

## Table of Contents:

Abstract.....	2
1.Introduction.....	2
2.Literature review.....	4
2.i.The World of Philately.....	4
Philately and stamps identification.....	4
Philately and Internet communities.....	5
2.ii.Data scraping.....	8
Background review and state-of-the-art techniques.....	8
Legal aspect of web scraping.....	10
2.iii.Content-Based Image Retrieval.....	11
Presentation.....	11
Keypoints detection.....	12
Features extraction.....	15
Data clustering.....	17
3.Technical achievements.....	21
3.i.Preliminary decisions.....	21
Organization overview.....	21
Programming language and framework.....	21
Tools selection.....	22
3.ii.Social network development.....	23
Implementation decisions.....	23
Database design.....	24
3.iii.Data scraping.....	26
Tool selection justification.....	26
Scraping experimentation.....	26
3.iv.Content based image retrieval.....	28
Preliminary test: Descriptors efficiency.....	28
Real condition tests.....	29
K-means experimentation.....	30
System integration.....	30
3.v.Dissertation conclusion.....	31
Technical achievement and original contribution.....	31
Evaluation and discussion.....	31
Conclusion.....	33
4.Critical appraisal.....	34
4.i.Research methodologies.....	34
4.ii.Development methodologies.....	35
4.iii.Legal, ethical and environmental issues.....	36
4.iv.Risk management.....	37
4.v.Professional development.....	39
4.vi.Further improvement.....	40
4.vii.Acknowledgment.....	41
5.References.....	42
Appendix A: Database design.....	45
Appendix B: Web application.....	46

# *A social network for philatelists*

*Camille Girardo – 12007518 Oxford Brookes University*

*Tutor: Prof. Hong Zhu*

---

## **Abstract**

This dissertation aim is to build a social network tailored for philatelists. An analysis of philately world with its particularities and its needs is performed, where current online solutions are assessed.

A literature review is carried out on web crawling and data scraping methodologies leading to the selection of a tool adapted to this project for automated retrieval of stamps pictures and metadata from other websites.

A second review covers Content-Based Image Retrieval specificities, with a description of each Bag of Visual Words stage and a comparison of several state-of-the-art descriptors extraction algorithms such as SIFT, SURF, ORB, BRISK and FREAK. Their default features detectors are also quickly explained. Different clustering methods such as k-means, AKM or HKM as well as common results enhancement techniques (weighing, stop words...) are inspected. Finally, an image matching feature will be integrated in the social network to help for stamps identification.

---

## **1. Introduction**

Social media are one of the most important evolution of the web over the last decade. A typical mistake for a new actor is to aim for a generalist target. The well established

competitors, composed of some of the biggest companies in the world makes it impossible. Even an innovative product with unique features would struggle for the success as a good amount of early-users is the key requirement for a massive adoption of the service. A smaller portion of the market need to be targeted and need to include services tailored to their needs with a more specialized set of features. This conclusion leads to the creation of this project: “A social network for philatelists”.

Philately, considered as the king of hobbies, is a well established discipline and with the rise of Internet a variety of historical actors tried to adapt their offer, but proposed solutions which does not necessarily aim to simplify philatelists quest for knowledge. Several competitors aimed to conquer the social network field, however the lack of innovative competitor leads to a fragmented landscape.

The majority of well-established actors on this field aims to reference stamps by its metadata. However, a frequent issue in the practice of philately is the encounter of an unknown stamp. This project will offer an innovative feature for this type of situations: a Content-Based Image Retrieval system dedicated to stamps identification. This particularity is a key for the success of the project as it provide an alternative to an historical search pattern based on the use of catalogs.

A good understanding of the philately world is necessary to build an attractive social-network answering to the various needs of stamps collectors. The website should allow to create a profile where a stamp collection can be managed. A specific attention will be provided to catalogs as those are a powerful way to identify a stamp. Interaction between users and trading functionalities are also vital for the success of the project.

A literature review on Content-Based Image Retrieval including pictures clustering methodologies as well as an analysis of suitable image processing algorithms dedicated to image matching will be necessary to ensure an efficient implementation of the original idea.

Efficiency of this system can only be shown if an initial database is present. To answer this need, a website scraping script has been developed, aiming to retrieve information on specialized websites and fill stamps details in the solution's database.

## 2. Literature review

### 2.i. *The World of Philately*

#### **Philately and stamps identification**

Philately is not only a hobby centred on stamps collecting, but also a form of historical study. Since its introduction in UK(1840), stamps have been subjected to a wide spread over the world, turning it into a great window on societies evolutions over time. Thus, philatelists enthusiasm can easily be understood.

Philatelic literature is one of the most prolific in term of publications. Several ways exists to share knowledge, the most recent and promising being the digital publishing, with the example of Stamp Insider[1].

This need for an accurate and relevant information have to be understood when designing a tool dedicated to philatelists[2]. Due to the worldwide use of stamps with a lack of uniformity between countries regulations, as well as changes caused by evolution of societies, stamps possess a number of variable characteristics, such as the colour, the type of paper used or the printing method. Potential similarities between different stamps can cause issues for stamps identification, explaining the importance of a great range of referenced features.

Stamps identification would be difficult without a good resource referencing stamps and their characteristics specificities. Key actors for this matter are known as catalogs, those spread shortly after the adoption of stamps in postal services, gathering information on various stamps and assigning unique identifiers. A summarization of typically referenced characteristics on this kind of publication allows to identify important features for our own database model.

An issue with catalogs is the absence of a leader covering the totality of countries, philatelists are forced to refer to various competitors. 4 major actors can be found for this matter: “Michel”, “Scott”[3], “Stanley Gibbons”[4], “Yvert et Tellier”[5].

A second issue encountered with this indexing method is the fact that stamps collectors often are forced to use various editions of a same catalog as range and depth of information evolve between editions. Caused by the need to cover an ever growing range of newer stamps on a support of fixed size.

These two problematic can be issued with the use of Internet technologies and are another proof that this project is needed. A review of current competitors on the market need to be performed.

## **Philately and Internet communities**

A wide range of competitors revolving around the philately field can be found on Internet. Some of those being direct evolutions of historical actors, transposed to the Internet. Their business model is interesting as it takes into account technological advancement of society, even though social interaction does not seems to be a priority.

### *Online catalogues[4][5]:*

These websites are amongst the ones possessing the most filled databases of stamps, due to their history. Stamps differentiation is based on respective identification numbers. They frequently offer possibilities to buy and sell stamps. On the down side, interaction between users is not the priority and stamps trades cannot be performed without involving the website as a third-party actor. Their list of stamps tend to lack of recent elements. Moreover, the referencing is directly controlled by an entity, which gets in the way of a fast knowledge base expansion.

### *Stamp Dealers[6]:*

These sites aim is to buy and sell stamps, they can be focused on a specific area or period. Offers can be about specific items or a package of several random stamps. In a general matter, databases are less filled than those of other competitors, as the priority is to highlight current offers.

### *News websites[7]:*

The aim here is to provide news on recent stamps, the use is not commercial unlike the previous actors, but contain information different from any other type of sites. Unfortunately its lack of data about older stamps prevents it to be a reliable information source. This type of website could however be used as an additional source for data once this project's website will be available on Internet.

With the rapid spread of social networks and its impact on Internet, some communities started to grow around philately. Two kind of solutions can be observed:

*Social Network diverted from initial functionality[8]:*

Some communities were built around stamp trading on generalist social networks, therefore functionalities are not adapted and security matters are preoccupying as scam can't be prevented. This type of communities could be a potential target for expanding the user base of the project and stand as a proof of users interest for online philately.

*Community-based collection management websites[9][10][11]:*

Functionalities on this type of websites are similar to what this project aims to accomplish, however, none of them possess a “killer feature” to stand out of the crowd. Image matching for stamp identification and possibility to set up a private version of the website are not currently offered on leading solutions. These remains good sources for implementation ideas as well as potential scraping targets. It also provides a list of stamps characteristics frequently used to put in relation with catalogs choices before designing our own database organisation.

A case by case analysis of prominent alternatives can be found in the following table, only websites with interesting strength are taken into account.

Site	Strength	Weaknesses
Colnect[9]	Biggest database amongst community-based solution. Good sorting functionalities, with possibility to apply several filters. Good depth of information for each stamp Efficient display of research queries trading list	Random redirection to French version. No upload content standardisation Heavy, too much features. Not only dedicated to stamps Sign up pop-up
Philateca[10]	Minimalist (KISS), no fancy/heavy eye-candy. Possibility to sort by country, tags, serie, printer (not found in other	Small database No integration of catalogs systems No social interactions

	websites) Records on uploader and addition date kept (provides a good traceability and identification diligent users) Accessible search engine	Smaller database
Philatelim[11]	News on upcoming issues Really good depth of information Stamps displayed per series Possibility to switch from list view to images view	Really bad sorting methods Low amount of country covered No social interactions or transactions
allworldstamps[12]	Good quality for Stanley Gibbons content	Free version: no pictures Paying version Bad user interface Only stamps referenced in Stanley Gibbons
Arpinphilately [6]	Good quantity of stamps, good search methods Good general interface	Shipping only to US/Canada, not possible to sell or trade stamps, commercial purpose
stampsoftheworld[13]	Community-based, good amount/depth of information Possibility to sort “images required” stamps (great idea for building the database)	No trade or social functionality
Yvert[5]	Possibility to access every Yvert et Tellier catalogs online, with a paying option.	Limitation in term of countries No trade or sale possibility Low depth of information Only stamps referenced in Yvert etTellier catalogs

This leads some reflections on priorities to adopt for this project. We already decided to focus on easing the building of a network amongst philatelists, rather than developing a more mercantile approach. It induces the fortuitous advantage of

avoiding a duplication of well established competitors services and therefore a direct rivalry. An improved security around transaction would be a serious argument in favour of our solution over communities built on generalist social networks. Releasing the source code and allowing users to build their own private communities could help in this direction. On a software point of view, feedback on transaction should have an important influence on users reputation. An additional issue revealed during this market analysis is the fact that pictures are not always present on other websites, in such case those are referenced with its specific identifier in a catalog. This causes issues for the data scraping stage of this project. To prevent incompatibilities, the use of pictures will not be mandatory in this project's database.

## ***2.ii. Data scraping***

### **Background review and state-of-the-art techniques**

Data scraping is a technology widely used nowadays for automated data retrieval from a wide range of sources.

It has to be differentiated from parsing, as the process can look similar at first sight. However parsing is performed on a set of data destined to be handled by a machine, which means that it is frequently well structured and documented for this specific purpose. On the other hand, scraping is performed on data displayed to the end-user. There is therefore no effort made to ease the process, data are structured for a better display with a good-looking layout. It also means that a considerable amount of non-relevant information will have to be filtered.

It can be used with different data sources, such as files or websites. The former is useful for data mining or data reorganization on a set of documents, which would be off-putting otherwise, while the later is widely used on the Internet and known as web harvesting or web scraping. It consists in retrieving information from a website. Different purposes can be claimed, from data retrieval for tools migration where the technology doesn't provide an easier and "cleaner" way, to data collection from a third-party actor[14].

Web scraping is frequently used in collaboration with a crawler. This web indexing method consists on following every link in each visited page and travel around the limits defined by the user. Domain name boundaries are frequently used but in cases where accuracy is needed, a selection of URLs can be employed to guide the



crawling process, helped by the use of regular expressions[15].

In some conditions, it might be faster to perform the retrieval task manually depending on factors such as accessibility and quantity of information.

Automated retrieval has been a field of research over the last decade and various techniques have been developed for web harvesting in the context of a structured content. Three approaches are worth considering and seen as state-of-the-art[16][17].

#### *Manual wrapper generation:*

This method involves working on each target website separately, which is not optimal if data need to be retrieved from a wide amount of sources before being aggregated. Two common strategies can be considered:

→ A website extraction or reproduction tool when used with regular expressions for filtering can fit this purpose.

→ Specific programming languages are dedicated to information retrieval in tree structures such as HTML or XML, XQuery is a good example. This is particularly useful for dynamically generated content as the layout of the page is shared by each page. The page source code has to be analysed and the relevant information path can be used to target which data will be processed.

A clear advantage of this method is the control given to the user and the understandability of the process. With the possibility to react to unexpected behaviours from the website.

#### *Wrapper induction:*

This approach is widely spread nowadays and dedicated to dynamically generated content. The aim is to train a script with a set of random pages from the data source, where user labels desired features and the program will learn from it and generate extraction rules. Several systems exist such as Stalker[18][19] which is currently known as the most efficient on the market, its flexibility and low training needs being its main assets. Older methods[20] such as WIEN[21][22], Softmealy[23] or WL<sup>2</sup>[24] are also worth considering.

Once implemented, this option is perfectly suited for a user with little to no

background in web development due to the ease of relevant information selection. It is also a predilection choice if a great number of sources needs to be processed.

#### *Fully automated:*

The third approach is the most recent field of research and aim to fully automate the process[25], without the need for a training step. It is not currently as efficient as other methods and manual post-processing is needed to select the desired set of information.

The main field of research related to web harvesting for unstructured text is focused on machine learning and natural language processing, but is out of this research's scope as the type of information needed for the project are likely to be structured.

The selected solution, Scrapy[26], is well recognized in the professional field and fulfil our requirements, as explained in the technical section of this dissertation.

This library is not part of an automated method as the needed scraping task is not overly complex and can easily be performed with basic tools. Implementing a more intelligent method would only make the process heavier.

### **Legal aspect of web scraping**

Despite its convenient functionalities, data scraping is frequently source of controversy due to its legal aspect, even if a set of data is publicly available it might still be considered illegal to build an automated retrieval tool and numerous case law in the past can prove it.

Reasons are various and include a surcharge of the data load for targeted servers which could induce increased hosting costs for the company. The flow of queries can decrease the quality of services offered to other users, if performed too aggressively, commonly referred as trespass to chattels.

Building a competitor website with a scraped set of data can lead to a loss of revenue and is not considered as fair.

Several famous cases can be cited, for different types of situations such as airlines prices scraping for aggregation and comparison, or database duplication([27-30]). The closest case compared to our situation can be Cvent versus Eventbrite but

ended up on a settlement between the two parties [30]. The issue here is that most cases have taken place in US where legislation is different from European countries, therefore jurisprudence is difficult to assess in UK. A list of potential legal issues is listed on DMH Stallard's website[31]:

- Copyright infringement, depending on the type of scraped data, this doesn't apply to stamps pictures as copyright is only valid for pictures with an artistic value.

- Database right infringement, valid when all or a substantial part of a database is duplicated.

- Data protection legislation, aiming to protect personal data, therefore not related to this case.

- Breach of contract, if Terms of Services diligently forbid this type of access, enforceability is subject to interpretation depending on its display prominence[32].

- Computer misuse, originally built to prevent menace of hacks can be interpreted and applied to web scraping.

Most cases involve a recurring scraping of a target website, which can be seen as damaging for the targeted server. In our case the scraper is supposed to be launched only once and data is not going to be displayed to the public which in real life condition should be enough to avoid this type of issues.

Permissions have still been asked to targeted websites, in the context of this dissertation, in order to prevent any risk.

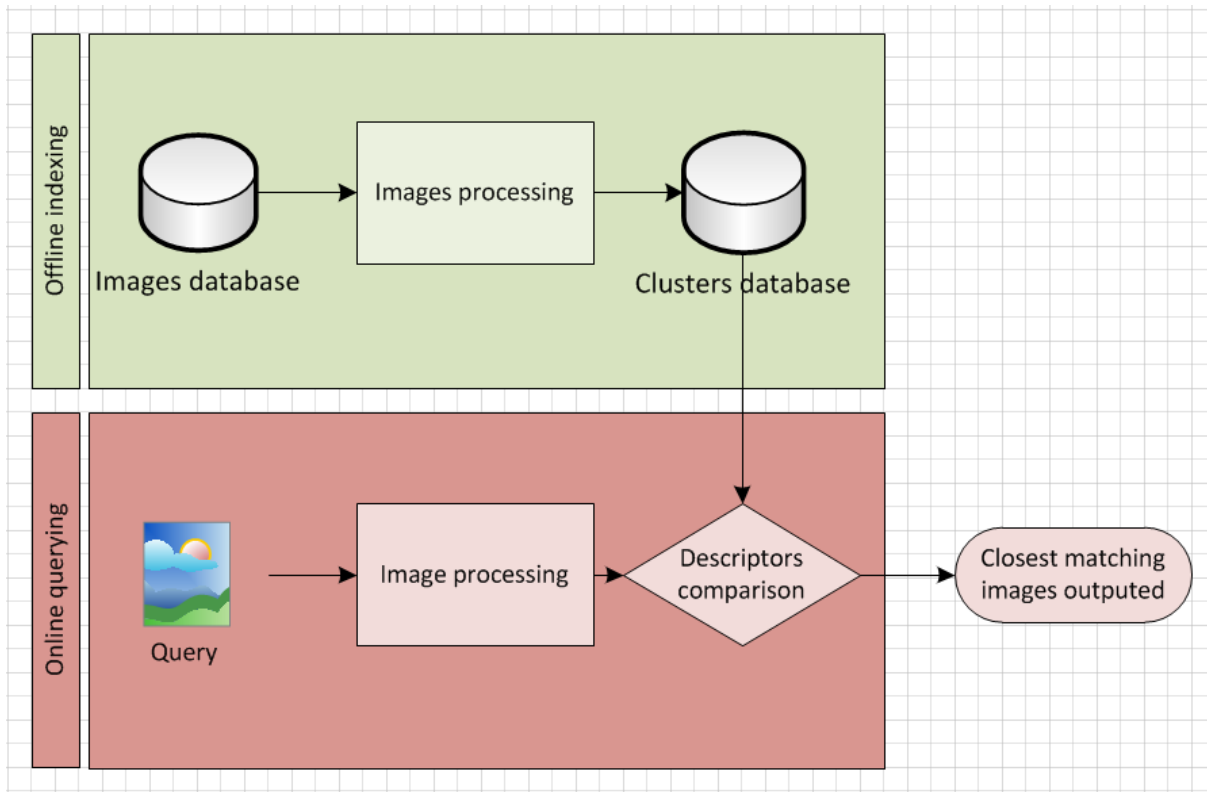
### ***2.iii. Content-Based Image Retrieval***

#### **Presentation**

Content-based image retrieval is a technique widely used on internet, some famous examples being shown as Google and Baidu reverse image search or TinEye. This technology consists on querying images by their low-level visual content[33] as opposed to metadata retrieval where keywords or tags are used.

To simplify the querying process, images descriptors can be extracted from an uploaded image and matched with indexed visual contents.

The most widely used representation method, Bag of Visual Words, has been selected for this project. Low level visual contents are detected and extracted from pictures, features are then quantized and clustered in a database.



*Fig. 1: Content-based image retrieval concept*

This indexing methodology relies on three separated steps explained in this review.

- Keypoints detection
- Descriptors extraction
- Data clustering

## Keypoints detection

Since mid 80's researchers greatly employed to improve features detection in images. Aims are various, from the improvement of computer vision exploited in the robotic field, to items and humans detection and recognition on different type of media such as pictures or videos. Several types of algorithms have been developed, the lowest level being area based matching subject to a pixel-by-pixel comparison, intolerant to scales variation or rotations[34]. Current state-of-the-art techniques are based on features detection.

A feature, when used in image processing, is known as an interesting point or area in a picture. Different concepts have been explored to provide an efficient answer to this problematic. The first thoughts on the matter involved investigation over edges

detection, where strong contrast on a small region where analysed, famous algorithms can be cited such as the historical Canny or the recent Sobel.

Further research headed for different types of features in order to improve the quality and usability of outputs. Corner detectors were a great improvement over edges as those can be used in a wider field of applications. These detectors are based on the same principles, but also aims to detect fast modifications on a curve, identifying those points as corners. This type of algorithm suffer from a lack of connectivity between features, which might be a handicap for a great outline of major level descriptors such as objects or texture, reducing the repeatability of features in case of images variation[35]. Regions or blob detectors are the newest stage of features detection, directly evolved from corners algorithms such as Harris, the concept is to select a wider area around corners in order to detect a region of same intensity, this methodology allow a better resistance to image alteration and ensure a better repeatability.

In the field of image matching, detectors directly depends from the selected features extraction algorithm. For this reason, only detectors used in the selected set of descriptors will be covered. None of them rely on edge detection but FAST[36], used in ORB and AGAST for BRISK and FREAK are corner detectors.

Both are based on SUSAN algorithm, where the brightness of a centre pixel is compared to its circular pixel neighbourhood in order to detect corners in an image. In this algorithm, the Bresenham's circle of diameter 3.4 pixels is used, making a 16 pixels segment to analyse.

FAST has been created with the aim to be a viable option for real-time applications, the main goal being low computational costs with a decent repeatability. According to the authors, experimentation gives ground-breaking results (7% of processing time compared to 300% with SIFT). These outstanding results are caused by the fact that FAST aims to analyse the fewest pixels possible around the centre pixel (the most frequently used version being 9 instead of the 16 required by SUSAN). On the other hand, this specificity makes the algorithm more sensible to noise. Despite this issue, FAST usually outperform older algorithms such as SUSAN or Harris[36][37][38], in both speed and repeatability, which makes it able to detect the same set of features after affine transformation, it quickly became popular after its release.

AGAST[39] shares the same base as FAST, the Accelerated Segment Test (AST)

algorithm, an improvement of the older SUSAN. According to the experimentations presented in the AGAST presentation publication, a large improvement in speed can be shown on several types of pictures when comparing FAST-9 to AGAST-7 or AGAST-5. The idea of AGAST is to build a generic binary decision tree for pixel colour intensity comparison needed for corners detection. In the case of FAST, this has to be built and trained for every new environment and slow the process down. It results a similar quality in detection with an improved computation time.

The field of blobs detection has been an interesting field of research, aiming to identify a region of similar textures in images in order to develop a strong repeatability in keypoints detection and improve scale variation insensibility.

Research started from a corner detector (Harris) and evolved publication after publication. Lindeberg worked on automatic scale selection and experimented with several methods such as Determinant of the Hessian and Laplacian. Mikolajczyk and Schmid refined these methods, creating scale invariant feature detector with high repeatability: Harris-Laplace Hessian-Laplace. Lowe then proposed a novel method known as Difference of Gaussian, aiming to improve processing speed compared to Laplacian of Gaussian (LoG).[43]

The Difference of Gaussian[40] is used by the SIFT algorithm, proposed by Lowe as an approximate of the LoG, it is the most robust of the algorithms presented in this review, but its computation cost is a huge drawback. It consists in building an image pyramid with resampling between each level in order to select key locations at maxima and minima of a DoG function[41]. Once the pyramid is built, each pixel from the bottom layer is compared to its neighbours, if a corner is detected, the same process is applied to the above layer until the top layer is reached. The detailed process is explained in SIFT specifications paper[40]. Additional processing is performed in the case of SIFT, keypoints are selected by isolating low contrast features, with the differences of neighbouring sample points method developed by Brown[42], and by eliminating keypoints close to parallel to an edge as those are easily victim of noise and therefore having a low repeatability.

The most computational expensive part of this algorithm is the building of the pyramid, which is not needed in SUSAN derivatives and can explain differences in robustness and computation speed.

Determinant of the Hessian is used in the SURF algorithm[43], the idea was born from the observation that Hessian-based detector are more stable than Harris-based. Blobs detection and scale selection are both done by the Hessian matrix as explained in Lindeberg publication[44], unlike in Hessian-Laplace where only the first stage rely on it. Integral images are used in concordance with this method to provide a scalable and fast blob-detection mechanism (integral images are used to detect the sum of intensity in a rectangular region of any size and only need the 4 extremities values). Unlike DoG where scale space representation is done by iteratively reducing image size, DoH, by its use of integral images with Hessian matrix only need an up-scaling of the filter size, making the process faster.

## **Features extraction**

Images descriptors are used to extract features from images, based on detected keypoints, these low-level visual content needed for image matching. Several descriptors are widely used, the most famous being SIFT, presented by David Lowe in 2004[40], as a continuation of his previous research on invariant features detection[41]. It quickly became a key actor on image matching, thanks to its enhanced efficiency compared to older state-of-the-art techniques, it is still one of the most frequently used algorithm nowadays despite its high computational cost, considered as a huge drawback for real-time applications. It generates a set of scale-invariant coordinates relative to local features, also called vectors. One of the specificity of this algorithm is the use of DoG for keypoint localization, as explained in the previous section, to detect an important density of interest points invariant to scale and orientation[40]. This density is particularly useful to detect small objects in the background.

GLOH(gradient location-orientation histogram)[44] provides a quiet similar type of vector as it belong to the same extractors family. It considers more spatial regions, making it more robust than its ancestor, but also slower to compute. Its use is reserved for niche applications where computation speed is not a priority.

Due to the growing demand for high-quality, high-speed features, most efforts made over the last decade aimed to improve the execution time while keeping a decent quality of descriptors. PCA-SIFT[45] and SURF[43] have been developed with this idea in mind. Unfortunately PCA-SIFT happen to significantly decrease the quality of

detection, the goal of the algorithm is to reduce the dimension of descriptors from 128 to 36, leading to some data loss, which explain the difference on robustness.

SURF[43] performed quite well and provide good-quality interest points, its authors claims for it to be more efficient and less sensitive to noise than its competitor but various experimentations shows it to be slightly less robust than SIFT. Its speed makes it an interesting algorithm for a multitude of projects[46][35]. It uses the determinant of Hessian matrix for blob detection and outputs the same type of descriptors as SIFT, but with 64 or 128 dimensions. It might be used with PCA to reduce dimensionality[47], but this choice impacts performance due to some additional computation-time.

BRIEF[48] has been produced with the will to provide an improvement to the dimensionality of descriptors, Calonder et al. idea was to build a short binary descriptor. Hamming distance have to be used instead of the state-of-the-art for SIFT-like descriptors: Euclidean distance, improving the distance computation time. The most important difference between the two types of descriptors is the use of lower level keypoints detector as less information on vectors are necessary to build a binary descriptor. The issue with BRIEF is its variance to scale and rotation change.

ORB is a novel algorithm aiming to assess this weakness[38], based on a FAST keypoints detector working with BRIEF for features extraction, hence the name: Oriented FAST and Rotated BRIEF. As said earlier those two algorithms are known for their good performances and low cost. Comparisons highlight the efficiency on computation cost of ORB over SURF and SIFT for most criteria and type of modification for an image. However, due to the use of BRIEF, this algorithm is sensible to in-plane rotation. A weakness has been highlighted for scale variation even if the use of FAST improve the quality of BRIEF alone. On the bright side ORB is two order of magnitude faster than SIFT and one faster than SURF, which is an asset for real-time applications.

BRISK [37] and FREAK [49] both uses AGAST, an improved version of FAST, for features detection. Their speed magnitude is slightly better than ORB, while showing to be invariant to scale modification and rotation. They were born from the need to improve performances, with the growth of the smart phones market, involving a decrease in memory compared to traditional computing. The aim is to evaluate some weighted Gaussians around keypoints, the particularity of FREAK is its retinal pattern disposition for corners detection, where a high density of points are assessed close



to the keypoint and decrease exponentially with the distance, in comparison BRISK rely on a constant rate independent to distance.

Various publications aims to compare and assess several algorithms and can give a good understanding of their assets and drawbacks[35][50-53].

Results often varies depending on if items are planar or more textured. For this reason, we need to test the most promising algorithms with a set of picture relevant to this project. From this set of publications a list of algorithms to assess can be identified.

SIFT is known for its robustness, but also for its high cost, it is going to be assessed in order to provide a good standard for our tests, but will most probably not be used for the final implementation.

The high cost issue is also valid for GLOH, improvement in robustness is not necessary in this case and GLOH will therefore not be considered.

As frequently said in these comparisons PCA-SIFT, despite having a good computing time is not robust enough for the type of application we are aiming to develop.

SURF is frequently outperformed by FREAK and BRISK both in accuracy and speed[50,51], however, it will be kept for real conditions tests due to its status of widely adopted state-of-the-art solution.

Even if ORB is more sensible to noise, scale and rotation issues, it can still be considered for this application as the type of images expected are scans or pictures of stamps, which should not suffer from too many transformations.

The final panel is composed of SIFT, SURF, ORB, BRISK and FREAK, the evaluation process is described in the methodology part of this dissertation.

## **Data clustering**

The last step of Bag of Visual Words is data clustering, the first action to accomplish is to generate a visual vocabulary. The method used for this matter is inspired by a common text retrieval strategy where several documents are processed with the aim to find highly descriptive words that can be used to build a vocabulary. In the context of image retrieval, this method is adapted with descriptors extracted from pictures and clustered into “visual words” of a vocabulary. Even though the semantic logic is lost, the method is still an efficient way to answer to this problematic.

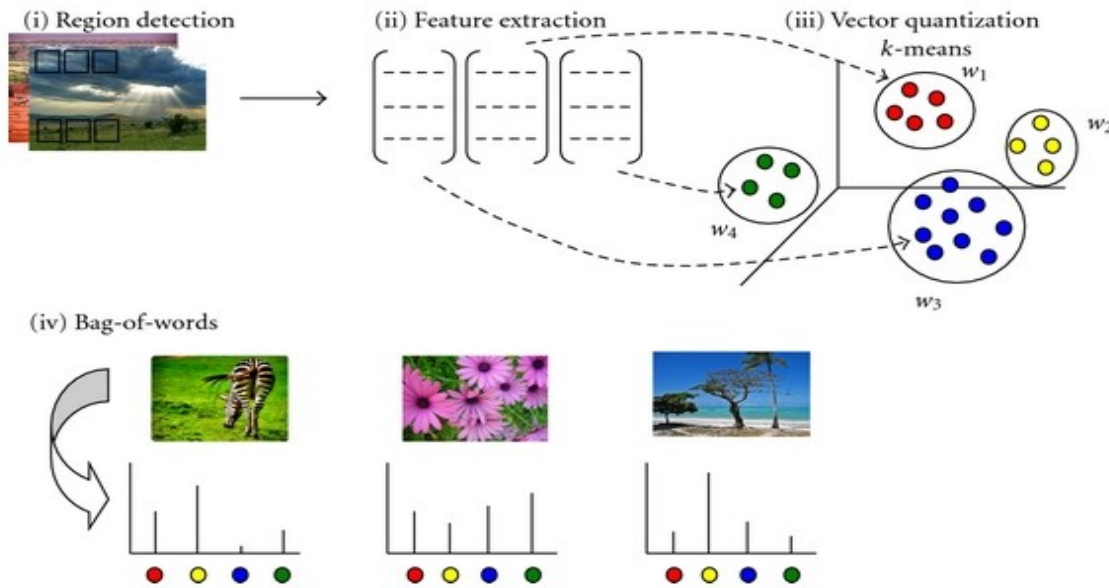


Fig. 2: Clustering process for Bag-of-Words[33]

Several algorithms can be used to perform this task, the historical one is known as  $k$ -means[33][57][55], also referred as Lloyd's algorithm.  $K$ -means analyses the set of descriptors and quantize those by clustering the closest looking descriptors in a visual word. The euclidean distance is used to determine the relation between descriptors. Two iterative steps composes this algorithm: assignment, where each descriptor is assigned to the closest looking cluster, and update step, where the mean of descriptors for each cluster is computed. The assignment cost can be represented as  $O(nk)$  where  $n$  is the number of descriptors and  $k$  the number of clusters.

Vocabulary size need to be specified by the user. A small vocabulary might lack of discriminative power, two different descriptors might be assigned to the same cluster, a big vocabulary will be too challenging and slightly altered descriptors might not be assigned to the good visual word (sensible to noise and blur). A survey on size choice in research papers shows experimentations distributed over a wide range of vocabulary size from 200 to 1 million[56]. This does not show a consensus on the question, tests will therefore be necessary before selecting an implementation.

This flat version of  $k$ -mean is efficient for small vocabularies but hardly scales when more visual words needs to be processed. Newer algorithms are based on this

approach but uses other technology on top of it such as approximate neighbour search or cluster hierarchy to increase the vocabulary size.

Approximate k-means (AKM) [57][58] substitute the use of nearest neighbour by approximate neighbour search during the assignment step with the aim to mitigate impact on descriptors close to different clusters. Several k-d trees are built and their conjunction are used to build the final tree. Studies shows that AKM is a good substitute to k-means as results looks similar for a fraction of the computing cost. This cost is  $O(n \log(k))$

Hierarchical k-means (HKM) [59] is based on cluster hierarchy, the process is easy to understand, the algorithm determine the  $k$  closest descriptors and create an average node with these  $k$  objects as children. The algorithm then start over, replacing the closest points by an average, the same goes until a tree is built. The user needs to select a threshold of distance and each identified closest neighbours with a distance lower than the threshold can be clustered in it. The issue as with other k-means is the need to indicate a threshold that will directly decrease efficiency if poorly chosen.

K-means can be easily computed when the descriptors are in the form of vectors such as for SIFT or SURF, because the Euclidean distance can be used to measure the distance and determine the average vector. However for binary descriptors (ORB, BRISK, FREAK), this metric is not suitable as it wont give a usable result. Instead the use of Hamming distance is advised, unfortunately this method alone doesn't provide a centroid usable as visual word. An interesting voting scheme is described by Grana C.[60] called "k-majority algorithm" adapted from the Lloyd algorithm. According to this research the processing time is really improved with this method compared to a typical k-means.

Once the vocabulary is built with the training sample of pictures, the quantization can be applied to the descriptors extracted from a set of pictures. It consists in assigning each feature to its closest matching visual word in the vocabulary. The most widely used algorithm is the nearest neighbour search where the aim is to find the closest-looking visual word compared to the local descriptor. Even if HKM and AKM need a specific adaptation to suit their tree organization.

Several techniques can be used in order to improve results or efficiency:

Different kind of weighing scheme can be implemented:

→ “tf” is the frequency of apparition of a visual word in an image, the goal is to identify more important words in a document.

→ “idf” (inverse document frequency) is a way to have an overview of the frequency of a word amongst the whole corpus, where the number of documents is put in relation with the number of occurrence of a specific word

→ “tf-idf” (term frequency-inverse document frequency) is the combination of tf and idf, subject to polemic as the efficiency of idf is not proven, especially for big vocabulary.

→ “binary weighing” consists of indicating presence or absence of a visual word with a 0 or a 1, discarding the number of occurrence making it less efficient for small vocabulary size where this factor is important[61][62]. This method is not commonly used in publications[33].

→ “soft weighing”, researchers formulated the observation that every other weighing scheme was using nearest neighbour search before calculating a weight and argued that this choice wasn't optimal “given the fact that two similar points may be clustered into different clusters when increasing the size of visual vocabulary.”[61]. From this observation, they introduced a new weighing scheme called “soft-weighing” where the top-N nearest visual word would be taken into account.

A stop list can be added afterwards, the idea is to remove the most frequently used visual word in order to speed up the comparison process. Ensuring a faster processing of data when searching for potential matches.

### **3. Technical achievements**

#### ***3.i. Preliminary decisions***

##### **Organization overview**

This dissertation is separated in 3 major activities and a simultaneous development had been considered as counter-productive, therefore prioritization of tasks had to be decided. The social network has been developed first as other phases needed to interact with it. This also gave an opportunity to get a feeling on Python and Django before working on more complex tasks. Web scraping was the second step as the literature review highlighted importance of a set of pictures for vocabulary training and image matching experimentations. An extended research period preceded each phase of development to improve understanding on potential issues and help requirements definition.

##### **Programming language and framework**

A website can be developed with several languages, both static and dynamic. Static languages are usually strongly and statically typed, implying declaration of every variable and involving compilation of the source code. It is an asset on software development as it detects errors earlier. A gain on performance is also noticeable as the program does not need to check variables at runtime. However, dynamic languages allow to produce a good quality code faster, it increases flexibility of the solution as variables and structural items can be created and modified during run time. This is more suitable for web development and this difference is widely understood by developers, which leads to a massive adoption of dynamic languages in the web field. Python is an asset as it can be used in an efficient way for scripting purpose as well as for web development, with the use of a dedicated framework. Moreover, with its specific code architecture deprived of curly brackets, Python provides a code syntax easy to read compared to its two most common competitors, PHP and Ruby. The diversity of libraries available covers a great range of the situations encountered for this project, from creating a web application to the writing of a scraping script as well as for images processing.

Python is commonly used in two different versions, an older one which is the most widely spread amongst professionals and a newer version. Preponderant cause the

new version lack of adoption is its incompatibility with commonly used libraries due to the modifications on the implementation of several key methods. Furthermore some famous server distributions such as Debian still provide the older version as a default choice. For the sake of compatibility and in order to ensure the completion of the project, the version used for this application is python 2.7.

Django is the leading framework for web-development with Python, it brings a good set of security-oriented features, from passwords encryption to forms verification, as well as a good visibility in the code.

The main information to know is that Django is a Model View Controller based framework, frequently referred as Model Template View as in this case the controller is managed by the framework. Models are used to interact with data sources, templates for content rendering and views are the link between other layers. The MVC design pattern is interesting for projects on which several developers are working[62], separation of tasks allow modification on a layer without impacting other parts of the software. Another asset is the diminution of code duplication, more frequent in statically written websites. The overall quality and versatility of the deliverable will be positively impacted by this framework.

## **Tools selection**

It had been decided to set a virtual machine as a web-server instead of hosting an actual machine on Brookes network as a full control on the machine eases the configuration process and access will be guaranteed even without a reliable Internet source. The main focus for tools selection is to prioritize the use of open-source solutions. It presents multiple advantages, the main one being the possibility to prepare the environment without having to wait for licenses availability. It also ensure a better security and durability as updates to major version are freely available. On this perspective, the web-server's operating system is a Linux, many distributions are worth considering, but a renowned one is a safest option. Debian has been selected for its stability and its presence in the professional field. Previous experiences with the deb package management system rather than rpm influenced this choice.

Several database engines are available, the two most famous open-source solutions being MySQL and PostgreSQL, those projects quality have merged over the last few years, but MySQL dependence to a company makes it less consistent to the free software philosophy. Due to Python scripting nature, complex IDE such as Eclipse

are not necessarily an asset, for this reason text-editors such as Sublime Text or Nano preferred throughout this dissertation.

### ***3.ii. Social network development***

#### **Implementation decisions**

As said previously, requirements were defined before starting the development phase, review on philately specificities and common issues helped to understand the needs this application will answer to.

On a general matter, this website principal aim should be the possibility to create and manage a stamp collection. This website will be community-based, stamps management features should therefore be made available in order to involve users in the building of the knowledge base.

An emphasis on catalogs had be considered as vital as those are the key for the current state-of-the-art stamp identification methodology. Several rules have been applied to the database for duplicates prevention. A stamp cannot be linked twice to the same catalog while the a catalog identifier can only be assigned once, users will be incited to update existing entries rather than create a duplicate. These types of restrictions could not have been set on other characteristics such as stamps names, series, country or year as several legitimate duplicates exists.

Concerning social interactions, users should be able to share their data with friends, two options were considered: a synchronous, where both user need to accept the relationship for a bound to be created or an asynchronous implementation, where a user can decide to share his information to someone, without a compulsory reciprocity. The second option seemed more logical for this type of application, working in simultaneity with a notification process to the other user in order to promote reciprocity of information sharing.

An important choice in the design of the website is the will to give freedom to users. A great example is the interest display feature. A possible implementation could have been to guide the process and force a user to give precise information on the quality and quantity in which he would be interested as well as which type of stamp he is willing to trade back. Instead, a user only need to select a set of stamps he would be interested in and the two philatelists can proceed to the transaction, it allows them to settle on the kind of operation they want to do, sale or trade. On this matter, the

social network is more of a way to get in contact with fellow philatelists and does not get in the way of actual transactions. In this context users should be able to chat on the website or through emails. Unfortunately, this solution is not perfect, as traceability of trades could have been an interesting feature to offer.

A common problematic with websites allowing transactions between users is the possibility of scam, to reduce this issue a rating mechanism with the possibility to comment on the proceeding of each transaction has been set up. Users can then access to a trade-experiences history of other stamps-collectors and can decide if he is worth trusting. Some screen-shots of the application can be found in appendix B.

## **Database design**

Building a good database is a key requirement as the reactivity of the whole application will depend from it.

Insurance of a good access and modification speed as well as prevention of information redundancy, potential cause of update or deletion incoherence, have to be taken into during the design process.

Various types of items will need to be stored and relevant information have to be identified for each of those, an analysis of competitors and catalogs had been necessary and highlighted the need for a wide range of different information for stamp characterization. A good understanding of philatelists needs helped to select what should be prioritized.

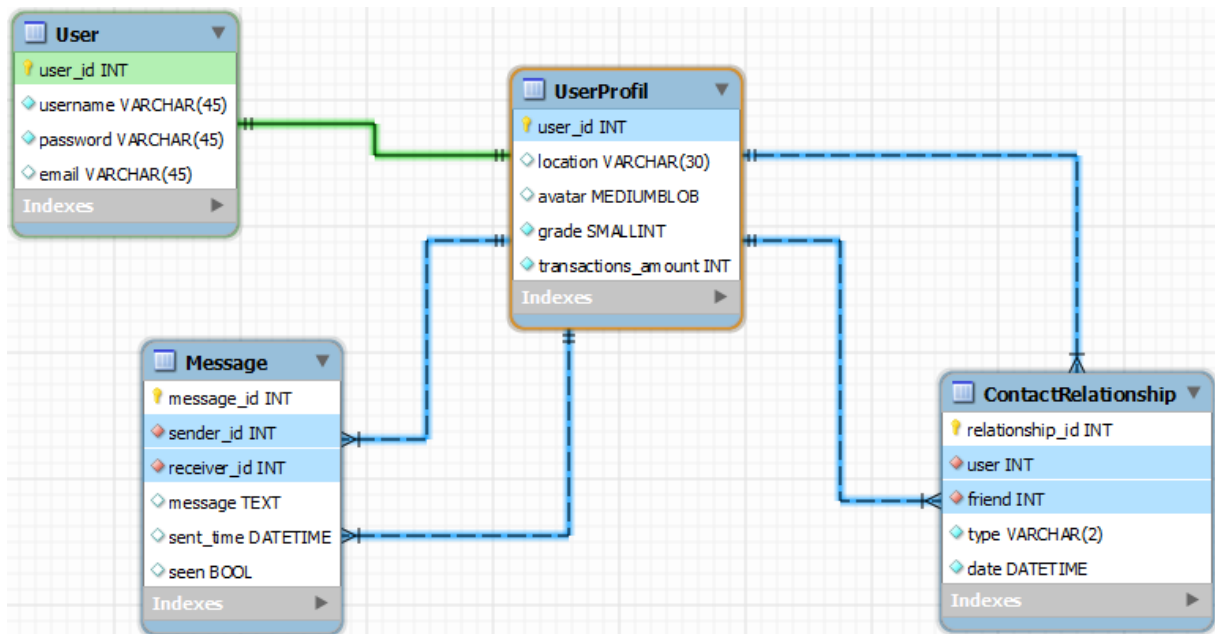
The choice had been made to centralize stamps information in a table common to every user, collection management is done with the help of a second table taking as key a user and a stamp. This choice helps prevention of systematic content duplication and will make the task of building a collection much faster for users.

As a social network, this website needs a table for users information. Due to the use of Django, an adaptation of the design is needed. In fact, Django comes with its own user management implementation including a table for users information, it is managed with specific functions and possess some security features such as passwords encryption. However, some important fields are missing. To sort this out, the creation of a second table is mandatory, bound with a one-to-one relationship.

Friendship is represented as a table with two foreign keys on user profiles in order to be adapted to the asynchronous solution presented earlier. Each row represent a unidirectional relationship, therefore if two users are both sharing their information

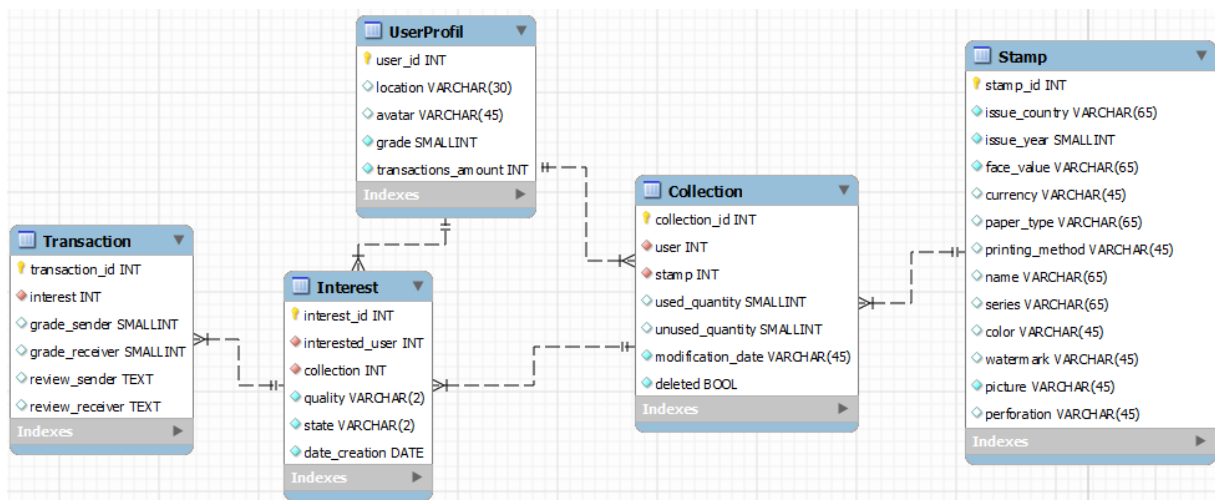


two rows will be necessary. As the possibility to converse with messages had to be implemented, a table was created to store discussions, with a boolean field to track if a message had been displayed already.



*Fig. 3: Database design for user interactions*

A choice had been made to create two tables for “interest” and “transaction”, where only one would have been necessary, the reason is that transaction informations are not needed when working with interests and the type of data stored might cause a delay in information display. The whole database design can be found appendix A.



*Fig. 4: Database design for transactions management*

### **3.iii. Data scraping**

#### **Tool selection justification**

As said in the research, the selected tool for data scraping is known as Scrapy, other alternatives are available in different programming language but for uniformity matters, using Python was a logical choice. We could have been tempted to gain some execution time by using a fast language such as C/C++, but the nature of a web scraper makes the difference insignificant as data retrieval is the most time-consuming part and processing time is not significant.

This specific tool has been selected for several reasons:

- Interfacing with Django allows direct interaction with the social network project and its databases, allowing the same level of data verification as for manually uploaded stamp. The process is speeded up as the use of an intermediary semi-structured file is not required.

- Data can be directly processed after the retrieval, allowing modification of information before its storage.

- Crawling and scraping functionalities using trustworthy and easy to learn technologies such as regular expressions and XPath.

- XPath ensure the correctness of retrieved data. Semi-automated retrieval systems, even if really efficient, give less control to the user. The level of complexity of the scraping task doesn't require a more powerful tool. Implementing a more intelligent method would only make the process heavier.

#### **Scraping experimentation**

I decided to launch a preliminary scrap in order to retrieve pictures needed for the training of my k-means during the image matching stage. The first step was to select the scraping source and the insight given by the market analysis helped in finding good databases amongst competitors. Investigation of a scraping feasibility on each potential target and ask for the web master permission. The main matter was to make sure stamps display pages were dynamically generated, ensuring a consistency in pages layout. The second point was to discover the optimal way to retrieve information without crawling through irrelevant pages, in order to avoid unnecessary traffic load on the website. It will be later used to guide the crawler throughout the website.

Colnect, a social network for stamp collection analysed during the market research, shows to be the most promising actor on Internet, this website possess not only a good architecture for social interactions, but also the biggest set of data amongst reviewed websites. For this reason, it had been selected as a priority source for data scraping. Some difficulties are worth noting: first, a question arose about which type of pages should be scraped, it would have been easy to select each stamp information page, but would have caused more load on servers. To reduce this issue, a deeper analysis of the website was necessary and showed that even if the website originally hide the possibility to display stamps by country when a certain quantity of stamps is reached on a specific country, making it unreachable for a crawler, the content of such page is dynamically generated, it is therefore possible to craft the right URL. This information is interesting as this type of page display stamps 6 by 6, which will greatly reduce computing costs for the target server.

Additionally, for some reason the crawler kept being redirected to the French version of the website, which caused errors with the XPath directions given as some points of reference were translated and therefore not recognized during the scraping step.

Due to the fact that this script is meant to be launched once, the solution has been to separate the crawler from the scraper, retrieve the URL to be scraped and store them in a file. By using the tool “sed”, occurrences of “fr” have been changed to “en” and the file was imputed to the scraping script.

In order to comply with the threshold of authorized content retrieval set by the website owner (5k stamps) and for an ethical scraping, the script had been stopped during its execution.

The most preoccupying issue with web scraping is the lack of control on the type of information in the target's database. The initial database for this project had strict, but close to reality normalization. However, in order to accept data from scraping, restrictions had to be lightened up. It raises questions on how to ensure the quality of data on this community-based website. Should more restrictions be applied, running the risk to set unrealistic rules potentially upsetting users willing to fill the database, or would it be better to accept less normalized data in order to comply with more users standards?

For this dissertation, the chosen option is the second, upon releasing this product,

the most important point will be to fill the database as fast as possible in order to motivate other users to join. Some modifications can be applied once the project is established, after an analysis of bothering recurring patterns a script can be used to modify data already present in the database in order to fit a stricter standard.

### ***3.iv. Content based image retrieval***

#### **Preliminary test: Descriptors efficiency**

The literature review related to features extraction algorithms allowed to identify a set of state-of-the-art solutions. However as we saw, algorithms efficiency is frequently dependant on pictures texture. In order to ensure a good selection of tools, some experimentation with different conditions related to the implementation specificities. An analysis of potential legitimate inputs from social network users showed that the algorithm should be able to manage several small modifications on a stamp picture. The two most likely modifications are the scale variation in case of a zoomed out picture submission and the blur caused if the picture is a close-up with a bad focus setting. Rotation, brightness and affine translation will be an issue for less powerful keypoints detectors, but should not be encountered too frequently in real-life conditions. The aim of this comparison is to highlight computation-cost efficiency of more recent algorithms while assessing quality of features detection.

A set of pictures with several variations had been taken representing various stamps in order to test efficiency of algorithms for each factor. To perform this set of tests, a library had been chosen, the renowned OpenCV, with its Python interface. It allows to easily experiment different algorithms and assess them in the same conditions, instead of looking for an original implementation of each method. Each stamp had an original picture, with a good view on its characteristics, and several other pictures implementing one or more of the alterations identified earlier.

A preliminary test had been performed where each picture of the pool was matched against each original picture, allowing to assess algorithms implementation efficiency on false positive and computation cost.

Computation times have been assessed, including detector and descriptor initialization, processing of the whole image base and matching of an image against the set. An average speed have been extracted from each iteration time, results can

be found in appendix C.

FREAK does not seem to be fully implemented as a crash of the script happened when initializing the detector. For this reason, it had been laid aside from this experimentation, however due to its close relationship with BRISK an estimation of the quality can be done. In case of outstanding results from this algorithm, another solution will be found to assess FREAK.

As matches are identified with a distance, a limit had to be set to preserve only the most consistent ones. After trying different values for each algorithm in order to retrieve the optimal limit for the identification of pictures consistency. The conclusion confirms information deduced from the literature review, SIFT is the most robust but also slower than corner detection based algorithms. A detail to be noted is the addition of SURF algorithm to the experimentation panel happened to be really disappointing as it revealed to be slower than SIFT. The hypothesis is a bad implementation of the algorithm. About ORB and BRISK, differentiation between relevant and irrelevant matches were slightly disappointing.

## **Real condition tests**

A real condition test was then designed, with the implementation of a bag of visual words mechanism, k-means was used for SIFT and its equivalent k-majority for binary shaped descriptors with the same values for quantity of iterations and vocabulary size. A set of scraped pictures had been used for vocabulary training in order to provide a good amount of features for clusters definition.

Several important points have to be assessed:

- Vocabulary generation time
- Robustness of matching
- Speed of matching in moderate sized databases (1k pictures)

Three tables had to be created for this purpose: a first one to store file names and assign an identifier, a second one where each row binds a picture to a visual word, several rows are needed to represent a full histogram. The last one is dedicated to histogram storing, each row contains an image histogram with the occurrence count of each visual word, stored as a pickled Python item generated after the quantization.

Matching relies on K-NN and is performed in two stages: first a list of images sharing several similar visual words is built and sorted in order to select only the most relevant. Then full histograms of each of these pictures are compared against the

queried picture and sort them by shortest distance.

## **K-means experimentation**

The use of k-means had been chosen for this project. The research revealed the presence of several criteria to set during the catalog initialization, directly impacting the implementation efficiency.

Amongst those, the number of visual words is the most important as it is the key for descriptors differentiation at later stages. The number of iteration also matters as it allows to refine visual words around existing descriptors. Finally, the number of images to use for training will provide a wider variety in descriptors. The issue to keep in mind is the computation time and a consensus need to be found.

For this assessment, the code created for the previous stage add been adapted to run several tests with a variations in those criteria.

## **System integration**

Once every important matter of this process had been assessed, it had to be integrated in the project. About database tables, only two are necessary as the relationship identifier/filename is already present on the form of the “stamps” table, list of visual words and histograms storing are still mandatory.

The functions developed for the test only need some slight modifications to be adapted to this design and can then be imported and called from the Django interface. The query interface needed to be adapted to accept images from a form as it used to select a picture already in the database.

The most noticed issue was the building of a bridge from a file located in the server memory to its OpenCV version because without this step image processing can not be performed. Uploaded image has to be stored temporary in a dedicated table before being processed by OpenCV.

This implementation plans the display of the 10 nearest matches, in order to be suited to cases where several stamps have a similar look with different colours.

An explanation of good practices should be set for users to understand the importance of a good representation of a stamp with as few defaults as possible, in order to praise to use of scanning and cropping, which should be the default method for optimal results.

### **3.v. *Dissertation conclusion***

#### **Technical achievement and original contribution**

Throughout the various stages of this dissertation, several objectives have been fulfilled. A social network implementation is proposed, with announced features such as creation of profiles and collection management. Interaction between users is enabled by the creation of an interest display feature allowing to get in contact with other philatelists. Transactions can be reviewed and commented to build a good reputation on the network. Community-based data gathering is supported by the addition of stamps management features.

Web scraping is explained and used on competitors, with their authorization, to build a typically sized database in order to show the solution on good working conditions.

Additionally a stamps identification system relies on the database to provide a Content Based image retrieval for the philately field, several methodologies have been assessed to ensure a quality of service. And literature review on indexing processes had been performed and a comparison between k-means algorithm and its equivalent dedicated to binary descriptors, k-majority.

#### **Evaluation and discussion**

The literature review highlighted the need for experimentation with various algorithms used in the image matching part of the project. Amongst those, the most important, impacting scalability and efficiency of the final product is the review of features extractors. The literature review highlighted a panel of algorithms to consider: SIFT, SURF, ORB, BRISK and FREAK.

The initial idea came from the observation of a high-computation costs on the two most used state-of-the-art methodologies, SIFT and SURF. Various researches highlighted this fact and claimed to provide a solution for high robustness and low computation features extractors algorithms. Proposed solutions needed to be assessed, as documentation on recent algorithms raised high expectations for a viable alternative to the famous SIFT-SURF duo.

The preliminary experience highlighted a high computation-time for SURF, that is not logical as this algorithm key argument is its speed efficiency compared to SIFT. The

hypothesis for these results is a bad implementation of the algorithm in OpenCV library, indeed, this experimentation's aim was to assess algorithms in a minimalistic environment, where results could not be altered by an external, unrelated, part of the code. Additionally, the code organization, which worked with a loop on identified algorithms was built in order to provide a similar environment for each iteration. For this reason SURF had been removed from assessed algorithm as SIFT was superior both in speed and matching quality.

SIFT results are outstanding for every type of variation, which is not the case for ORB and BRISK, frequently failing to recognize stamps when a different scale or a slight rotation was applied to the image. This was expected due to weaknesses inherent to these algorithms for this type of variation, but needed to be assessed in order to understand the scope of this issue.

Tests on real-life conditions, with the use of Bag of Words, happened to exacerbate features detection issues for ORB and BRISK. Frequently causing pictures of different scale or with a slight rotation to be ejected from the top 10 matches with a 250 images database. This problematic might also be caused by issues in the application of k-means to binary descriptors, known as k-majority, which is not currently widespread on Internet and therefore not covered in comparison publications. Flaws in the implementation are also an eventuality.

SIFT, on the other hand, performed quiet well for images matching. Its computation cost issue is lowered due to the fact that only one image needs to be processed before being matched to other pictures in the database. This issue is only preoccupying for addition of a large dataset of images to the database. On the bright side, the user does not expect an answer from this computation, which is different than for image matching queries. This stage can be performed locally if it reveals to be a handicap for user experience.

It is important to understand the need for a versatile algorithm as potential users are not photographers which might result in various variation on the representation of a stamp. Also, if the process of image preparation for matching is too complicated, users might just reject the feature, for this reason and for this specific case of use it has been decided to use SIFT for the final version, despite its computation cost on image addition to the database.



Concerning indexing methodologies, experimentations had been made with k-means as it provides a good quality vocabulary during the quantization process with an affordable initialization time for a one-time event. The literature review did not allow to evaluate a trend on certain factors selection such as vocabulary size, images training panel size and number of iterations. Tests had to be made in order to analyse best matching quality with a variation of these factors, where a good compromise between efficiency and speed had to be assessed.

A survey on size choice in research papers shows experimentations distributed over a wide range of vocabulary size. Experimentations confirmed the initial theory: a small vocabulary might lack of discriminative power, while a big vocabulary will be too precise and slightly altered descriptors might not be recognized and assigned to a wrong visual word (sensible to noise and blur).

Experimentations of image matching with high number of images (over 10k) showed to be really slow and resources demanding on the server. It can be easily explained by the structure of the implementation, each visual word present in the query image involve the retrieval of large set of data from the words table which is then sorted and processed. A good area of research for improvement is the implementation of a supervised learning models applied to images classification, such as SVM or Naïve Baye.

## **Conclusion**

This dissertation gave the opportunity to cover a wide range of subjects, from the development of a social network, for which philately's key particularities had to be understood. An adequate answer to typical stamps collectors needs in term of stamp identification had to be proposed. Alternatively, a solution to develop trading amongst the social network community had to be provided. Databases had been filled with the help of web scraping, the question of legal issues arose from the research.

Content-Based image retrieval systems had been pictured as an optimal solution for stamps identification and had therefore been implemented, experimentations with various algorithms showed the superiority of SIFT for this specific field of application. K-means had been used for images descriptors quantization, various parameters had been assessed to ensure a good balance between speed and results quality.

More time would have been necessary for an optimal application as some minor

issues still need to be assessed before a release, however key requirements are present in the current solution.

## **4. Critical appraisal**

### ***4.i. Research methodologies***

A specificity of this dissertation subject is the diversity of researches needed on the three different area of work, involving the creation of a website and the selection of algorithms or methodologies, various research methods had to be adopted for an efficient and adapted gathering of information.

The main concern for the social network was to suit philatelists needs and provide the right type of information and features. This matter could only be solved with a better understanding of the philately world for general matters, such as stamps identification methods. The second step was to study competitors to understand what had been done and deduce important requirements for this project as well as interesting ideas for stamps display and users interactions.

The research methodology for the image matching feature was different, the first step was to research publications covering the whole process, leading to the discovery and understanding of Content Based Image Retrieval and its state-of-the-art method: Bag of Visual Words. The type of resources is wide, it can take form of a website, a book or an actual research paper. Once the basics are understood, the main focus is to retrieve survey on currently used techniques, this step is important to determine historical actors and potential successors, several sources are necessary for more subjectivity. Publications where algorithms are compared to each other also helps for the selection. A list of most interesting algorithms can then be studied more efficiently by retrieving the original publication where specificities are explained. The final step consist in searching for implementation examples in order to make the development task easier.

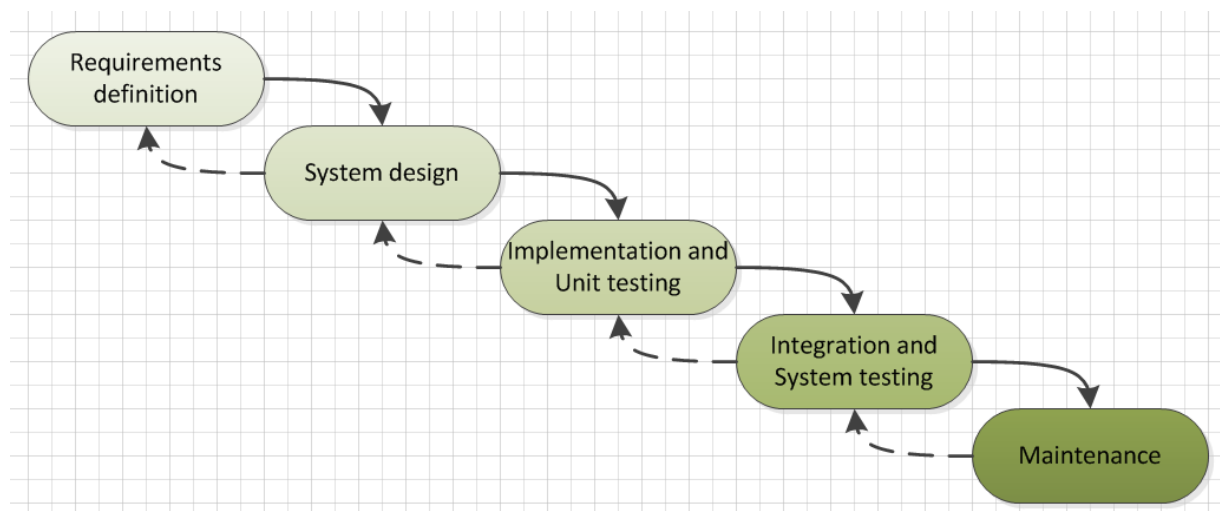
With the same idea, web scrapping principles and alternatives had to be understood before actually reviewing solutions, generalists websites or books are suited for this

matter. A survey on possible methods is also an asset for this part of the research in order to give more insight on the discipline before the selection of a method. Once done, searching for implementation or frameworks ensure an efficient work.

#### **4.ii. Development methodologies**

The waterfall model is a basic, but strong design process for project management.

It relies on a clear separation of each specific step of a project. Each stage being the foundation of the next one, as shown in the figure 5, additional steps such as feasibility study can be added for a project of bigger size, where costs and profit can be compared but this is not the case here as the project has no commercial purpose and does not aim to be profitable. This method is interesting as it ensure a good understanding of the objectives before actually starting the development phase. In an unsupervised environment it is quiet common that the development is started before knowing important requirements and testing phase is often underestimated.



*Fig.5: Waterfall method principle*

This method has some known issues, which are not necessarily incompatible with this project. Waterfall is known to be weak against requirement changes, in this case those are defined after an extended research phase in which potential modifications might have already been considered. Moreover as this project does not involve a customer, I have more freedom to decide whether or not a feature should be modified and when it is more suitable to do so.

A reason for the selection of this methodology is its simplicity, it is better to use a basic method rather than aim for a complex one that could be misunderstood and might delay the project completion.

Concerning the development of the social-network, the priority was to provide functionality before implementing a good user interface, in order to ensure a fast improvement of proposed features.

For web scraping and image matching feature, the initial development had been performed as standalone in order to avoid bugs and incompatibilities with the current environment. After making sure of the correct implementation of both pieces of software, those had been integrated to the working solution. It allowed a better focus on errors as less vectors of potential issues were present at the same time.

#### ***4.iii. Legal, ethical and environmental issues***

Intellectual property is one of the most common legal issue encountered in a project. In our case there are two specific situations we should be aware of, the first one is related to websites scraping. We want to be able to retrieve images and information from other websites, however, this may be unauthorized in the terms of use. Laws are still unclear about the liability of such terms and fair use of data may be a good argument in our favour, nevertheless, it would be more professional to make sure we follow the restrictions of authors before using their data. Only specific sources should be targeted and in the event of a clearly indicated restriction on information copy, an exceptional authorization can be asked through email.

This issue is worth considering as the initial idea was to use retrieved data to duplicate content on a competitor website, which is rather unethical. The proposed solution is to use scraped data to train and test the efficiency of the image matching algorithm in order to prove the viability of the concept.

However, the data should not be released as part the website, it could instead rely on two type of data sources: data extracted from the purchase of an existent database and community generated content. A question remains on the liability of the website if an external user decides to perform a scraping and automates the upload process.

The second scenario is bound to the users uploads, as those are stored in our server. We could face some legal issues if one of those is the intellectual property of another entity. To handle this issue, a message should be displayed to users explaining this

issue and preventing the upload of copyrighted resources. This is a common way of defence used by several other actors, another important point is to make sure it is easy to report this kind of issue in order to provide an efficient support.

A specific issue arose with the selection of a features extractor, the chosen algorithm is currently patented in US[63]. It is not as preoccupying in U.K. as patents filled in America are not valid there, but as this application might be used in US, the question needs to be raised. It is however highly unlikely that an academic project would be sued for the use of such an algorithm.

The tools selected for this project are open-source as it provides a better security and more control on the final product. As some licenses forces inheritance on the final product (such as the GPLv2), a specific attention will be given to the choice of the final deliverable licence. In any case, the project will be released as open-source in order to facilitate the spread of private websites dedicated to stamps collecting for small communities. This will also ensure a better durability for the project.

#### ***4.iv. Risk management***

During the proposal, several risks have been foreseen in order to plan an appropriate response in case of occurrence, those concerns the process of the project and have been rated by probability and potential impact, the list can be seen in this table:

<b>Lack of good sources for website scraping</b>	<b>Prob</b>	<b>Imp</b>
Chosen language and framework not suited for social media website	0.1	4
Database engine unable to manage a big enough amount of data	0.2	3
Incompatibilities of python 3 with desired framework, tools and web servers due to lack of maturity	0.2	5
Bad understanding of philatelists community needs	0.3	3
Tools and libraries not as efficient as expected	0.3	5
Lack of good scraping sources	0.4	3
Database design not adapted to social media	0.2	4
Incompatibility with famous web browsers	0.5	4

Methodology not suited for the project	0.3	2
Unrealistic amount of tasks to achieve	0.2	4

The first three potential risks have been avoided with a safe selection of tools and were therefore not preoccupying.

The second bloc represents three risks where more knowledge was required, with a deeper understanding of tasks to be achieved. Those revealed to be insignificant preoccupations after a correct time spent on the research.

The third block was important to keep in mind during the development of the solution, being aware of these potential risks helped to avoid underestimate the importance of these tasks and the specific care that should be brought.

The methodology selection was an issue in this project as I tried to apply a misunderstood project management methodology (v-model) during the first part of the project realization. This had been highlighted during the presentation and required a reaction. The waterfall model had been mentioned during this presentation, after a comparison of those two models, I concluded that waterfall was adapted to this project as it is safer to use a basic method rather than aim for a complex one that could be misunderstood and might delay the project completion.

However even by assessing those risks, the time planned in the Gantt chart had been underestimated, mostly due to the third area of research complexity where preliminary background review was not deep enough to correctly estimate required time for the assessment of various algorithms and methods used for image matching and data indexing. This lead to a rush on the last section of the project where more time would have been necessary to produce a quality content. I particularly regret not spending time experimenting on image content classifiers such as SVM or Naïve Bayes.

Some risks have been determined during the project process and concerns the future of the product:

*Rushed release:*

The release should not happen without certitudes on the web application stability and quality. Consequences would be fatal to the adoption of the service. A possible countermeasure would be to publish the social network

with an access restricted to a short panel of beta-testers in order to identify and fix potential bugs.

#### *Use of a patented descriptors*

This issue has been assessed during the project and the choice to prioritize patent-free algorithms had been made in order to prevent this potential threat.

#### *Illegal use of scraped data*

As explained earlier, scraped data should not be used in the final version of the application as it would be a violation of several laws. This risk should not be underestimated as it could cause the closure of the service and more serious consequences for its maintainer.

#### *Lack of interest for the project:*

This project will need a good amount of advertisement to build a user base, several philatelists communities have been identified on Internet and should be notified of this social network release.

To preserve enthusiasm for the solution, improvement suggestions and bug notifications should be frequently taken into account.

### **4.v. Professional development**

This dissertation has been a great way to develop some essential qualities both in the professional and academical world. It has been a good opportunity to experiment each step of a project, from how to conduct a research and perform a critical review able to highlight important points related to the project, to the redaction of this academic thesis. Several stages are usually underestimated by freshly-graduated workers such as the risks assessment as a way to foresee potential issues or the importance of time management for a long project, with the presence of frequent milestones for an early detection of delay.

Additional knowledge had been developed due to the particularities of my subject, the most influential for my career is the choice made to learn Python in the course of this dissertation. It will definitely be an asset as the type of position I envisage for the future, system and network administration, usually require experience in this scripting language, especially if it involves UNIX systems. Experimenting with Django and Scrapy also gives me more versatility in the type of tasks I can accomplish.

The different area of research explored throughout this dissertation gave me a better understanding on technologies such as indexing methods used in search engines,

various ways to improve results consistency and computation time and the different steps involved in image matching, as applied in object recognition in videos or images and computer vision, these might not directly impact my career but are still interesting to understand if a similar project is encountered.

Finally, data compilation using web scraping is an astounding technology with endless possibilities of applications, getting to understand stakes related to this technology as well as the controversy related to its use is an asset as its use is worth considering even as a hobby.

I think that most of all, the opportunity given to get involved on each stage of a project including software development will allow to get more prepared to the professional world with a better understanding of the whole process, an employee will be more efficient if he knows what is at stake and which choices led to the assignation of a task.

#### ***4.vi. Further improvement***

The final product sources will be released on an online platform dedicated to this purpose in order to facilitate the creation of small private communities around the product, this is a way to provide an answer to potential scam issues and to ensure an increased lifespan. Alternatively the product can be installed on a web server. If some users are interested by the platform, a support can be considered for new features development and bugs patches. Creating a community around this project would also be a great way to get recognized in the open-source field and to be more credible for future projects.

Further development should focus on a data modification control mechanism in order to avoid harmful infringement to the database integrity.

An interesting idea to help scaling to a big quantity of data would be to combine the use of metadata to the image matching features, it could be used as pre-filtering and would reduce computation time and required database queries. Alternatively a colour histogram matching feature could be built on top of the features detection process as none of the assessed algorithm takes into account coloured pictures, multiple examples of similar stamps with different colours can be found in stamps databases, especially concerning standard yearly issues.



An issue detected with the current user interface of the social network is the high demand in user action for the submission of a stamp to the image matching system. The user needs to take a picture, scan his stamp or find a version of the stamp on internet, before being able to upload it by browsing through his other files. An interesting idea to reduce this process would be to develop a smart phone application able to scan and extract interesting descriptors from a stamp before accessing the social network database in order to compare the matches and display the closest looking stamp. Due to the ever increasing computing power on new mobile devices and the recent discoveries in optimized image processing algorithms reducing the computing costs while preserving a decent robustness on detection, this idea can easily be implemented, numerous publication on internet shows the use of such platforms of development for this field of application.

#### **4.vii.      *Acknowledgment***

I would like to thanks my tutor, Prof. Hong Zhu, for this dissertation idea which gave me the opportunity to work on such interesting fields of Computer Science. His guidance also helped me to redefine my priorities throughout this dissertation milestones which greatly helped me to improve proposed content.

I am grateful to Amir Wald, Daniel Gil Horvat and Andreas Liverdos, owners of respectively: colnect.com, philateca.com and philatelism.com for their authorization to retrieve information on an automated way on their website. Thanks to their kindness I managed to gain experience with scraping technologies on different environments.

I would also like to thanks my family and friends for their patience and moral support during the redaction of this thesis.

## 5. References

- [1] Stamp Insider (2013) Magazine [Online] Available at: <http://www.stampinsider.org/>
- [2] S. Phillips, Stamp Collecting: A guide to modern philately, London, Stanley Gibbons Limited, 1983
- [3] J. Pettway, (2006) John W. Scott and the Evolution of the Scott Catalogue [Pdf]. Available at: <http://www.stampclubs.com/news/knoxville/nlaug06-2.pdf>
- [4] Stanley Gibbons (2013) Home page [Online] Available at: <http://www.stanleygibbons.com>
- [5] Yvert et Tellier (2013) Timbres [Online] Available at: <http://www.yvert.com/c-1-timbres.aspx> (French)
- [6] Arpin philately (2013) Collectible stamps [Online] Available at: <http://www.arpinphilately.com/products/stamps/index.html>
- [7] Stampnews (2013) Stamp collecting news for Beginners and Philatelists [Online] Available at: <http://www.stampnews.com>
- [8] Facebook (2013) Philately stamps collecting group [Online] Available at: <https://www.facebook.com/groups/129649707051024/>
- [9] Colnect (2013) Stamps [Online] Available at: <http://colnect.com/en/stamps>
- [10] Philateca (2013) Home page [Online] Available at: <http://www.philateca.com>
- [11] Philately (2013) Home page [Online] Available at: <http://www.philately.com>
- [12] Stanley Gibbons (2013) Home page [Online] Available at: <http://www.allworldstamps.com>
- [13] Andrew Alison (2013) Stamps of the world [Online] Available at: <http://www.stampsoftheworld.co.uk/>
- [14] K. Pol, N. Patil, S. Patankar, C. Das, "A Survey on Web Content Mining and Extraction of Structured and Semistructured Data", First International Conference on Emerging Trends in Engineering and Technology, 2008. p 543-546
- [15] S.K Malik, S.A.M. Rizvi, "Information Extraction using Web Usage Mining, Web Scrapping and Semantic Annotation", In 2011 International Conference on Computational Intelligence and Communication Networks (CICN), 2011, p. 465-469
- [16] B. Liu, "Editorial: Special Issue on Web Content Mining" In ACM SIGKDD Explorations Newslett. Volume 6 Issue 2, December 2004, p.1-4
- [17] Y. Zhai, B.Liu, "Extracting Web Data Using Instance-Based Learning", In Web Information Systems Engineering – WISE, 2005, p. 318-331
- [18] Muslea, I., Minton, S., Knoblock, C.: A hierarchical approach to wrapper induction. In: AGENTS '99: Proceedings of the third annual conference on Autonomous Agents. (1999)
- [19] Muslea, I., Minton, S., Knoblock, C.: Active learning with strong and weak views: A case study on wrapper induction. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-2003). (2003)
- [20] Knoblock, C.A., Lerman, K., Minton, S., Muslea, I.: Accurately and reliably extracting data from the web: a machine learning approach. (2003)

- [21]Kushmerick, N.: Wrapper induction for information extraction. PhD thesis (1997) Chairperson-Daniel S. Weld
- [22]Kushmerick, N. Wrapper Induction: Efficiency and Expressiveness. *Artificial Intelligence*, 118:15-68, 2000.
- [23]Hsu, C.N., Dung, M.T.: Generating finite-state transducers for semi-structured data extraction from the web. *Information Systems* 23 (1998)
- [24]Cohen, W., Hurst, M., Jensen, L.: A Flexible learning system for wrapping tables and lists in html documents. In: *The Eleventh International World Wide Web Conference WWW-2002*. (2002)
- [25]Chang, C.H., Lui, S.C.: Iepad: information extraction based on pattern discovery. In: *WWW '01: Proceedings of the 10th international conference on World Wide Web*. (2001) 681-688
- [26] Scrapy (2013) Welcome to Scrapy [Online] Available at: <http://www.scrapy.org/>
- [27] D.J. Cosby, "American Airlines, Inc vs. Farechase, Inc" Temporary injunction, 67<sup>th</sup> district court, Tarrant County, TX (2003)
- [28] H.L. Hupp, "Ticketmaster Corp. v. Tickets.com, Inc.", supra2000 U.S. Dist. Lexis 12987 (August 2000)
- [29] L.M. Brinkema, "Cvent, Inc vs. Eventbrite, Inc", Memorandum opinion - US District Court, Alexandria, Virginia, 2010
- [30] Imperva (2011) Detecting and Blocking Site Scraping Attacks [Pdf] Available at: [http://www.imperva.com/docs/WP\\_Detecting\\_and\\_Blocking\\_Site\\_Scraping\\_Attacks.pdf](http://www.imperva.com/docs/WP_Detecting_and_Blocking_Site_Scraping_Attacks.pdf)
- [31] DMH Stallard (2013)Data scraping [Online] Available at: [http://www.dmhstallard.com/site/home/data\\_scraping\\_jennings](http://www.dmhstallard.com/site/home/data_scraping_jennings)
- [32] Electronic Frontier Foundation (2013) Chilling effects: Frequently Asked Questions (and Answers) about Linking [Online] Available at: <http://www.chillingeffects.org/linking/faq.cgi#QID596>
- [33] C.F. Tsa, "Bag-of-Words representation in Image Annotation: A Review" In *ISRN Artificial Intelligence*, Volume 2012, Article 3768
- [34] Babbar, G., Bajaj, P., Chawla, A. & Gogna, M., "Comparative Study Of Image Matching Algorithms" *International Journal of Information Technology and Knowledge Management*, 2010, Volume 2, No. 2, pp. 337-339
- [35] G. Maridalia, "A comparative study of three image matching algorithms: SIFT, SURF, and Fast" M.S. Thesis, Dept. Civ. and Env. Eng., Utah State Univ., 2011
- [36] E. Rosten, T. Drummond, "Machine learning for high speed corner detection", In *9th European Conference on Computer Vision*, Volume 3951 (2006), p 430-443
- [37][BRISK]
- [38] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, "ORB: An efficient alternative to SIFT or SURF" In *International Conference on Computer Vision*, Barcelona (2011) p. 2564-2571
- [39]E.Mair, G.D.Hager, D.Burschka, M. Suppa, and G.Hirzinger. Adaptive and generic corner detection based on the accelerated segment test. In *Proceedings of the European Conference on Computer Vision(ECCV)*, 2010
- [40] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 60, 2 (2004), pp. 91-110
- [41] David G. Lowe, "Object recognition from local scale-invariant features," *International Conference on Computer Vision*, Corfu, Greece (September 1999), pp. 1150-1157
- [42] Brown, M. and Lowe, D.G. 2002. Invariant features from interest point groups. In *British Machine Vision Conference*, Cardiff, Wales, pp. 656-665
- [43] Bay, H., A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)" *Computer Vision ECCV 2006*, Vol. 3951. *Lecture Notes in Computer Science*. p. 404-417

- [44] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, p. 1615–1630, 2005.
- [45] Ke, Y., and R. Sukthankar, “PCA-SIFT: A more distinctive representation for local image descriptors *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2004) p506-513
- [46] Luo. J, Oubong G., “A comparison of SIFT, PCA-SIFT and SURF”, *International Journal of Image Processing (IJIP)*, Vol. 3, No. 4. (2009), pp. 143-152
- [47] Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching” *IEEE Conference on Computer Vision and Pattern Recognition*, (2007) p1-8
- [48] Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. *Computer Vision–ECCV 2010*, p. 778–792, 2010.
- [50] levgen Khvedchenia (2011) “Feature descriptor comparison report” [Online] Available at: <http://computer-vision-talks.com/2011/08/feature-descriptor-comparison-report/>
- [51] levgen Khvedchenia (2012) “A battle of three descriptors: SURF, FREAK and BRISK” [Online] Available at: <http://computer-vision-talks.com/2012/08/a-battle-of-three-descriptors-surf-freak-and-brisk/>
- [52] C. Schaeffer, “A Comparison of Keypoint Descriptors in the Context of Pedestrian Detection: FREAK vs. SURF vs. BRISK”, Academic project, Dept Comput. Sci., Stanford University, CA, 2012
- [53] E. Oyallon, J. Rabin, “An analysis and implementation of the SURF method, and its comparison to SIFT”, unpublished
- [54] J.A. Hartigan. “Clustering algorithms” John Wiley & Sons, Inc. (1975)
- [55] Hartigan, J. A.; Wong, M. A. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm". *Journal of the Royal Statistical Society, Series C* 28 (1): 100–108. JSTOR 2346830
- [56] Y. G. Jiang, J. Yang, C. W. Ngo, and A. G. Hauptmann, “Representations of keypoint-based semantic concept detection: a comprehensive study,” *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 42–53, 2010.
- [62] J. Liu, “Image retrieval based on Bag-of-Words model”, arXiv preprint (1304.5168), 2013
- [57] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.
- [58] J. Wang, J. Wang, Q. Ke, G. Zheng, S. Li, “Fast Approximate K-means via Cluster Closure” In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, p. 3037-3044
- [59] David Nistér and Henrik Stewénus. “Scalable recognition with a vocabulary tree” In *CVPR* (2006), p. 2161–2168.
- [60] C. Grana, D. Borghesani, M. Manfredi, R. Cucchiara. “A fast approach for integrating ORB descriptors in the bag of words model” In *Proc. SPIE 8667, Multimedia Content and Mobile Devices*, 866709 (2013)
- [61] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR*, pages 494–501, 2007.
- [62] A. Holovaty, J. Kapan-Moss. (2013). *The Django Book – Chapter 5: Models (2.0)*[Online]. Available at: <http://www.djangobook.com/en/2.0/chapter05.html>
- [63] D.G. Lowe, “Method and apparatus for identifying scale invariant features in an image and use of same for locating an object in an image” U.S. Patent 6711293 B1, March 23, 2004.

## Appendix A: Database design

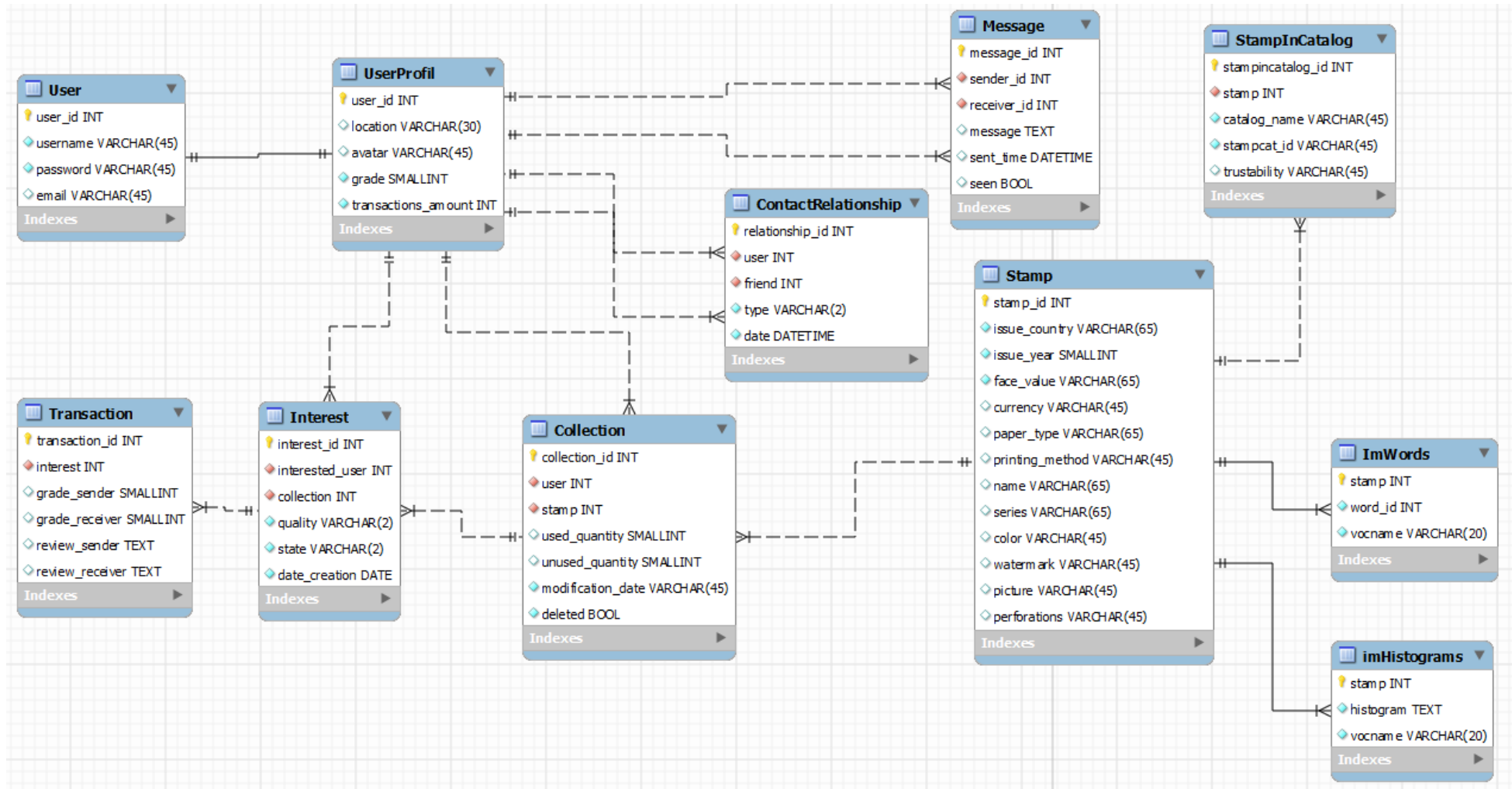


Fig. 6: Database design

## Appendix B: Web application

MyPhilately - A social network for stamps lovers!

Profile Messages Friends Interests Collection Upload Finder Stamps database Logout

Profile

Username: thesis  
Location: France  
Email: thesis\_myphilately@yopmail.com  
Grade: 52(5) [Edit](#)

Search for an user  
Username:  [Search](#)

Pending request  
[Camille - July 12, 2013](#) [Accept](#) [Deny](#)

Recent additions

Name	Series	Used Stamps	Unused Stamps	Addition Date
Peak Pahia, Bora Bora	Oceanie	0	4	Sept. 27, 2013
Views	Oceanie	1	0	Sept. 27, 2013
Sales tax collectable	Oceanie	1	1	Sept. 27, 2013
Overprint SELF GOVERNMENT	King George VI and Views, optd	0	2	Sept. 27, 2013
E F O 1915	Oceanie	0	1	Sept. 27, 2013
Channel Fataoua	Oceanie	0	1	Sept. 27, 2013
Red Cross	Oceanie	1	0	Sept. 27, 2013
Tahitian	Oceanie	0	1	Sept. 27, 2013

Your interests

Pending interest

Name	Series	Face Value	Interested User	State	Accept
Sales tax collectable	Oceanie	10	Epsilon	PE	<a href="#">Confirm / Cancel</a>
Channel Fataoua	Oceanie	5French centime	Epsilon	PE	<a href="#">Confirm / Cancel</a>
Tahitian	Oceanie	10French centime	Omega	PE	<a href="#">Confirm / Cancel</a>

Fig.7: Profile page

MyPhilately - A social network for stamps lovers!

Profile Messages Friends Interests Collection Upload Finder Stamps database Logout

Your MailBox

User	Message	Time sent
<a href="#">Omega</a>	Oh, sure, go ahead and tell me what could interest you!	July 11, 7:07 p.m.
<a href="#">Camille</a>	Hello Thesis! How are you?	July 12, 2:15 p.m. New message(s)
<a href="#">Presentation</a>	Hello Thesis, could you share your info with me? Cheers	July 12, 2:50 p.m.

Fig.8: Messages page

MyPhilately - A social network for stamps lovers!

Profile Messages Friends Interests Collection Upload Finder Stamps database Logout

Profile

Username: thesis  
Location: France  
Email: thesis\_myphilately@yopmail.com  
Grade: 52(5) [Edit](#)

Search for a friend  
Username:  [Search](#)

Shared

Username	Addition date	Type
<a href="#">Epsilon</a>	June 28, 2013	AC <a href="#">Delete</a>
<a href="#">Omega</a>	June 28, 2013	AC <a href="#">Delete</a>
<a href="#">Sigma</a>	June 28, 2013	AC <a href="#">Delete</a>
<a href="#">Lambda</a>	June 27, 2013	AC <a href="#">Delete</a>

Friends

Username	Addition date	Type
<a href="#">Camille</a>	July 12, 2013	AC
<a href="#">Presentation</a>	July 12, 2013	AC
<a href="#">Lambda</a>	June 28, 2013	AC
<a href="#">Omega</a>	June 28, 2013	AC
<a href="#">Epsilon</a>	June 27, 2013	AC

Fig.9: Friends management page

**MyPhilately** - A social network for stamps lovers!

Profile Messages Friends **Interests** Collection Upload Finder Stamps database Logout

History

State	User	Grade	Name	Series	
Received	Epsilon	100	C.E.P.T.- Children's Play	C.E.P.T.- Children's Play	<a href="#">See review</a>
Sent	Epsilon	100	Peak Pahia, Bora Bora	Oceanie	<a href="#">See review</a>
Sent	Epsilon	100	Red Cross	Oceanie	<a href="#">See review</a>

Next Current : 1 Previous

Trade completed

State	User	Grade	Review	Name	Series
Received	Epsilon	100	<a href="#">Review</a>	Vikings	
Sent	Omega	N/A	Reviewed	E F O 1915	Oceanie
Sent	Omega	N/A	<a href="#">Review</a>	Sales tax collectable	Oceanie

Ready to trade

Type	User	Name	Series	
ToReceive	Omega	Anniversary of Victory	King George VI - Commemoratives	<a href="#">Confirm</a> / <a href="#">Cancel</a>
ToReceive	Omega	Overprint	Overprint	<a href="#">Confirm</a> / <a href="#">Cancel</a>
ToReceive	Omega	10th Anniversary of Liberation	Oceanie	<a href="#">Confirm</a> / <a href="#">Cancel</a>
ToSend	Omega	Peak Pahia, Bora Bora	Oceanie	<a href="#">Confirm</a> / <a href="#">Cancel</a>

Your interests

Name	Series	Current Owner
Uniforms	Uniforms	Epsilon
U.I.T.		Epsilon
Type Groupe	Oceanie	Omega
Heraldry	Heraldry	Omega
Peak Pahia, Bora Bora	Oceanie	Omega

Pending interest

Name	Series	Interested User
------	--------	-----------------

Fig.10: Trades management

**MyPhilately** - A social network for stamps lovers!

Profile Messages Friends Interests Collection Upload Finder **Stamps database** Logout

Search criteria

Search:  Country of issue ☐

Stamps list

Name	Series	Issue Country	Issue Year	Face Value	Catalogs
<a href="#">Peak Pahia, Bora Bora</a>	Oceanie	French colonies and territories	1955	13	YT: FR-OCE PA32
<a href="#">10th Anniversary of Liberation</a>	Oceanie	French colonies and territories	1954	3	YT: FR-OCE PA31
<a href="#">Channel Fataoua</a>	Oceanie	French colonies and territories	1929	60French centime	YT: FR-OCE TA14
<a href="#">Sales tax collectable</a>	Oceanie	French colonies and territories	1948	30	YT: FR-OCE TA19
<a href="#">Sales tax collectable</a>	Oceanie	French colonies and territories	1948	10	SN: FR-OCE J18 YT: FR-OCE TA18
<a href="#">Maori</a>	Oceanie	French colonies and territories	1929	3	YT: FR-OCE TA17
<a href="#">Nafea Faa, ipoipo</a>	Oceanie	French colonies and territories	1953	14	YT: FR-OCE PA30
<a href="#">75th Anniversary of the Universal Postal Union (UPU)</a>	Oceanie	French colonies and territories	1949	10	YT: FR-OCE PA29
<a href="#">Channel Fataoua</a>	Oceanie	French colonies and territories	1929	50French centime	YT: FR-OCE TA13
<a href="#">Channel Fataoua</a>	Oceanie	French colonies and territories	1929	30French centime	YT: FR-OCE TA12


Fig.11: Stamps list

MyPhilately - A social network for stamps lovers!
Profile
Messages
Friends
Interests
Collection
Upload
Finder
Stamps database
Logout

Common information
Oceanie - Peak Pahia, Bora Bora
French colonies and territories - 1955
Face value: 13 F

Advanced information
Printing method:
Color: Black blue

Catalogs
YT: FR-OCE PA32



Ownership
Unused quantity:
Used quantity:

Fig.12: Stamp display

MyPhilately - A social network for stamps lovers!
Profile
Messages
Friends
Interests
Collection
Upload
Finder
Stamps database
Logout

Epsilon' review
Grade: 100
Review:
Good trade, thanks

Common information
Oceanie - Red Cross
French colonies and territories - 1915
Face value: 10+5French centime

Advanced information
Printing method: 1

thesis' review
Grade: 100
Review:
Good job




Fig.13: Trade review





Fig.14: Example image matching features

```
myphilately=# select count(*) from stamps_stamp;
count
-----
12132
(1 row)
```

Fig.15: Result web scraping

## Appendix C: Preliminary tests report

	Average		SIFT	ORB	BRISK	
SIFT	8.7818213872	golf-similar2.jpg	1813	377	122	
SURF	12.117571592	golf-scalemod.jpg	1073	107	66	
ORB	0.6973971639	golf-gauss2.jpg	972	366	72	
BRISK	1.3671344348	golf-rot.jpg	975	127	64	
<i>Fig.16: Average time</i>		golf-rot-gauss.jpg	705	140	37	
		golf_similar-unzoom.jpg	516	177	10	
		golf_unzoom.jpg	504	85	24	
		golf-smallrot.jpg	178	132	17	
		golf_unzoom2.jpg	156	37	3	
		golf_small.jpg	153	39	1	
		golf_rot135.jpg	154	48	5	
		golf_rot90+blur.jpg	15	12	2	
		random1.jpg	10	19	16	
		random2.jpg	0	6	2	
		random3.jpg	0	35	4	
		random4.jpg	3	11	4	
		random5.jpg	32	5	10	
		random6.jpg	7	16	12	
		random7.jpg	1	26	6	
		random8.jpg	20	7	12	
		random9.jpg	86	23	65	
		random10.jpg	6	22	2	
		random11.jpg	4	15	16	
		random12.jpg	4	25	4	
		random13.jpg	19	37	17	
		random14.jpg	6	12	6	
		random15.jpg	6	27	14	
		random16.jpg	1	27	1	
		random17.jpg	11	17	10	
		random18.jpg	7	16	23	
		random19.jpg	9	22	15	
		random20.jpg	51	15	68	
		random21.jpg	4	31	0	
		random22.jpg	4	12	11	
		random23.jpg	67	13	34	
		random24.jpg	17	28	24	
		random25.jpg	3	26	4	
		random26.jpg	32	20	14	
		random27.jpg	40	20	56	
		random28.jpg	20	29	25	
		random29.jpg	4	25	8	
		random30.jpg	73	12	78	
		random31.jpg	0	15	2	
		random32.jpg	24	2	2	
		random33.jpg	4	18	9	
		random34.jpg	31	32	46	

*Fig.17. Typical results for preliminary tests*