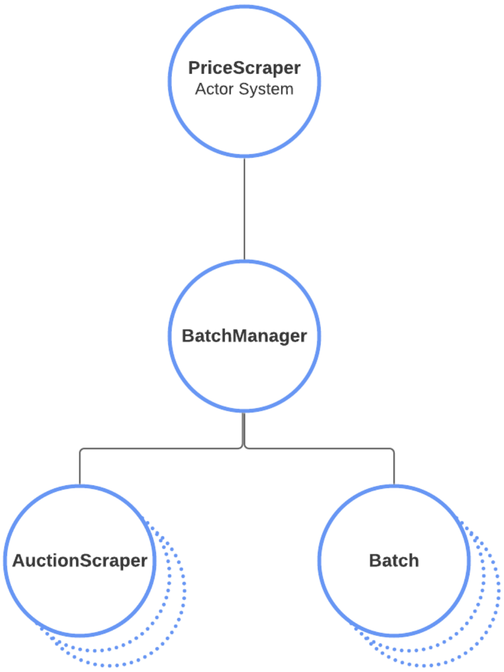
PriceScraper

Table of Contents

1. PriceScraper

The **PriceScraper** is a process that allows to extract auctions along with their characteristics, based on a list of **Batch Specifications**. The price scraper doesn't create isolated auctions, instead it creates what we call a **Batch** that contains a list of auctions, this list of auctions are those auctions that are found on a web page at the time of scraping. Such a web page is for example a page containing the list of stamps for an url provided in a Batch Specification (named listing page url later).

We have decided to use Scala, Akka, CQRS/ES for the implementation, below is an overview of the actor system at the time of writing :



The different components of this actor system are :

PriceScraper: it's the entry point of the application and it creates the actor system and the BatchManager

BatchManager: it communicates with the AuctionScraper to provide the BatchSpecifications and it creates the Batch actors upon reception of messages from the AuctionScraper

AuctionScraper: it communicates with the BatchManager to obtain the BatchSpecifications and it scraps the urls provided in the BatchSpecifications, then it provides the scraped data to the BatchManager so that it can create the Batch.

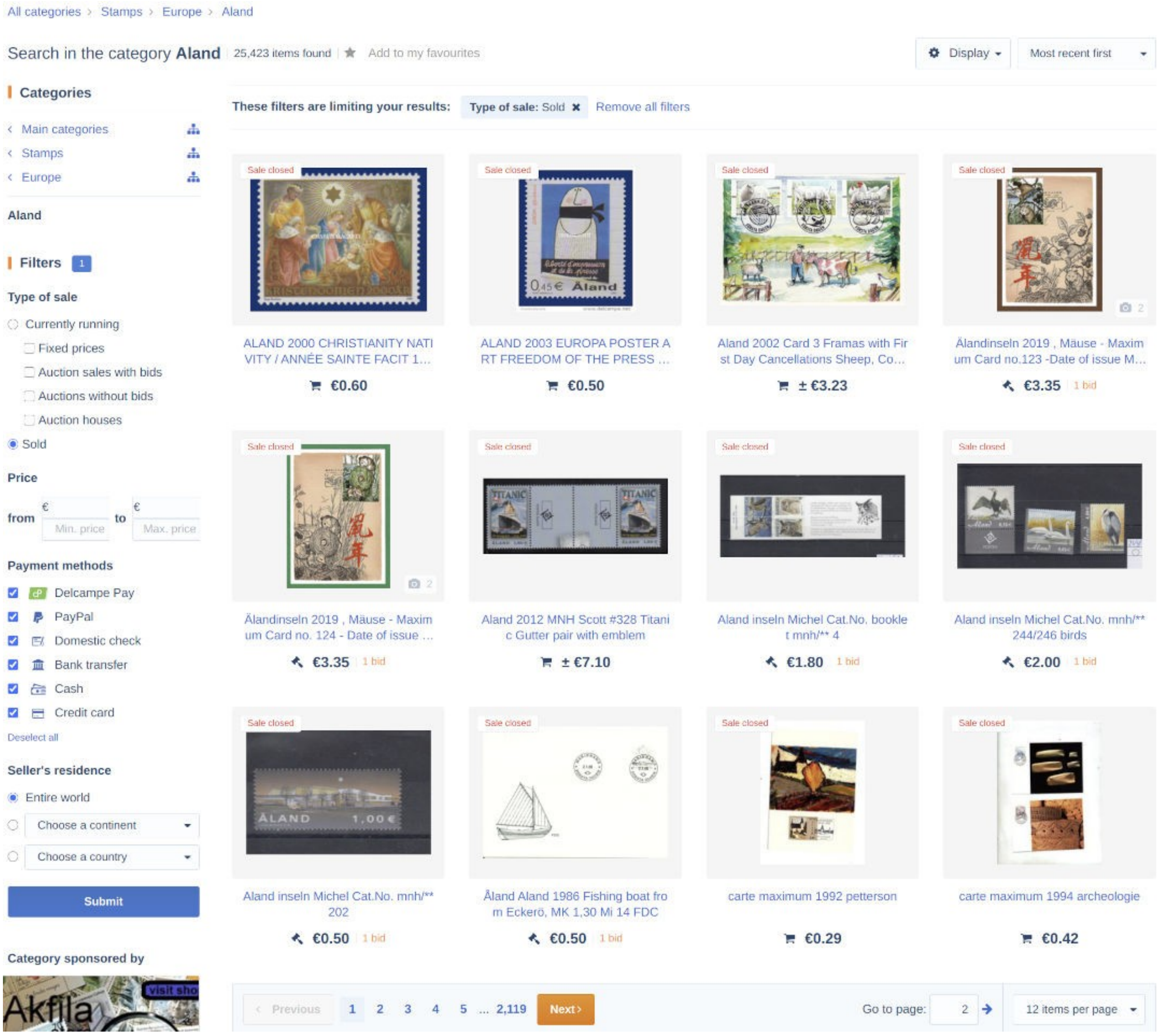
A **BatchSpecification** mostly provides a **Listing Page URL** for a **Provider** along with some characteristics for the auctions that we expect to find on this Listing Page URL :

- the **family identifier** of the auctions for that URL (mandatory)
- the **country identifier** of the auctions for that URL (optional)
- the **topic identifier** of the auctions for that URL (optional)
- a range of year during which the object related to the auction has been produced
  - the **start year** (optional)
  - the **end year** (optional)
- a **condition identifier** (optional) for the object related to the auction (mandatory, at least one condition)
- the **URL of the last auction** that has been processed for the Listing URL (optional)

The **Provider** is a value that allows the batch extractor to better identify the web site from which it is going to extract the auctions, i.e. Delcampe, Ebay, ...

The **Listing Page URL** is an URL that allows to access a list of auctions who have some common characteristics as described in the BatchSpecification.

Below is a screenshot showing an example of what can be found on a Listing Page Url :



All categories › Stamps › Europe › Aland

Search in the category **Aland** 25,423 items found ★ Add to my favourites

Display | Most recent first

Categories

‹ Main categories
‹ Stamps
‹ Europe

Aland

Filters 1

Type of sale

○ Currently running
□ Fixed prices
□ Auction sales with bids
□ Auctions without bids
□ Auction houses
● Sold

Price

from € Min. price to € Max. price

Payment methods

☑ Delcampe Pay
☑ PayPal
☑ Domestic check
☑ Bank transfer
☑ Cash
☑ Credit card

Deselect all

Seller's residence

● Entire world
○ Choose a continent
○ Choose a country

Submit

Category sponsored by

These filters are limiting your results: Type of sale: Sold ✖ Remove all filters

Sale closed
ALAND 2000 CHRISTIANITY NATIVITY / ANNÉE SAINTE FACIT 1...
€0.60

Sale closed
ALAND 2003 EUROPA POSTER ART FREEDOM OF THE PRESS ...
€0.50

Sale closed
Aland 2002 Card 3 Framas with First Day Cancellations Sheep, Co...
± €3.23

Sale closed
Ålandinseln 2019 , Mäuse - Maximum Card no.123 -Date of issue M...
€3.35 1 bid

Sale closed
Ålandinseln 2019 , Mäuse - Maximum Card no. 124 - Date of issue ...
€3.35 1 bid

Sale closed
Aland 2012 MNH Scott #328 Titanic Gutter pair with emblem
± €7.10

Sale closed
Aland inseln Michel Cat.No. booklet mnh/** 4
€1.80 1 bid

Sale closed
Aland inseln Michel Cat.No. mnh/** 244/246 birds
€2.00 1 bid

Sale closed
Aland inseln Michel Cat.No. mnh/** 202
€0.50 1 bid

Sale closed
Åland Aland 1986 Fishing boat from Eckerö, MK 1,30 Mi 14 FDC
€0.50 1 bid

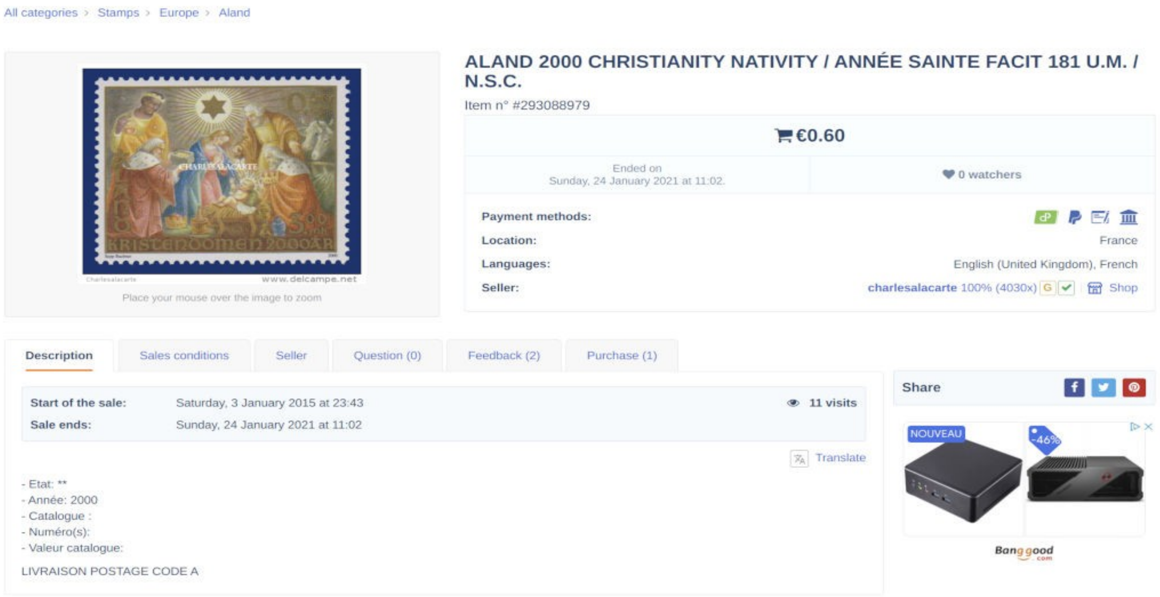
Sale closed
carte maximum 1992 petterson
€0.29

Sale closed
carte maximum 1994 archeologie
€0.42

‹ Previous 1 2 3 4 5 ... 2,119 Next›

Go to page: 2 → 12 items per page

And another screenshot showing what can be found on the Auction Url for the first auction appearing on the Listing Page



All categories › Stamps › Europe › Aland

Place your mouse over the image to zoom

ALAND 2000 CHRISTIANITY NATIVITY / ANNÉE SAINTE FACIT 181 U.M. / N.S.C.

Item n° #293088979

€0.60

Ended on
Sunday, 24 January 2021 at 11:02.

0 watchers

Payment methods:	
Location:	France
Languages:	English (United Kingdom), French
Seller:	charlesalacarte 100% (4030x) G ✓ Shop

Description | Sales conditions | Seller | Question (0) | Feedback (2) | Purchase (1)

Start of the sale:	Saturday, 3 January 2015 at 23:43	11 visits
Sale ends:	Sunday, 24 January 2021 at 11:02	

Translate

- Etat: **
- Année: 2000
- Catalogue :
- Numéro(s):
- Valeur catalogue:

LIVRAISON POSTAGE CODE A

Share

The AuctionScraper will be processing each of this auctions and it will extract the following information :

- title
- image thumbnail URL
- image URL
- type (fixed price or bidding)
- starting price with the currency
- ending price with the currency (price it was sold)
- number of bids (this value is set to 1 for a fixed price auction)
- list of bids (bidder pseudo, bidder price with the currency)
- seller pseudo
- seller country

The scraping process consists of the following steps :

1. the AuctionScraper asks to the BatchManager for the next BatchSpecification that has to be processed, for the Provider that this AuctionScraper is related to
2. the BatchManager answers to the AuctionScraper with the BatchSpecification that has to be updated for the Provider provided by the AuctionScraper
3. the AuctionScraper opens the BatchSpecification's Listing Page Url
   - 3.1. if the Listing Page Url is opened successfully
     - 3.1.1. the AuctionScraper extracts the Auction Urls found on that page
     - 3.1.2. the AuctionScraper extracts the details of the Auctions found on each of this Auction Url
     - 3.1.3. the AuctionScraper sends a message to the BatchManager with a Batch containing the BatchSpecificationID and the list of Auctions it has extracted
     - 3.1.4. if the page number equals 1, the AuctionScraper sends a message to the BatchManager with the Auction Url of the first Auction appearing on the Listing Page
     - 3.1.5. the AuctionScraper *increments the page number* for the Listing Page Url
     - 3.1.6. the AuctionScraper goes to 3
   - 3.2. if the Listing Page Url is not opened successfully the process stops

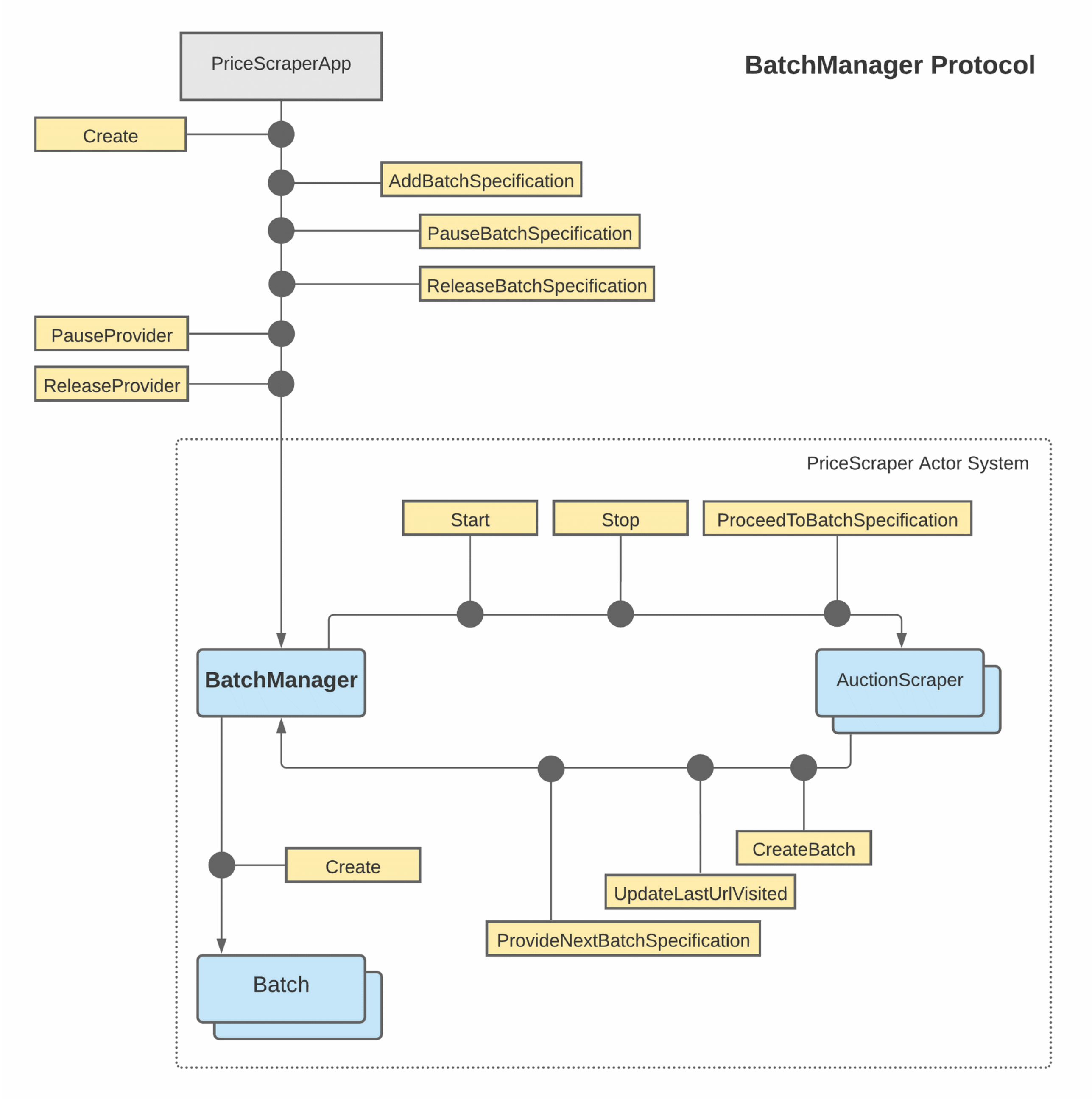
1.1. BatchManager Actor

The BatchManager is an actor that is created by the PriceScraper application when it starts and it has the following responsibilities :

- it creates and starts the AuctionScraper actors for the different providers
- it holds and manages the list of BatchSpecifications

1.1.1. Protocol

The BatchManager protocol is depicted in the schema below and detailed in the later paragraphs.



1.1.1.1. Create

After creating the BatchManager actor the PriceScraper application sends the **Create** message to the BatchManager so that it can initialize itself. During its initialization process the BatchManager creates the AuctionScraper actors for each provider and sends the **Start** message to each of this AuctionScraper actors.

1.1.1.2. AddBatchSpecification

Upon reception of this Command the BatchManager checks that no other BatchSpecification exists with the same name and adds it to the list of BatchSpecifications it holds.

1.1.1.3. PauseBatchSpecification

Upon reception of this Command the BatchManager looks for an existing BatchSpecification with the provided name and marks it as paused (a paused batch specification will not generate any new Batch, until it is released).

1.1.1.4. ReleaseBatchSpecification

Upon reception of this Command the BatchManager looks for an existing BatchSpecification with the provided name and releases it by updating the pause indicator to false, allowing this BatchSpecification to be included in upcoming scraping schedule.

1.1.1.5. PauseProvider

Upon reception of this Command the BatchManager looks for all the BatchSpecification whose provider equals the provided provider, and marks this BatchSpecifications as paused.

1.1.1.6. ReleaseProvider

Upon reception of this Command the BatchManager looks for all the BatchSpecification whose provider equals the provided provider, and releases this BatchSpecifications by updating their pause indicator to false.

1.1.1.7. Start

After creating the AuctionScraper actors the BatchManager sends the **Start** message to each AuctionScraper actor it has created (one per provider). This allows the AuctionScraper to initialize and be ready to handle other messages.

1.1.1.8. CreateBatch

Upon reception of this Command the BatchManager creates a Batch actor with the provided list of auctions and BatchSpecification details, and it sends the Create command to this Batch actor.

1.1.1.9. UpdateLastUrlVisited

Upon reception of this Command the BatchManager looks for an existing BatchSpecification with the provided BatchSpecification ID and updates the lastUrlVisited field with the provided value.

1.1.1.10. ProvideNextBatchSpecification

Upon reception of this Command the BatchManager looks for a BatchSpecification that needs to be updated and whose provider equals the provided provider. If more than one

BatchSpecification meets this criteria then the **ProceedToBatchSpecification** is sent to the AuctionScraper with the BatchSpecification that has been updated the least recently.

A BatchSpecification needs to be updated if the timestamp corresponding of its updatedAt plus its intervalSeconds field is greater than the current timestamp.

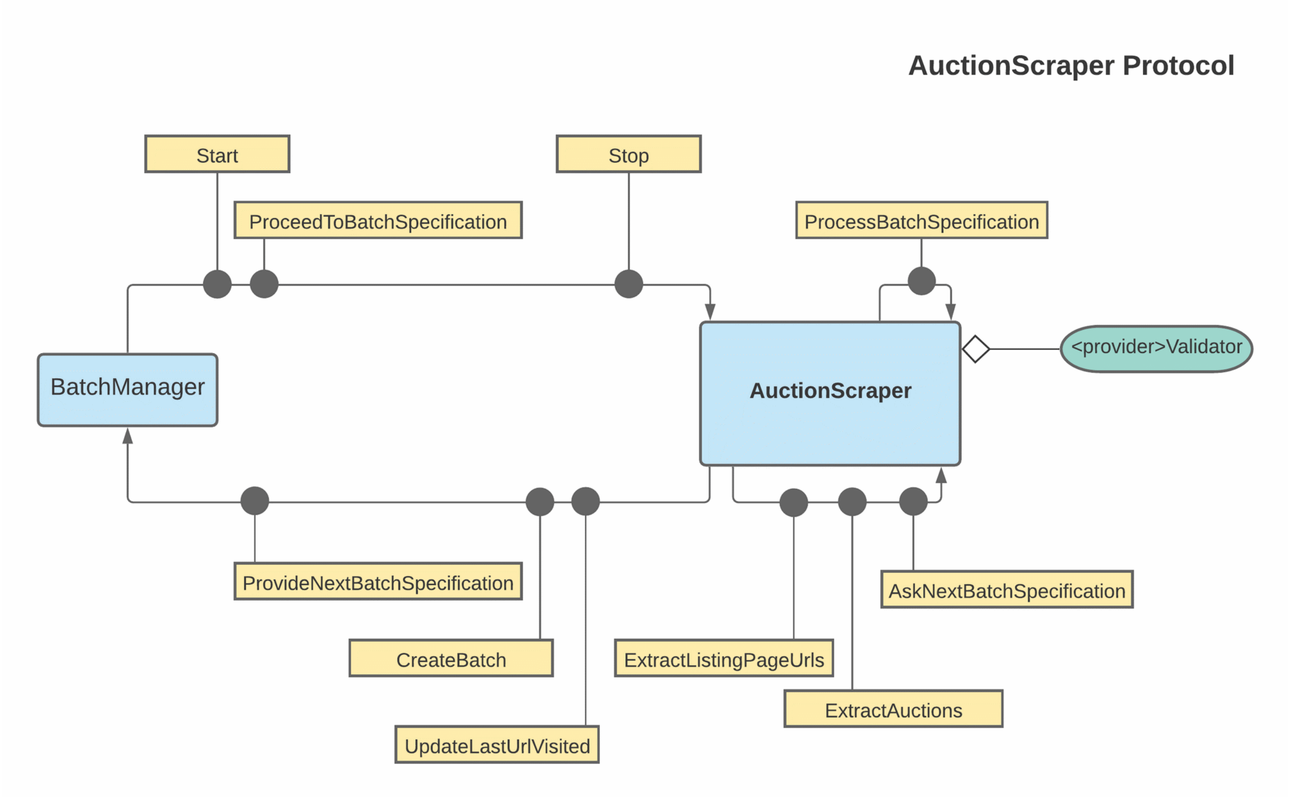
1.2. AuctionScraper Actor

The AuctionScraper is an actor that is created by the BatchManager actor when it starts and it has the following responsibilities :

- it asks the BatchManager for the next BatchSpecification to process (for the provider to whom the AuctionScraper is related to).
- it processes the BatchSpecification returned by the BatchManager and it extracts the Auctions for each of the valid pages of the BatchSpecification's Listing Page Url. The AuctionScraper processes only the Auctions that are new since the last extraction for the BatchSpecification.
- it sends to the BatchManager a request to create a Batch with the list of Auctions it has extracted and it informs the BatchManager to update the last visited url of the BatchSpecification.

1.1.1. Protocol

The AuctionScraper protocol is depicted in the schema below and detailed in the later paragraphs. Each AuctionScraper has a **validator** object that is specialized in the analysis of the HTML code of a specific provider.



1.2.1.1. Start

Upon reception of this Command the AuctionScraper starts sending to himself the **AskNextBatchSpecification** command and is ready to process other Commands.

1.2.1.2. AskNextBatchSpecification

Upon reception of this Command the AuctionScraper sends the **ProvideNextBatchSpecification** to the BatchManager (is then expects to receive the **ProceedToBatchSpecification**, or it timeouts and sends the **AskNextBatchSpecification** to himself).

1.2.1.3. ProvideNextBatchSpecification

The AuctionScraper sends this Command to the BatchManager to obtain the next batch specification to process for the provider it holds.

1.2.1.4. ProceedToBatchSpecification

Upon reception of this Command the AuctionScraper sends the ProcessBatchSpecifiction to himself.

1.2.1.5. ProcessBatchSpecification

The AuctionScraper sends this Command to himself upon reception of the ProceedToBatchSpecification from the BatchManager.

1.2.1.6. ExtractListingPageUrls

The AuctionScraper sends this Command to himself so that it can extract the list of auction urls found on the Listing Page.

1.2.1.7. ExtractAuctions

The AuctionScraper sends this Command to himself so that it can extract the auction details based on the auction urls extracted from the Listing Page during the ExtractListingPageUrls message processing.

1.2.1.8. CreateBatch

The AuctionScraper sends the CreateBatch message to the BatchManager for each Listing Page is has scraped. This message contains a list of Auctions extracted from the Listing Page Url of the BatchSpecification provided by the BatchManager through the ProceedToBatchSpecification message.

1.2.1.9. UpdateLastUrlVisited

The AuctionScraper sends the UpdateLastUrlVisited to the BatchManager for the first page of a Listing Page provided by the BatchManager through the ProceedToBatchSpecification message.