Testing in a scraping application

Within the core of Datlinq / Roamler we have a database filled with out of home foodservice locations in Europe. These are first of all, all horeca locations, restaurants, hotels, bars etc. But also zoos, petrol stations, minimarkets etc.

To keep the database up to date we try to collect and process as much data as possible about these locations. A big part of that is data we scrape the html from a lot of platforms from the web, like google, tripadvisor, delivery sites. After scraping we need to parse the html and store the results in a semi structured way. To that end we build an internal scraping service / application.

The application consist of two steps

- Extract the html of the website
- Parse the html to our own defined schema

One of the challenges we are facing in this whole process is testing. Testing within a data application always provides it's own set of challenges, in this case the complication is that the website we are scraping and parsing can change without us actively knowing it. Making for example our parsing logic incorrect.

We would like you to describe what kind of tests you would want to implement to test the application and more specifically how would you address the data testing.

- Which methods or frameworks would you use?
- In what phase of the development would you test what?
- Which tests would you implement in the production application?
    - With production we mean which tests / checks would you implement in the process where we continuously pick up data from websites.