

# 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](mailto: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

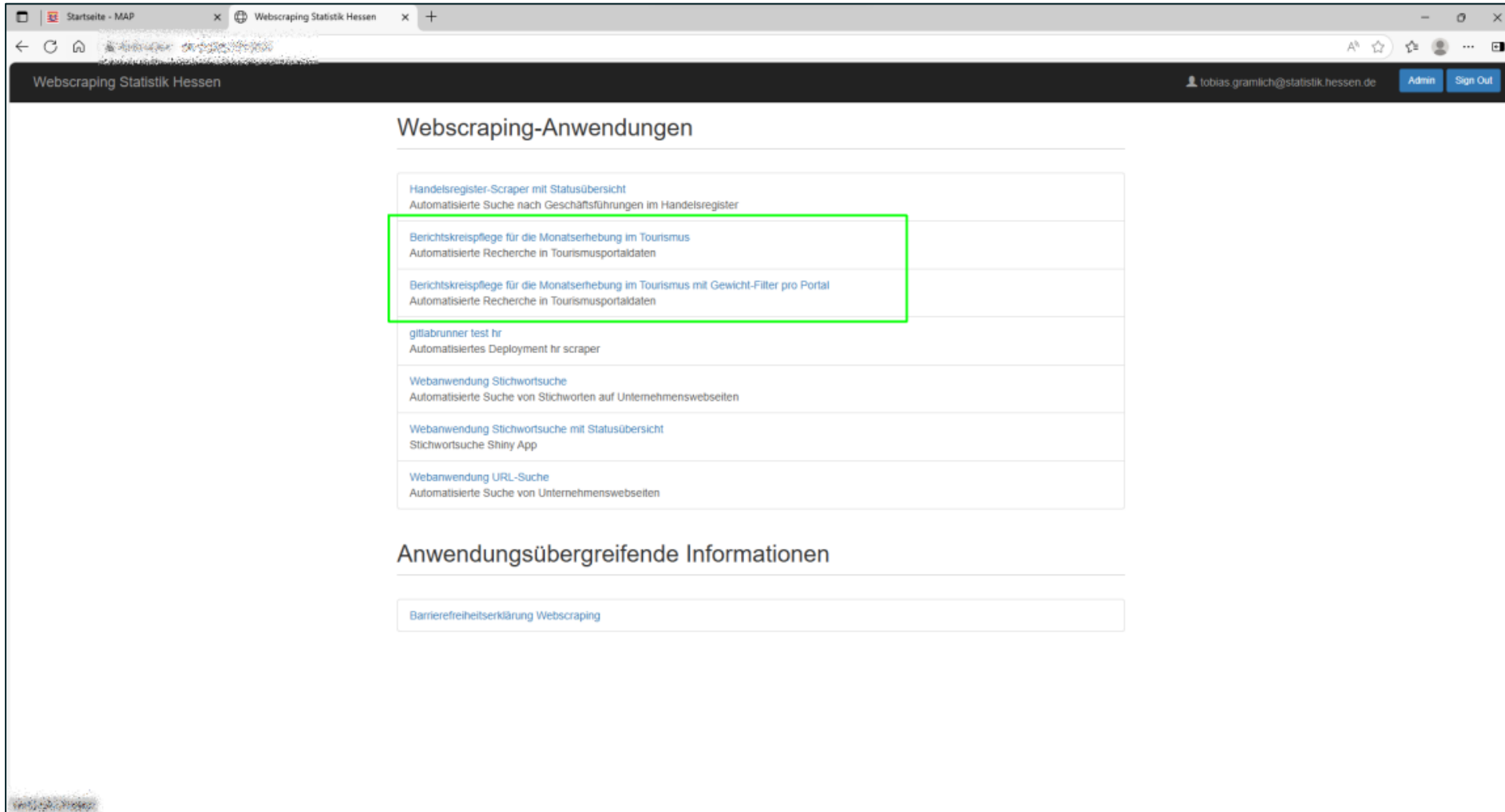
## (Production) Systems of Statistical Scraping at Statistics Hesse

- (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 necessary between different network areas due to data protection / information security requirements

# GUI: Access point for all applications



Startseite - MAP   Webscraping Statistik Hessen

l Tobias Gramlich   Admin   Sign Out

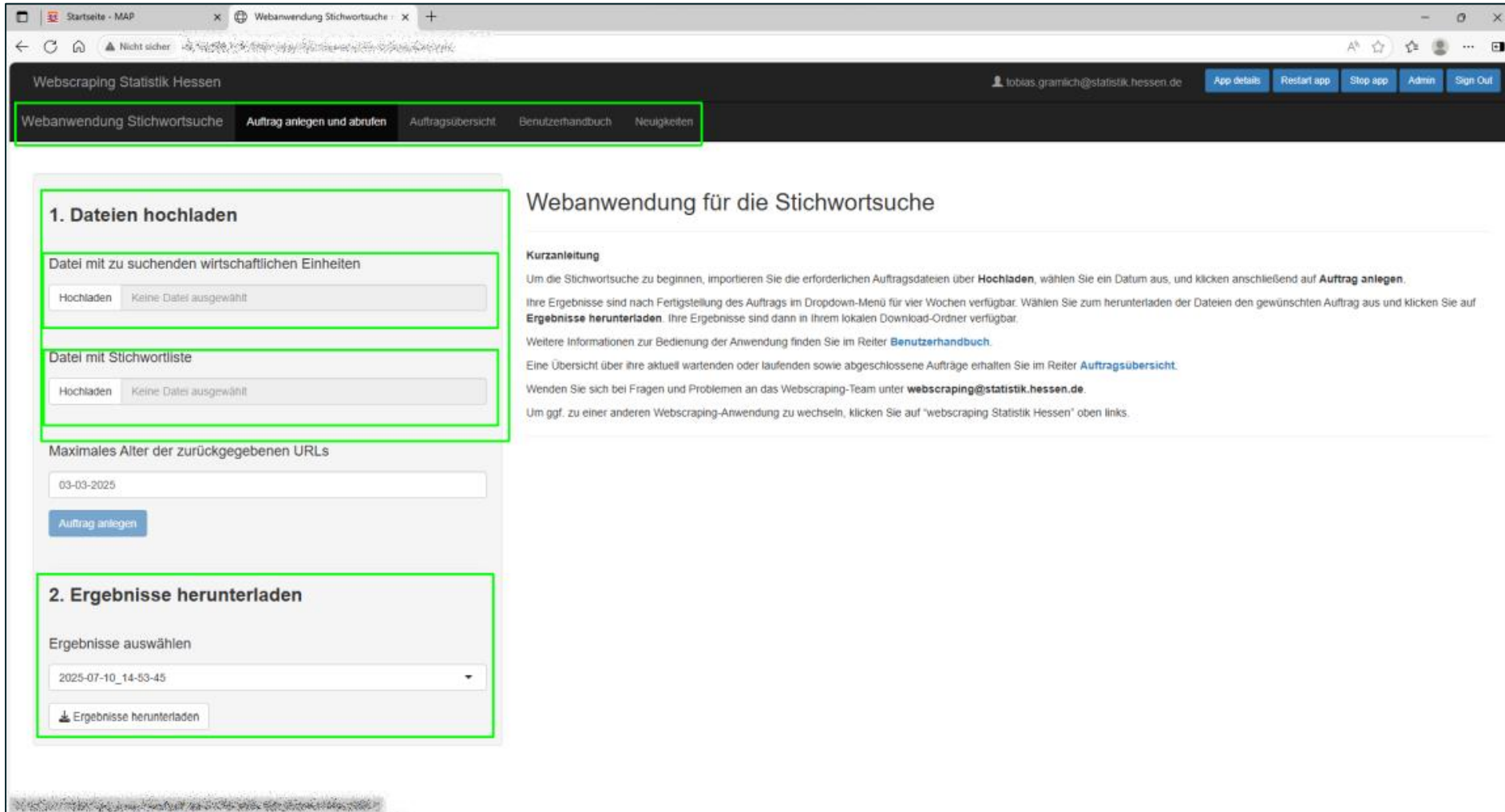
## Webscraping-Anwendungen

<a href="#">Handelsregister-Scraper mit Statusübersicht</a> Automatisierte Suche nach Geschäftsführungen im Handelsregister
<a href="#">Berichtskreispflege für die Monatserhebung im Tourismus</a> Automatisierte Recherche in Tourismusportaldateien
<a href="#">Berichtskreispflege für die Monatserhebung im Tourismus mit Gewicht-Filter pro Portal</a> Automatisierte Recherche in Tourismusportaldateien
<a href="#">gitlabrunner test hr</a> Automatisiertes Deployment hr scraper
<a href="#">Webanwendung Stichwortsuche</a> Automatisierte Suche von Stichworten auf Unternehmenswebseiten
<a href="#">Webanwendung Stichwortsuche mit Statusübersicht</a> Stichwortsuche Shiny App
<a href="#">Webanwendung URL-Suche</a> Automatisierte Suche von Unternehmenswebseiten

## Anwendungsübergreifende Informationen

<a href="#">Barrierefreiheitserklärung Webscraping</a>
--

# GUI: Providing input files, parameters (1)



Webanwendung Stichwortsuche

Auftrag anlegen und abrufen Auftragsübersicht Benutzerhandbuch Neuigkeiten

### 1. Dateien hochladen

Datei mit zu suchenden wirtschaftlichen Einheiten

Hochladen Keine Datei ausgewählt

Datei mit Stichwortliste

Hochladen Keine Datei ausgewählt

Maximales Alter der zurückgegebenen URLs

03-03-2025

Auftrag anlegen

### 2. Ergebnisse herunterladen

Ergebnisse auswählen

2025-07-10\_14-53-45

Ergebnisse herunterladen

### Webanwendung für die Stichwortsuche

**Kurzanleitung**

Um die Stichwortsuche zu beginnen, importieren Sie die erforderlichen Auftragsdateien über **Hochladen**, wählen Sie ein Datum aus, und klicken anschließend auf **Auftrag anlegen**.

Ihre Ergebnisse sind nach Fertigstellung des Auftrags im Dropdown-Menü für vier Wochen verfügbar. Wählen Sie zum Herunterladen der Dateien den gewünschten Auftrag aus und klicken Sie auf **Ergebnisse herunterladen**. Ihre Ergebnisse sind dann in Ihrem lokalen Download-Ordner verfügbar.

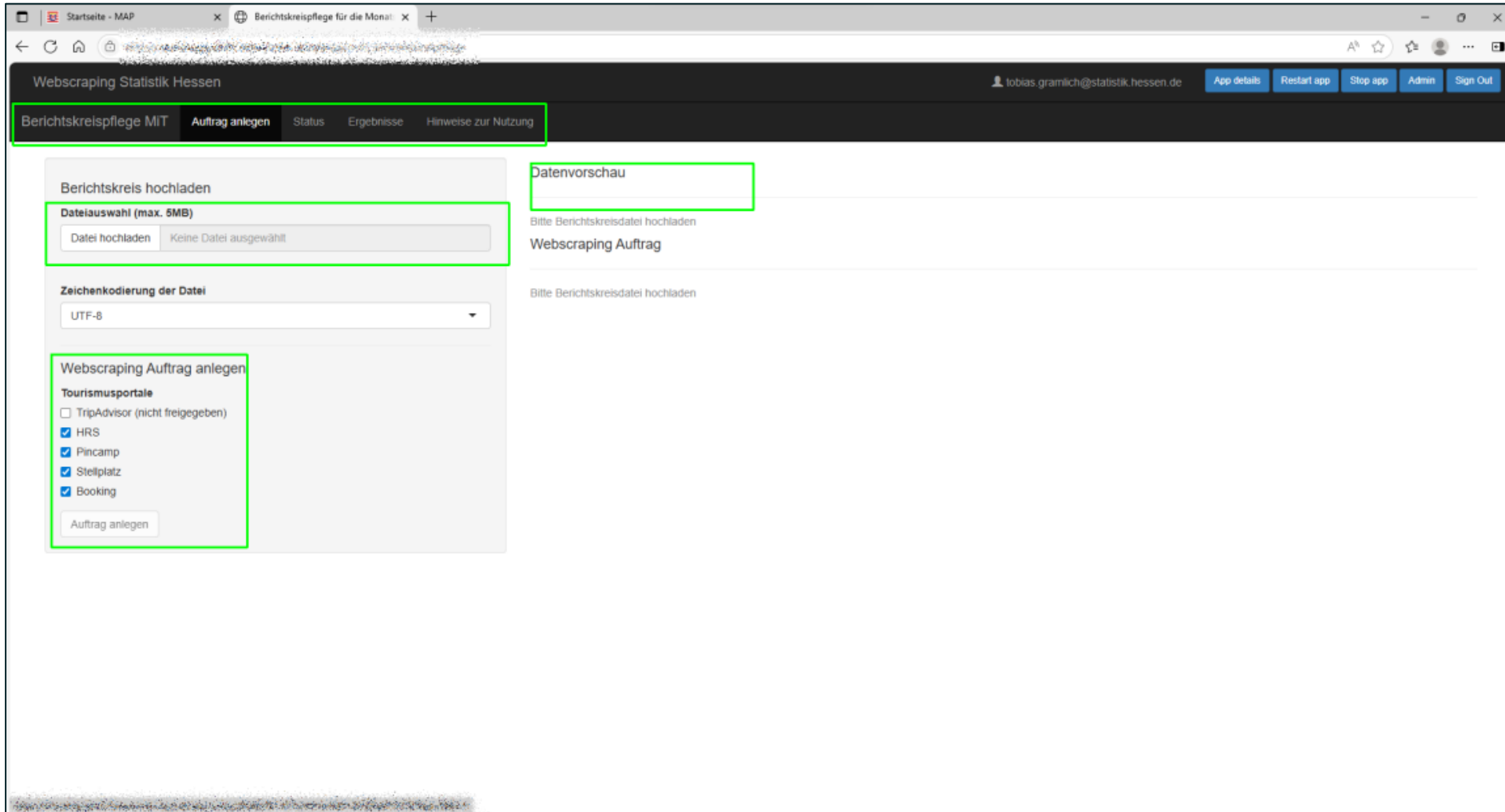
Weitere Informationen zur Bedienung der Anwendung finden Sie im Reiter [Benutzerhandbuch](#).

Eine Übersicht über Ihre aktuell wartenden oder laufenden sowie abgeschlossene Aufträge erhalten Sie im Reiter [Auftragsübersicht](#).

Wenden Sie sich bei Fragen und Problemen an das Webscraping-Team unter [webscraping@statistik.hessen.de](mailto:webscraping@statistik.hessen.de).

Um ggf. zu einer anderen Webscraping-Anwendung zu wechseln, klicken Sie auf "webscraping Statistik Hessen" oben links.

## GUI: Providing input files, parameters (2)



Startseite - MAP x Berichtskreispflege für die Monat x +

Web scraping Statistik Hessen

tobias.gramlich@statistik.hessen.de App details Restart app Stop app Admin Sign Out

Berichtskreispflege MIT Auftrag anlegen Status Ergebnisse Hinweise zur Nutzung

Berichtskreis hochladen

Dateiauswahl (max. 5MB)

Datei hochladen Keine Datei ausgewählt

Zeichenkodierung der Datei

UTF-8

Webscraping Auftrag anlegen

Tourismusportale

☐ TripAdvisor (nicht freigegeben)

☒ HRS

☒ Pincamp

☒ Stellplatz

☒ Booking

Auftrag anlegen

Datenvorschau

Bitte Berichtskreisdatei hochladen

Webscraping Auftrag

Bitte Berichtskreisdatei hochladen

# GUI: Provding results (1)

Webanwendung Handelsregister-Scraper

**1. Datei hochladen**

Datei wählen

Hochladen Keine Datei ausgewählt

Auftrag anlegen

**2. Ergebnisse herunterladen**

Auswahl Anrede

☐ Anrede automatisch ermitteln

Ergebnisse auswählen

Ergebnisse herunterladen

**Webanwendung für den Handelsregister-Scraper**

**HINWEIS:** Die Webanwendung Handelsregister-Scraper wurde aktualisiert. Informationen zu den Änderungen finden Sie unter dem Reiter [Neuigkeiten](#).

**Kurzanleitung:**

Um die Recherche zu beginnen, importieren Sie eine Auftragsdatei über **Hochladen** und klicken anschließend auf **Auftrag anlegen**.

Ihre Ergebnisse sind nach Fertigstellung des Auftrags im Dropdown-Menü für zwei Wochen verfügbar. Wir bieten eine automatisierte Ermittlung der Anrede anhand des Vornamens an. Setzen Sie dafür einen Haken bei **Anrede automatisch ermitteln**. Wählen Sie zum herunterladen der Ergebnisdateien den gewünschten Auftrag aus und klicken Sie auf **Ergebnisse herunterladen**. Ihre Ergebnisse sind dann in Ihrem lokalen Download-Ordner verfügbar.

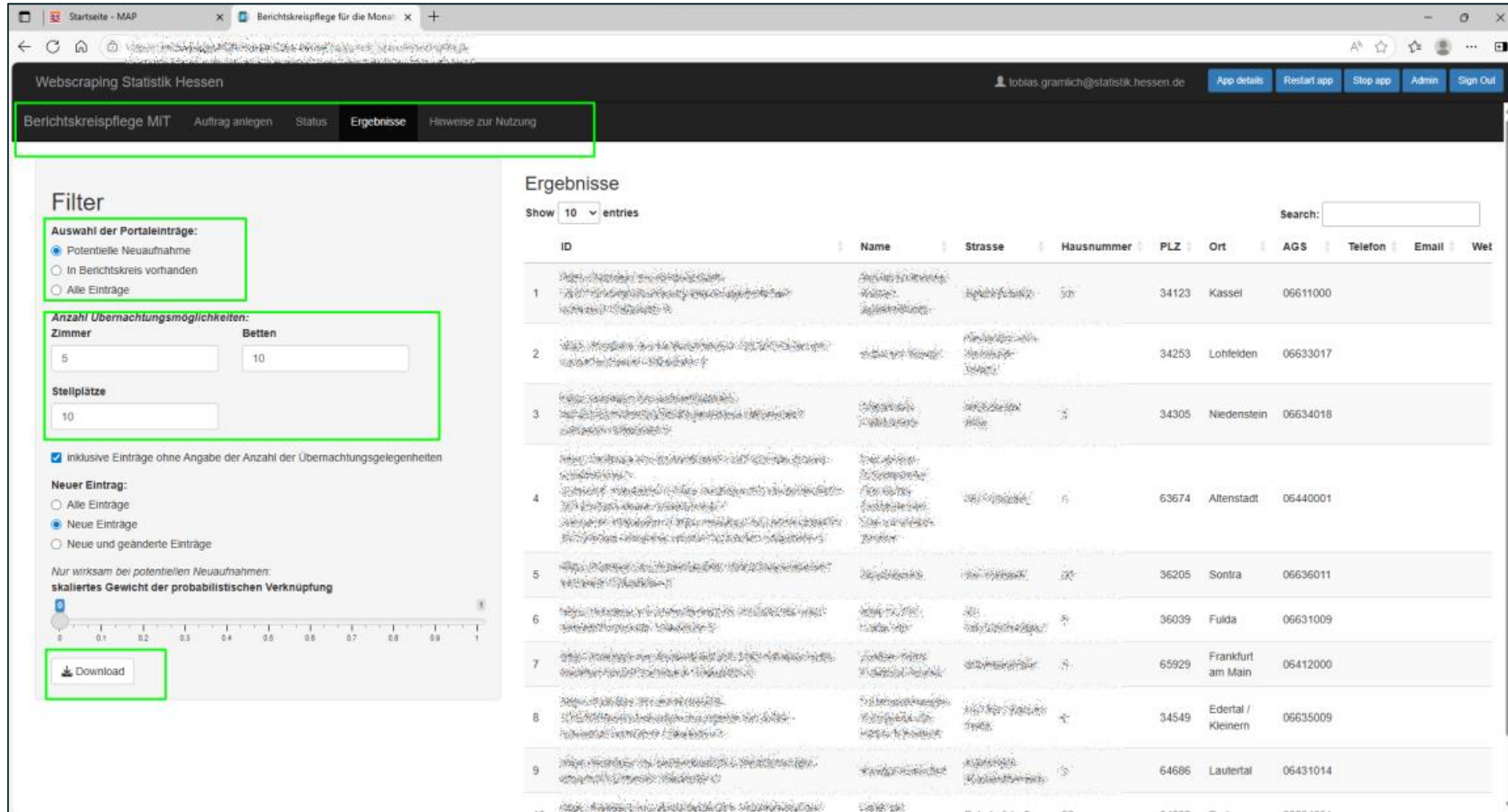
Eine Übersicht über Ihre aktuell wartenden / laufenden Aufträge sowie Ihre vergangenen Aufträge finden Sie unter dem Reiter [Auftragsübersicht](#).

Weitere Informationen zur Bedienung der Anwendung finden Sie in den Reitern [Benutzerhandbuch Datenbereitstellung](#) und [Benutzerhandbuch Webanwendung](#).

Wenden Sie sich bei Fragen und Problemen an das Webscraping-Team unter [webscraping@statistik.hessen.de](mailto:webscraping@statistik.hessen.de)



## GUI: Provding results (2)



The screenshot displays the 'Web scraping Statistik Hessen' application. The navigation bar at the top includes links for 'Berichtskreispflege MIT', 'Auftrag anlegen', 'Status', 'Ergebnisse' (highlighted), and 'Hinweise zur Nutzung'. The 'Ergebnisse' section shows a list of 10 entries with columns for ID, Name, Strasse, Hausnummer, PLZ, Ort, AGS, Telefon, Email, and Web. The filter sidebar on the left includes sections for 'Auswahl der Portaleinträge', 'Anzahl Übernachtungsmöglichkeiten' (with input fields for Zimmer, Betten, and Stellplätze), and 'Neuer Eintrag' (with radio buttons for 'Alle Einträge', 'Neue Einträge', and 'Neue und geänderte Einträge'). A 'Download' button is located at the bottom of the filter sidebar.

**Filter**

**Auswahl der Portaleinträge:**

- ☒ Potentielle Neuaufnahme
- ☐ In Berichtskreis vorhanden
- ☐ Alle Einträge

**Anzahl Übernachtungsmöglichkeiten:**

Zimmer:  Betten:

Stellplätze:

☒ inklusive Einträge ohne Angabe der Anzahl der Übernachtungsgelegenheiten

**Neuer Eintrag:**

- ☐ Alle Einträge
- ☒ Neue Einträge
- ☐ Neue und geänderte Einträge

Nur wirksam bei potentiellen Neuaufnahmen:  
skaliertes Gewicht der probabilistischen Verknüpfung

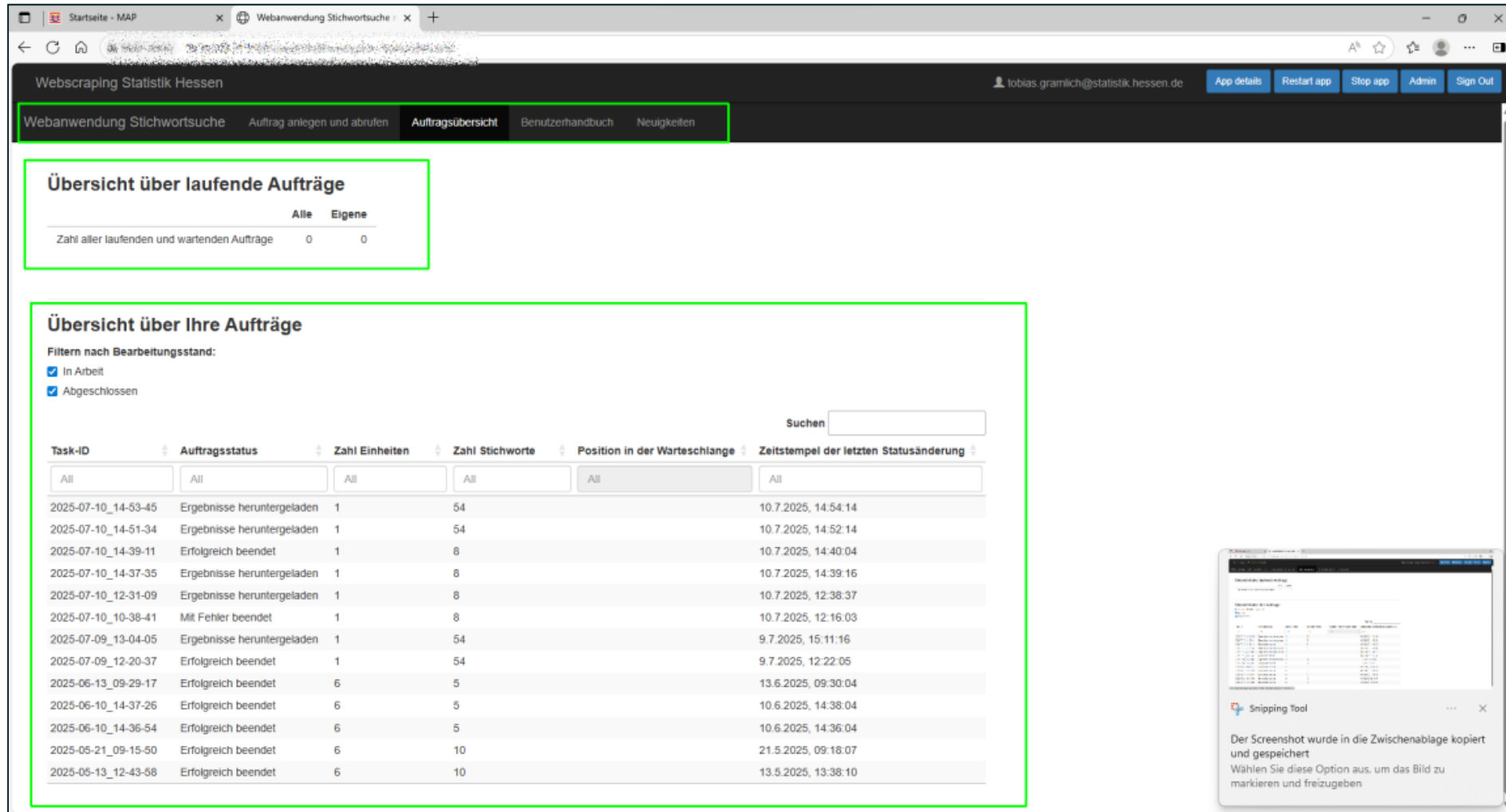
**Ergebnisse**

Show 10 entries

Search:

ID	Name	Strasse	Hausnummer	PLZ	Ort	AGS	Telefon	Email	Web
1	...	...	...	34123	Kassel	06611000			
2	...	...	...	34253	Lohfelden	06633017			
3	...	...	...	34305	Niederstein	06634018			
4	...	...	...	63674	Altenstadt	06440001			
5	...	...	...	36205	Sontra	06636011			
6	...	...	...	36039	Fulda	06631009			
7	...	...	...	65929	Frankfurt am Main	06412000			
8	...	...	...	34549	Edertal / Kleinern	06635009			
9	...	...	...	64686	Lautertal	06431014			

## GUI: overview for users



**Webanwendung Stichwortsuche**

Navigation: Auftrag anlegen und abrufen | **Auftragsübersicht** | Benutzerhandbuch | Neuigkeiten

**Übersicht über laufende Aufträge**

Alle | Eigene

Zahl aller laufenden und wartenden Aufträge: 0 | 0

**Übersicht über Ihre Aufträge**

Filtern nach Bearbeitungsstand:

- ☒ In Arbeit
- ☒ Abgeschlossen

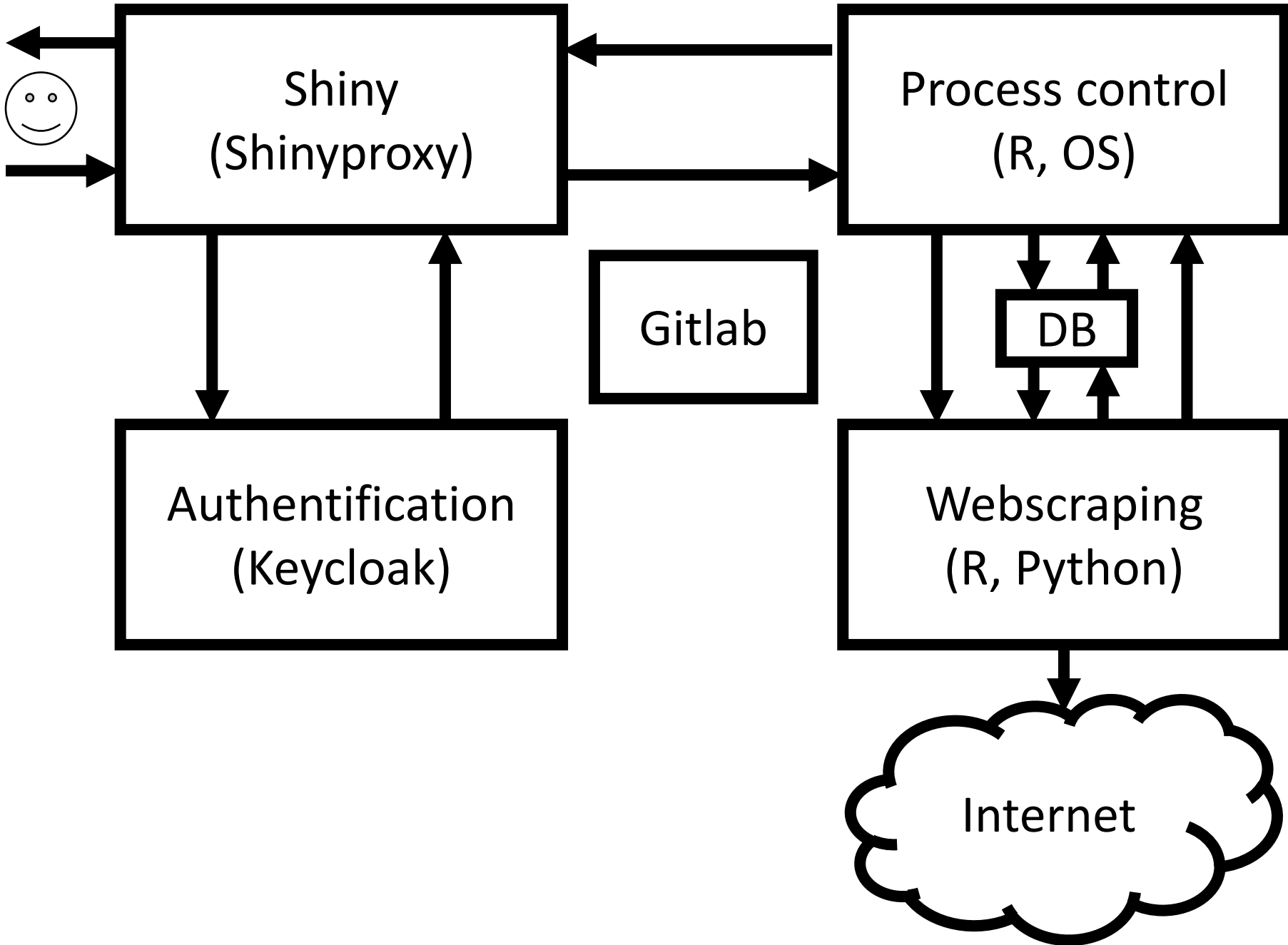
Suchen:

Task-ID	Auftragsstatus	Zahl Einheiten	Zahl Stichworte	Position in der Warteschlange	Zeitstempel der letzten Statusänderung
2025-07-10_14-53-45	Ergebnisse heruntergeladen	1	54		10.7.2025, 14:54:14
2025-07-10_14-51-34	Ergebnisse heruntergeladen	1	54		10.7.2025, 14:52:14
2025-07-10_14-39-11	Erfolgreich beendet	1	8		10.7.2025, 14:40:04
2025-07-10_14-37-35	Ergebnisse heruntergeladen	1	8		10.7.2025, 14:39:16
2025-07-10_12-31-09	Ergebnisse heruntergeladen	1	8		10.7.2025, 12:38:37
2025-07-10_10-38-41	Mit Fehler beendet	1	8		10.7.2025, 12:16:03
2025-07-09_13-04-05	Ergebnisse heruntergeladen	1	54		9.7.2025, 15:11:16
2025-07-09_12-20-37	Erfolgreich beendet	1	54		9.7.2025, 12:22:05
2025-06-13_09-29-17	Erfolgreich beendet	6	5		13.6.2025, 09:30:04
2025-06-10_14-37-26	Erfolgreich beendet	6	5		10.6.2025, 14:38:04
2025-06-10_14-36-54	Erfolgreich beendet	6	5		10.6.2025, 14:36:04
2025-05-21_09-15-50	Erfolgreich beendet	6	10		21.5.2025, 09:18:07
2025-05-13_12-43-58	Erfolgreich beendet	6	10		13.5.2025, 13:38:10

Snipping Tool

Der Screenshot wurde in die Zwischenablage kopiert und gespeichert

Wählen Sie diese Option aus, um das Bild zu markieren und freizugeben

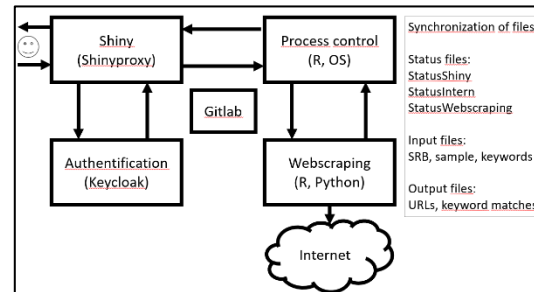
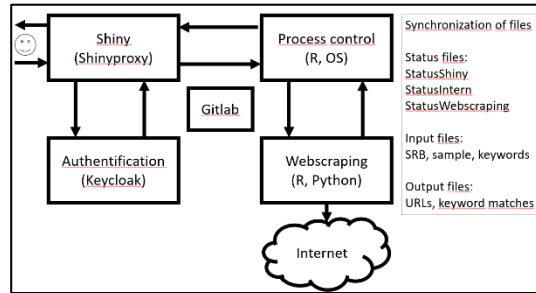
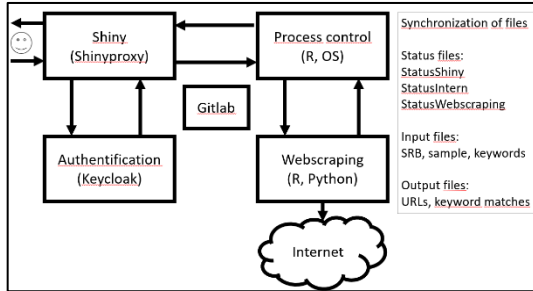
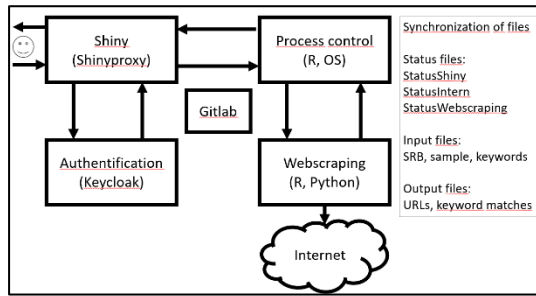


Synchronization of files

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](mailto:tobias.gramlich@statistik.hessen.de)**

**E-Mail: [webscraping@statistik.hessen.de](mailto:webscraping@statistik.hessen.de)**

**<https://statistik.hessen.de>**