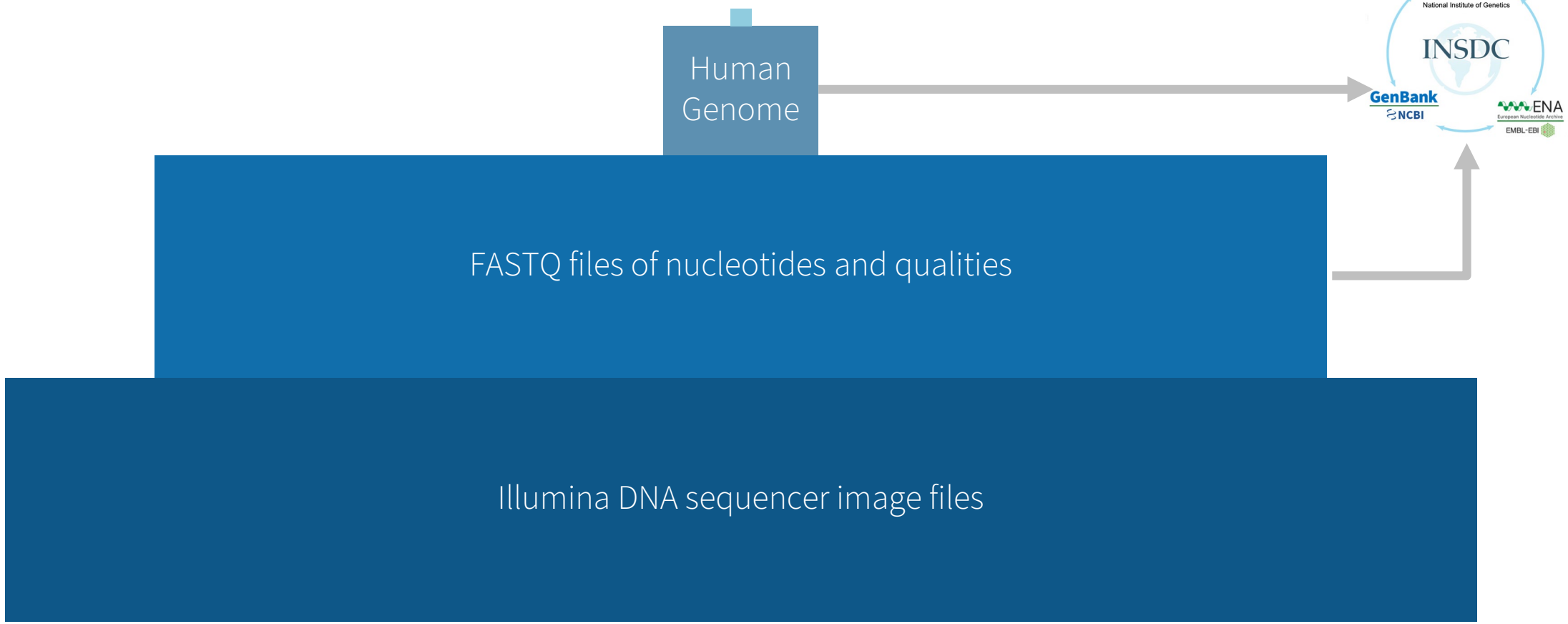
Presentation by Thomas Rattei

Data in Life Sciences

- Huge data collection in short time:
  - Biological and medical imaging
  - Real-time microscopy
  - DNA and RNA Sequencing
  - Chemical analytics
- Interpretation with machine learning and artificial intelligence
  - Transformation of data into features
  - Real-time access to large data for training
  - Compute-intensive learning

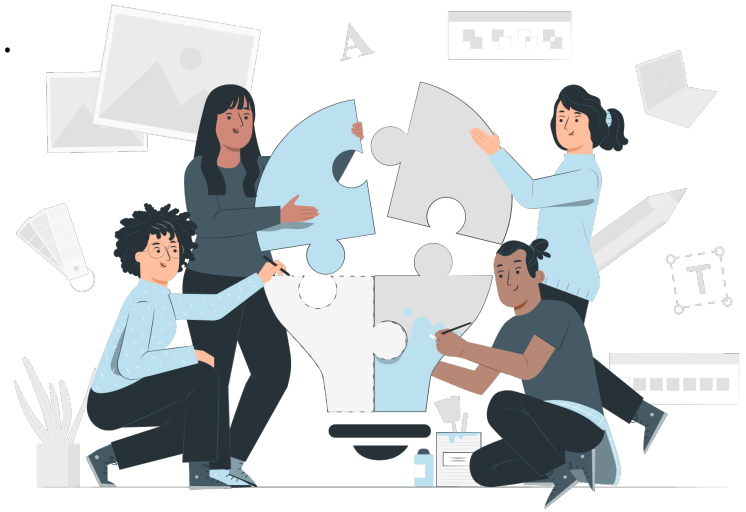
Example: Genome data

1 % of the human genome differ between individuals



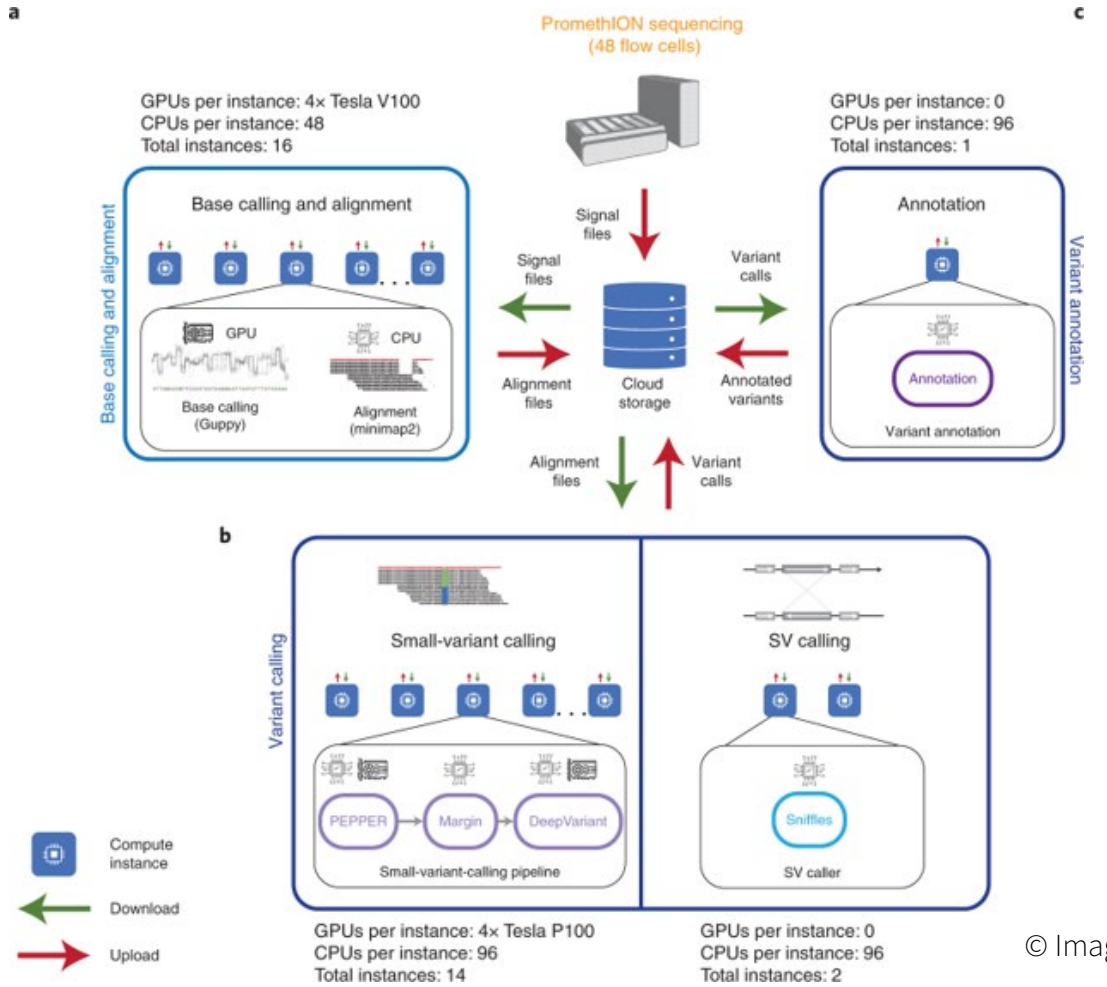
Future needs: data-intensive projects

- FWF Cluster of Excellence „**Microbiomes Drive Planetary Health**”
  Centre for Microbiology and Environmental Systems Science et al.
- **ERC Advanced Grants**
  - Manuel Zimmer
  - Michael Wagner
- **Wittgenstein Awards**
  - Christa Schleper
  - Michael Wagner



© Image by storyset on Freepik

Data acquisition, storage and processing



PromethION 48 Specification:

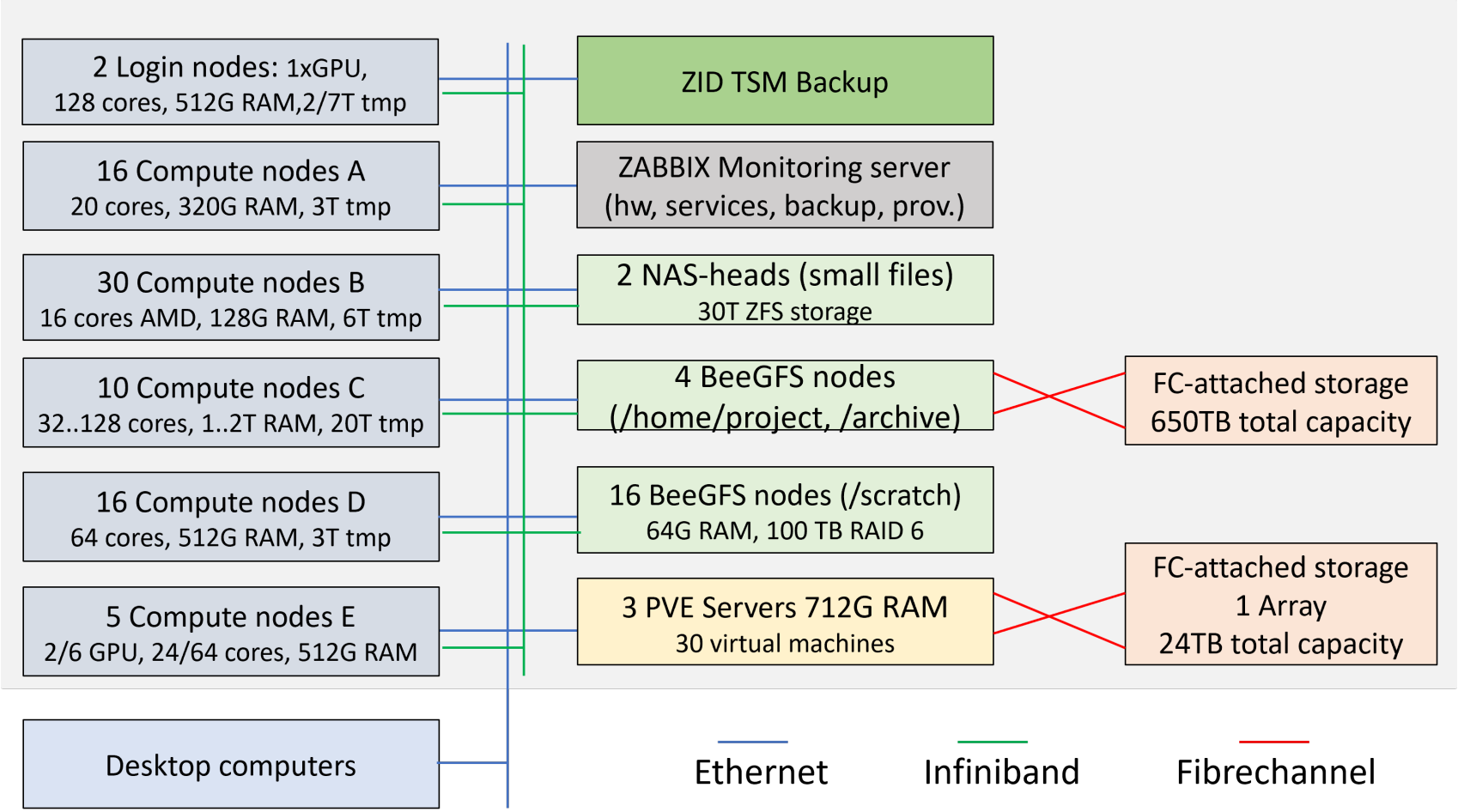
Sequencer:
- H 190 x W 590 x D 430 mm, 25 kg

Data Acquisition Unit:
- H 440 x W 178 x D 470mm, 25 kg

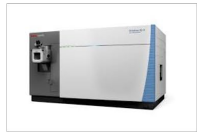
Compute spec:
- 64 TB SSD Storage
- 384 GB RAM
- Intel CPU
- 4 x GV100 basecall accelerators

© Image by *Goenka et al., Nature 2022*

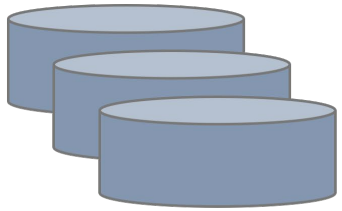
Now: Life Science Compute Cluster



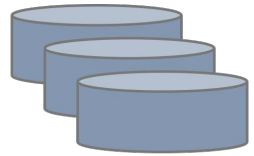
Next: DataLife project



MedUni



Vienna Biocenter



Arsenal

UZA/Chemistry

Funding by BMBWF in “Digital Infrastructures” initiative. Total costs 2.5 mio, ministry contribution 1.8 mio.

Data creation sites

- Ahrends group Metabolomics/Lipidomics (Chemie 1090)
- BioOptics Facility (Max Perutz Labs DBG VBC5)
- EDGE (UZA2 1090)
- EM Facility (IST Austria)
- EM Facility (VBCF)
- FACS Facility (Max Perutz Labs)
- Gerner group Proteomics (Chemie 1090)
- Joint Microbiome Facility / BSF (Anna Spiegel Building CeMM)
- Joint Microbiome Facility / MedUni (AKH 5P)
- Joint Microbiome Facility / CMESS (UBB)
- MassSpec Facility (Max Perutz Labs / VBCF)
- Menche group (Max Perutz Labs (DBG VBC5))
- NGS Facility (VBCF / BSF Anna Spiegel Building CeMM)
- Pharmazie (UZA2 1090)
- Strukturbiologie (Max Perutz Labs (DBG VBC5)
- Zanghellini group Metabolic Modelling (Chemie 1090)
- Zimmer group (UBB)

Data processing sites

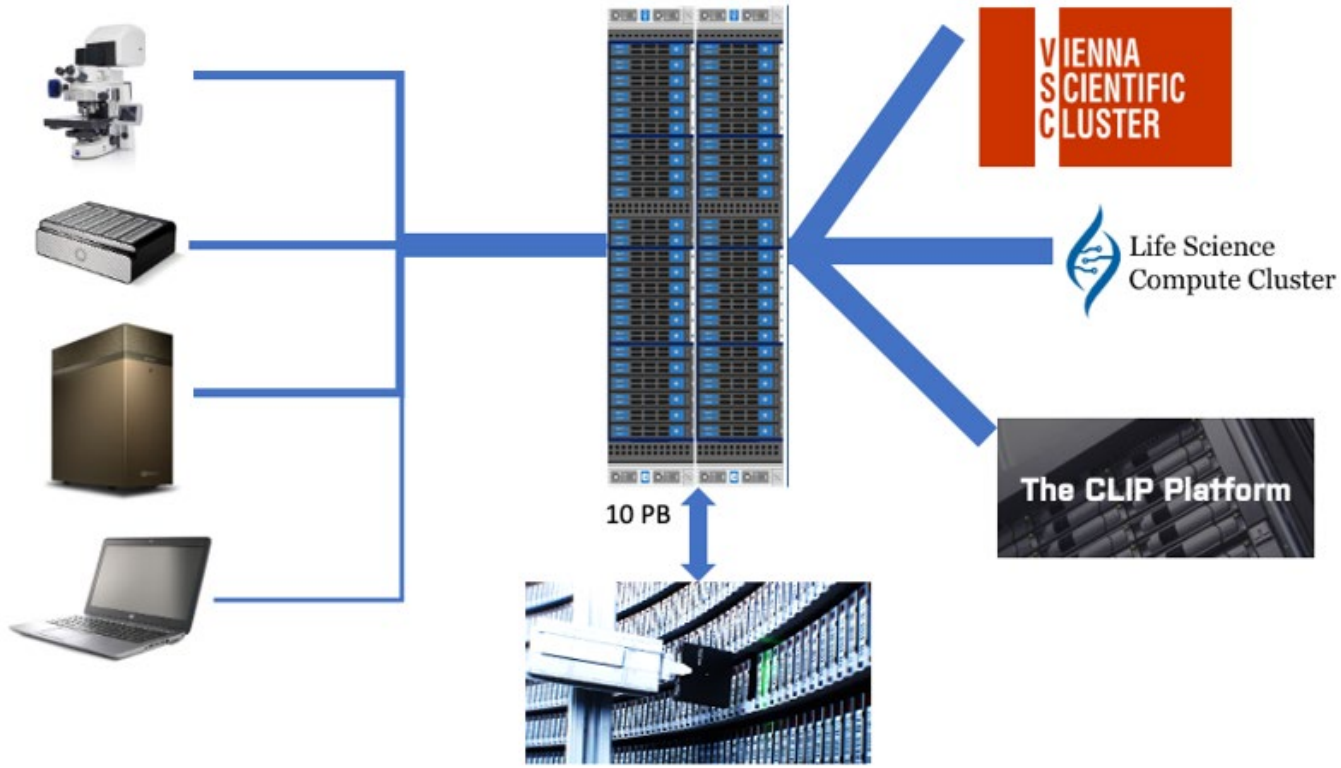
- Vienna Scientific Cluster
- CLIP/CBE of ÖAW
- CIBIV-Cluster at Max Perutz Labs
- Life Science Compute Cluster of University of Vienna (LeWi, CMESS, Chemistry and JMF)
- MedUni Wien Computing facilities
- Laboratory of Computational Biophysics high-performance cluster (Max Perutz Labs)
- Dorothy-Cluster (Max Perutz Labs, Structural Biology)

universität wien

IT4S IT for Science FORUM

High-performance data storage

- GPFS Storage with IBM Active File Management (AFM), partially redundant
- Storage nodes with Flash-Disk ratio of 1:10



Network infrastructure

Data creation sites <-> Storage

- Fast(er) LAN connections

Storage <-> Data processing sites

- Direct dedicated networks with multiple 100 Gbit/s Ethernet RDMA links

Side effects:

- Newly created redundant 100 Gbit/s between UBB and Arsenal via NIG and University Main Building
- Redundant 100 Gbit/s between Medical University resp. MeduniBackup and Arsenal (via NIG, via Campus3A/University Main Building)

DataLife work/time plan

1. Concept (2023)
2. Planning (2023/24)
3. Realization (2023/24)
4. Operation (2025/26)

Two additional FTE:

- Technician (ZID)
- Data Scientist (Vienna Biocenter)

© Image by storyset on Freepik

Data management and harmonisation

Training of user community needed:

- Open Science
- FAIR principles
- Data lifecycles and storage economy
- Data standards

Help by:

- 2 data stewards
- LiSC Servicedesk

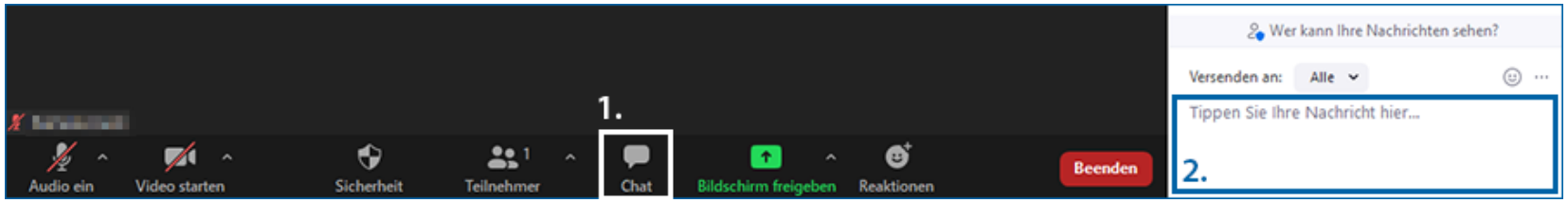
© Image by storyset on Freepik

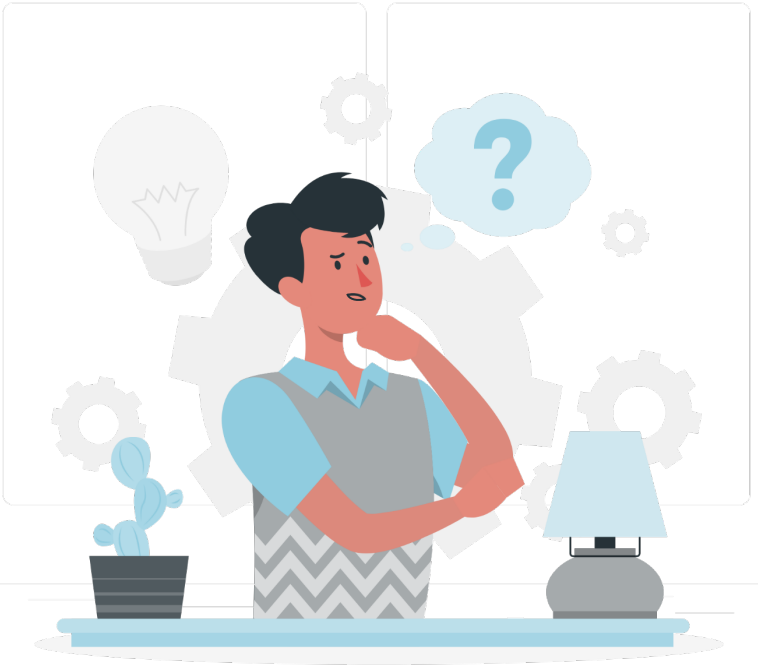
Outlook and further impacts

- Contribution to national and international infrastructure:
  - European Open Science Cloud
  - EURO-Bioimaging
  - ELIXIR
- Formation of ELIXIR Austria node
- Contact: thomas.rattei@univie.ac.at

Questions about DataLife – Data Infrastructure for Life Sciences?

- Please ask your questions as a chat comment.





© Image by storyset on Freepik

universität wien

IT4S IT for Science FORUM

IT services of the ZID for science and research

- Supporting projects in all stages – from development to execution
- Overleaf – Collaborative Online LaTeX Editor

universität wien

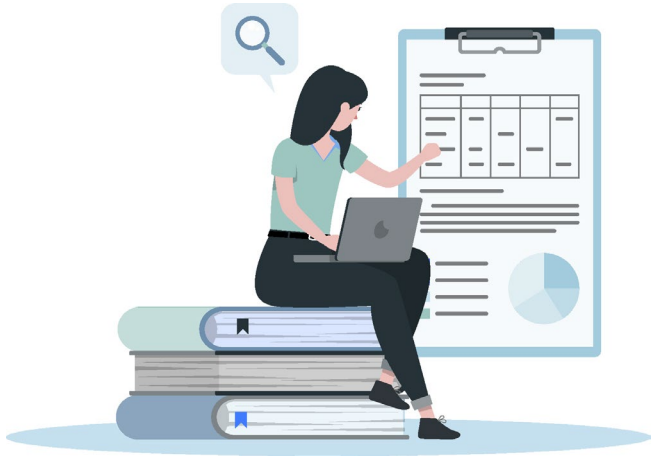
IT4S IT for Science FORUM

Supporting projects in all stages – from development to execution

Presentation by Ulrich Kiermayr

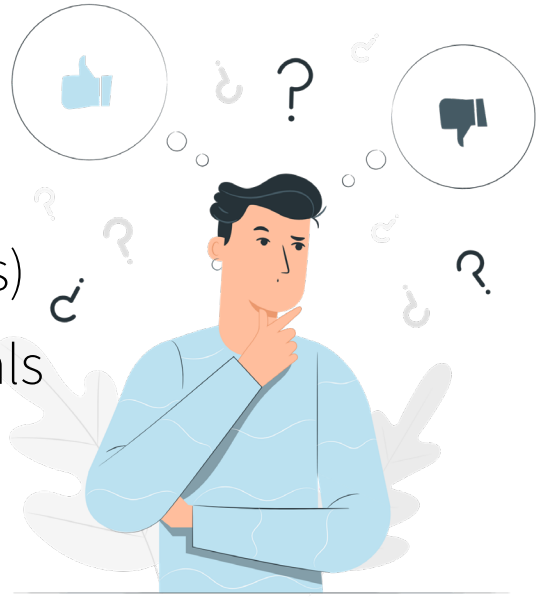
What we can offer

- Support in the project planning phase regarding the needed infrastructure
- Evaluation of options to meet project requirements
- Contact point for external support
- Contracts to use in procurement phase
- Technical support in the realisation of projects
- Monitoring and debugging in case of problems



Why this is important

- Vienna University's campus network is a complex system
- Constraints we need to meet
- Different options to achieve the goals of the project
- Additional services can be offered by the ZID
- Changes to the infrastructure require time (delivery times, cabling, setup times)
- Changes cost (a lot of) money, has to be taken into account in project proposals
- The infrastructure has to be managable on a large scale



© Image by storyset on Freepik

What our infrastructure offers today

- 1 Gbit to every system
- 10 Gbit between locations
- 100 Gbit uplink to the internet (including other ACOnet participants like VSC)
- Redundancy in the core network (including several large university locations)

© Image by storyset on Freepik

What is currently in the making

- Connecting large locations (UBB, UZA …) via 100 Gbit to the core network
- Continuous upgrade to access infrastructure to allow more bandwidth inhouse
- Options for dedicated links between services like VSC

What can be done

- Increase of bandwidth
- Links to other locations
- 10 Gbit to dedicated systems
- Serverhousing for dedicated systems

Bear in mind: this needs money, people and time – so early planning and budget availability is key to a successful project.

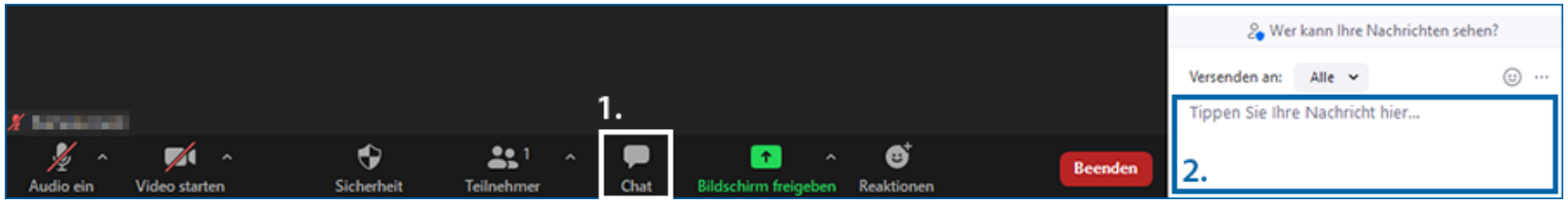
© Image by storyset on Freepik

Contact

- Via Servicedesk: **Request about the data network**
- Via service e-mail address: **digital.zid@univie.ac.at**
- Directly: ulrich.kiermayr@univie.ac.at, christian.kracher@univie.ac.at

Questions about Supporting projects in all stages?

- Please ask your questions as a chat comment.



universität wien

IT4S IT for Science FORUM

Overleaf – Collaborative Online LaTeX Editor

Presentation by Eva Karall

Why LaTeX and how does it work?

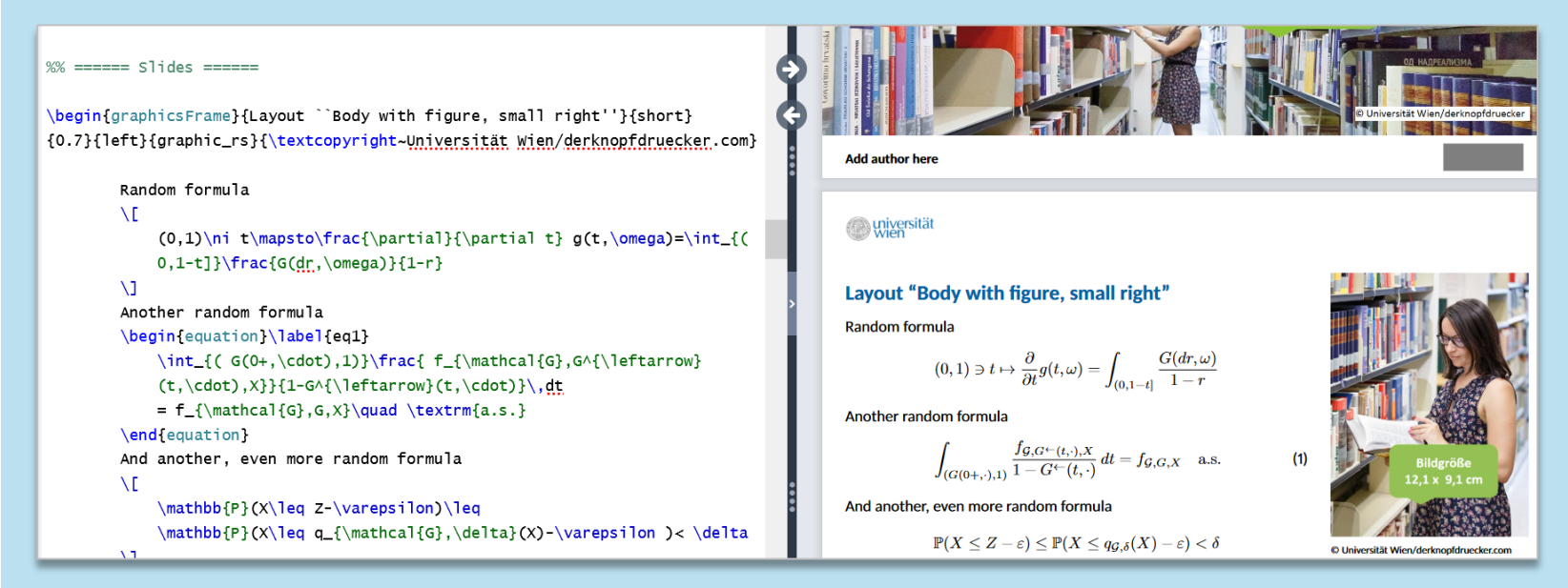
- Developed by Leslie Lamport in the 1980s, LaTeX came to its name: Lamport TeX
- A markup language that facilitates the production of well-formatted documents

```
This is a simple math expression \(\sqrt{x^2+1}\) inside text.
```

This is a simple math expression $\sqrt{x^2+1}$ inside text.

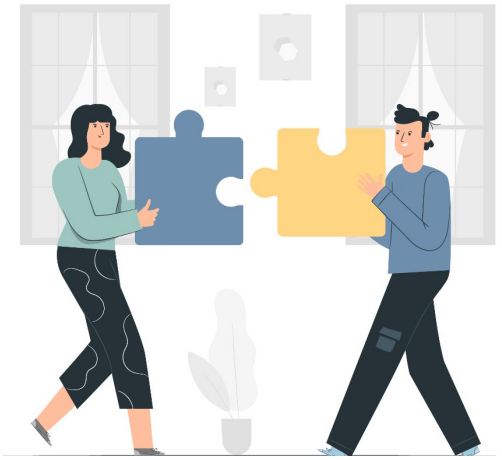
What is Overleaf?

- Overleaf is an online LaTeX editing tool that
  - allows you to create LaTeX documents directly in your web browser
  - “compiles” your LaTeX automatically to show you the results



Which license do I get?

- **Students** of the University of Vienna receive an **Overleaf Free license**:
  - Unlimited number of projects
  - Collaborating in real time with another person
  - Using templates



© Image by storyset on Freepik

- **Employees** of the University of Vienna receive an **Overleaf Professional license free of charge.** In addition to the features of the Free license, this enables:
  - tracking changes in real time
  - unlimited number of authors for documents
  - full version history
  - synchronization with e.g. Git/GitHub

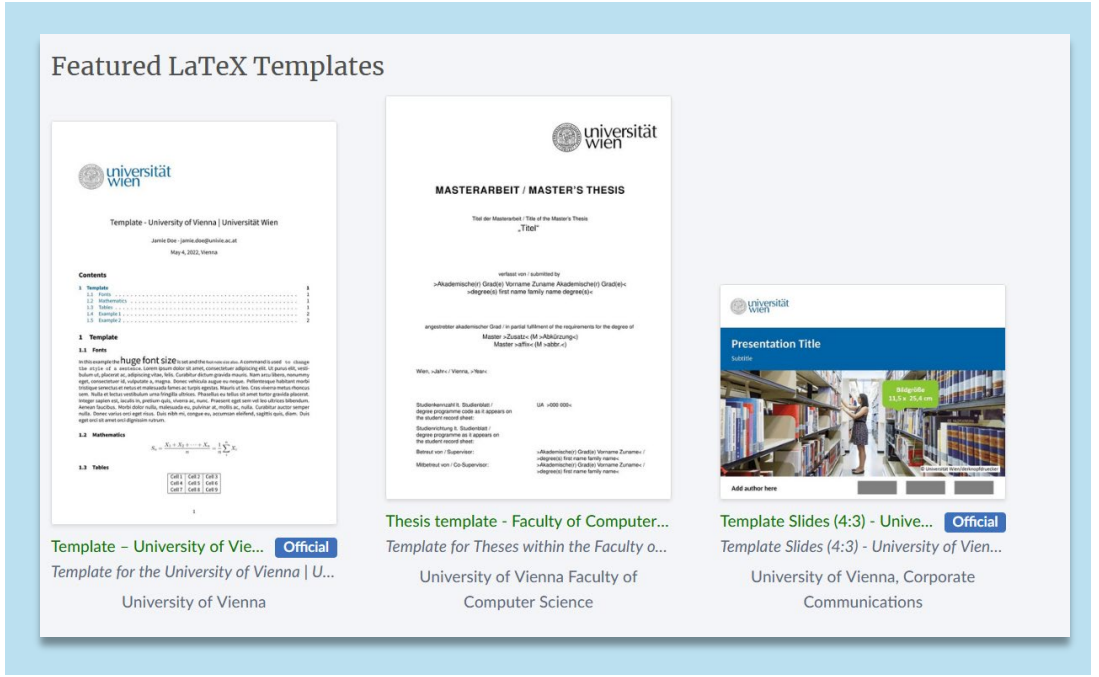
University of Vienna on Overleaf

- To start using Overleaf go to www.overleaf.com/edu/univie
- **How to get your license**:
  - Click on the button *Log in through your institution* below
  - Log in via weblogin with your u:account
  - Start new Overleaf project
    - starting from scratch
    - uploading your own files
    - or using one of the many templates available

© Image by fullvector on Freepik

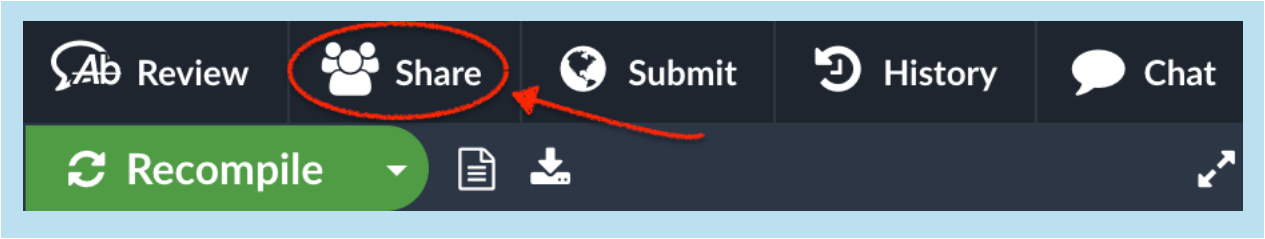
Templates

- **Gallery** with thousands of templates
- Anyone can submit projects for possible inclusion in the gallery
- Templates available for thesis, CV or submission to **scientific journals** and conference, e.g. the IEEE] and Springer
- Official **University of Vienna templates** available



Sharing

- Every document you create on Overleaf is private by default
- Two ways to share your work with collaborators:
  - By private invitation
    - Private invitations allow you to invite selected, named collaborators to access your projects
    - the number of named collaborators you can invite on each project depends on your plan
  - By link-sharing
    - Link-sharing allows you to share your projects via secret links for viewing, commenting and editing

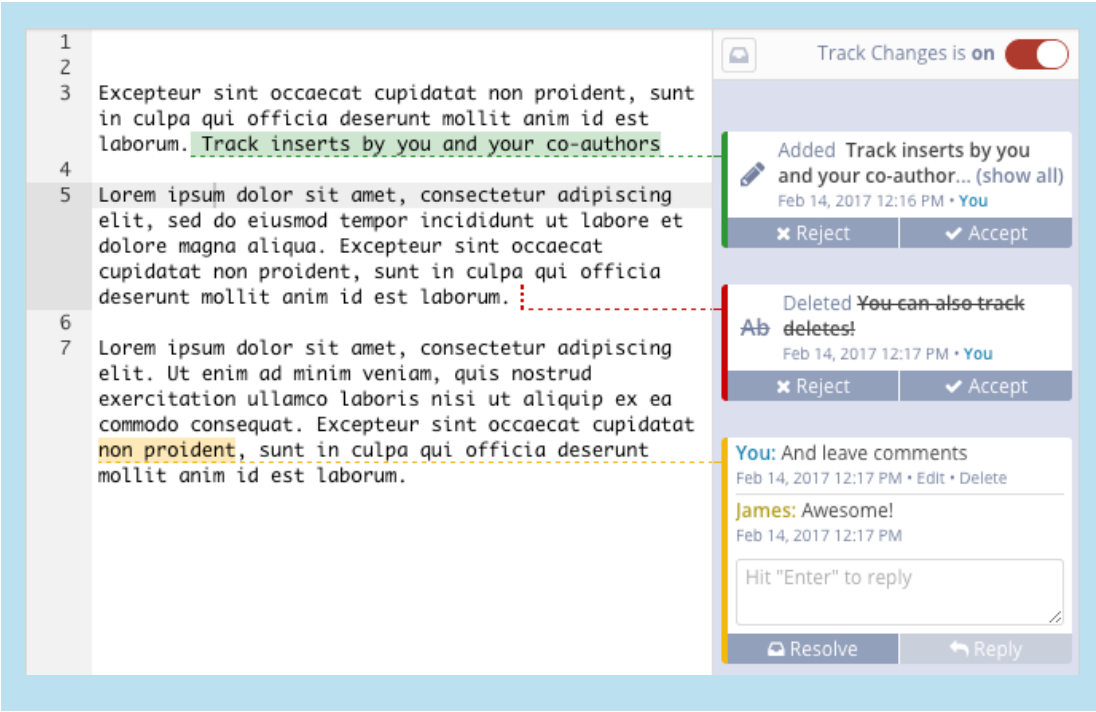


Screenshot: www.overleaf.com/learn/how-to/Sharing_a_project

Collaboration

- Single master version of each document
- Simultaneous collaborative editing

- Discussing within Overleaf
- Tracking all changes
- Keeping one to-do list
- Accepting and rejecting changes

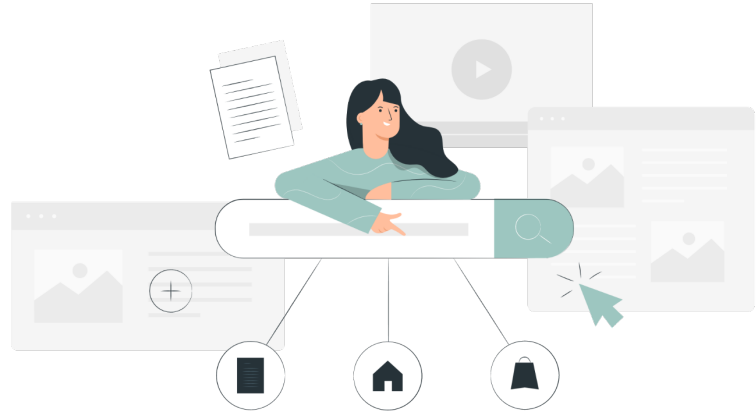


Screenshot: https://www.overleaf.com/blog/track-changes-and-comments-in-latex-2017-03-09

More information

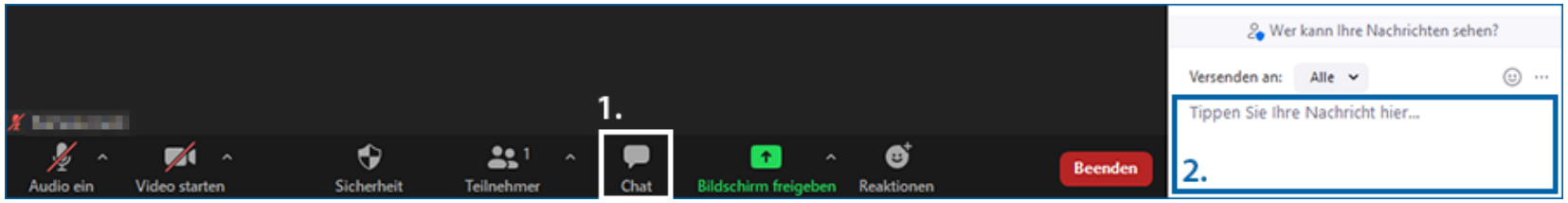
- Service page – zid.univie.ac.at/en/overleaf/
- University of Vienna on Overleaf – www.overleaf.com/edu/univie

- **Instructions and support via Overleaf**
  - Documentation of Overleaf
  - Helpful how-to guides
  - Technical support via contact form or e-mail to support@overleaf.com



Questions about Overleaf – Collaborative Online LaTeX Editor?

- Please ask your questions as a chat comment.



Outlook

Thank you for your attention!