

# 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

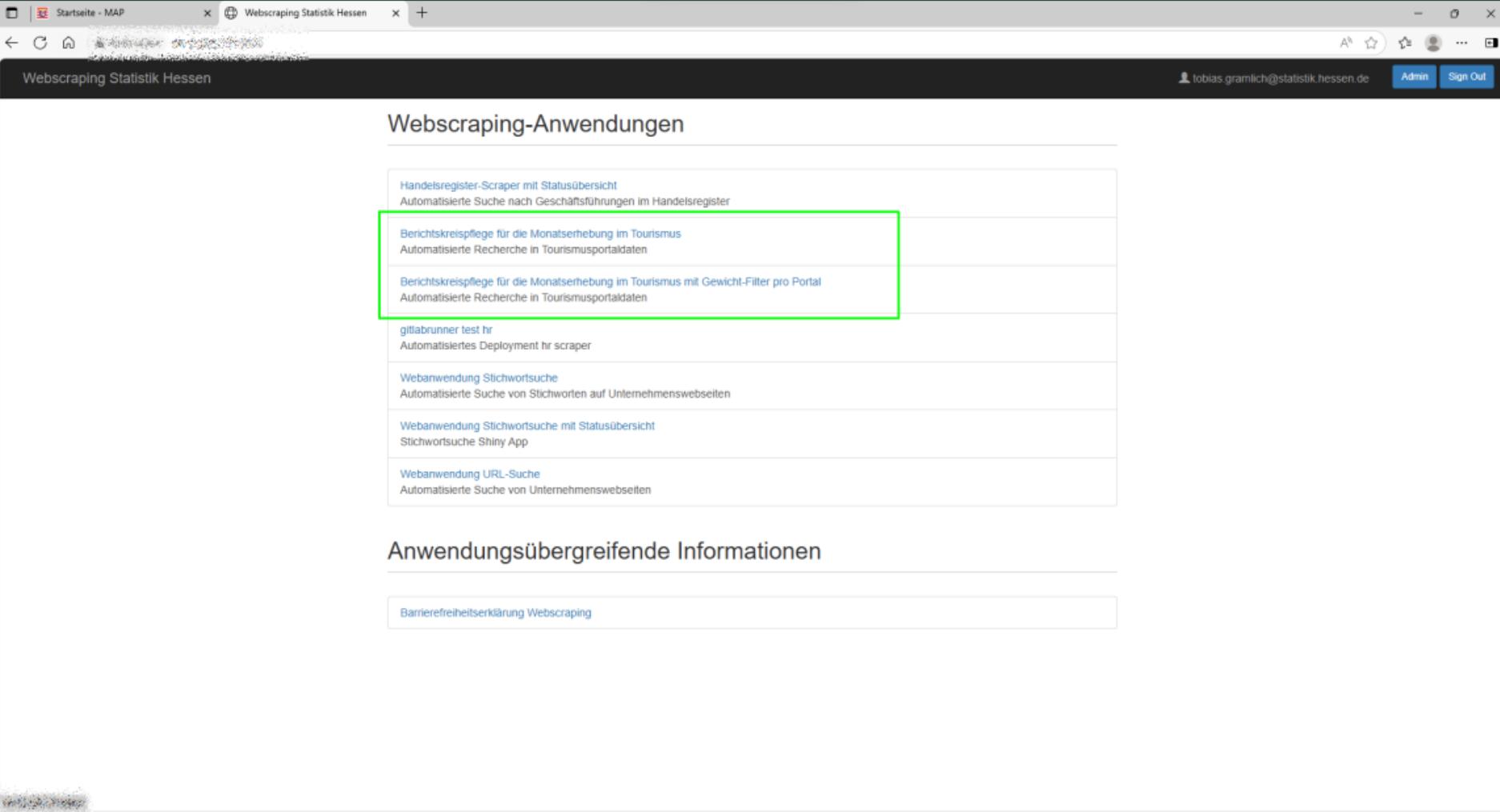
## (Production) Systems of Statistical Scraping at Statistics Hesse

- (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



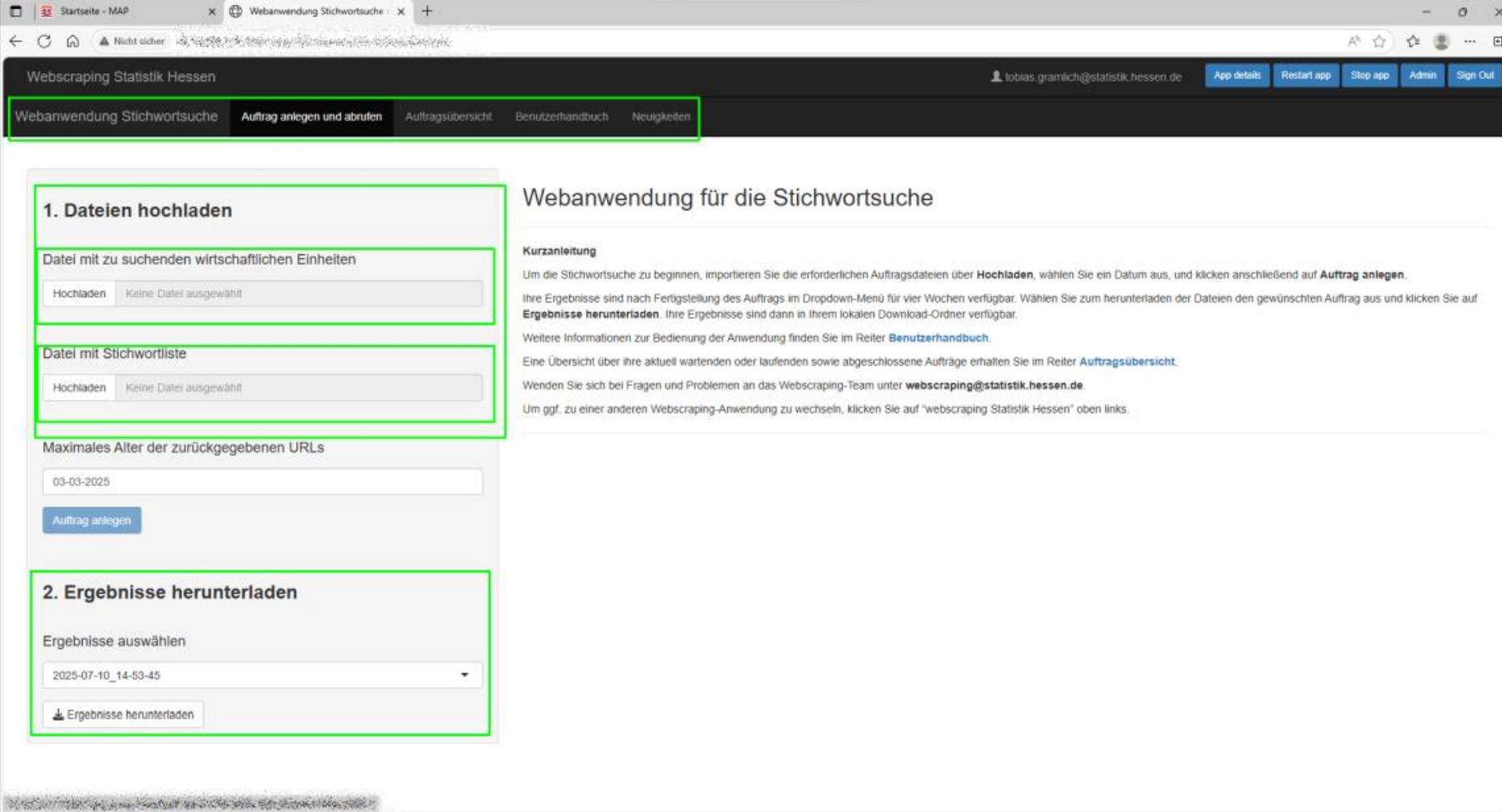
The screenshot shows a web-based application interface titled "Webscraping Statistik Hessen". The main heading is "Webscraping-Anwendungen". Below it is a list of applications:

- Handelsregister-Scraper mit Statusübersicht  
Automatisierte Suche nach Geschäftsführungen im Handelsregister
- Berichtskreispflege für die Monatserhebung im Tourismus  
Automatisierte Recherche in Tourismusportaldaten (highlighted)
- Berichtskreispflege für die Monatserhebung im Tourismus mit Gewicht-Filter pro Portal  
Automatisierte Recherche in Tourismusportaldaten
- gitlabrunner test hr  
Automatisiertes Deployment hr scraper
- Webanwendung Stichwortsuche  
Automatisierte Suche von Stichworten auf Unternehmenswebseiten
- Webanwendung Stichwortsuche mit Statusübersicht  
Stichwortsuche Shiny App
- Webanwendung URL-Suche  
Automatisierte Suche von Unternehmenswebseiten

Below this section is another heading: "Anwendungsübergreifende Informationen".

- Barrierefreiheitserklärung Webscraping

# GUI: Providing input files, parameters (1)



The screenshot shows the "Webanwendung Stichwortsuche" (Search Application) interface. The top navigation bar includes links for "Webanwendung Stichwortsuche" (highlighted with a green box), "Auftrag anlegen und abrufen", "Auftragsübersicht", "Benutzerhandbuch", and "Neuigkeiten". On the right, there's a user profile for "lambert.gramlich@statistik.hessen.de" and buttons for "App details", "Restart app", "Stop app", "Admin", and "Sign Out".

**1. Dateien hochladen**

- Datei mit zu suchenden wirtschaftlichen Einheiten**: A file upload field showing "Hochladen" and "Keine Datei ausgewählt".
- Datei mit Stichwortliste**: A file upload field showing "Hochladen" and "Keine Datei ausgewählt".

**Maximales Alter der zurückgegebenen URLs**: A date input field set to "03-03-2025".

**Auftrag anlegen**: A blue button.

**2. Ergebnisse herunterladen**

- Ergebnisse auswählen**: A dropdown menu showing "2025-07-10\_14-53-45".
- Ergebnisse herunterladen**: A blue button with a download icon.

**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.

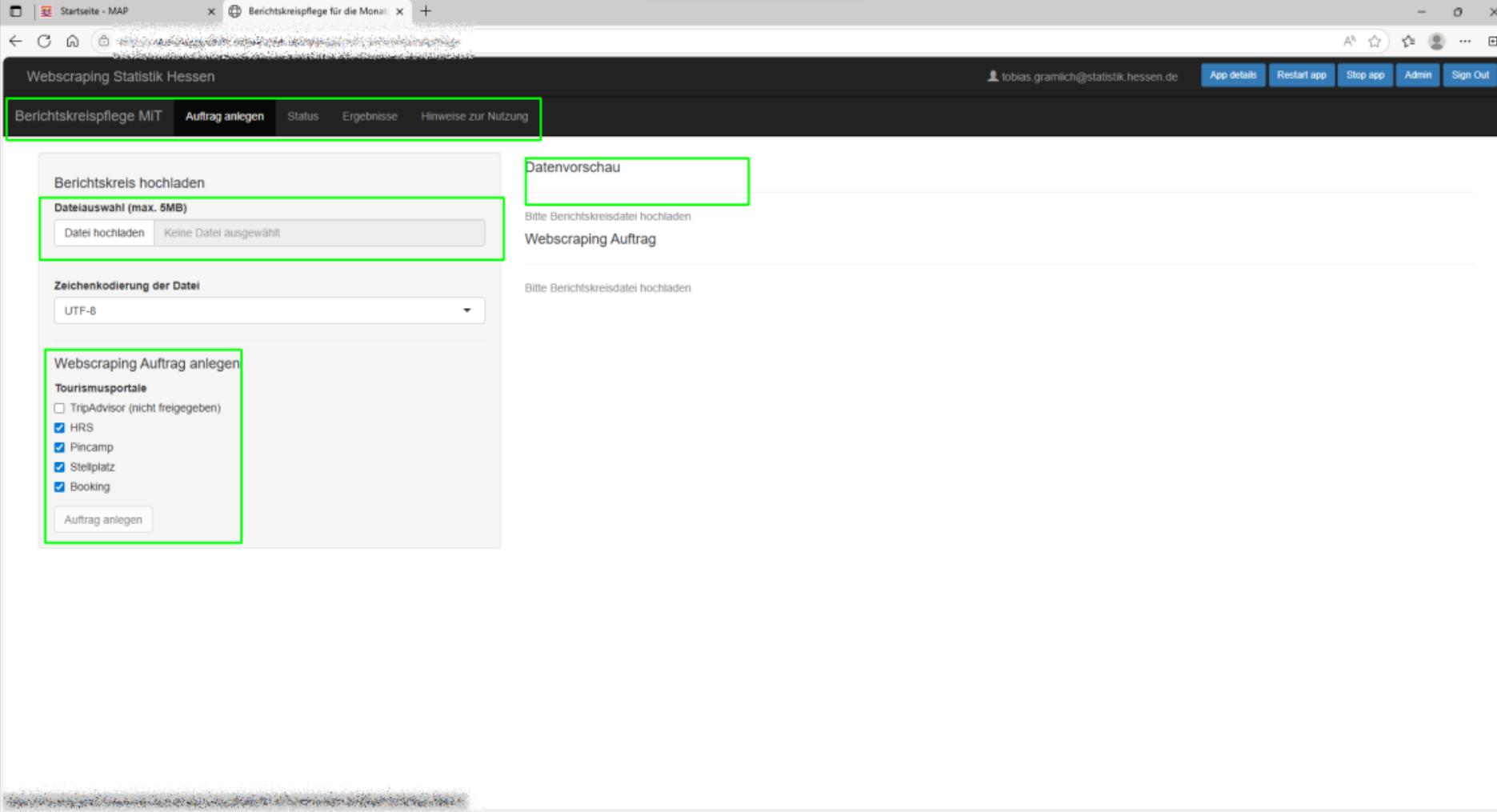
Weitere Informationen zur Bedienung der Anwendung finden Sie im Reiter **Benutzerhandbuch**.

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)



The screenshot displays a web-based user interface for managing scraping tasks. At the top, there's a navigation bar with tabs: 'Berichtskreispflege MIT' (highlighted with a green box), 'Auftrag anlegen', 'Status', 'Ergebnisse', and 'Hinweise zur Nutzung'. On the right side of the header, there's a user profile and several blue buttons: 'App details', 'Restart app', 'Stop app', 'Admin', and 'Sign Out'.

The main area contains two main sections:

- Berichtskreis hochladen:** A section for uploading a report district file. It includes a file input field labeled 'Dateiauswahl (max. 5MB)' with the placeholder 'Datei hochladen' and 'Keine Datei ausgewählt'. Below it is a dropdown menu for 'Zeichenkodierung der Datei' set to 'UTF-8'.
- Datenvorschau:** A section for previewing the data. It has a placeholder text 'Bitte Berichtskreisdatei hochladen'.

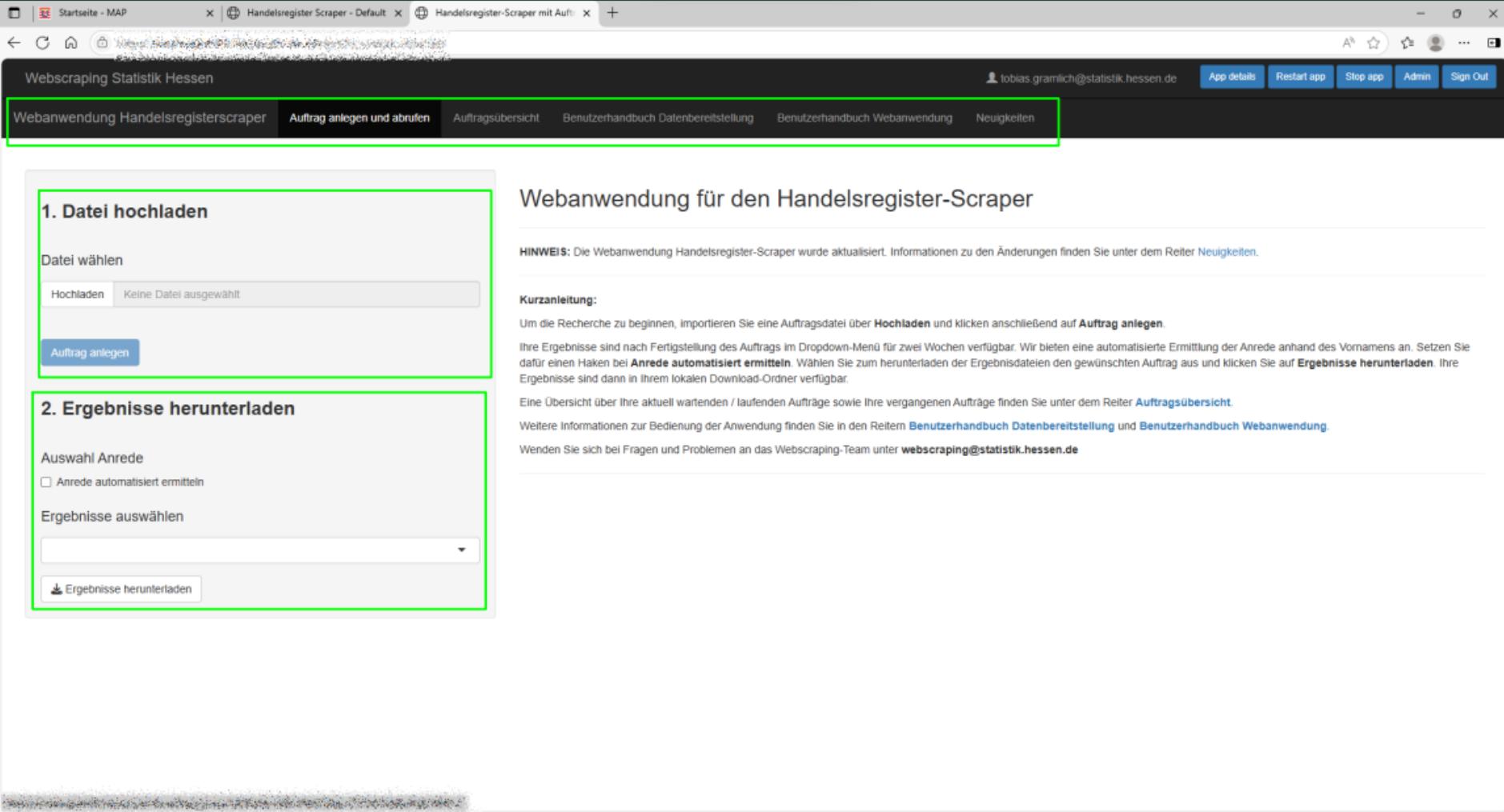
On the right side of the main area, there's another section titled 'Webscraping Auftrag' with a placeholder text 'Bitte Berichtskreisdatei hochladen'.

**Webscraping Auftrag anlegen:** This section is enclosed in a green box and contains a list of tourism websites with checkboxes:

- TripAdvisor (nicht freigegeben)
- HRS
- Pincamp
- Stellplatz
- Booking

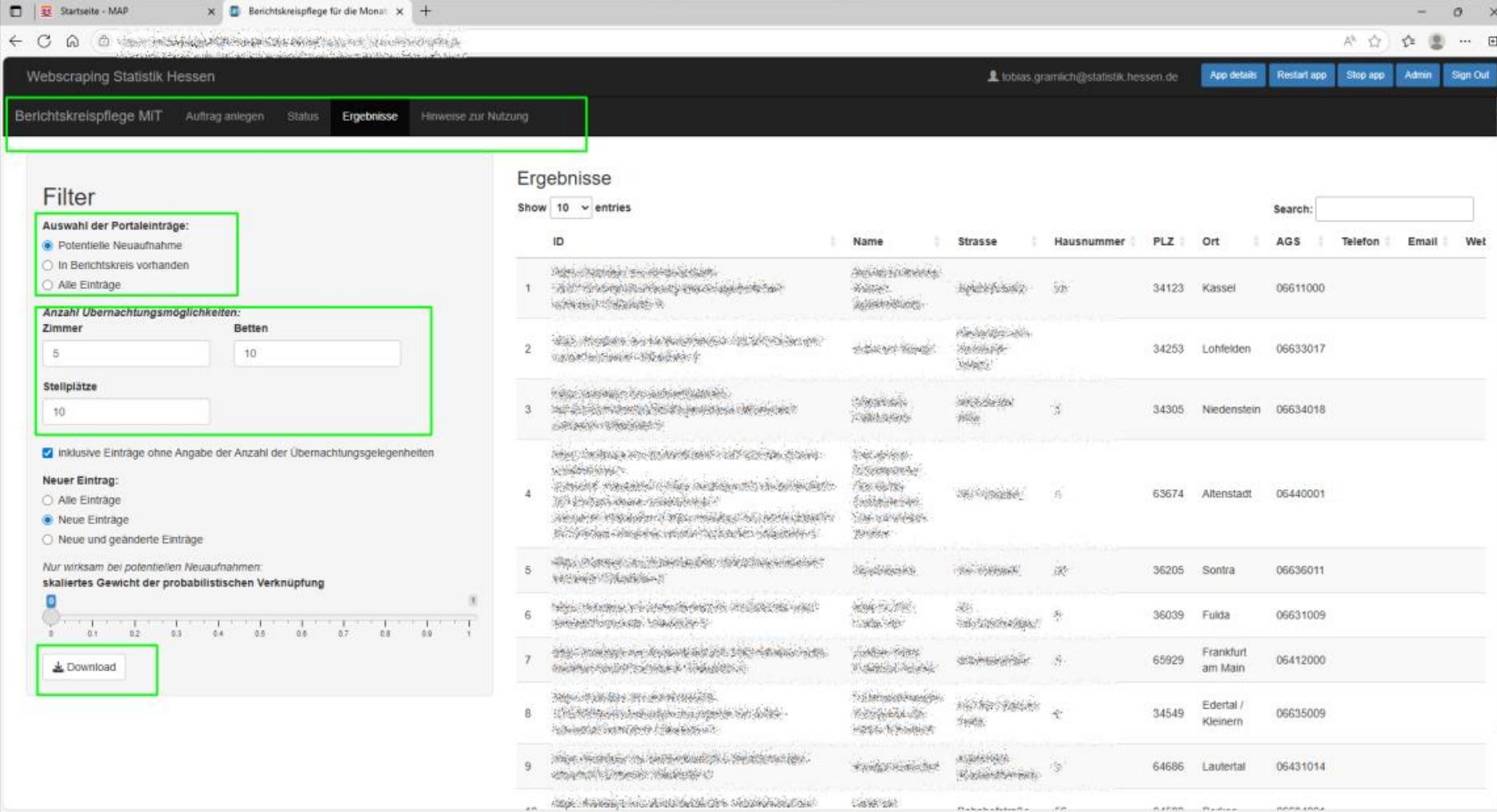
A button labeled 'Auftrag anlegen' is located at the bottom of this section.

# GUI: Providing results (1)



The screenshot shows the "Webscraping Statistik Hessen" application interface. The top navigation bar includes links for "Webanwendung Handelsregisterscraper" (highlighted with a green box), "Auftrag anlegen und abrufen", "Auftragsübersicht", "Benutzerhandbuch Datenbereitstellung", "Benutzerhandbuch Webanwendung", and "Neuigkeiten". The main content area is titled "Webanwendung für den Handelsregister-Scraper". It features two main sections: "1. Datei hochladen" (highlighted with a green box) and "2. Ergebnisse herunterladen" (highlighted with a green box). The "1. Datei hochladen" section contains a file input field labeled "Datei wählen" with the placeholder "Hochladen Keine Datei ausgewählt" and a blue "Auftrag anlegen" button. The "2. Ergebnisse herunterladen" section contains a checkbox "Anrede automatisiert ermitteln", a dropdown menu for selecting results, and a blue "Ergebnisse herunterladen" button. A note at the top right states: "HINWEIS: Die Webanwendung Handelsregister-Scraper wurde aktualisiert. Informationen zu den Änderungen finden Sie unter dem Reiter Neuigkeiten." Below it, a "Kurzanleitung:" provides instructions for starting the search by uploading a request file and clicking "Auftrag anlegen". Further details about results availability and download are provided.

## GUI: Providing results (2)



The screenshot shows a web-based application interface for reporting. The top navigation bar includes links for 'Startseite - MAP', 'Berichtskreispflege für die Monate', and other administrative functions like 'App details', 'Restart app', 'Stop app', 'Admin', and 'Sign Out'. The main content area is divided into two sections: 'Filter' on the left and 'Ergebnisse' on the right.

**Filter Section:**

- Auswahl der Portaleinträge:**
  - Potentielle Neuaufnahme
  - In Berichtskreis vorhanden
  - Alle Einträge
- Anzahl Übernachtungsmöglichkeiten:**

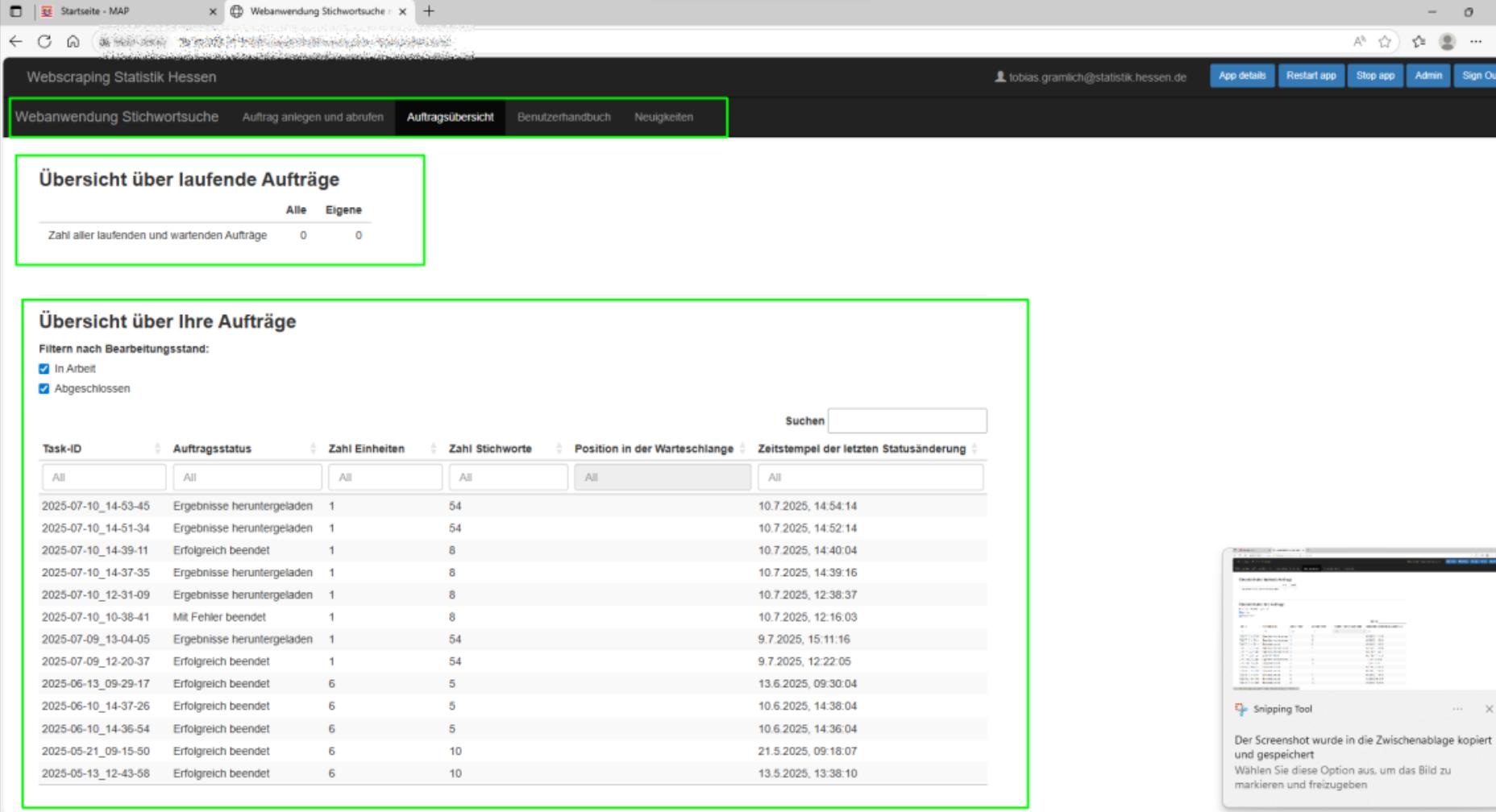
Zimmer	Betten
5	10
- Stellplätze:**  
10
- Inklusive Einträge ohne Angabe der Anzahl der Übernachtungsgelegenheiten
- Neuer Eintrag:**
  - Alle Einträge
  - Neue Einträge
  - Neue und geänderte Einträge
- Nur wirksam bei potentiellen Neuaufnahmen:**  
skaliertes Gewicht der probabilistischen Verknüpfung  
A slider scale from 0 to 1, currently set at 0.
- Download** button

**Ergebnisse Section:**

Show 10 entries

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

# GUI: overview for users



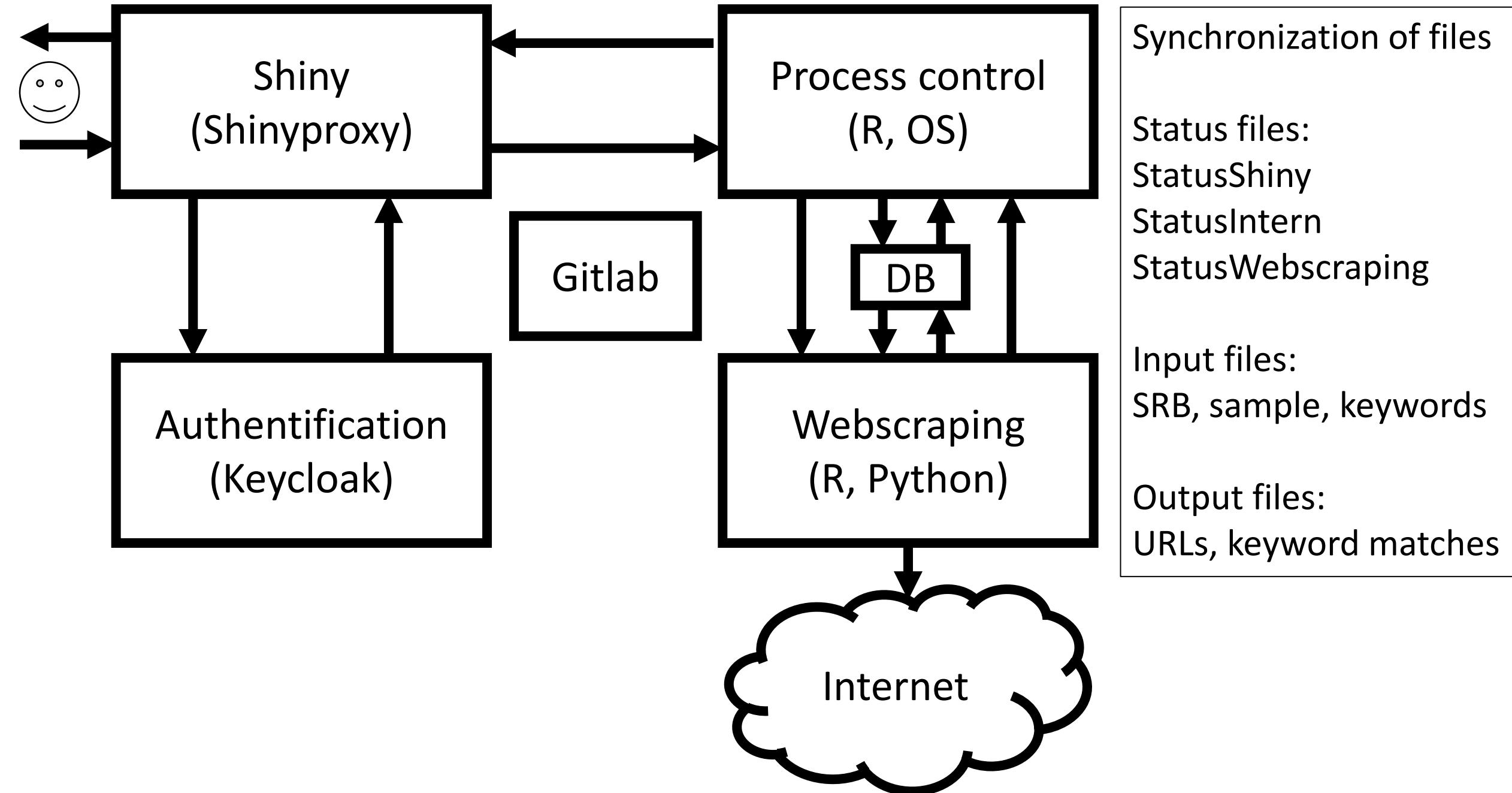
The screenshot displays the 'Webscraping Statistik Hessen' application interface. At the top, there is a navigation bar with tabs: 'Webanwendung Stichwortsuche' (highlighted with a green border), 'Auftrag anlegen und abrufen', 'Auftragsübersicht' (highlighted with a green border), 'Benutzerhandbuch', and 'Neuigkeiten'. On the right side of the header, there is a user profile icon and links for 'App details', 'Restart app', 'Stop app', 'Admin', and 'Sign Out'.

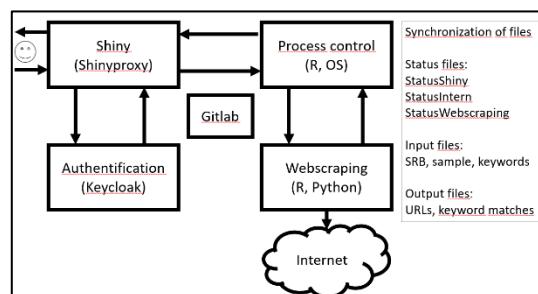
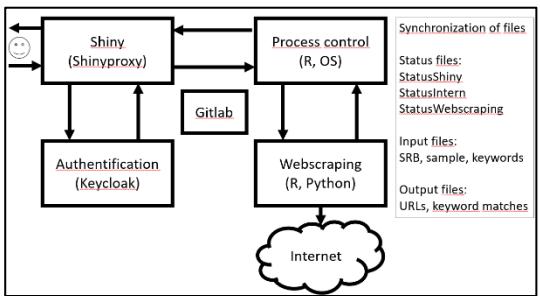
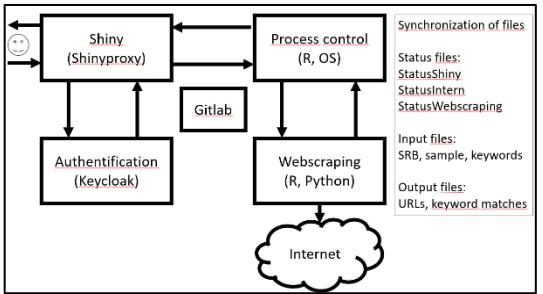
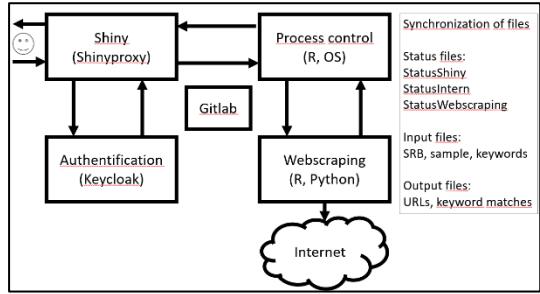
The first section, 'Übersicht über laufende Aufträge', contains a summary table with two rows: 'Alle' and 'Eigene'. The 'Alle' row shows 'Zahl aller laufenden und wartenden Aufträge' as 0. The 'Eigene' row also shows 0. Below this, there is a table titled 'Übersicht über Ihre Aufträge'.

The 'Übersicht über Ihre Aufträge' section includes a filter for 'Bearbeitungsstand': 'In Arbeit' (checked) and 'Abgeschlossen' (checked). It features a search bar labeled 'Suchen' and a table with columns: Task-ID, Auftragsstatus, Zahl Einheiten, Zahl Stichworte, Position in der Warteschlange, and Zeitstempel der letzten Statusänderung. The table lists various tasks with their details.

A small window titled 'Snipping Tool' is overlaid on the bottom right, showing a screenshot of the application's interface. The text in the snipper says: 'Der Screenshot wurde in die Zwischenablage kopiert und gespeichert. Wählen Sie diese Option aus, um das Bild zu markieren und freizugeben.'

Task-ID	Auftragsstatus	Zahl Einheiten	Zahl Stichworte	Position in der Warteschlange	Zeitstempel der letzten Statusänderung
All	All	All	All	All	All
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





Separate environments:

Development

Test

Preproduction

Production

May differ in requirements regarding data protection, information security, access restrictions, resources (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>**