Hessisches Statistisches Landesamt

STATISTIKHESSEN

HESSEN

Statistical Scraping at Statistics Hesse

Statistical Scraping Interest Group SSIG
1st Meeting, Vienna, Statistik Austria

September 16-17th, 2025

Tobias Gramlich, Statistik Hessen, Wiesbaden, Germany
Tobias.gramlich@statistik.hessen.de

(Production) Systems of Statistical Scraping at Statistics Hesse

- (1) Tourism statistics, hotel booking portals
  - Scraping of hotel booking portals (4 portals relevant for Germany)
  - Record linkage with sampling frame: identification of potential new units to include to the sampling frame
  - Coverage: Germany / federal states; hotels and camping sites
- (2) Commercial Trade Registry:
  - (Born out of necessity: no API available / accessible)
  - General purpose, mainly statistical business registry
  - Name of the owner, also status of the business, new / changed address, assist classification of economic activity

(Production) Systems of Statistical Scraping at Statistics Hesse

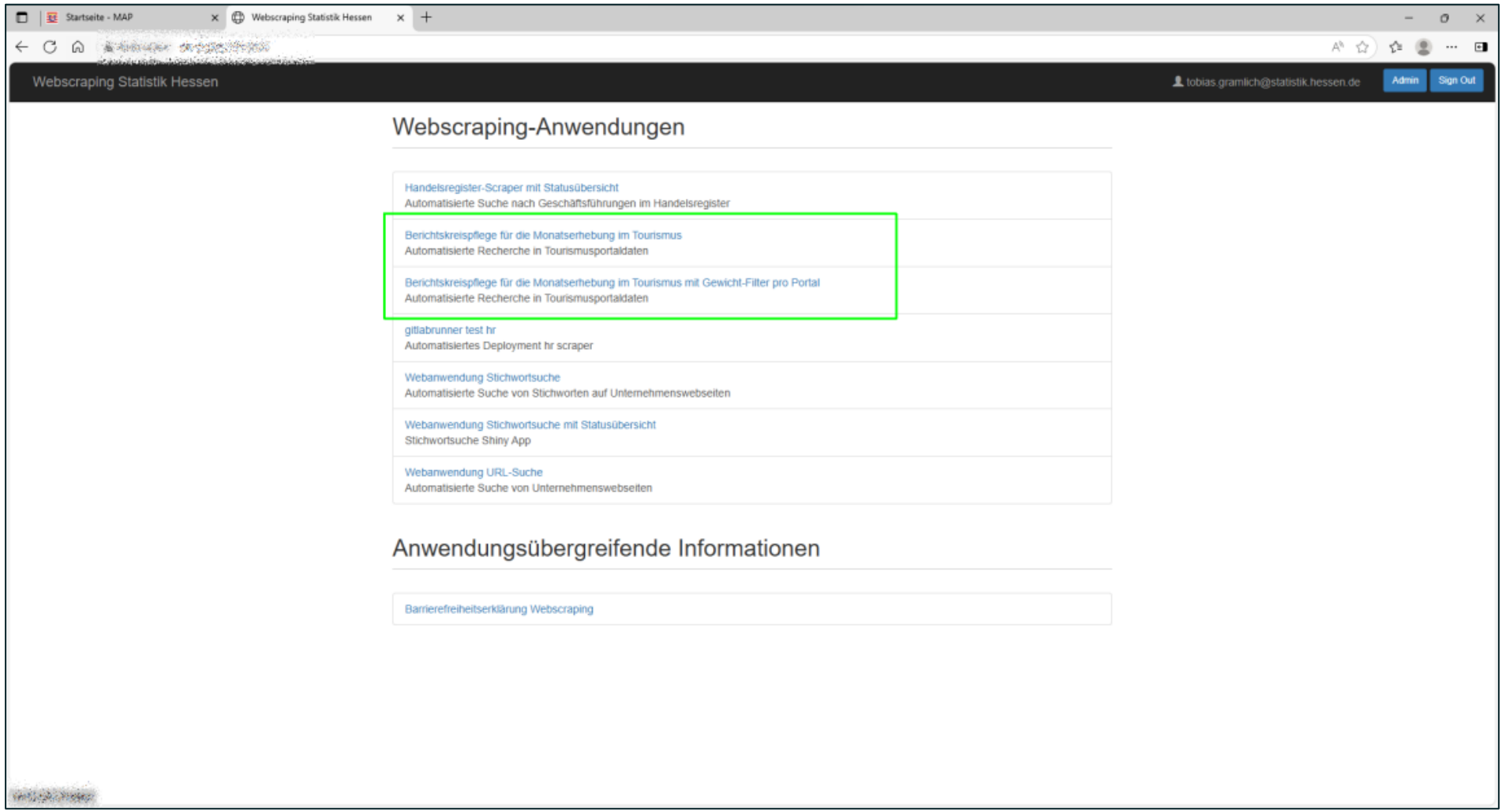
- (3) URL finding
  - For units from the statistical business registry (SBR)
  - Uses Alphabet / Google, 10.000 searches / day
  - Links URLs to SBR IDs; saves websites‘ texts of linked URLs
  - General purpose: SBR, …, 1st step for / input to keyword search
- (4) Keyword search
  - For units from the statistical business registry (SBR)
  - Provide IDs of SBR and list of keywords (with limits for both lists)
  - Search websites‘ texts of units from (3) for matches / nonmatches of keywords
  - Assist classification of economic activity, clarify / identify target population

(Production) Systems of Statistical Scraping at Statistics Hesse

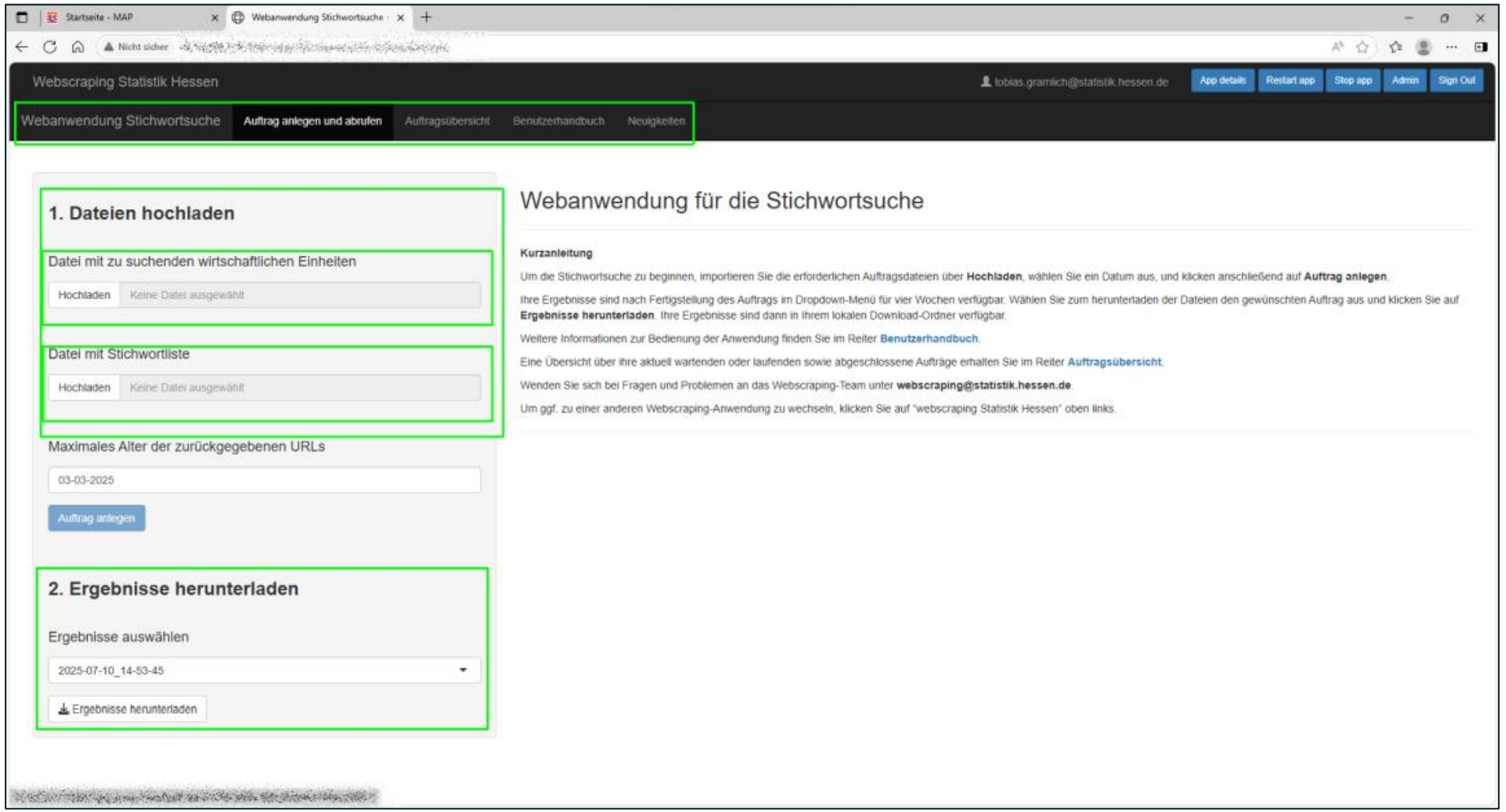
- One access point for all webscraping applications
- In use / can be used by all 15 statistical offices
- Registration of every single user and activation of every single application for every user

- Separate environments for development, test, preproduction, production
- R, Shiny, Shinyproxy, R, Python, Kafka, Keycloak
- A lot of synchronization of status files, input and output files necesary between different network areas due to data protection / information security requirements

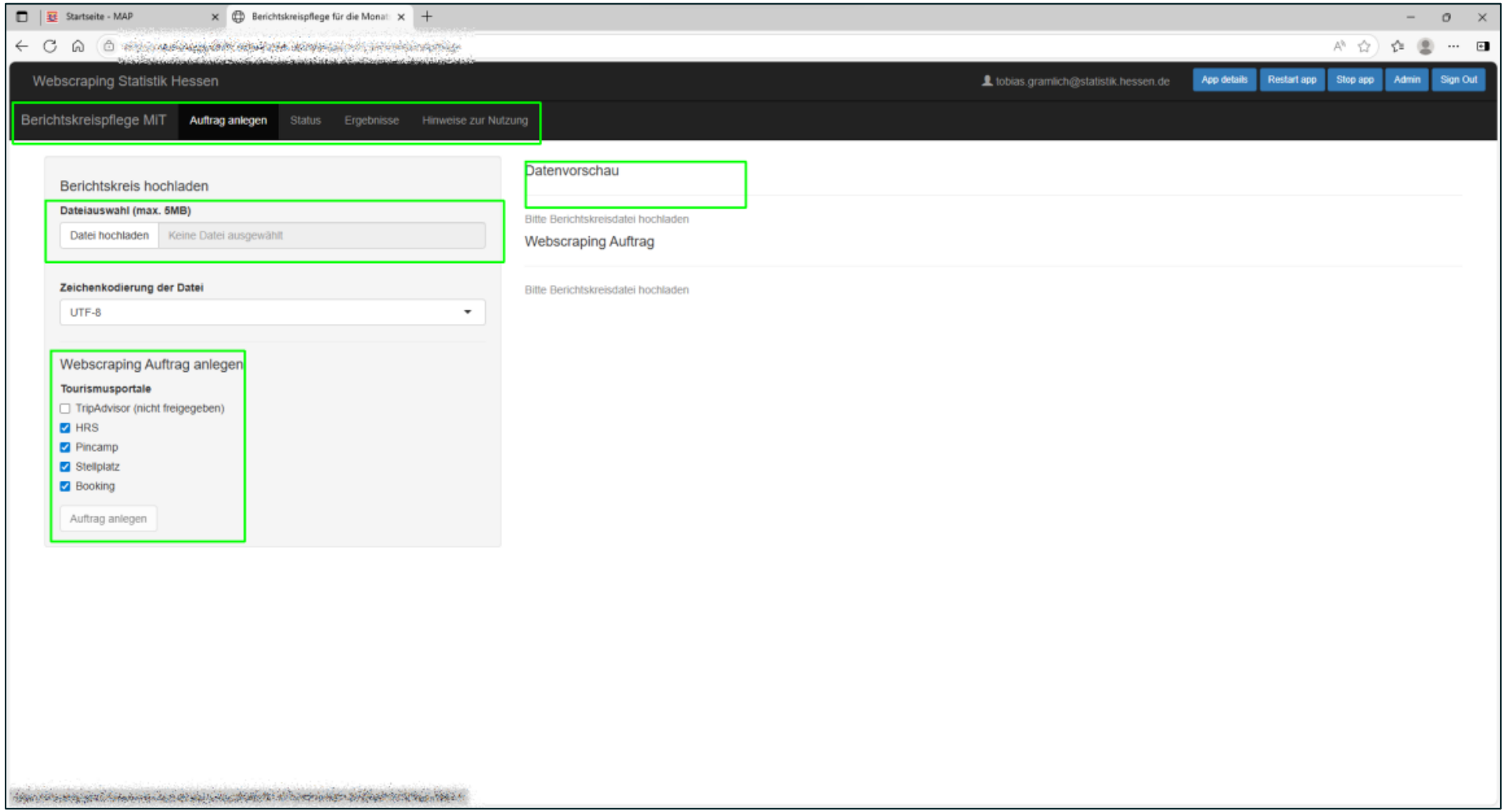
GUI: Access point for all applications



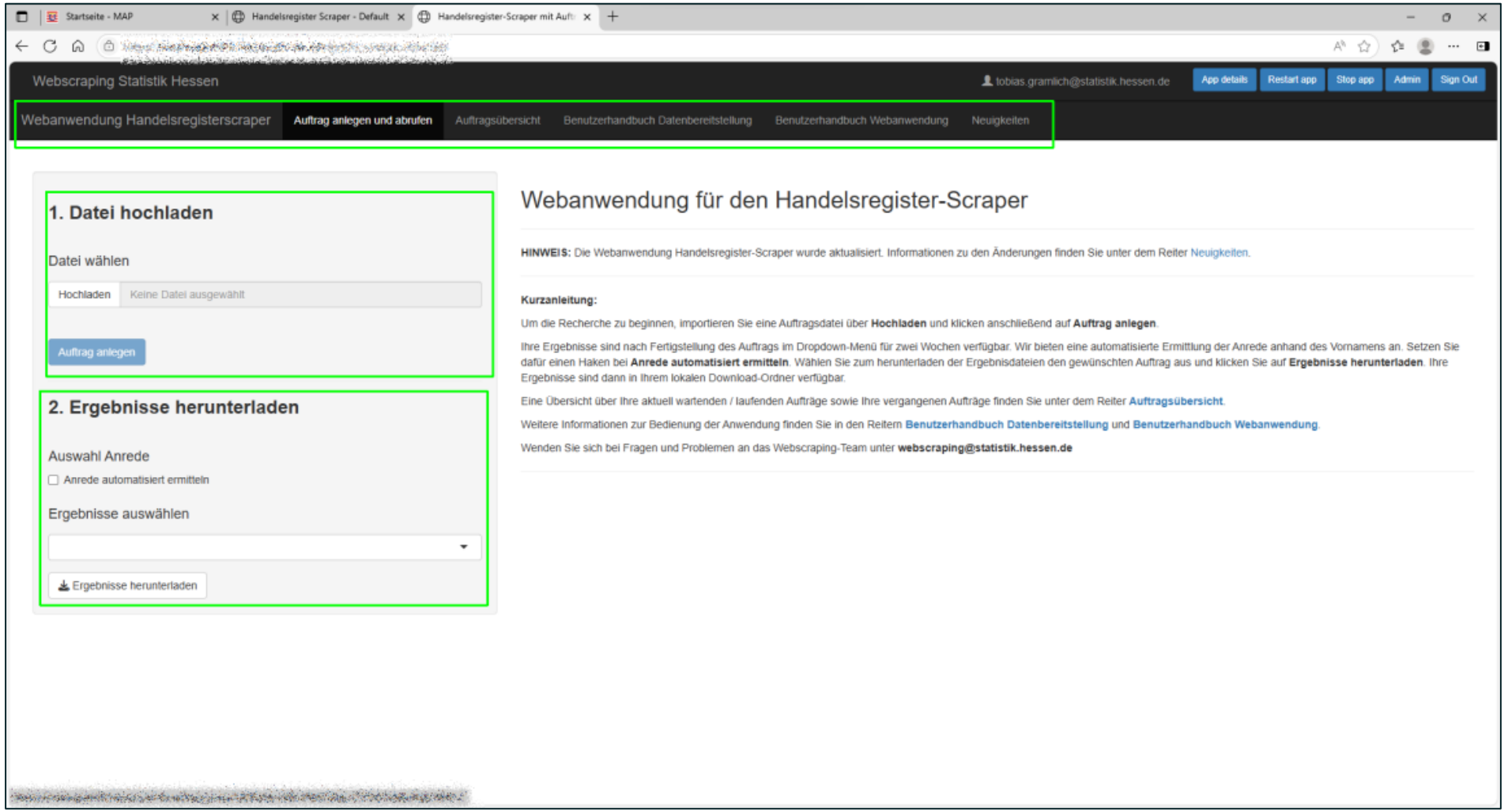
GUI: Provding input files, parameters (1)



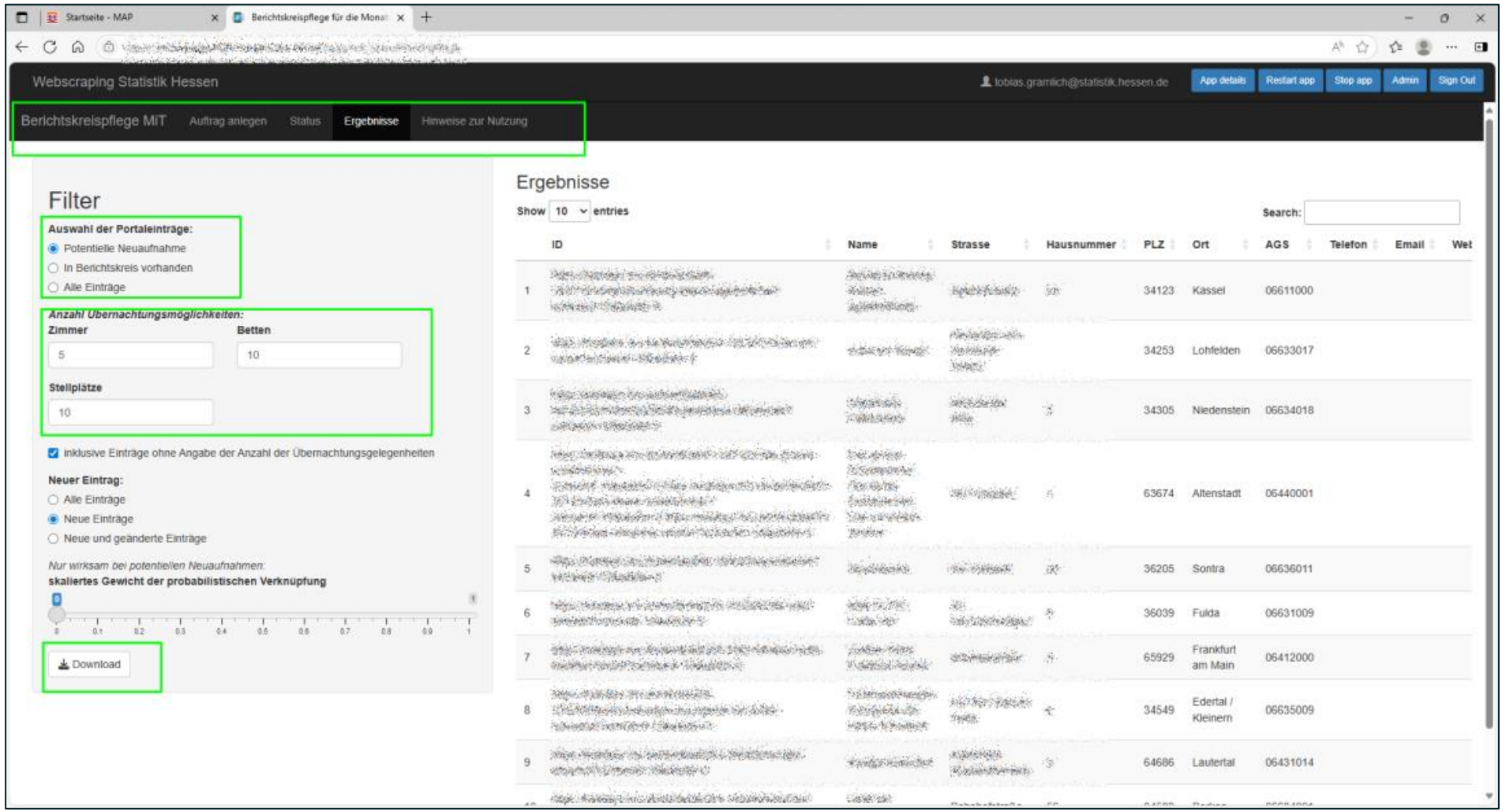
GUI: Provding input files, parameters (2)



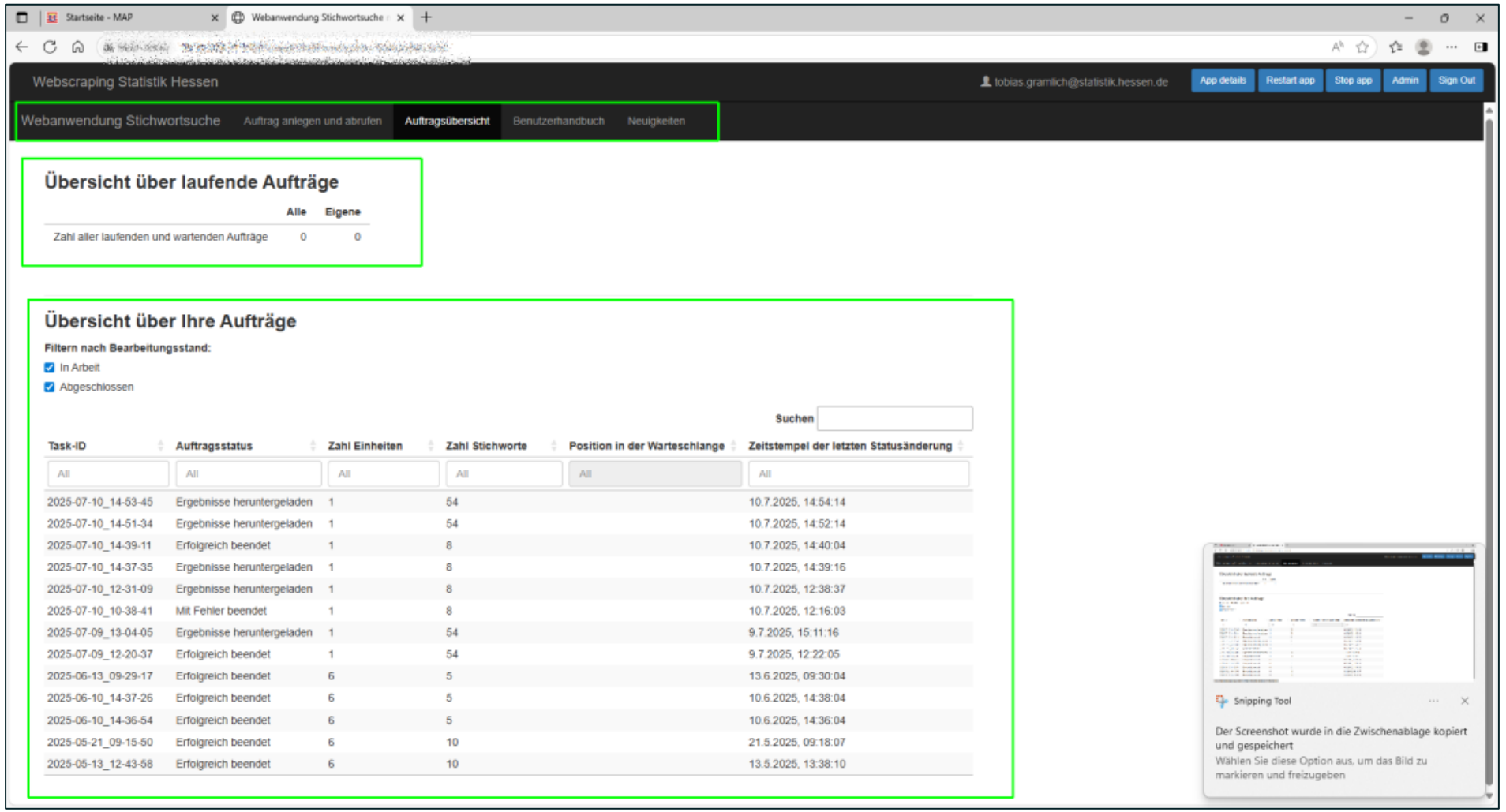
GUI: Provding results (1)

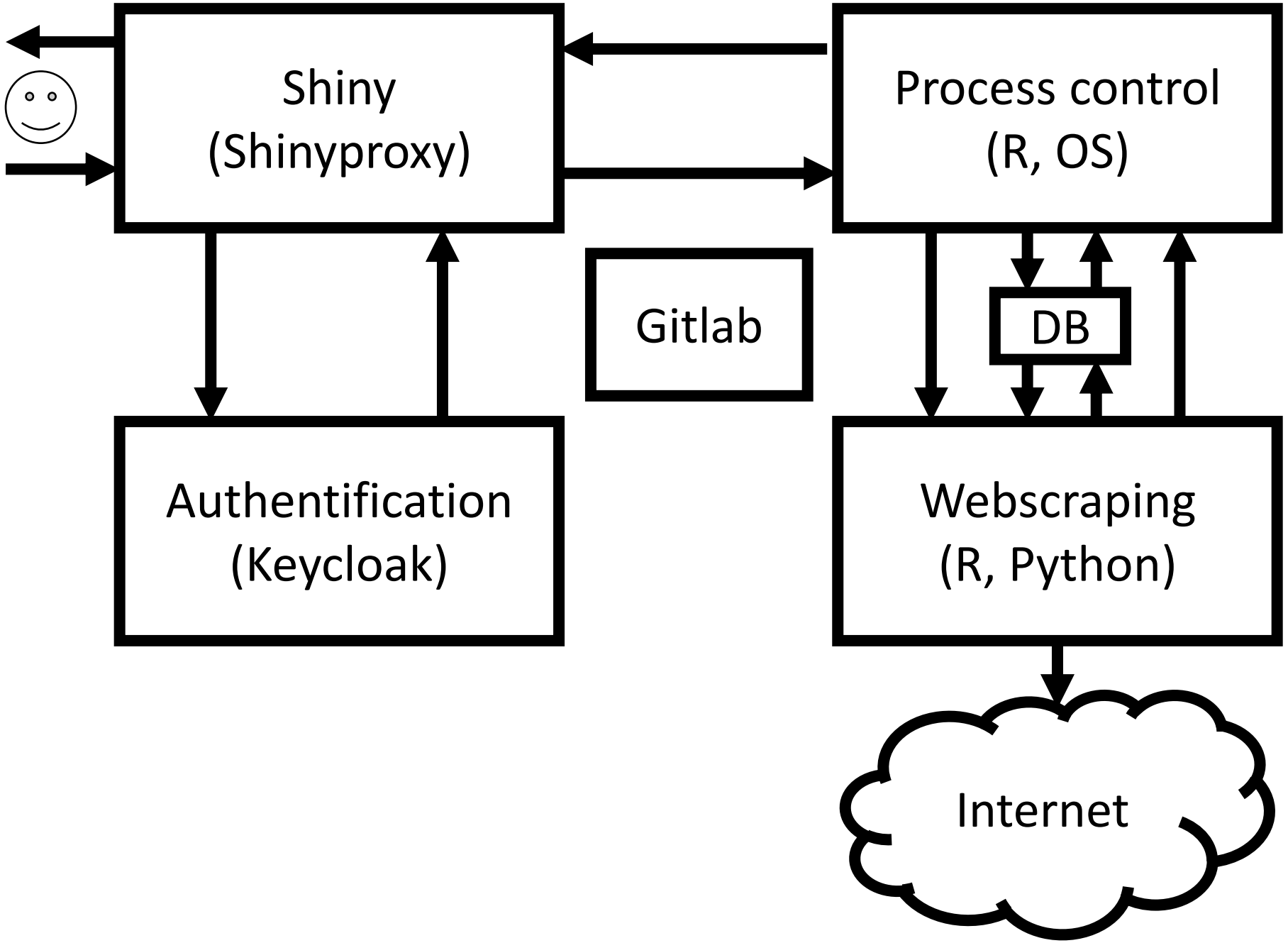


GUI: Provding results (2)



GUI: overview for users



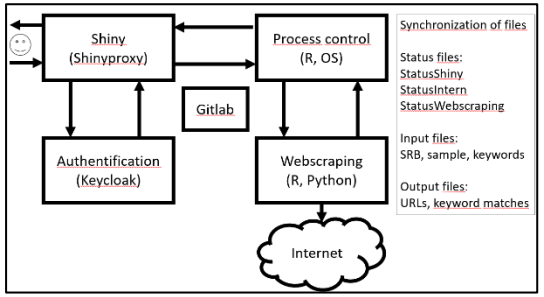


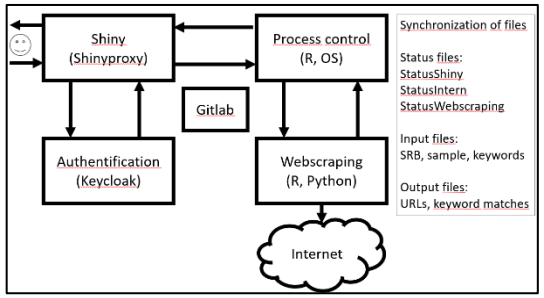
Synchronization of files

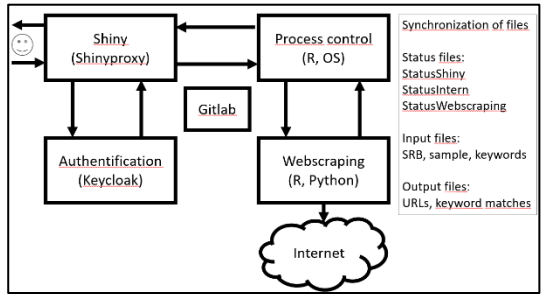
Status files:
StatusShiny
StatusIntern
StatusWebscraping

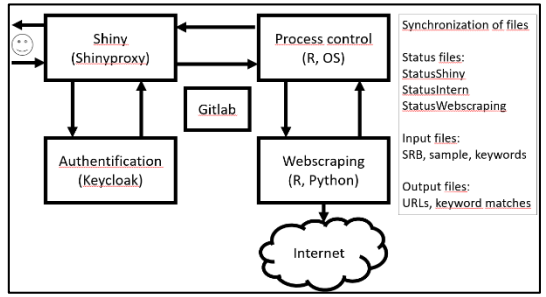
Input files:
SRB, sample, keywords

Output files:
URLs, keyword matches









Separate environments:

Development
Test
Preproduction
Production

May differ in requirements regarding data protection, information security, access restrictions, ressources (CPU, RAM, disk space)

Usage, user accounts

- Accounts for at least one user from each of the 15 statistical offices in Germany (offices of the German Länder as well as from Destatis)
- More than 120 registered users (with different activity of offices / users)

Next steps

- First of all: keep things running
- Stabilize production rather than new developments
  - Small improvements, bug fixes
  - User management
  - Ressource management
  - (Improve linking URLs to SBR units using improved ML / LLM)
  - (Updating recorded URLs?)
  - (Find and extract imprint information?) (=> officially provide imprint template?)
  - (Directly answer requests from other systems | provide API / feed APIs?)

Contact

Tobias Gramlich

Hessisches Statistisches Landesamt

Phone: +49 (0)611 845

E-Mail: tobias.gramlich@statistik.hessen.de

E-Mail: webscraping@statistik.hessen.de

https://statistik.hessen.de