

University of Strasbourg

Master 1 CSMI

Development of algorithms for automating
web data extraction processes



Tetrao

Céline Van Landeghem

2020-2021

Contents

1	Introduction	2
2	Presentation of Tetrao	3
2.1	The start-up Tetrao	3
2.2	The process of Tetrao	3
2.3	The development of Tetrao	4
3	Context and objectives	5
4	Working tools	6
5	The investment funds	7
6	The process to the final product	10
6.1	Extraction of the data	10
6.2	Annotation of extracted data	13
6.3	Artificial intelligence models	15
7	The two main tasks	16
8	The analyses	18
8.1	Search for issuers	18
8.2	Writing the analysis	19
8.3	Verification	20
8.4	The interest of the analysis	20
8.5	Examples of analyses	21
9	Extraction algorithms	26
9.1	The tools	26
9.2	Organization of the project	26
9.3	The extraction files	29
9.4	Executing the extraction files	29
9.5	The API Earnestnet	30
9.6	An example of an extraction script	36
9.7	Specific cases	42
9.7.1	Use of JavaScript	42
9.7.2	Hidden ISIN codes	43
9.7.3	List of shares on several pages	44
9.7.4	Several profiles to implement	45
10	Conclusion	47

1 Introduction

Nowadays, information on the web is available in large quantities, which makes it difficult and complex to process. Thus, with the appearance of the web, users have benefited from access to multiple information from different sources : it's the era of Big Data.

Artificial intelligence techniques are used to identify and extract information from the content of web pages. However, the possible changes in the structure of websites often make it difficult to collect information automatically. One tool used to solve this problem is intelligent extraction. It is a data acquisition process based on optical character recognition.

It is in this context that I did a three-month internship in the start-up Tetrao located in Luxembourg. My missions consisted in analyzing websites and extracting the data they contain.

First, I will present the start-up, its methods and processes, insisting on its particularities, especially for the investment funds project. Then I will explain the missions carried out during this internship before showing my work.

2 Presentation of Tetrao

2.1 The start-up Tetrao

The start-up Tetrao was founded in 2014. Its goal is to develop and industrialize a technological innovation based on artificial intelligence. This type of technology allows training models being capable of automating tasks that, until now, had to be done manually. Thus, its field of activity is the automation of processes on the Internet, particularly in the business management and finance sectors.

Tetrao has more than 30 employees working in three locations in Europe. The company's head office is located in Luxembourg. It gathers the majority of the employees and a good part of the annotators, as well as the team of researchers and engineers in artificial intelligence. It is on this site that the development of AI algorithms takes place.

The other two sites are located respectively in Bras-sur-Meuse in France and Vilanova in Spain. The main activity in Spain is the maintenance of software and the development of data extraction programs for websites. The French site is mainly dedicated to annotators, whose mission is to analyze and understand all types of documents.

2.2 The process of Tetrao

Tetrao uses a very powerful process, which has been adapted over time. It includes the extraction of any data from the Internet as well as artificial intelligence models.

For the extraction algorithms, Tetrao relies on the visual aspect of the Web pages, unlike the "scrapping" method. This method is a classic extraction technique used by most of the Tetrao's competitors. It consists in analyzing the source code, which means that it does not adapt well to the changes of the website. The technique used by Tetrao, which simulates human behavior to identify, read and understand complex information from websites and documents, is much more resilient.

The artificial intelligence models are then used to process these collected pages and documents. After being trained, they are able to identify specific metadata without manual help.

As proof of the relevance of this process, Tetrao was the winner of the BNP Paribas "International Hackathon" in 2017. The startup won against 160 other fintechs, offering their solution to facilitate the opening of a business bank account using artificial intelligence. In addition, Tetrao was listed among the ten

most promising fintechs in Europe in the "The Fintech50 2018 Power List".

2.3 The development of Tetrao

This process has considerable versatility, allowing Tetrao to diversify further. Thus, the process is used and adapted for different projects, including the collection of investment funds, the collection of green bonds and the completion of Editus databases, the historical directory in Luxembourg.

Regarding the collection of investment funds, Tetrao's process provides a clear view of the market that secures the relationship between investors and the funds. The artificial intelligence models are able to reconcile and update many funds on a daily basis. All official documents on each fund are gathered to have a global view. The slightest change in content is quickly taken into account. In the case of a conflict, a warning is triggered and forces the user to manually choose which information on the fund is correct or incorrect.

The start-up allows Editus to complete and update its database, collecting names, phone numbers, email addresses of companies found on the Web.

The company has been expanding rapidly for a few months now, the team is growing more and more. One of the main reasons is the arrival of the Luxembourg Stock Exchange as a shareholder of the company.

3 Context and objectives

During my internship, I mainly worked on data extraction of the investment fund industry, one of Tetrao's main activities.

As I already mentioned in the previous section, Tetrao trains an artificial intelligence model being capable of first updating the information extracted from a fund on a daily basis. Then to analyze them by detecting immediately the slightest change.

The main goal of this project is to increase the number of extracted shares, i.e. financial products gathered by a fund. Currently, Tetrao collects the data of more than 120,000 shares from different countries, such as Luxembourg, France or Germany.

However, the whole process until the realization of this model is long and includes several steps. These are, for example, website searches, implementation of extraction algorithms or annotations.

The strength of Tetrao is that the extractions do not depend on the HTML code of the website, but only on the visual representation of the pages. The technical code can be modified without having an influence on the visual of the page. By using the "scrapping" method, the extraction script has to be rewritten after each such modification whereas the method used at Tetrao still works.

Tetrao has developed the Application Programming Interface Earnestnet. This API, used in extraction scripts, allows browsing the website like a human being and building the visual representation of any type of document, whether it is a PDF document or a capture from a website.

To do this, the API only uses the technical code of a website or PDF document to extract the position and sometimes the font of all the words. Once this data is available, artificial intelligence models are used to group the words into lines, the lines into paragraphs and to detect other elements, like the links and icons.

The set of these elements is called the mask of the page, so its visual representation. The artificial intelligence models also create a virtual page of this mask allowing to annotate the different elements and train the models used as final product.

The objective of my internship is to learn to perform the website searches and implement extraction scripts. To understand these two tasks I first had to familiarize myself with the structure of investment funds as well as the form of their documents and website. Then I had to learn the basics of the Scala programming language and the Earnestnet API used for the extraction algorithms.

4 Working tools

The goal of expanding share coverage requires good cooperation from all project members. In order to guarantee this, Tetrao uses various working tools.

Firstly, meetings are regularly organized. They often concern a specific theme. In this case, there are one or two leaders who are responsible to explain the topic to the other members of the group. Examples of topics are the structure of investment funds, the functionalities of Gitlab or the steps of extraction scripts. During my internship I attended about ten such meetings.

We also add short meetings, several times a week, to present the last progress. Each member has to explain the work he is doing, talk about the problems he has encountered and propose solutions. This allows the team leader to have an overview of the tasks already done and to be done.

Secondly, we use the version management tool Gitlab, allowing to well organize the code as well as keeping an history on all the modifications. All the files, i.e. the extraction scripts, are put in a single repository to which the whole team has access. Gitlab's features, such as "Issues" and "Milestones", allow to have a view on the tasks that have to be done and to easily distribute them among the members. The concrete organization of the project using Gitlab will be explained later in the report.

Third, each member must send weekly reports to the company manager. These reports should summarize all the tasks carried out during the week, as well as the problems and difficulties faced. For the latter, it must be explained what efforts were made to find a solution and if it was successful.

To communicate with other team members, the "Telegram" application is used to send messages to a single person or a whole group. This tool is mainly used to chat with employees in Spain.

5 The investment funds

To better understand the project on collecting the data of investment funds, I will first explain their structure.

Investment funds are "Collective Investment Undertakings", abbreviated as "CIU". Investment funds are public or private companies whose objective is to gather the savings of many investors in the same pot in order to make them grow.

Funds use the savings, called assets, that are entrusted to them in different ways, to make a profit. The assets are used to buy either shares, bonds or other financial products. While the main objective of the fund is of course to make a profit, many funds also aim to support innovative or thematic projects, such as ecological ones.

The funds allow clients to benefit from professional advice and to have access to a large choice of investment products. Thus, the risk of losing money is minimized and the chance of gain is increased.

The fund manager is called an asset manager. His job is to manage the money that investors have invested in the fund and to make it profitable.

An investment fund is composed of several levels. It can have several sub-funds. Each sub-fund may comprise several shares. The shares of the same sub-fund can be distinguished by their currencies, their investor categories, their income treatment or their risk and return characteristics.

Let's take the example of the Irish management company "IFSL International Limited" [7]:

ICAV Fund of Funds

Marlborough Adventurous					
Marlborough Balanced					
Marlborough Cautious					
Marlborough Defensive					
Name	Sedol	Nav	Valuation Date/Point	More Info	
Marlborough Defensive Class X Acc EUR (Hedged)	BHNDWB9	10803	13/07/2021 - 12.00	More Details →	
Marlborough Defensive Class X Acc GBP (Hedged)	BHNDW64	10421	13/07/2021 - 12.00	More Details →	
Marlborough Defensive Class X Acc USD (Hedged)	BHNDWCO	10922	13/07/2021 - 12.00	More Details →	
Marlborough Defensive Class Y Acc EUR (Hedged)	BHNDWF3	-	-	More Details →	
Marlborough Defensive Class Y Acc GBP (Hedged)	BHNDWD1	10594	13/07/2021 - 12.00	More Details →	
Marlborough Defensive Class Y Acc USD (Hedged)	BHNDWG4	-	-	More Details →	
Marlborough Defensive Class Z Acc EUR (Hedged)	BHNDWJ7	10898	13/07/2021 - 12.00	More Details →	
Marlborough Defensive Class Z Acc GBP (Hedged)	BHNDWH5	10594	13/07/2021 - 12.00	More Details →	
Marlborough Defensive Class Z Acc USD (Hedged)	BHNDWK8	11016	13/07/2021 - 12.00	More Details →	

Figure 1: Sub-funds of the company "IFSL International Limited"

This investment fund is composed of four sub-funds: "Marlborough Adventurous", "Marlborough Balanced", "Marlborough Cautious" and "Marlborough Defensive". This last sub-fund includes several shares, differentiated by their currency and investor category.

Depending on the investor's profile, an investment fund may contain more or less sub-funds and shares. The profile is characterized by the country, the language used on the website and the type of investor. The types are differentiated into "institutional", "retail" and "professional". At Tetrao, the investor profile is given in the form :

country - type - language

Each share is identified by a code, called the International Securities Identification Number, abbreviated as ISIN. It is a sequence of twelve letters or numbers, the first two letters indicate the country code.

Let's consider the first share "Marlborough Defensive Class X Acc EUR (Hedged)" in our example. The ISIN code is given by :



Figure 2: ISIN code of "Marlborough Defensive Class X Acc EUR (Hedged)"

Information about the individual shares is, in most cases, published on the asset manager's website. This documentation provides information on the objectives and strategy of the shares, their performance and an assessment of the risk. The documentation is supported by PDF files. These are standardized documents that provide additional and more detailed information.

For the "Marlborough Defensive Class X Acc EUR (Hedged)" share, the asset manager's website publishes the following documents :

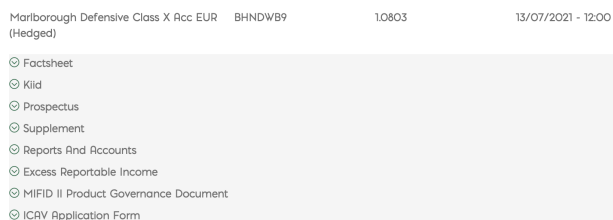


Figure 3: Documents of "Marlborough Defensive Class X Acc EUR (Hedged)"

The most important documents are the KIID, the prospectus, the factsheet and the management reports. The KIID, Key Investor Information Document, is a document standardized at European level. It usually consists of two pages containing all the information the investor needs. The prospectus is the reference document for an investment fund. It contains, often on hundreds of pages, all the information in detail. The factsheet provides additional information to the KIID, such as data on the performance of each share. This document is regularly updated by the investment funds, up to once a week.

When extracting data from the investment fund industry, Tetrao tries to retrieve the ISIN code for each share, as well as the maximum amount of information from the asset manager's web page and the PDF files.

6 The process to the final product

In this section I will briefly explain all the steps necessary to achieve the final product.

Firstly, all web pages and documents containing important data on investment funds are extracted. Then on these extracted files numerous metadata are annotated. Finally, these annotations are used to train artificial intelligence models.

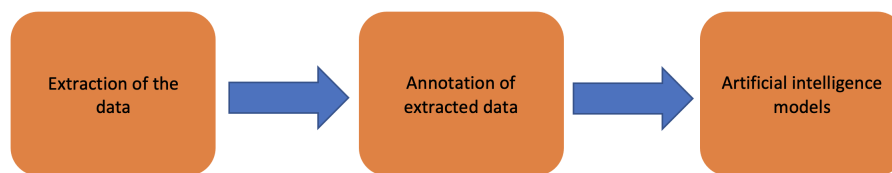


Figure 4: Process to the final product

6.1 Extraction of the data

First, a theoretical description of the website of each management company has to be written. This description is called "analysis" and is published on Gitlab. Its purpose is to give information on how to extract data.

The description is divided into two parts. The first part describes the list task. This task collects in a list the ISIN code of all the shares found on the site. The second part details the node task, which extracts the information of each share of the list.

Then, this analysis is used to implement the extraction algorithms, executing the two tasks. For these algorithms, we use the programming language Scala, as well as the Application Programming Interface Earnestnet. Scala is used by Tetrao for all its applications, mainly because of the simplicity of the code. All these extraction files are also available on Gitlab.

The result of the extraction code is visible on Stratego, the central server of Tetrao. Stratego contains, for each extraction, two workflows. These workflows are named by the name of the source, often accompanied by the investor profile used.

6 THE PROCESS TO THE FINAL PRODUCT

Name	Repository Type	Repository Path
FUNDI.LAFRANCAISE_LU_EN	extraction	/fundi/funds_extraction
FUNDI.ALLIANZ_LU_EN	extraction	/fundi/funds_extraction
FUNDI.AXA_LU_EN	extraction	/fundi/funds_extraction
FUNDI.ROBECO_LU_EN	extraction	/fundi/funds_extraction
FUNDI.PICTET_LU_EN	extraction	/fundi/funds_extraction
FUNDI.SCHROEDERS_LU_EN	extraction	/fundi/funds_extraction
FUNDI.SYCOMORE_LU_EN	extraction	/fundi/funds_extraction
FUNDI.MANDARINE_LU_EN	extraction	/fundi/funds_extraction
FUNDI.FIDELITY_LU_EN	extraction	/fundi/funds_extraction
FUNDI.HENDERSON_LU_EN	extraction	/fundi/funds_extraction
FUNDI.BLACKROCK_LU_EN	extraction	/fundi/funds_extraction

Figure 5: Some workflows visible on Stratego

By opening the information of a workflow of the first task, we can observe if we succeeded in creating the list of shares or if it is in error. In case of success, we can observe the total number of ISIN codes found. This allows us to be sure that it corresponds to the number we were expecting.

As the number of shares on a site may change over time, this task is performed once or twice a month, to constantly update the list.

For the second task, the Stratego server makes it possible to observe for which elements of the list virtual pages and documents have been collected and for which elements this has not been done. In the last case, the reason for the error is published.

Considering the workflow "FUNDI.ABN_AMRO_FR_PRO_EN". For the 200 items in the list, the node task has been done for 172 of them. The other 28 are in error :

Workflow 139

ID	139
Project ID	1
Policy ID	3
Name	FUNDI.ABN_AMRO_FR_PRO_EN
Repository Type	extraction
Repository Path	/fundi/funds_extraction
Repository Class Name	eu.tetrao.extraction.api.ExtractionRepository
Blocked	false
Disabled	false
Created At	2019-10-21 11:39:00
Updated At	2021-06-22 14:30:00

TASKS Total: 271 Enabled: 200

Executions of Group #986 [20210716171522](#)

Total	Todo	Done	Error
200	0	172	28

Figure 6: Results for workflow "FUNDI.ABN_AMRO_FR_PRO_EN"

6 THE PROCESS TO THE FINAL PRODUCT

By opening the details of the "Done" and "Error" items, we can see the ISIN code of the items in question :

Execution List (20)

Workflow ID	Task ID	Task Group Name	Exec ID	Exec State
139	1153401	LU2075255683	33245595	Done
139	549117	LU1470609113	33245567	Done
139	549098	LU1481555755	33245538	Done
139	549052	LU1165253440	33245599	Done
139	549041	LU1716254202	33245574	Done
139	549069	LU3321433091	33245608	Done
139	585774	LU1470912551	33245620	Done
139	585770	LU1490518025	33245616	Done
139	549101	LU1890764682	33245618	Done
139	585756	LU1890801662	33245563	Done

Figure 7: Part of "Done" ISINs

Execution List (20)

Workflow ID	Task ID	Task Group Name	Exec ID	Exec State
139	549191	FR0013297882	33245648	Error
139	549149	FR0015622076	33245640	Error
139	752254	LU1329548730	33245652	Error
139	585789	LU1890844336	33245651	Error
139	549189	FR0013251239	33245647	Error
139	549131	FR0010138370	33245637	Error
139	549188	FR0013219102	33245646	Error
139	549125	LU1329548144	33245583	Error
139	585768	LU1890844096	33245650	Error
139	585766	LU1890803361	33245649	Error

Figure 8: Part of "Error" ISINs

Considering an ISIN code in the "Done" section. It can be seen that the Stratego server regroups the directories containing all the screenshots taken, as well as the extracted documents :

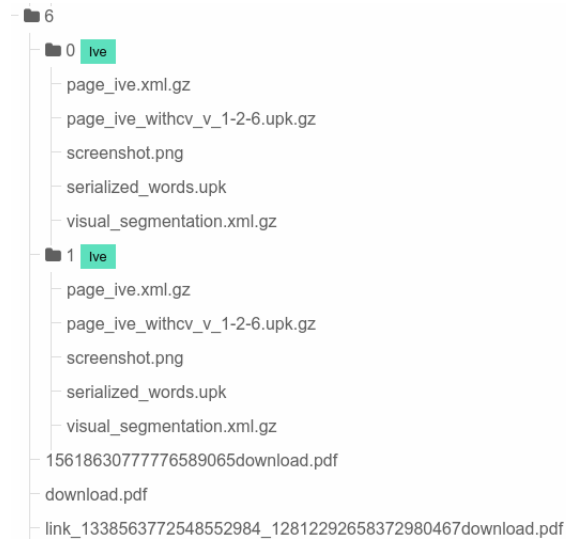


Figure 9: Part of the extracted documents for an "Done" ISIN

For an ISIN code in the "Error" section, we can observe why the code execution did not work. In our case, we failed to close one pop-up correctly :

```

Result Text:
error
java.lang.RuntimeException: Unable to handle the popup
    at eu.tetrao.extractions.fryn.abnamro.AbnAmroTask.handle_disclaimer_js$1(AbnAmroBase.scala:183)
    at eu.tetrao.extractions.fryn.abnamro.AbnAmroTask.fund_extract(AbnAmroBase.scala:208)
    at eu.tetrao.extractions.ExtractionFund.extract(ExtractionFund.scala:45)
    at eu.tetrao.extractions.ExtractionFund.extracts(ExtractionFund.scala:27)
    at eu.tetrao.extractions.fryn.abnamro.AbnAmroTask.extract(AbnAmroBase.scala:146)
    at eu.tetrao.extraction.api.ExtractionTask.run(ExtractionRepository.scala:233)
    at eu.tetrao.stratego.api.v2.packagesStgTaskRunner._run(API.scala:78)
    at eu.tetrao.stratego.api.v2.packagesStgTaskRunner._run(API.scala:84)
    at eu.tetrao.extraction.api.ExtractionTask._run(ExtractionRepository.scala:226)
    at java.base/jdk.internal.reflect.GeneratedMethodAccessor7.invoke(Unknown Source)
    at java.base/jdk.internal.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.base/java.lang.reflect.Method.invoke(Method.java:566)
    at eu.tetrao.stratego.client.services.repository.creators.FolderRepositoryCreatorStgTaskRunnerFacade.run(FolderRepositoryCreator.scala:193)
    at eu.tetrao.stratego.client.services.TaskRunners.run(TaskRunner.scala:52)
    at eu.tetrao.stratego.client.actors.ClientVirtualActor.sanonfunrun_task$2(ClientVirtualActor.scala:180)
    at scala.util.Try$.apply(Try.scala:213)
    at eu.tetrao.stratego.client.actors.ClientVirtualActor.sanonfunfuture_with_timeout$2(ClientVirtualActor.scala:299)
    at scala.concurrent.Future$.sanonfunapply$1(Future.scala:659)
    at scala.util.Success$.sanonfunmap$1(Try.scala:255)
    at scala.util.Success.map(Try.scala:213)
    at scala.concurrent.Future$.sanonfunmap$1(Future.scala:282)
    at scala.concurrent.impl.Promise$.liftedTree$1$1(Promise.scala:33)
    at scala.concurrent.impl.Promise$.sanonfuntransforms$1(Promise.scala:33)
    at scala.concurrent.impl.CallbackRunnable.run(Promise.scala:64)
    at java.base/java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1128)

```

Figure 10: Result of an "Error" ISIN

All these errors are analyzed regularly, in order to correct the scripts and improve the extraction code. If an error is detected then an issue is created on Gitlab, detailing the problem. A developer can then take care of it. Due to the variability of the extracted data, this task is performed every day.

After checking the result on the Stratego server, the extracted data is sent to the Cognita platform to be annotated.

6.2 Annotation of extracted data

Annotation is a very important step for the proper functioning of machine learning models.

To make these annotations, we use the Cognita platform. This platform groups all the data extracted from a management company into different corpuses. A corpus is a directory that gathers a specific type of documents. In total, we distinguish between nine corpuses, for example, one corpus for prospectuses and one for KIIDs.

6 THE PROCESS TO THE FINAL PRODUCT

The corpuses for the company "ABN_AMRO" and for the investor profile fr-pro-en are given by :

ABN_AMRO_FR_PRO_EN - Brochure - pro
ABN_AMRO_FR_PRO_EN - DICI - pro
ABN_AMRO_FR_PRO_EN - Documents bruts - pro
ABN_AMRO_FR_PRO_EN - Prospectus - pro
ABN_AMRO_FR_PRO_EN - Supplement Prospectus
ABN_AMRO_FR_PRO_EN - Web - pro
ABN_AMRO_FR_PRO_EN - brochure_hebdo
ABN_AMRO_FR_PRO_EN - presentation_esg
ABN_AMRO_FR_PRO_EN - rapport_esg

Figure 11: The different corpuses

For each corpus separate tasks are defined. The tasks indicate which information should be annotated. We distinguish between a good hundred of these tasks. Each one is named by the document used and the information that interests us.

Examples of tasks are DICI - Sub-fund - Name or DICI - Share - Currency. For these, the name of the sub-fund and the currency of the share found on the DICI must be annotated respectively.

On Cognita, the annotators can directly make a rectangle around the searched data, before validating and going to the next task.

To give an example of an annotation, consider the task DICI - Sub-fund - Name. The annotator puts a rectangle on the searched data, so on the name of the sub-fund :

Informations clés pour l'investisseur  LA FINANCIÈRE DE L'ÉCHIQUIER

Ce document fournit des informations essentielles aux investisseurs de cet OPCVM. Il ne s'agit pas d'un document promotionnel. Les informations qu'il contient vous sont fournies conformément à une obligation légale, afin de vous aider à comprendre en quoi consiste un investissement dans ce fonds et quels risques y sont associés. Il vous est conseillé de le lire pour décider en connaissance de cause d'investir ou non.

ECHIQUIER ARTIFICIAL INTELLIGENCE - Action B (ISIN : LU1819480192)
Compartiment de la SICAV Echiquier Fund gérée par La Financière de l'Echiquier

Objectifs et politique d'investissement

Echiquier Artificial Intelligence est un compartiment dynamique recherchant la performance à long terme à travers l'exposition sur des valeurs de croissance des marchés internationaux. En particulier, le compartiment cherche à investir dans des valeurs qui développent l'Intelligence Artificielle et/ou des valeurs qui en bénéficient.

sur les notations proposées par les agences. Les titres obligataires concernés sont des titres réputés « Investment grade », à savoir notes au minimum BBB- par Standard & Poor's ou équivalent ou considérés comme tels par l'équipe de gestion.

Les instruments financiers à terme, négociés ou non sur des marchés

Figure 12: Annotation of the name of the sub-fund

It also happens that the searched data is not found on the document. In this case, the annotator must invalidate the task before sending it.

All these sent tasks, called "samples", are the basis of the artificial intelligence models. It is very important that the tasks are done correctly, otherwise the model will not be able to recognize the data.

6.3 Artificial intelligence models

The annotations are used to train an artificial intelligence model. They have two objectives: first, to train the models, to give training data to the models which will be able to extract these data much faster. Secondly, to find the data that have not been found by the existing models.

The artificial intelligence models learn little by little until they find 90% of the data automatically. They become functional after a few thousand annotations have been saved.

Tetrao uses two types of artificial intelligence models, generic models and specific models.

Generic models are associated to a specific task. They are used to find a single data in a set of documents. This data can for example represent the name of the share or the asset manager. These models, being deployed on the documents of several management companies, are very complex. They only work to extract data from documents that are similar regardless of the source, thus for standardized documents, such as the KIID.

For all other data coming directly from a web page, Tetrao uses specific models. Therefore, Tetrao has a specific model for each management company website that is extracted. As a result, Tetrao has more than 1000 of these in operation.

The main goal of Tetrao is to deploy perfectly accurate models.

The different steps, I worked on during my internship, are explained in detail in the following sections.

7 The two main tasks

In order to extract the investment fund data, two types of tasks must be performed. The list task and the node task.

First, we need to perform the list task. The goal is to find all the shares that are on the website of an asset manager and to store them in a CSV file. As the shares are identified by their ISIN code, the list contains, for each share, this ISIN code on one line, as well as the URL of the page on which the information of the concerned share is published. In some cases, to facilitate the extraction code, additional details are added, such as the name of the sub-fund or the name of the share. The data from this list is then used in the node task.

A part of the list of our example "IFSL International Limited" [7] from the previous part is given by :

isin	url	compartment_name	share_name
IE00BMT7HC60	https://ireland.marlbroughfunds.com/funds/IFSL%20International/	Marlbrough Balanced	Marlbrough Balanced Class V Acc GBP
IE00BN6HJB94	https://ireland.marlbroughfunds.com/funds/IFSL%20International/	Marlbrough Cautious	Marlbrough Cautious Class W Acc USD
IE00BHNDWK81	https://ireland.marlbroughfunds.com/funds/IFSL%20International/	Marlbrough Defensive	Marlbrough Defensive Class Z Acc USD
IE00BMT7HH30	https://ireland.marlbroughfunds.com/funds/IFSL%20International/	Marlbrough Cautious	Marlbrough Cautious Class V Acc USD
IE00BMT7HH16	https://ireland.marlbroughfunds.com/funds/IFSL%20International/	Marlbrough Cautious	Marlbrough Cautious Class V Acc GBP
IE00BHNDVW39	https://ireland.marlbroughfunds.com/funds/IFSL%20International/	Marlbrough Adventurous	Marlbrough Adventurous Class Z Acc USD
IE00BHNDWJ76	https://ireland.marlbroughfunds.com/funds/IFSL%20International/	Marlbrough Defensive	Marlbrough Defensive Class Z Acc EUR
IE00BHNDWL98	https://ireland.marlbroughfunds.com/funds/IFSL%20International/	Marlbrough Balanced	Marlbrough Balanced Class X Acc GBP

Figure 13: Part of the list of "IFSL International Limited"

The node task is performed for each line of the list. Its purpose is to collect all the information published for each share present in the list. The data collected during the list task is used to access the specific page dedicated to the share. This page is scrolled through, in order to take screenshots and to download all the important PDF files. The extraction algorithms scroll through the pages and don't use one big viewpoint. This avoids being detected by the website and getting blocked.

By running the node task locally for the "Marlbrough Defensive Class X Acc EUR (Hedged)" share, we get the directory :

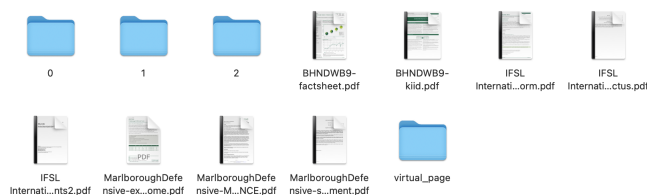


Figure 14: Node task for "Marlbrough Defensive Class X Acc EUR (Hedged)"

7 THE TWO MAIN TASKS

In the sub-directories "0", "1" and "2" are located the different screenshots of the website which are concatenated to form a virtual page. During this concatenation, artificial intelligence models are used to remove repeating paragraphs. The virtual page has therefore the same aspect as the web page and corresponds to the way a human perceives the page.

The virtual page of this example is given by :

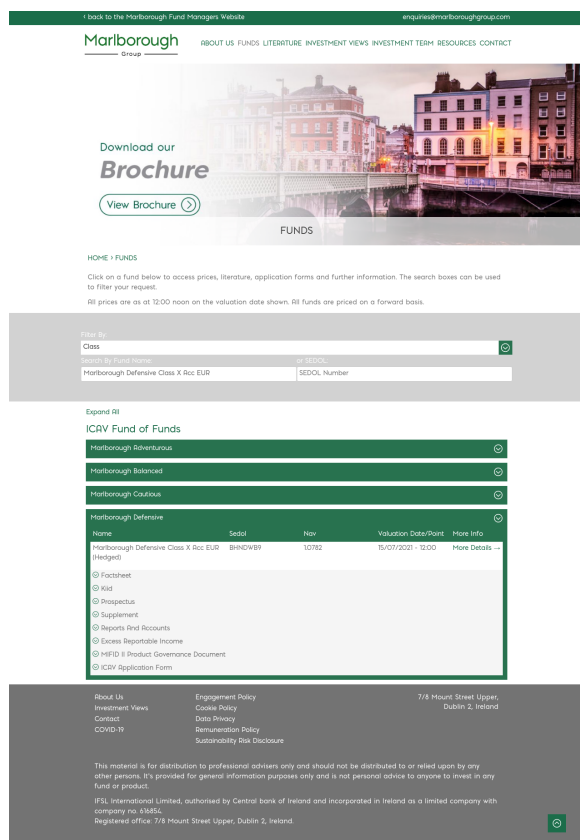


Figure 15: The virtual page

The purpose of this virtual page is mainly to simplify the work of annotators. It is easier for a human being to annotate a document without repeating paragraphs.

The realization of these two tasks is explained in detail in the section "Extraction algorithms".

8 The analyses

The analyses allow the theoretical description of the two tasks. They detail a precise methodology to be carried out by the extraction algorithm. Thus, the developer only has to apply the different steps explained in the analysis.

The process of analysis contains different steps. First, it is necessary to search the web browser to find appropriate sources. Then, these sources must be analyzed by writing a detailed description. Finally, in the case of lack of information found for a source, there is a verification step, to analyze the reason.

8.1 Search for issuers

The analysis process begins with the search for issuers. Data sources, i.e. websites that provide access to the information of the shares, must be found.

Tetrao uses Excel files to know for which management companies it does not yet extract any or a minority of shares. Tetrao gets these files, for example, from the Luxembourg Stock Exchange.

The files are separated by country. They show for the management companies the number of ISIN codes grouped in the funds and the number, percentage of ISIN codes that are collected by Tetrao.

Let's take the example of the file dedicated to Germany :

Management Company	Nbr ISINs	Coverage Tetrao	Coverage Tetrao previous period	Change	Missing	% Coverage
Helaba Invest Kapitalanlagegesellschaft mbH	412	0			-412	0%
Union Investment Institutional GmbH	404	0			-404	0%
BayernInvest Kapitalverwaltungsgesellschaft	175	0			-175	0%
Société Générale Securities Services GmbH	141	0			-141	0%
Warburg Invest AG	109	0			-109	0%
Amundi Deutschland GmbH	83	0			-83	0%
Siemens Fonds Invest GmbH	33	0			-33	0%
Union Investment Institutional Property GmbH	27	0			-27	0%
First Private Investment Management KAG mbH	21	0			-21	0%
Alte Leipziger Trust Investment-Gesellschaft mbH	19	0			-19	0%
Lazard Asset Management (Deutschland) GmbH	19	0			-19	0%
Monega Kapitalanlagegesellschaft mbH	105	1			-104	1%
IntReal International Real Estate Kapitalverwaltungsgesellschaft mbH	91	1			-90	1%
WestInvest Gesellschaft für Investmentfonds mbH	7	1			-6	14%
HAUJCK & AUFHÄUSER FUND SERVICES S.A.	7	1			-6	14%
LRI INVEST S.A.	6	1			-5	17%
MEAG MUNICH ERGO Kapitalanlagegesellschaft mbH	139	2			-137	1%
LYXOR FUNDS SOLUTIONS S.A.	19	2			-17	11%
DWS Grundbesitz GmbH	6	2			-4	33%
Union Investment Real Estate GmbH	8	3			-5	38%
IPCONCEPT (LUXEMBURG) S.A.	31	4			-27	13%
Vertas Investment GmbH	21	4			-17	19%
Deka Immobilien Investment GmbH	28	5			-23	18%
AXA Investment Managers Deutschland GmbH	37	9			-28	24%
BlackRock Asset Management Deutschland AG	38	18			-20	47%

Figure 16: File of german management companies

Each file has a second page where it is possible to visualize the ISIN codes for each management company.

For the management company "Lazard Asset Management (Deutschland) GmbH", it can be noticed, that none of the 19 ISIN codes are collected yet. These codes are given by :

ID	ISIN	Jur	Management Company
DEA0XVV	DE000A0DLNJ3	DE	Lazard Asset Management (Deutschland) GmbH
DEA16Q4	DE000A0DLNK1	DE	Lazard Asset Management (Deutschland) GmbH
DEA1AMR	DE0005647986	DE	Lazard Asset Management (Deutschland) GmbH
DEA1E1C	DE0005319016	DE	Lazard Asset Management (Deutschland) GmbH
DEA1KF6	DE000A0H1FW8	DE	Lazard Asset Management (Deutschland) GmbH
DEA1MBQ	DE000A0RHKT6	DE	Lazard Asset Management (Deutschland) GmbH
DEA1RTJ	DE000A0RHKS8	DE	Lazard Asset Management (Deutschland) GmbH
DEA1WKG	DE0006231327	DE	Lazard Asset Management (Deutschland) GmbH
DEA220Z	DE000A0RHKW0	DE	Lazard Asset Management (Deutschland) GmbH
DEA3OX0	DE0005555841	DE	Lazard Asset Management (Deutschland) GmbH
DEA5F41	DE0006231384	DE	Lazard Asset Management (Deutschland) GmbH
DEAL8V2	DE0005647846	DE	Lazard Asset Management (Deutschland) GmbH
DEAL8V4	DE0005647978	DE	Lazard Asset Management (Deutschland) GmbH
DEAL8V7	DE0005648174	DE	Lazard Asset Management (Deutschland) GmbH
DEAL8V1	DE0006231491	DE	Lazard Asset Management (Deutschland) GmbH
DEAL98T	DE0009849471	DE	Lazard Asset Management (Deutschland) GmbH
DEAL9QY	DE000A14UUB0	DE	Lazard Asset Management (Deutschland) GmbH
DEAL9R2	DE000A14UUA2	DE	Lazard Asset Management (Deutschland) GmbH
DEALUB8	DE000A14UUC8	DE	Lazard Asset Management (Deutschland) GmbH

Figure 17: ISIN codes for "Lazard Asset Management (Deutschland) GmbH"

Often management companies have their own website listing the details of the different shares. However, there are also management companies that do not publish this information on their own website or others that do not have a specific website. In this case, these ISIN codes are tried to be found on other websites, which is not always possible.

8.2 Writing the analysis

This step concerns the management companies not yet processed. After the search for an issuer, a detailed analysis of this website must be made. The objective is to make a theoretical description of the two tasks as well as all the obstacles that can be encountered on the site, such as the pop-up to accept cookies and the choice of the investor profile.

Each analysis is published on Gitlab as an issue with a specific template. First, general information about the management company is given, as well as the website used. It includes the name, the code, the URL, the chosen investor

profile and the number of shares that should be extracted.

Next, we detail how to close the cookie pop-up and how to choose the investor profile we want to use. We continue with the description of the two tasks.

Finally, we describe where on the site are documents that may be interesting to download but are not obligatory in the set of information extracted for a share. An example of such a document is the ESG presentation sheet of the management company, explaining the environmental, social and governance factors.

8.3 Verification

This step concerns the management companies for which only a part of the shares is extracted. This is a more complex process, because the lack of shares can have several reasons.

Either the extraction algorithm for the source in question is badly written. The program then does not collect all the ISIN codes that are on the source's website. As the expected number of shares is indicated during the analysis, it can easily be compared with the number obtained for the list task on Stratego.

Either not enough investor profiles are used. Depending on the investor profile chosen, the website will not necessarily propose the same number of shares. It is very important to check if all the different shares are available by browsing the main profiles.

To give an example, let's take the management company "LRI INVEST S.A.". Using the Excel file from the German sources, it can be seen that only one of the six ISIN codes are extracted. Based on the analysis already carried out for this source, it can be noticed that only the lu-retail-de profile is considered. By looking at other profiles, with Germany as country, the missing shares are present on the website. These new profiles must therefore be added to the analysis.

Either there is no website publishing information on the ISIN code in question. It is therefore impossible to extract this data.

8.4 The interest of the analysis

Performing these analyses is important for several reasons.

First, this description allows to know the different steps performed by the extraction algorithm without having to understand the code. This is especially

advantageous for beginners, who can look at the code while knowing what the different functions do, which makes it easier to understand them.

Secondly, the person in charge of writing the extraction program does not have to identify the obstacles on the site, nor the location of the data, which saves a lot of time.

Thirdly, in the case where extraction errors are due to changes in page structure, these changes are easily noticeable.

8.5 Examples of analyses

The structures of the web pages from one management company to another can differ greatly. Depending on the different cases, the analysis contains more or less information.

Let's first consider, the ideal structure of a website to perform an analysis. In this case, the site displays a table listing all the shares of the management company, with a link for each share to a specific page that contains only the details of the share in question. The "one page per share" principle facilitates the extraction algorithms.

Let's take the example of the management company "Lazard Asset Management (Deutschland) GmbH" [8]. After searching in the web browser, I found a site that publishes all the shares of the Excel file and others that are not listed in the file. As the goal at Tetrao is to extract data from as many ISINs as possible, we collect them all.

The first window that appears when opening the site is the choice of the investor profile :

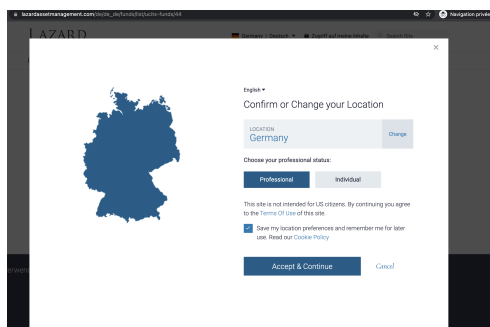


Figure 18: Pop-up : Choice of the investor profile

It is necessary to select the country, as well as the type of profile. After analyzing the number of shares published on the site for the two types of profile, I noticed that the profile "professional" should be chosen to recover a maximum of shares. Since this is the default investor profile, it is only necessary to close this window.

After closing it, there is a cookie pop-up. It is important to analyze how to close it. In this case, we simply need to click on "Accept Cookies".

After closing the pop-up, the table containing the ISIN code for each share appears directly. The URL to the specific pages of the shares is displayed, as well as recoverable, by putting the cursor on the name of the funds.

By moving the cursor over the name "Lazard Developing Markets Equity Fund", its URL appears at the bottom of the page :

Fondsname	Anleihekategorie - Typ	Assetklasse	Auflegung	List. Jahr	Performance (%)						Stand
					1J	3J	5J	10J	S.A.		
Lazard Actions Euro	S - FR0013300035	Aktien	02-Jan-2018	-	-	-	-	-	-	-	-
Lazard Alpha Euro	I - FR0010828913	Aktien	11-Mai-2005	-	-	-	-	-	-	-	-
Lazard Alpha Europe	R - FR0011034131	Aktien	29-Mrz-2012	-	-	-	-	-	-	-	-
Lazard Developing Markets Equity Fund	A Acc USD - IE00B4948049	Aktien	27-Okt-2010	2.58	37.21	11.97	13.66	2.63	2.53	30-Jun-2021	
Lazard Emerging Markets Core Equity Fund	A Dist USD - IE00B91N5K51	Aktien	07-Aug-2014	3.01	35.43	7.66	-	-	5.49	30-Jun-2021	
Lazard Emerging Markets Equity Advantage Fund	A Acc USD - IE00B2159905	Aktien	13-Okt-2020	-	-	-	-	-	-	-	
Lazard Emerging Markets Equity Fund	A Dist USD - IE00B1L6AF22	Aktien	23-Mrz-2007	9.99	40.02	6.15	7.46	2.00	3.79	30-Jun-2021	
Lazard Emerging Markets Managed Volatility Fund	A Dist EUR Hedged - IE00BLP6C60	Aktien	-	-	-	-	-	-	-	-	
Lazard Emerging World Fund	B Dist USD - IE0005022946	Aktien	25-Nov-1996	7.91	45.97	13.17	13.13	4.56	6.44	30-Jun-2021	
Lazard Equity SRI	PC EUR - FR0000003918	Aktien	28-Sep-2016	-	-	-	-	-	-	-	
Lazard European Equity Fund	B Dist EUR - IE0005060367	Aktien	11-Jul-1996	12.98	27.53	8.54	9.10	8.15	7.63	30-Jun-2021	
Lazard European MicroCap	Accumulation EUR - DE000A0H1FW8	Aktien	01-Jun-2006	-	-	-	-	-	-	-	
Lazard Global Equity Franchise Fund	A Acc USD - IE00BYR8PK92	Aktien	30-Jun-2015	16.11	47.71	9.23	11.64	-	10.82	30-Jun-2021	
Lazard Global Listed Infrastructure Equity Fund	A Dist EUR Hedged - IE00B4552M33	Aktien	18-Jun-2013	7.99	11.06	4.10	7.05	-	9.50	30-Jun-2021	

Figure 19: URL of "Lazard Developing Markets Equity Fund"

In order to perform the first task, we need to add all the ISIN codes and the URLs to the list.


For the node task, we access the URL of the ISINs. Using the virtual page we can observe that we will have to go through several tabs, like "Jahresrenditen" or "Renditen im Kalenderjahr", to collect all the information. In addition, we need to download all the standard documents listed in the "The investment funds" section.

Lazard Developing Markets Equity Fund

Der Lazard Developing Markets Equity Fund versucht, über einen vollständigen Marktzyklus starke relative Erträge zu erwirtschaften, indem er in Unternehmen mit nachhaltigen Gewinnwachstum zu attraktiven Bewertungen investiert. Der Fonds investiert in der Regel in Aktienwerte von Unternehmen, die sich in Ländern befinden, die im MSCI Emerging Markets Index enthalten sind mit einer Marktkapitalisierung von über USD 300 Mio. und einer ausreichenden Liquidität.

Morningstar Kategorie[®]
Global Emerging Markets Equity

Morningstar Style[™]



ECM Results Letter (English) / (French) / (German) / (Italian) / (Spanish)
Prospectus Documentation (English) / (French) / (German) / (Italian) / (Spanish)
The EA Share Classes for the Lazard Global Equity Franchise Fund are closed to all investors with effect from 31 March 2021.

16-Feb-2020

Informationen per Anteilklasse

A Acc USD - IE00B4W4B049

NAV (US\$) 12,5792

% Change -2,12

US\$ Change -0,2727

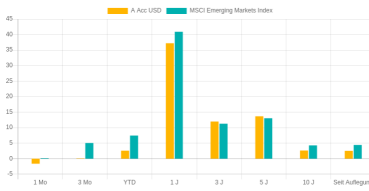
[Historisches NAV](#)
Datenstand 08-Jul-2021

<p>Primärer Vergleichsindex MSCI Emerging Markets Index</p> <p>Ausschüttungsdatum April und Oktober</p> <p>Mindestanlage US\$ 250.000</p> <p>Verwaltungsvergütung p.a. (%) 1,00%</p> <p>ISIN IE00B4W4B049</p>	<p>Verwaltetes Vermögen € 116,9 million</p> <p>Auflegungsdatum 27-Okt-2010</p> <p>Maximaler Ausgabeaufschlag 3,00%</p> <p>Ticker LZDMUA ID</p> <p>WKN A1JAXR</p>
--	---

* AUM Datenstand : 30-Jun-2021

Performance-Rückblick

Jahresrenditen
 Renditen im Kalenderjahr
 Statistiken



Type	1 Mo	3 Mo	YTD	1 J	3 J	5 J	10 J	Seit Auflegung
A Acc USD	-1,62	0,11	2,58	37,21	11,97	13,66	2,63	2,53
MSCI Emerging Markets Index	0,17	5,05	7,45	40,90	11,27	13,03	4,28	4,42

30-Jun-2021 | Alle Werte, sofern nicht anders ausgewiesen, sind annualisiert und in US\$ | Performance Inception: 26-Okt-2010
Die Performance wird net after Gebühren (und Steuern) dargestellt.

Allokation

Sektor
 Geografie
 Marktkapitalisierung

MSCI Emerging Markets			
Sektor	Lazard (%)	MSCI Emerging Markets	Übergewicht/ Untergewicht
Financials	24,71	17,78	+6,93
Information Technology	24,60	20,42	+4,18
Consumer Discretionary	16,54	17,58	-1,04
Communication Services	13,60	11,26	+2,34
Industrials	10,18	4,89	+5,29
Materials	4,65	8,42	-3,77
Energy	2,18	5,03	-2,85
Consumer Staples	0,93	5,62	-4,69
Health Care	0,62	5,04	-4,42
Utilities	0,61	1,95	-1,34
Real Estate	0,00	2,01	-2,01
Cash	1,37	0,00	+1,37

30-Jun-2021 | Alle Werte, sofern nicht anders ausgewiesen, in US\$. Allokationen basieren auf einem repräsentativen Portfolio, das die geplante Investition für ein vollständiges diskretionäres Portfolio darstellt. Allokationen sowie Titelnamen sind Änderungen vorbehalten.

Fact Sheet

Verwenden Sie das Dropdown-Menü, um weitere Anteilsklassen mit Fact Sheet anzuzeigen.

Frühere Fact Sheets


A Acc USD - IE00B4W4B049

This video is either unavailable or not supported in this browser

Error Code:


OK

Manager Overview



Outlook on Emerging Markets

Our emerging markets equity reports expect a stronger recovery in emerging markets in the second half of the year. They looked for increasing capital expenditures, particularly in developed markets, to be a boon to emerging markets equities and also...



Emerging Markets Monitor

What's happening in the developing world? View our EM Monitor to stay up to date on all things emerging markets.

Steuerdaten

Fondsreporting

Wie Sie investieren können

Um in den Fonds zu investieren, sollten Sie folgende Dokumente heruntergeladen und aufmerksam lesen.

- [KIID Dokument](#)
- [Verkaufsprospekt](#)
- [SFDR Addendum to the Prospectus](#)
- [Manager's SFDR Disclosure](#)
- [Prospektergänzung \(Deutschland\)](#)
- [Fondsaufstellung](#)
- [Prospektergänzung](#)
- [Jahresbericht](#)
- [Halbjahresbericht](#)
- [Gründungsurkunde](#)
- [Satzung](#)
- [Ordermodalitäten](#)

Kontakt

Kontaktieren Sie noch heute Ihren persönlichen Experten

Figure 20: Virtual page of "Lazard Developing Markets Equity Fund"

This analysis must be described using the specific template. The analysis of this source has the following form :

- Name of the management company : Lazard Asset Management
- Code : lazardassetmanagement
- Profile : de-inst-en
- URL : https://www.lazardassetmanagement.com/de/en_uk/funds/list/ucits-funds/44
- Language : en
- Number of shares : 176

Cookies : you will need to close the "Our Cookie Usage" cookie by clicking on "Accept Cookies".

Identification of the investor profile : you will have to close the pop-up "Confirm or Change your Location" by clicking on "Accept & Continue".

List of shares : We have a list of funds. A line can contain more than one ISIN which are in a sub-list that appears by clicking on the triangle next to the ISIN. To recover the URL, it will be necessary to make the cursor on the name of the fund. Click on the triangle and choose all the other ISINs in this sub-list.

Share data collection :

- you will have to go through the different sections of the page (in German) of the share (Performance-Rückblick, Allokation, ...). These sections are not available for all shares.
- you will have to download all the documents in pdf format that are in the rectangle "Wie Sie investieren können". (KIID, Verkaufsprospekt, Jahresbericht, Halbjahresbericht, ...).

Optional documents : go to the "Sustainable Investing" tab and then "Our Approach" and download all the documents in pdf format under "Policy Document Quick Links" (Annual Sustainable Investment Report, Our ESG Policy, Our Global Governance Principles, etc).

Figure 21: Analyse of "Lazard Asset Management"

Let's then consider the source web pages that do not have this structure. Not having a specific page per ISIN code complicates data extraction. In these cases, different shares of the same sub-fund, or even all sub-funds of a fund, have the same URL. Since the goal of the collection task is to extract only the information dedicated to a single ISIN code, one is forced to add additional parameters to the list, such as the name of the share or the name of the sub-fund. This allows to consider only the data of the share in question during the extraction and not all the information found on this URL. This addition of parameters should be quoted when describing the list task in the analysis.

Let's consider the web page of the management company "Corum Butler" [9]. For each sub-fund there is a specific URL where all the different shares are listed :

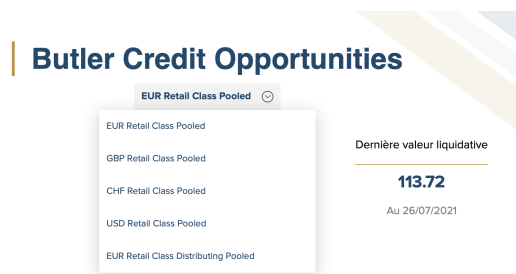


Figure 22: Website of "CorumButler"

Thus, when extracting, we need to know the name of the share to be able to click on the corresponding tab and collect only its information.

Finally, there are also web pages of management companies that only publish the standard documents of the shares and no additional information. In this case, the list task is not performed. We simply download all the PDF files present on the page without referencing them to an ISIN code. This is done manually, in the annotation step.

To give an example, we can use the website of the management company "Merrion Capital Investment Managers Limited" [10]:

MULTI ASSET			
Fund	Class A	Class B	Class C
Merrion Managed Fund	Download	Download	Download
Merrion Ethical Fund	Download	Download	Download
Multi Asset 30 Fund	Download	Download	Download
Multi Asset 50 Fund	Download	Download	Download
Merrion Balanced Fund	Download	Download	

Figure 23: Website of "Merrion Capital Investment Managers Limited"

Finding only standardized documents, it is enough to download all of them, without referencing them to a share.

During my internship I performed around 80 analyses. These analyses were mainly for German and Irish sources. At the beginning of my internship, I also made some annotations. This task especially helped me to know what information we want to extract on the shares, which facilitates writing analyses.

9 Extraction algorithms

In this chapter, I will present the extraction algorithms. To realize them, we use, at Tetrao, the visual information of the Web pages. This allows us to analyze them in the same way as a human being would. This particularity allows the algorithms to be very robust to the changes of the sites, because they do not depend on the HTML code of the website.

I will first about the tools, the organization of this project, the files and their execution before detailing the algorithms.

9.1 The tools

The extraction algorithms are written in the programming language "Scala", a universal programming language designed to allow concise and simple code. It is based on the Java language, and thus inherits its libraries and its virtual machine. To familiarize myself with Scala, I did some exercises of the "Euler" project[12].

We use the editor "IntelliJ" intended for the development of computer software based on the Java technology. This editor allows to open a project in Gitlab and to directly make specific actions, like creating a new branch or making a push of modifications.

All the tools and methods needed to perform the extractions are in the Earnestnet project. It includes its own Application Programming Interface and its own Chromium browser version. The Chromium browser version contains some modifications compared to the usual versions. When downloading a PDF document, for example, there is no window asking for the save directory.

9.2 Organization of the project

For a long time, the extractions were mainly implemented by the employees of the site in Spain. However, during my internship, I worked with two other interns on this project. Because of this expansion of the team, we had to set up a common organization for both sites to have an overview of the status of the extractions of all management companies. This organization is based on the "Scrum" method and uses Milestones on Gitlab. The project is named "funds-extraction".

First, for each management company, the analysis is done. This is put in an issue, named "analysis issue", labeled with the label "Source".

Then, "task issues" are created when there is a new extraction to be performed or a modification to be made to an extraction already implemented. The task in question is explained in the "task issue" and the "analysis issue" of the corresponding source is referenced. By "grafting" tasks to the "analysis issue", a history is kept in its comments.

The "task issues" are labeled with different labels describing the task or type of modification that needs to be performed. The most frequent labels are : "new extraction", "extraction-errors" and "new profile" respectively for a new extraction to be performed, errors in the code or a new profile to be added.

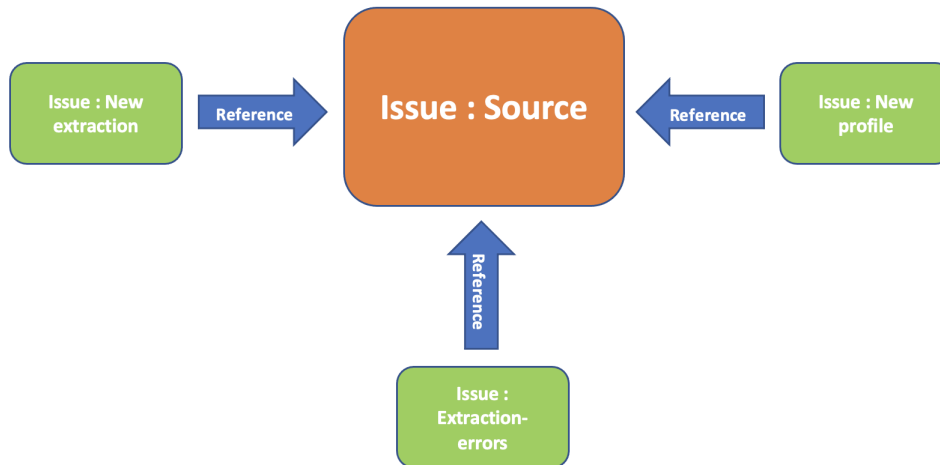


Figure 24: Organization of the Issues

Pending "task issues" are added to a Milestone, called "extraction-backlog". The issues in this Milestone are ready to be started, but are not yet scheduled or assigned to a developer.

At the beginning of each week a "Sprint" Milestone is created. All issues that are supposed to be processed during that week are added to it. This Milestone has three columns: "Unstarted issues", i.e. unassigned issues, "Ongoing issues", issues that are assigned and therefore in progress and "Completed issues", tasks that have been completed.

Below is an example of the Milestone "extraction-week-30-2021", i.e. the sprint of the thirtieth week of the year 2021 :

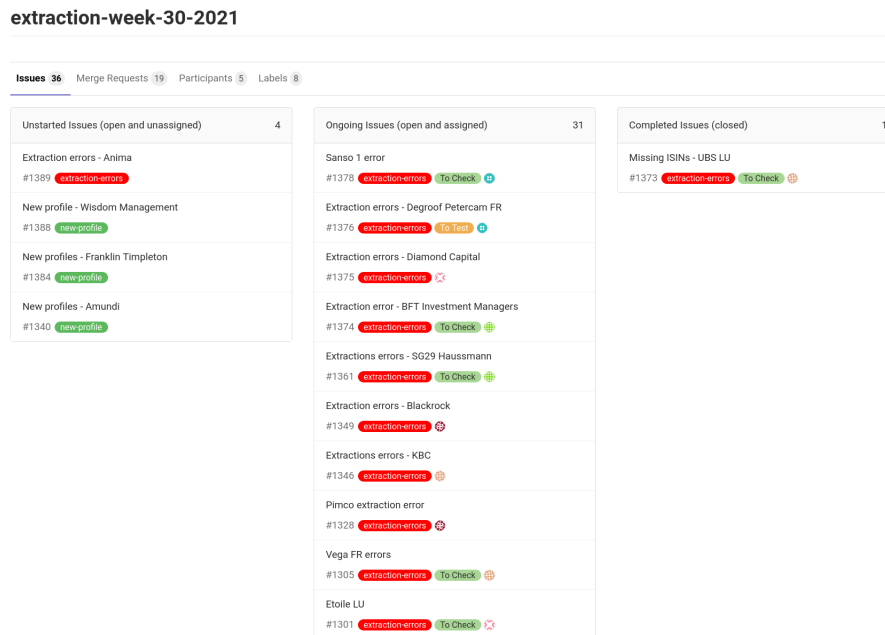


Figure 25: Organization of the "Sprint" Milestone

In an ideal case, at the end of the week, all the issues are closed. However, it is possible that some tasks are still in progress. These are put in the next "Sprint" Milestone. Very rarely, some issues are still not assigned. They are, depending on their importance in relation to the "extraction_blacklog" Milestone, put back into the latter or into the next sprint. The "Sprint" Milestone is closed once all issues are in the "Completed issues" column.

Each developer can choose task issues from the "Unstarted issues" column and assign them to himself. For each task, he creates a new branch from the "develop" branch, the pre-production branch. He works only on this branch to make the modifications and additions. Once the task is implemented, he makes a push of these modifications and creates a merge request from his branch to the "develop" branch. When this merge request is accepted, the extraction script is integrated in production.



Figure 26: Organization of the branches

When creating the merge request, the developer assigns the task issue to the integrator and add the label "to test". This label shows that the script is in the test phase. Once the integrator has tested the extraction code and accepted the merge request, he reassigns the issue to the developer and replaces the "to test" label with the "to check" label to indicate that the script is in the verification phase. The developer can then close the issue after checking the result on the Stratego server.

9.3 The extraction files

To create a new extraction file we use the template "TemplateExtraction". It automatically creates all the elements present in each script.

Firstly, it creates a trait. It is an interface defining all the methods used in the classes of the file. This file contains two classes, one for the list task and one for the node task.

Secondly, it creates an object for each class. In Scala, everything inside the object will be executed. Each object contains a variable named "task" taking as parameters the data needed to perform the main tasks. These parameters are, for example, the URL of the site, the investor profile or the ISIN code of a share. An example is given later on.

The purpose of these objects is to test and verify the results of the two classes locally. This ensures that the algorithms work properly before putting them into production.

9.4 Executing the extraction files

There are two different ways to use the Chromium browser : manual or virtual opening.

When opening manually, the browser can be seen continuously during the execution. This allows to observe the different steps and the state of the page on which the program stopped. The advantage is to better understand the reason for crashes. However, during this opening, the system will essentially focus on

the browser, which makes it impossible to work on the computer during this time, which can become long.

During the virtual opening, there is no direct view of the extraction. It can be done without the browser appearing. Thus, it is possible to work while the extraction is taking place. In addition, this virtual mode runs the program in the same way as the Stratego server. It is therefore important to check that the code is working properly using this mode.

9.5 The API Earnestnet

I will now explain the basic objects and methods used during the extraction algorithms. Due to the diversity of the structure of the management companies' web pages, the API also includes many other functions. These are only useful in specific cases and are therefore used less often. Some are shown in the sections "An example of an extraction script" and "Specific cases".

The central element of this API is the "agent". Its role is to communicate with the different elements, for example the words, links or images of the web page. As Tetrao relies on the visual aspect of the pages, the agent connects to the Chromium browser, loads a URL and essentially interacts with the visual representation of these elements. In addition, it is the agent that allows to browse the pages and to retrieve screenshots.

To better understand the roles of the agent, we will consider the web page of the management company "Lazard Asset Management (Deutschland) GmbH" [8], for which the analysis is in the section "Examples of analyses".

The agent has several specific methods. At the beginning of each extraction code, the agent has to load a URL. The method takes as a parameter a string "URL" corresponding to the URL.

```
1 var page = agent.get(URL)
```

Then, we have to give the agent the order to browse this page to get the visual representation of the elements and to make the screenshots. This order is given by the following method :

```
1 page = agent.snapshot()
```

Different page navigation modes can be used. We distinguish between the default mode and the "single page" mode. These two modes are defined by the following methods :

```
1 agent.set_default_page_mode()
2 agent.set_single_page_mode()
```

The default mode considers the whole page, it can communicate with all the elements of the page regardless of their location. This means even with those that are not in the browser window without scrolling. The agent takes screenshots of the entire page and returns the virtual page.

The "single page" mode considers only the part of the web page that is visible in the browser window. This mode is mostly used for closing cookie pop-ups or disclaimers. The button to close the pop-up is always located in the visible part of the page. When opening the web page of the management company "Lazard" in "single page" mode, the returned capture contains the whole disclaimer :

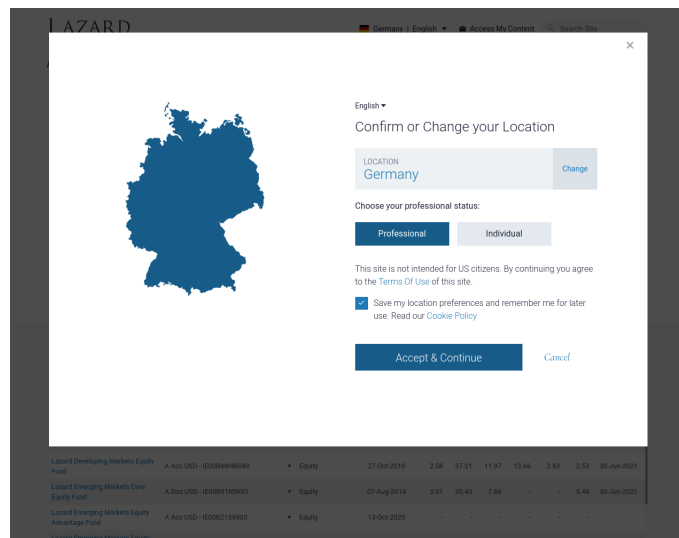


Figure 27: Pop-up of a disclaimer

Only this capture is returned by the agent and not a virtual page.

Finally, it is also the agent that allows to click on the different elements of the page. Let's note "element" the element that allows to close the pop-up disclaimer "Confirm or Change your Location", so the button "Accept & Continue". The command :

```
1 page = agent.click(element)
```

allows to click on this button and close the pop-up disclaimer.

The visual representation is stored in a variable named "page". Each time the agent redefines this variable, the page corresponds to the representation of the newly opened page. It can therefore no longer find the elements of the old representations.

To visualize all the elements of the page, we must consider its mask. The mask allows to retrieve any element present on the page.

We have a tool, named ComputerVisionManager, allowing to display the set of elements. These are framed with a defined color code, which makes it possible to identify and distinguish the different types of elements.

The command to frame all the elements of the page is :

```
1 ComputerVisionManager.plot(page)
```

The display for a part of the page is then given by the following image :

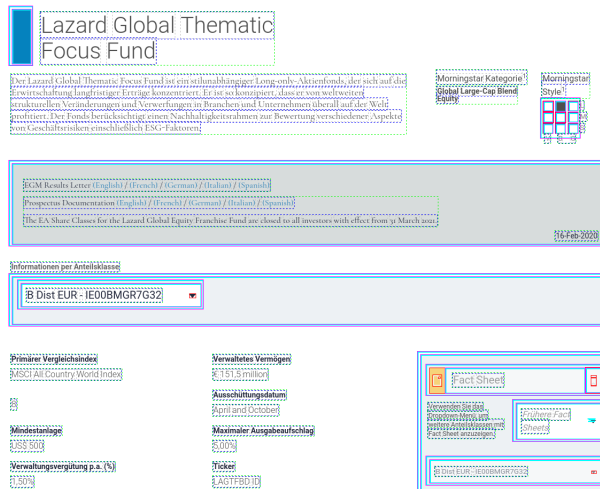


Figure 28: One part of the ComputerVisionManager of the page

Among the elements, the most often used are lines, framed in blue, words, framed in gray, icons, in red and links. However, the mask also includes selects, images and paragraphs.

We can recover a sequence containing all the elements of the same type :

```
1 val lines = page.ive_mask.lines
2 val words = page.ive_mask.words
3 val icons = page.ive_mask.icons
4 val links = page.ive_mask.links
```

An element of such a sequence contains several pieces of information, such as its position on the page or the text it contains.

Often we want to retrieve a specific element of the page. For this, the Earnestnet API has different functions. These functions allow either to specify the position of the element on the page, or to put a filter on the elements. The filtering methods differ slightly depending on the type of element.

To specify the location of the element, we can use four methods returning a single element each time :

```
1 val uppermost_line = page.ive_mask.lines.uppermost
2 val lowermost_line = page.ive_mask.lines.lowermost
3 val rightmost_line = page.ive_mask.lines.rightmost
4 val leftmost_line = page.ive_mask.lines.leftmost
```

These are the uppermost, lowermost, most at right and most at left lines of the page respectively.

These methods are often used in combination with the filtering functions. The latter allow to distinguish the elements following their text for words, lines and links, their URL for links or following their signature for icons. The three commands are respectively given by :

```
1 val filter_text = page.ive_mask.words.text_matching()
2 val filter_url = page.ive_mask.links.url_matching()
3 val filter_sig = page.ive_mask.icons.signature_matching()
```

The "text_matching" and the "url_matching" take as a parameter either a string representing the text we want to find, or a regex. A regex, also called regular expression describes, according to a precise syntax, a set of possible strings.

The "signature_matching" takes as parameter a signature of an icon. It is a long string of characters made of numbers and letters.

The ComputerVisionManager can also be used to frame a single element of the page. Thus it allows to check if, after using the position specification and

filtering methods, the right element has been found. To give an example, we look for the uppermost line on the page that contains the regular expression of an ISIN code :

```

1 val isin_reg = "[A-Z]{2}[0-9A-Z]{9}[0-9]".r
2
3 val ISIN = page.ive_mask.lines.text_matching(isin_reg).uppermost
4
5 ComputerVisionManager.plot(
6   page,
7   Map(
8     "line" -> Seq(ISIN)
9   ),
10  Map.empty[String, IVEMask],
11  "debug"
12 )

```

We frame the line found with the ComputerVisionManager :

The screenshot shows the Lazard Global Thematic Focus Fund page. The ISIN code 'B Dist EUR - IE00BMGR7G32' is highlighted with a red box in the 'Informations per Anteilklasse' section. The page includes a header with the fund name, a description of the fund, Morningstar category and style information, and various financial metrics and links.

Figure 29: One element detected with the ComputerVisionManager

In some cases, these position specialization and filtering methods are not sufficient to find the element that we want to find. This is the case when the searched word is often repeated on the page or is located in the middle. It is then possible to restrict the mask used before searching for the element.

Let's consider the following page :

The screenshot shows the Lazard website's 'UCITS Funds' page. It features a navigation menu with 'About', 'Investments', 'Funds', 'Sustainable Investing', and 'Research & Insights'. Below the navigation, there is a section titled 'UCITS Funds' with a brief introduction and a search filter for 'Asset Class'. The filter options are 'All', 'Equity', 'Fixed Income', 'Alternatives', and 'Multi-Asset'. The main content is a table of fund performance data.

Fund Name	Share Class	Asset Class	Inception Date	YTD	1 Yr	3 Yr	5 Yr	10 Yr	SI	As of Date
Lazard Actions Euro	S - FR0013300035	Equity	02-Jan-2018	-	-	-	-	-	-	-
Lazard Alpha Euro	I - FR0010828913	Equity	11-May-2005	-	-	-	-	-	-	-
Lazard Alpha Europe	R - FR0011034131	Equity	29-Mar-2012	-	-	-	-	-	-	-
Lazard Developing Markets Equity Fund	A Acc USD - IE00B4W48049	Equity	27-Oct-2010	2.58	37.21	11.97	13.66	2.63	2.53	30-Jun-2021
Lazard Emerging Markets Core Equity Fund	A Dist USD - IE00B91NSK51	Equity	07-Aug-2014	3.01	35.43	7.66	-	-	5.49	30-Jun-2021
Lazard Emerging Markets Equity Advantage Fund	A Acc USD - IE00B2159905	Equity	13-Oct-2020	-	-	-	-	-	-	-
Lazard Emerging Markets Equity										

Figure 30: A part of a list of shares

The word "Lazard" occurs frequently on this page, so it is impossible to find a particular word with the previous methods.

Suppose we want to retrieve the name of the first share in the "Fund Name" column. One method is to restrict the top and bottom of the mask to the first ISIN code found.

To do this, we set the top and bottom of the new mask to the top, respectively the bottom, of the position of this ISIN code :

```

1 val lazard_restrict = page.ive_mask.copy.restrict_to(
2   top = page.ive_mask.words.text_matching(isin_reg).uppermost.top,
3   bottom = page.ive_mask.words.text_matching(isin_reg).uppermost.bottom
4 )

```

The "restrict" method is applied directly to the indicated mask. It is important to always make a "copy" of this mask, otherwise we cannot longer recover the whole page.

We finally look for the line containing "Lazard" in this new mask :

```
1 val lazard = lazard_restrict.lines.text_matching("Lazard").leftmost
```

It is possible to restrict the mask further by also specifying the "right" or "left".

9.6 An example of an extraction script

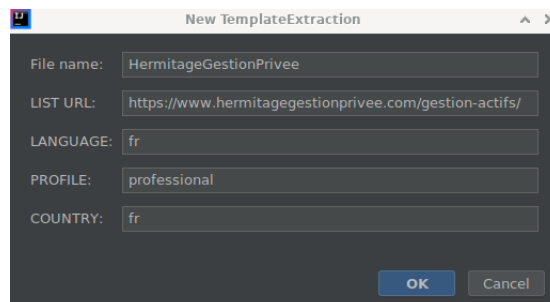
To better understand the steps of an extraction script, we will analyze them one by one based on a analysis. To do this, we will consider the management company "Hermitage Gestion Privée" [13].

First, we will create a new file using the extraction template. The first information of the analysis :

- Name of the management company : Hermitage Gestion Privée
- Code : hermitagegestionprivee
- Profile : fr-pro-fr
- URL : <https://www.hermitagegestionprivee.com/gestion-actifs/>
- Language : fr
- Number of shares : 2

Figure 31: First part of the analysis of "Hermitage Gestion Privée"

allows to know the coordinates necessary to use this template. The following fields are filled in :



The image shows a dialog box titled "New TemplateExtraction" with a dark background. It contains five text input fields, each with a label and a value:

- File name: HermitageGestionPrivee
- LIST URL: <https://www.hermitagegestionprivee.com/gestion-actifs/>
- LANGUAGE: fr
- PROFILE: professional
- COUNTRY: fr

At the bottom right, there are two buttons: "OK" and "Cancel".

Figure 32: Coordinates for the extraction template

As described in the "The extraction files" section, this template automatically fills most of the two objects used to test the list and node tasks locally.

The completely filled list object is given by :

```
object HermitageGestionPriveeListTask extends App with ExtractionTester with HermitageGestionPriveeTask {
  val task = Map(
    "url" -> Set("https://www.hermitagegestionprivee.com/gestion-actifs/"),
    "profile" -> Set("professional"),
    "country" -> Set("fr"),
    "language" -> Set("fr")
  )

  val path = Paths.get( first = "/tmp/extraction_test")
  val result: v2.StgResult = run(path, list_class_name, task, running_chrome = true)
  generateCSV_with_xml(result)
}
```

Figure 33: Object to test the list task

The last lines of these objects allow to execute the corresponding class considering the parameters of the variable "task". In addition, they allow to indicate the directory where the results will be saved and how to run the Chromium browser.

Then, opening the URL, we can see the appearance of a cookie. In the analysis it is noted :

Cookies : you have to close the cookie pop-up by clicking on "Accept".

Figure 34: Information about the pop-up cookie

The pop-up cookie appears every time we open the URL, so it must be closed several times in our extraction script. For this reason, we define a "close_cookies" function in the trait :

```
1 def close_cookies(page)(implicit agent): IVEDocument = {
2   page.ive_mask.words.text_matching("Accept").lowermost_option match {
3     case Some(tab) => agent.click(tab)
4     case None => current_page
5   }
6 }
```

This function takes as parameter the current page and returns, thanks to the "match", either the page obtained without cookies if it was present or the same page in the opposite case.

The agent is given as an implicit variable. When it is not given as a parameter to the function, Scala will implicitly give its value.

Then, it is necessary to choose the investor profile :

The screenshot shows a web form with two dropdown menus: 'Profil d'investisseur' (set to 'INVESTISSEUR PRIVÉ') and 'Domicile' (set to 'FRANCE'). Below these is a large text box containing a disclaimer in French. At the bottom of the form, there is a checkbox labeled 'Informations légales lues et acceptées' and a blue button labeled 'ACCEPTER'.

Figure 35: Choice of the investor profile

According to the analysis, the steps to be performed are :

Identification of the investor profile : In the drop-down menu " Profil d'investisseur" you must select " Investisseur professionnel ". Then click on " Informations légales lues et acceptées" and on "ACCEPTER".

Figure 36: Description of the identification of the investor profile

We create a "select_profile" function, defined in the trait :

```

1 def select_profile(page)(implicit agent): IVEDocument = {
2   val select_profile = page.ive_mask.selects
3     .find(_.text_options.contains("Investisseur prive")).get
4   val option_select_profile = select_profile.text_options
5     .indexOf("Investisseur professionnel")
6   agent.select_option(select_profile, option_select_profile)
7   val tmp = agent.click(page.ive_mask.lines
8     .text_matching("Informations legales lues").lowermost)
9   agent.click(tmp.ive_mask.words.text_matching("ACCEPTER").lowermost)
10 }

```

The Earnestnet API allows to identify the "selects" present on the page. The function starts by searching for the select in question by putting a filter on the text of the default option, lines 2-3. Then we define the index of the option we want to select, lines 4-5. Finally, we order the agent to select this option, line 6. At the end, the function clicks on the two necessary buttons to validate this choice and returns to the new page.

Now we can start with the list task. This one is implemented in the corresponding class. The analysis shows us the next steps :

List of shares : First, you will have to close the pop-up cookie and identify the investor profile. Then, in the list of the "Nos fonds communs de placement" section you will have to collect all the ISIN codes, as well as the URLs by going on the ISIN.

Figure 37: Description of the list task

```
1 agent.set_single_page_mode()
2
3 var page = agent.get(task.url)
4 page = close_cookies(page)
5
6 agent.set_default_page_mode()
7 page = agent.snapshot()
8
9 page = select_profile(page)
10
11 val isin_reg = "[A-Z]{2}[0-9A-Z]{9}[0-9]".r
12 page.live_mask.words.text_matching(isin_reg).foreach{ isin =>
13   page = agent.click(isin)
14   val share_url = page.url
15
16   val params_generated = List(
17     "group_name" -> isin.text,
18     "isin" -> isin.text,
19     "url" -> share_url,
20     "country" -> task.country.get,
21     "language" -> task.language.get,
22     "profile" -> task.profile.get,
23     "class_name" -> node_class_name
24   )
25
26   val result_xml = generate_xml_result(params_generated)
27   result.add_text("tasks", s"${result_xml}")
28
29   page = agent.go_back()
30 }
```

First, on lines 1-4, we set the "single page" mode to load the URL of our task and close the cookie. Then, on lines 6-9, we set the default mode, start with the screenshots and choose the investor profile. After defining the regular expression of an ISIN code, line 11, we look for all the words on the page that contain this regex, line 12, and for each of them we perform several tasks.

We click on the element in question, line 13 and we recover the URL of the page then opened, line 14. The following commands allow us to enter the different parameters, the text of the element, i.e. the ISIN code, as well as the URL and other information, in the list. Finally, line 29, we return to the previous page, before starting again with these steps, until we have processed all the ISIN codes found on the page.

By running the object, we obtain the csv file containing the list. We can observe that it gathers well the information of the two shares we are supposed

to find.

This information is used to complete the node object :

```
object HermitageGestionPriveeNodeTask extends App with ExtractionTester with HermitageGestionPriveeTask {  
  
  val task = Map(  
    "url" -> Set("https://www.hermitagegestionprivee.com/fonds/luxe-low-cost-leaders-i/"),  
    "isin" -> Set("FR0013313640"),  
    "profile" -> Set("professional"),  
    "country" -> Set("fr"),  
    "language" -> Set("fr")  
  )  
  
  val path = Paths.get( first = "/tmp/extraction_test")  
  
  run(path, node_class_name, task, running_chrome = true)  
}
```

Figure 38: Object to test the node task

Finally we have to implement the node task. The analysis says :

Share data collection : you will have to close the pop-up cookie and browse the page and download the documents in the "DOCUMENTS DISPONIBLES" section.

Figure 39: Description of the node task

We start by defining a method "download_documents" in the trait. Its role is to download the different documents.

Since we want to collect the documents from the "DOCUMENTS DISPONIBLES" section, we restrict the mask to this section :

```
1 val document_restrict = document_page.ive_mask.copy.restrict_to(  
2   top = document_page.ive_mask.lines  
3   .text_matching("DOCUMENTS DISPONIBLES").lowest.bottom  
4 )
```

We look for all the links that are located in this new mask. To be sure to find only the links that interest us, we add a filter on their URL, in this example we choose the string ".pdf" :

```
1 val document_links = document_restrict.links  
2   .url_matching(".pdf")  
3   .toList
```

Finally the function downloads all the links stocked in the "document_links" variable.

Then we can define the node class :

```
1 agent.set_single_page_mode()
2
3 var page = agent.get(task.url)
4 page = close_cookies(page)
5
6 agent.set_default_page_mode()
7 page = agent.snapshot()
8
9 if (page.live_mask.words.text_matching(task.isin.get).isEmpty) {
10     throw new IsinNotFoundException
11 }
12
13 check(page, "core", lines_to_check_map)
14
15 val documents = download_documents(page, task.cached_urls)
16 result.add_result_ive(page, tag = "documents",
17     ExtractionAttachment(documents))
```

The beginning of this class is done exactly the same way as for the list task. Then, there are two steps of verification.

First, lines 9-11, we check that the ISIN code of our task is present on the page we are browsing. Then, line 13, we examine the presence of different words on the page. These words have been previously defined in the "lines_to_check_map" variable.

In case of failure, the reason is indicated on Stratego for the ISIN code in question. Both checks are exceptions. This means that on Stratego the tasks are in error but execution is still maintained. This makes it possible to quickly identify small changes on the page, without losing the information on the shares.

Finally, the documents are downloaded and saved, together with the page containing the important information, in a directory. This is the folder that is sent to the Stratego server.

The last line of the terminal, after the execution, is given by :

```
1 StgResult(success, Map(), List(/tmp/extraction_test/0/mappings.xml,
2     /tmp/extraction_test/0/2), ArrayBuffer())
```

It gives the number of the directory or directories sent to the server. This allows to check if all the information has been extracted.

Each extraction script includes these steps. Depending on the Web page, the implementation of these steps can be more or less complicated.

9.7 Specific cases

In this part I will show different examples of extraction and site structure that I encountered during my internship. These examples are quite frequent and can make the algorithms more difficult.

9.7.1 Use of JavaScript

Sometimes the visual representation of the elements does not allow to perform all the necessary actions. It is then necessary to slightly modify the HTML code of the Web page. To do this, we use JavaScript.

Let's consider the website of the management company "Corum Butler" [9]. For each sub-fund all the shares it contains are on the same URL. We must therefore select the different shares to retrieve the information.

However, the agent cannot detect the unselected shares because it is not a "select". In fact, if we examine the HTML code of this element, we can see that it is a so-called hidden class :

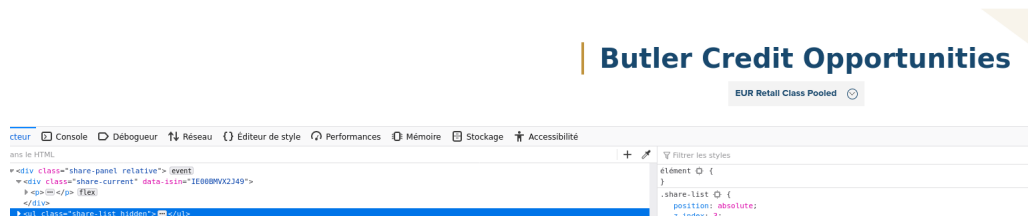


Figure 40: Example of a hidden class

This means that all elements of this class remain hidden until they are selected after dragging the cursor over the visible part. It is only by removing this "hidden" parameter that all shares can be inspected no matter where the cursor is located :

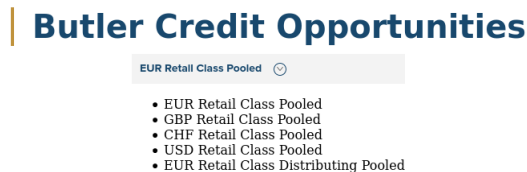


Figure 41: Removing the hidden class

In this way, the different names of the shares are detected in the visual representation of the page.

We will have to implement the command that allows us to make this change in the HTML code. To indicate to the agent that we are using JavaScript, we must use the following method of the API :

```
1 agent.exec_javascript()
```

This method takes as a parameter a string corresponding to the JavaScript command that we want to execute. In our example, we execute :

```
1 agent.exec_javascript("document
2 .getElementsByClassName('share-list hidden')[0].className='');" )
```

This command looks for the first class in the HTML code called "share-list hidden" and replaces its name to get rid of the "hidden" parameter.

9.7.2 Hidden ISIN codes

Some sites have the desired "one page per share" structure. However, in the initial list, the ISIN code is hidden, so it cannot be directly collected. An example is the website of "Bankinter" [14] :

▼ BANKINTER - MULTIFUND A (EUR)		6.37%	EUR	MEDIO	0.84%	★ ★ ★ ★ ★
MIXTO RENTA VARIABLE						
ISIN	COMISIONES ¹	COMISIONES ANUALES		MÍNIMO INICIAL		
LU1496043081	Suscripción 0,00% Reembolso 0,00%	Gestión 0,36% Depositoria 0,00% Distribución 0,48%		10.00 EUR		
VER FICHA →	VER VALORES LIQUIDATIVOS →	CONTRATAR ESTE FONDO				
> BANKINTER AHORRO ACTIVOS EURO - R		-0,37%	EUR	REDUCIDO	0.6%	NO VALORADO CONTRATAR →
> BANKINTER AHORRO RENTA FIJA FI - R		-0,08%	EUR	MUY REDUCIDO	0.55%	★ ★ ★ ★ ★ CONTRATAR →

Figure 42: Website of "Bankinter"

We have to click on the name of the share so that its ISIN code becomes visible.

To implement the list task, the idea is to first retrieve all the names of the shares, then put them in a list and finally click on them to get the necessary information. This sounds easy, but we have to be careful about the order of the names in the list.

When we click on the first name on the page, the window that appears shifts the position of the names that follow. Thus, when the agent clicks on the second name, it clicks on the initial position of it, which is no longer the line corresponding to the name. The window containing more information does not appear and the ISIN code is not recoverable.

The solution is to fix the order of the names, from the lowest name on the page to the highest. This way, after a click, the position of the following elements in the list have not changed.

The corresponding code is given by :

```

1 val shares_list = page.ive_mask.copy.restrict_to(
2   top = page.ive_mask.words.text_matching("NAME").uppermost.bottom,
3   bottom = page.ive_mask.lines.text_matching("PURCHASE").lowermost.bottom
4 ).lines.text_matching("BANKINTER").sorted_vertically.reverse.toList

```

We collect all the share names by doing a "restrict" and a "text_matching" of "Bankinter", then we sort these elements vertically and reverse the order.

9.7.3 List of shares on several pages

It happens that the list of shares is on several pages that we have to go through. Often, to go from one page to another we have to click on an arrow or the number of the next page. This is the case for the management company "Sabadell" [15]:

AMUNDI FUNDS ASIA EQUITY CONCENTRATED - R2 EUR (G) Código ISIN LU1882443130	EUR	99,48 € 6/8/2021	Art. 8
AMUNDI FUNDS ASIA EQUITY CONCENTRATED - R2 USD (G) Código ISIN LU1882443113	USD	71,56 US\$ 6/8/2021	Art. 8

(1) SFDR: Sustainable Finance Disclosure Regulation - Reglamento sobre la divulgación de información relativa a la sostenibilidad

1 2 3 4 5 ... 21 >

Resultados por página: 20

Figure 43: Website of "Sabadell"

To manage this situation, we must first recover the number of pages, in this case 21. As the number of pages can vary, we must make a "restrict" on the

lines that surround it and search in this new mask for the word that is furthest to the right :

```
1 var next_page_icon_restrict = page.ive_mask.copy.restrict_to(  
2   top = page.ive_mask.lines  
3   .text_matching("Sustainable Finance Disclosure Regulation")  
4   .lowermost.bottom+10,  
5   bottom = page.ive_mask.lines.text_matching("Resultados por")  
6   .lowermost.bottom+30,  
7   right = page.ive_mask.lines.text_matching("Resultados por")  
8   .lowermost.left,  
9   left = page.ive_mask.lines  
10  .text_matching("Sustainable Finance Disclosure Regulation")  
11  .lowermost.left  
12 )  
13 val number_pages = next_page_icon_restrict.words.rightmost.text.toInt
```

The "+10" and "+30" add ten and thirty pixels respectively to the initial position. The location on the page of pixel coordinates (0,0) is the upper left corner.

Then, we must "number_pages - 1" times click on the arrow to go through all the pages in order to extract the necessary information.

```
1 (1 to number_pages).foreach { i =>  
2  
3   next_page_icon_restrict = page.ive_mask.copy.restrict_to(  
4     top = page.ive_mask.lines  
5     .text_matching("Sustainable Finance Disclosure Regulation")  
6     .lowermost.bottom+10,  
7     bottom = page.ive_mask.lines.text_matching("Resultados por")  
8     .lowermost.bottom+10,  
9     right = page.ive_mask.lines.text_matching("Resultados por")  
10    .lowermost.left-100,  
11    left = page.ive_mask.words.text_matching("1").lowermost.left  
12  )  
13  
14  next_page_icon_restrict.icons.rightmost_option match {  
15    case Some(next_page) => page = agent.click(next_page)  
16    case None =>  
17  }  
18 }
```

The idea is again to restrict the mask and look for the rightmost icon. Finally we click on this icon.

9.7.4 Several profiles to implement

Often several profiles have to be implemented for one management company. The structure of the sites of the different profiles remains the same, however, they may use different URL or language, resulting in slight differences in the extraction.

Taking as example a source for which we must consider the profiles gb-inst and fr-inst-fr. We start defining tasks for each profile used. If we inspect the

sites corresponding to the URLs, we can see that the first one is in English and the second one in French. As we want to write only one extraction algorithm, we have to implement methods to select the elements in the right language.

We create a map containing for each language, the keys, a second map that groups the elements we need in the extraction code :

```
1 val list_profile_map = Map(  
2   "gb" -> Map(  
3     "codeISIN" -> "ISIN code",  
4     "pastPerf" -> "Past performance is not a guide",  
5   ),  
6   "fr" -> Map(  
7     "codeISIN" -> "Code ISIN",  
8     "pastPerf" -> "Les performances passees ne prejudent",  
9   ),  
10 )
```

Then, we get the right one, by giving the language as a parameter :

```
1 val selected_profile_map = list_profile_map(task.language.get)
```

Thus, we can simply use the elements of this map by entering the corresponding key :

```
1 val isin_mask = page.ive_mask.copy.restrict_to(  
2   bottom = page.ive_mask.lines  
3     .text_matching(selected_profile_map("pastPerf")).uppermost.top,  
4   left = page.ive_mask.lines  
5     .text_matching(selected_profile_map("codeISIN")).uppermost.left  
6 )
```

During my internship I implemented or modified about 35 extraction scripts.

10 Conclusion

I did my three-month internship as an intern in automation of data extraction for investment funds in the start-up Tetrao in Luxembourg. During this time, I was able to develop my IT skills acquired during my studies, while being confronted to the working world of a fast growing start-up.

This internship was very enriching for me, because it allowed me to discover the field of engineering, more exactly the field of data extraction. Before my internship, I had little knowledge about this project, and I was able to benefit from important theoretical and technical contributions and to acquire new tools.

Moreover, this internship brought me a first professional experience. I was confronted with the difficulties encountered in the daily life of a company and I could observe the methods of organization of a team of engineers. This experience also allowed me to reinforce many skills such as versatility and teamwork.

Finally, I would like to thank Christian Gillot, the Tech Founder and Chief Executive Officer, and Laurent Cherpitel, the Chief Financial Officer, for welcoming me and integrating me into the Tetrao team.

I would also like to thank my tutor Etienne Rigaud, the engineer Stanislas Barbillon and the software developer Albert Razquin who trained and supervised me throughout my internship and who allowed me to perform many interesting tasks. I would like to express my gratitude to them for the explanations and advices they gave me during the missions I was able to participate in.

References

- [1] Rigaud Etienne. *Développement de modèles d'apprentissage automatique pour la compréhension de documents*. Tetrao, 2020.
- [2] Michels Théo. *Assistant-ingénieur*. Tetrao, 2021.
- [3] Tetrao.
https://wiki.tetrao.eu/wiki/index.php/Main_Page.
- [4] Thierry Labro. *Comment Tetrao va disrupter l'industrie des fonds*.
<https://paperjam.lu/article/comment-tetrao-va-disrupter-in>, 18.12.2019.
- [5] Thierry Labro. *La Bourse s'invite chez Tetrao, un win-win intelligent*.
<https://paperjam.lu/article/bourse-s-invite-chez-tetrao-wi>, 26.01.2021.
- [6] Mehdi Ouchallal. *Comment créer un fonds d'investissement?*.
<https://www.legalplace.fr/guides/creer-fond-investissement/>, 06.04.2021.
- [7] IFSL International Limited.
<https://ireland.marlboroughfunds.com/funds/IFSL%20International/>.
- [8] Lazard Asset Management GmbH.
<https://www.lazardassetmanagement.com>.
- [9] Corum Butler.
<https://www.corumbutler.com/fr/bond-funds/butler-credit-opportunities/>.
- [10] Merrion Capital Investment Managers Limited.
<https://cantorfitzgerald.ie/kids/>.
- [11] Xavier Niveau. *Le langage Scala*.
http://igm.univ-mlv.fr/dr/XPOSE2011/le_langage_scala/index.html#1.
- [12] Project Euler.
<https://projecteuler.net>.
- [13] Hermitage Gestion Privée.
<https://www.hermitagegestionprivee.com/gestion-actifs/>.
- [14] Bankinter.
https://bancaonline.bankinter.com/fondos/buscador_fondos.xhtml.
- [15] Sabadell.
<https://www.sabadellassetmanagement.com/es/tip>.