The Language Technology Group

Johannes Graën

Monday 8th May, 2023

What We Do

The Language Technology Group offers services in the areas of

1. Natural Language Processing
2. Corpora and corpus search
3. Application development
4. Assistive technology
5. Basic IT services

Data Processing / Data Life Cycle

Acquisition

- Lab data
- Web scraping
- Corpus archives

Processing

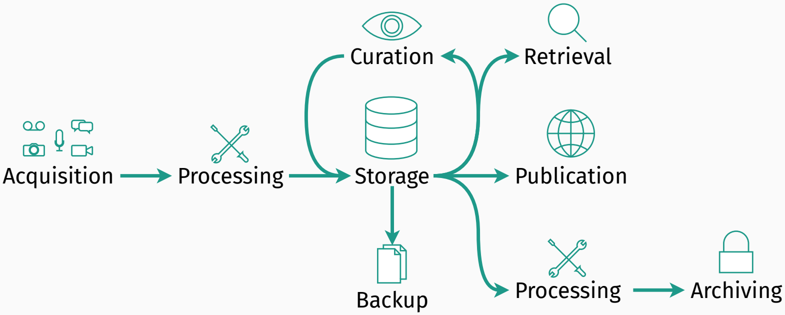
- Conversion
- Cleaning
- Consolidation

Curation

- Custom-made tools

Retrieval

- Query engines



Publication

- Headless web applications

Backup

- Incremental Backups

Archiving

- Long-term archiving in SWISSUbase

Bigger Projects

- Lexique étymologique de la Galloromania médiévale > 5000h
- Linguistic Corpus Platform > 1223h
- Swissdox@LiRI > 882h
- Répertoire critique des manuscrits littéraires en ancien occitan > 880h
- SWISSUbase (long-term archiving) > 582h
- Syntaktischer Atlas der deutschen Schweiz > 554h
- Video Analysis (multimodal corpora) > 351h
- Mappatura dell'Italo-Romanzo Antico > 250h
- Federal Archive (anonymization) > 236h
- Stimmen der Schweiz > 181h
- Evolving paradigms > 168h
- DSI Health Challenge = 150h
- Support NCCR EvoLang > 128h
- Eye-tracking data conversion = 119h
- Chintang verb senses = 118h
- Artificial moderator/facilitator > 100h
- Teaching corpus for Slavic languages = 100h

Internal Working Groups

Natural Language Processing (NLP)

- Application of state-of-the-art NLP techniques to research questions
- Coaching and workshops for academic and commercial customers
- 9 members (2.0 FTE)

Application Development (AD)

- Development of sustainable web applications for data curation and dissemination
- Databases; visualizations; technology transfer
- 5 members (3.4 FTE)

IT Infrastructure (IT)

- Setup and maintenance of our distributed infrastructure; IT and Network Coordinators
- Hosting of data and tools for customers
- 4 members (0.65 FTE + paid services)

Coordination & Administration (C&A)

- Project management & resource planning; reporting and billing; contracts
- 3 members (0.85 FTE)

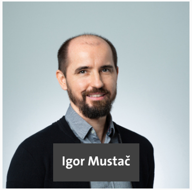
Who’s Who

Ahmet Uluslu

Bernard Schroffenegger

Daniel McDonald

Gerold Schneider


Igor Mustač

Johannes Graën

Jonathan Schaber

Klaus Rothenhäusler

Nikolina Rajović

Stefan Bircher

Teodora Vuković

Tilia Ellendorff

starting June 2023

Jean-Philippe Goldman

starting Q3 2023

N.N.

starting Q3/Q4 2023

N.N.

Customers & Collaborations

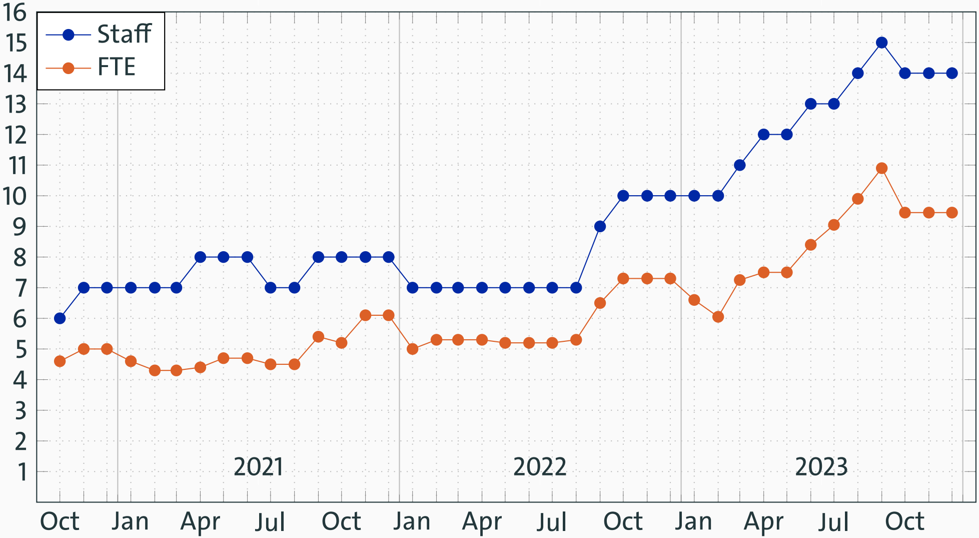
University of Zurich

- Center of Dental Medicine
- Department of Business Administration
- Department of Communication and Media Research
- Department of Comparative Language Science
- Department of Computational Linguistics
- Department of Film Studies
- Department of Geography
- Department of German and Scandinavian Studies
- Department of History
- Department of Informatics
- Department of Political Science
- Department of Psychology
- Department of Slavonic Languages and Literature
- Department of Social Anthropology and Cultural Studies
- Digital Society Initiative
- English Department
- Institute of Art History
- Institute of Biomedical Ethics and History of Medicine
- Institute of Education
- Institute of Evolutionary Medicine
- Institute of Romance Studies
- NCCR Evolving Language
- Research Center for the Public Sphere and Society
- URPP Language and Space

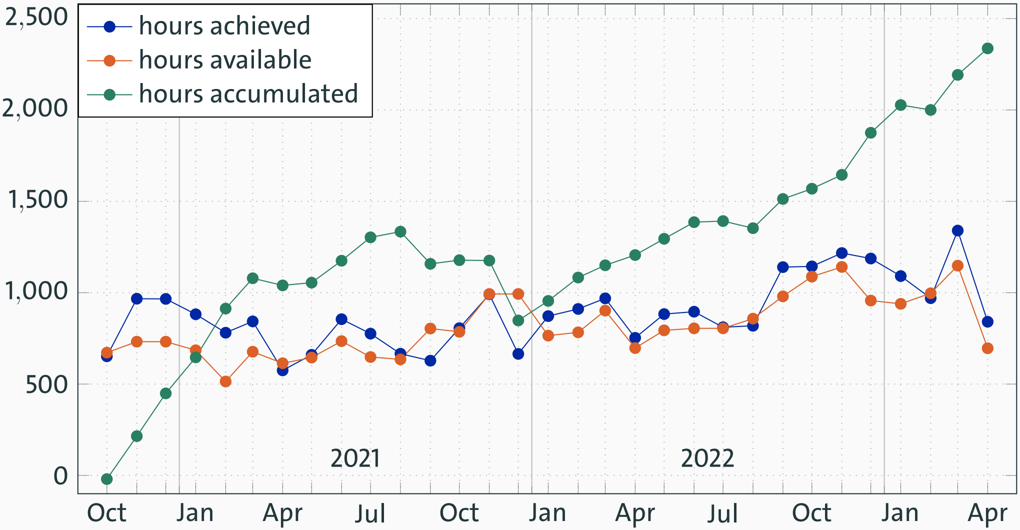
Other institutions & companies

- Agroscope (agricultural research)
- Center for Religion, Economy and Politics
- ETH Zurich
- Federal Office of Communications
- Free University of Bozen-Bolzano 🇮🇹
- Helsana (health insurance)
- Innovista (innovation mentoring)
- Peoplegeist (AI for manufactoring)
- PRODAFT (cyber threat intelligence)
- Swiss Federal Archives
- Swiss National Bank
- University College London 🇬🇧
- University of Applied Sciences of the Grisons
- University of Basel
- University of Bern
- University of Lausanne
- University of Lucerne
- University of Lyon 🇫🇷
- University of Queensland 🇦🇺
- Zurich University of Applied Sciences
- Zurich University of Teacher Education

Staff and Full Time Equivalents (FTE)



Working Hours



SWOT Analysis

Strengths

- Staff with complementary skills and experiences
- Stimulating environment
- Strong personal commitment; intrinsic motivation of group members

Weaknesses

- Keeping knowledge and skills up-to-date requires continuous education
- Often more work than work force available due to high demand
- Underestimation of the administrative effort, which needs to be covered by fees (non-disclosed overhead)

Opportunities

- Language Technology is increasingly relevant for many fields
- Many potential users/customers for tools we build (at UZH, in CH, and internationally, e.g. within CLARIN)
- Closeness to research

Threats

- Many competitors (in Zurich) for skilled technical staff
- Fluctuating demand, often short-term requests

University of Zurich[UZH]

Linguistic Research Infrastructure – Language Technology Group

IT infrastructure

Requirements, options, solutions

Johannes Graën

Monday 8th May, 2023

Requirements

LiRI and LiZZ have considerable requirements on IT in all dimensions

- Storage space (lab data, corpus collections, customer data, ...)
- Computational power (parallel data processing, training language models, ...)
- Memory (language models, databases, applications, caching, ...)
- IO/bandwidth (databases, IR systems for media, ...)

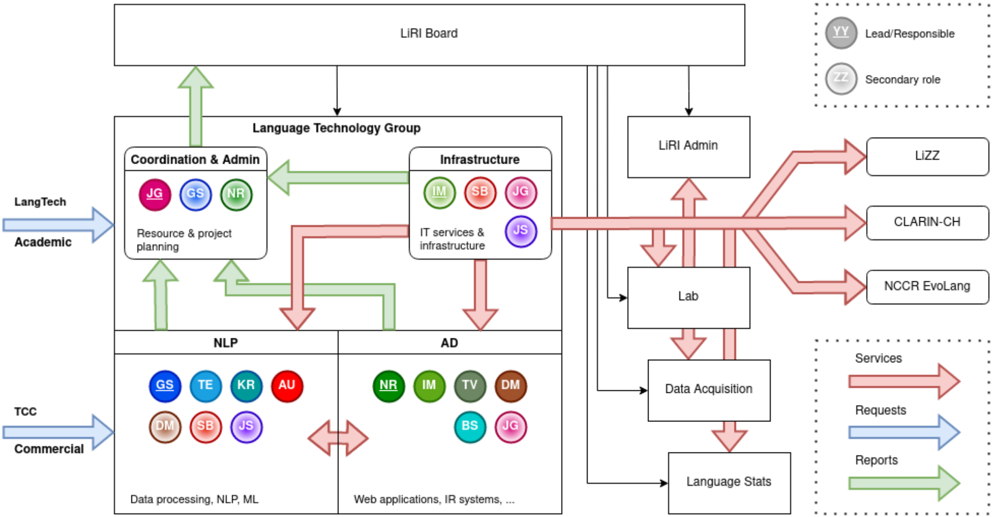
Services Offered

- Web applications (with databases and headless, in-house and third-party)
- NLP pipelines (e.g. anonymization) on CPU and GPU via CLI and API
- Structured data in DBMS (distributed and stand-alone)
- Communication platforms (Wiki, Chat, ...)
- Media streaming and analysis
- Storage space
- *Proxy for IT and network-related tasks (not part of the infrastructure)*

Additionally, for our own purposes

- Integrated storage for project data
- Admin and maintenance tools (time tracking, system and service monitoring, backup)

IT Services Demanded by LiZZ and LiRI



Resource Available

ScienceCloud (Science IT at UZH)

- Virtual machines based on OpenStack and Ceph with CPUs and (few) GPUs
- Unreliable IO throughput due to distributed and shared file system
- Legal restrictions

Central IT at UZH

- Virtual machines based on VMware
- Better IO, but limited access and limited performance
- IT security restrictions

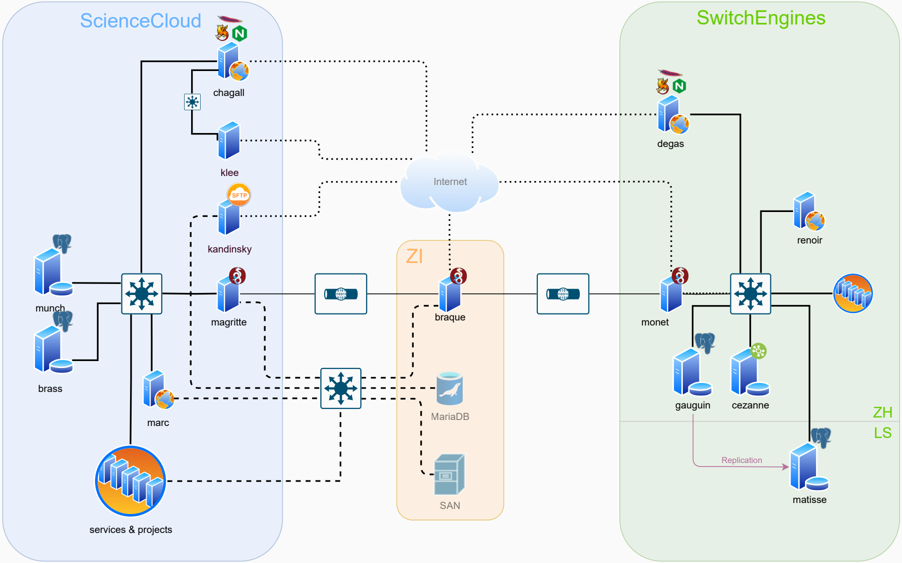
SWITCHengines (SWITCH)

- Virtual machines based on OpenStack and Ceph and S3 (only CPUs)
- Same technical limitations as for ScienceCloud

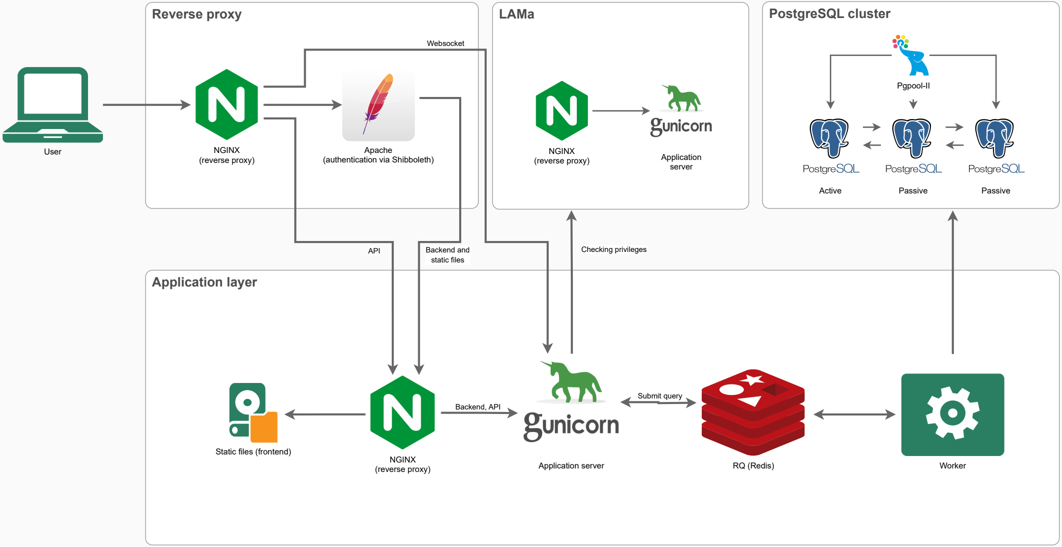
Other Services by Central IT & SWITCH

- IP Address Management (DNS)
- Authentication service for TLS wildcard certificates
- Resource registry for interfederated authentification (eduGAIN)
- Central redundant SAN (only available in central data center)
- ...

Infrastructure (Up to Now)



Distributed Applications (LCP as example)



New Infrastructure at Central Data Center

1. Redundant NFS file servers on SAN with failover
   - "Infinite" amount of storage space with very fast access (redundant fibre channel)
   - Access rights, quotas and backups managed by us
2. Virtualization cluster with three nodes and failover scheme
   - PostgreSQL cluster for high availability and load distribution
   - Six application servers with alternating failover destination
   - Several small management and service machines with hot failover