Troisème partie

# 3. Analyse des besoins des utilisateurs

Dans le but de mieux cerner les besoins des utilisateurs en matières de description des manuscrits arabes, un questionnaire a été utilisé. Les résultats de cette enquête ont permis de définir les métadonnées nécessaires pour l'accès aux manuscrits dans une base de données. Notre but est de rassembler le plus possible d'idées qui nous permettent de créer les métadonnées les plus pertinentes et les plus compréhensibles possibles. Pour atteindre notre objectif, nous avons distribué personnellement dans une conférence à Venise des copies de questionnaires à des collègues qui travaillent dans le domaine des manuscrits arabes. Egalement, nous avons distribué des copies pendant un mois à la BNF (Bibliothèque Nationale de France) par l'entremise de Madame la Directrice de la division des manuscrits orientaux et de Madame Geneviève Guesdon. Nous avons aussi diffusé le questionnaire sur l'Internet par l'intermédiaire du groupe de discussion Liste-Middle East.

Pour identifier les réponses, nous les avons numérotées (Q1, Q2, etc.) sans aucun ordre particulier.

A partir des 15 réponses que nous avons reçues, nous voudrions, dans ce chapitre, analyser les réponses selon les aspects suivants :

- Une étude brève sur les catégories d'utilisateurs (leur âge, leur nationalité et leur profession) pour montrer que, malgré le faible nombre de réponses, la qualité de ceux qui ont répondu est réelle. Tout cela pour montrer la pertinence de notre conclusion.
- L'étude des manuscrits vue par les utilisateurs du point de vue de la typologie, de la codicologie, de la paléographie et de la mise en page.
- La recherche d'information manuelle : les problèmes et les perspectives.
- La numérisation des manuscrits. Ce qu'on attend d'une recherche électronique de document à partir des manuscrits numérisés.

### 3.1.1. Les utilisateurs

| Réponse | Nationalité | Profession | Age de l'utilisateur | | | | | Sexe | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 20-30 | 31-40 | 41-50 | 51-60 | 60- | F | M |
| Q1 | Française | Professeur d'université | | | X | | | | X |
| Q2 | Américaine | Bibliothécaire | | | X | | | | X |
| Q3 | Grecque | Maître de conférences et arabisant | | | X | | | X | |
| Q4 | Canadienne | Erudite | | | | X | | | X |
| Q5 | Palestinienne | Chercheur, étude islamique | X | | | | | | X |
| Q6 | Tunisienne | Directeur de recherche | | | | | X | | X |
| Q7 | Belge | Maître de conférences | | X | | | | | X |
| Q8 | Anglaise | Conservateur de bibliothèque | | X | | | | | X |
| Q9 | Allemande | Bibliothécaire, spécialisés en littérature arabe | | | X | | | X | |
| Q10 | Française | Directeur de recherche | | | | X | | | X |
| Q11 | Marocaine | Maître de conférences | | | X | | | | X |
| Q12 | Française | Chercheur CNRS | | | | X | | | X |
| Q13 | Italienne | Chargé d'enseignement | | X | | | | | X |
| Q14 | Française | Etudiante | X | | | | | X | |
| Q15 | Française | Maître de conférences | | | | X | | | X |

**Tableau n°. 19 : L'ensemble des réponses au questionnaire**

A propos du tableau ci-dessus, on peut faire les remarques suivantes :

- A partir de notre échantillon, nous avons un assemblage assez riche de nationalités : onze nationalités différentes, dont cinq françaises (4 homme et une femme). Il y a aussi une allemande, un américain, un belge, un britannique, une grecque, un italien, un marocain, un palestinien et un français d'origine tunisienne.
- Treize sur quinze ses trouvent dans la tranche d'âge de 31 à 60 ans, ce qui indique que la plupart ont des expériences assez riches dans le domaine de manuscrits.
- Par conséquent, leur profession ainsi que leur expérience donne une base très riche pour établir la description nécessaire à l'accès aux manuscrits numérisés.

### 3.1.2. Les études de manuscrits

| Réponse / Domaine d'intérêt | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 | Q13 | Q14 | Q 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| L'édition de texte | X | X | X | X | X | X | X | X | | | X | X | X | | X |
| La consultation de texte | X | X | | X | | X | X | X | | X | | | X | | X |
| L'histoire des collections | | X | | X | X | | X | X | | X | | | X | | |
| La paléographie | | X | | | | X | X | X | X | | | X | | | |
| La codicologie | | | | | | | X | X | X | | | X | X | | |
| L'histoire de l'art | | X | | | | | | | | | | | | X | |

**Tableau n°.20 : Les domaines d'intérêt dan l'étude codicologique.**

Du tableau ci-dessus, on peut pointer les domaines qui intéressent le plus les utilisateurs dans leur étude des manuscrits arabes:

- La majorité de notre échantillon (12) sont intéressée par l'édition de textes manuscrits ;
- En deuxième lieu vient la consultation de textes manuscrits (9) ;
- En troisième lieu, l'histoire des collections (7) ;
- En quatrième lieu, la paléographie(6) ;
- En cinquième, la codicologie (5) ;
- En sixième et dernière étape vient l'histoire de l'art (2).

### 3.1.2.1. Autres centres d'intérêt lors de l'étude d'un manuscrit

| **Centres d'intérêt lors de l'étude d'un manuscrit** | | **Première priorité** | **Deuxième priorité** | **Troisième priorité** | **Quatrième priorité** |
|---|---|---|---|---|---|
| *Le frontispice* | | Q1, Q6, Q7, Q8, Q13, Q14 | Q5 | Q9 | |
| *L'illustration* | | Q5, Q9 | Q8, Q13, Q14 | | |
| *Le décor dans le texte* | | | Q1, Q7, Q9 | Q8, Q13, Q14 | |
| Le décor de la reliure | | | | Q1, Q7, | Q9, Q14 |
| Autres | Q3 | Suite de textes dans un manuscrit donné | | | |
| | Q4 | Le contenu du texte | | | |
| | Q5 | L'édition du texte et la focalisation du manuscrit concerné sur la terre sainte, la Palestine, et plus précisément Jérusalem. | | | |
| | Q8 | Pas d'un point de vue d'histoire de l'art, mais seulement pour en repérer la provenance et la date ; aussi pour des raisons de calligraphie (usage des enluminures)<br>"Not from an art history point of view, but only as indicators for provenance and date; also scribal function (use of illumination)" | | | |
| | Q10 | Le contenu scientifique du document | | | |
| | Q12 | Histoire du texte, l'histoire et la technique de fabrication du livre, paléographie | | | |
| | Q13 | Les incipit des textes | | | |
| | Q14 | Le texte, la légende des images | | | |
| | Q15 | Le contenu en relation avec la pensée arabe classique | | | |

**Tableau n°.21 : Autres centres d'intérêt lors de l'étude d'un manuscrit**

Concernant l'étude du contenu des manuscrits, on peut distinguer deux catégories différentes :

La première catégorie comprend les aspects qui ont été considérés comme prioritaires par les utilisateurs et que nous avons pris comme base. En deuxième catégorie, il y a les autres aspects ajoutés par les personnes interrogées.

- Pour la première catégorie, on constate que les *frontispices* prennent la première place dans l'intérêt de l'utilisateur, peut-être à cause de leur richesse au niveau

couleur aussi bien qu'au niveau du style. L'*illustration* et le *décor dans le texte* viennent à la deuxième et à la troisième place, c'est une indication que ces deux aspects sont aussi importants dans l'étude d'un manuscrit et que nous avons besoin de les prendre en considération dans la création de métadonnées. Bien que le *décor de la reliure* vienne à la quatrième place, c'est-à-dire en dernière priorité dans notre échantillon, cela représente encore un pourcentage de 27% de l'échantillon ; nous ne pouvons donc pas le marginaliser non plus.

- Dans la deuxième catégorie, on trouve huit aspects nouveaux ajoutés par les utilisateurs. On peut les regrouper selon les catégories suivantes :
  - Le contenu du texte est mentionné dans les réponses Q3, Q4, Q5 Q10, Q14 et Q15. Dans ces réponses, le contenu des textes est considéré suivant différents point de vue, le contenu scientifique du document, les aspects philosophiques (exemple : La pensée arabe classique), la légende des images. Egalement pour l'auteur de la réponse Q5, qui est aussi intéressé par le contenu du texte mais surtout en ce qui concerne Jérusalem.
  - L'histoire du texte et la technique de fabrication
  - L'incipit des textes.
  - L'histoire de l'art comme des indicateurs de provenance, et pour dater le manuscrit comme l'indique Q8 (Not from an art point of view, but only as indicators for provenance and date; also scribe function (use of illumination) Pas d'un point de de l'histoire de l'art, mais seulement pour en repérer la provenance et la date ; aussi pour des raisons de calligraphie (usage des enluminures)

#### 3.1.2.2. La catégorie du manuscrit

Les manuscrits arabes, comme nous l'avons dit dans le chapitre qui concerne la description des manuscrits arabes, sont classés en deux catégories : les manuscrits arabo-islamiques et les manuscrits arabo-chrétiens avec, pour chacun, ses propres caractéristiques. Notre but ici est de savoir quel pourcentage est intéressé par l'un ou

l'autre de deux catégories, afin de prendre cela en considération dans la création des métadonnées nécessaires.

| Question / Catégorie de mss | Q 1 | Q 2 | Q 3 | Q 4 | Q 5 | Q 6 | Q 7 | Q 8 | Q 9 | Q10 | Q11 | Q12 | Q13 | Q14 | Q15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Arabo-Islamique | X | X | | X | X | X | X | X | X | X | X | X | X | X | X |
| Arabo-Chrétiens | | X | X | | X | | | | | | | | X | | |
| **Tableau n°.22 : La catégorie du manuscrits** | | | | | | | | | | | | | | | |

Dans le tableau ci-dessus, on voit que 14 utilisateurs sur 15 sont intéressés par les manuscrits arabo-islamiques, alors que seulement quatre (27%) manifestent un intérêt pour l'autre catégorie ( arabo-chrétiens). Parmi les quatre dernières réponses, il y en a une (Q3) qui dit un intérêt pour les manuscrits arabo-chrétiens seulement. Les trois autres (Q2, Q5 et Q13) sont concernés par les deux catégories.

### 3.1.2.3. La typologie des manuscrits

| Typologie / Réponse | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 | Q13 | Q14 | totale |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Coranique** | | X | | X | | X | | X | X | | | | | | 5 |
| **Autres textes religieux** | X | X | | X | | | X | X | | | X | | X | | 7 |
| **Scientifiques** | | | | | X | | | X | X | | | X | X | X | 6 |
| **Littéraires** | | X | X | | X | | X | X | X | | | X | X | X | 9 |

| | | |
|---|---|---|
| **Autres** | Q2 | Documentaire |
| | Q6 | Méthodologie juridique musulmane au moyen age et plus particulièrement dans l'Andalousie et le Maghreb |
| | Q8 | Toutes les typologies mentionnées ci-dessus pour le besoin de catalogages |
| | Q10 | Philosophiques |
| | Q11 | Historiques |
| | Q12 | Grammaticales |
| | Q15 | Mystique et philosophie arabo-islamique |

**Tableau no 23 : La typologie des manuscrits**

- Neuf réponses concernent le domaine de la *littérature*, ce qui à mon avis est très logique car les manuscrits sont les sources premières qui englobent toutes les richesses de la littérature arabe.
- Sept réponses concernent *l'autre texte religieux* tels que le *hadith*, le *tafsir* etc.
- Malgré sa richesse surtout en médecine, pharmacie, astrologie, etc, le texte scientifique occupe la troisième place dans la priorité des réponses (6 réponses)
- Le texte coranique lui-même se situe en dernière place (5 réponses)

En revanche, on peut ajouter d'autres domaines qui sont proposés par les répondants, comme les suivants :

- Juridique musulman en Andalousie et au Maghreb (proposé par Q6) ;
- La mystique et la philosophie islamique, selon Q15 ;
- L'histoire par Q11 ;
- La philosophie Q10 ;
- La grammaire de la langue arabe (Q12) ;
- Et finalement pour Q2 le documentaire

### 3.1.2.4. La période historique des manuscrits

En réponse à cette question, nos interlocuteurs, nous ont donné deux catégories de données différentes : La première est purement historique, par période, et la deuxième à la fois sujet et période.

| La période Historique | Numéro de questionnaire |
|---|---|
| Islamique | Q4 |
| IV°-X° siècles | Q11 |
| VI°-VII° | Q6 |
| Médiévale Islamique (VII-XV) | Q1, Q8 |
| VIII°-XVI° (700-1500) | Q9 |
| X°-XX° | Q7 |
| XII°-XIV° | Q6 |
| XII-XV | Q15 |
| Toutes les périodes | Q14 |
| La codicologie, IX°-XVI° | Q12 |
| Jérusalem islamique et pré-islamique | Q5 |
| Paléographie (toutes les périodes) | Q12 |
| Les textes grammaticaux IX°-XVI | Q12 |

**Tableau n°. 24 : La période historique proposée**

- La première catégorie : les dix réponses de Q4, Q11, Q6, Q1, Q8, Q9, Q7, Q6 Q14 et Q15, spécifient bien les périodes d'intérêt, sans donner aucune indication sur le domaine.
- La deuxième catégorie : deux de nos réponses (Q5 et Q12) spécifient bien leur domaine d'intérêt par rapport à la période historique.
    - Pour Q5, en tant que palestinien, il est bien évidemment intéressé par tous les manuscrits spécialisés sur Jérusalem pendant la période islamique et pré-islamique.
    - Alors que Q12 est intéressé par la codicologie et les textes grammaticaux entre le IX°-XVI° siècles et en ce qui concerne la paléographie, à toutes les périodes.

### 3.1.2.5. Les objectifs des chercheurs pour l'étude de manuscrits

| Reponses / Objectif de recherche | | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 | Q13 | Q14 | Q15 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Comparer les titres de plusieurs manuscrits | | **X** | | **X** | | | **X** | **X** | **X** | **X** | **X** | | **X** | **X** | | **X** | **10** |
| La formation du nom de l'auteur dans plusieurs manuscrits | | | | | | **X** | **X** | **X** | X | | **X** | **X** | **X** | **X** | | **X** | **9** |
| Les colophons de plusieurs manuscrits | | **X** | | | **X** | | **X** | X | **X** | | **X** | **X** | **X** | **X** | **X** | **X** | **11** |
| Les noms des copistes de plusieurs manuscrits | | **X** | | | **X** | **X** | **X** | **X** | **X** | | | | **X** | **X** | **X** | **X** | **11** |
| Autres | Les écritures | | | | | | | | **Q12** | | | | | | | | |
| | La variante textuelle | | | | | | | | **Q4, Q12** | | | | | | | | |
| | Les incipit | | | | | | | | **Q13** | | | | | | | | |
| | Comparaison entre la mise en page et les images | | | | | | | | **Q10** | | | | | | | | |
| | Identification de l'auteur | | | | | | | | **Q11** | | | | | | | | |

**Tableau n°. 25 : Les objectifs des chercheurs pour l'étude de manuscrits**

Le but de cette question est de pointer les objectifs des utilisateurs lors de leur étude des manuscrits.

- Onze réponses sont plus concernées par la comparaison entre les colophons dans plusieurs éditions d'un même manuscrit. Cette partie des manuscrits est très essentielle pour trouver des informations sur le nom du copiste, le lieu où il a effectué son travail et la date si elle est indiquée.
- Onze aussi des réponses concernent la comparaison entre les noms des copistes et surtout par la façon dont le même nom est transcrit dans les différents manuscrits.
- Neuf réponses sont plutôt concernées par la comparaison des noms d'auteurs.
- Alors que dix d'entre elles sont intéressées par la comparaison entre les titres donnés à un même manuscrit dans des copies différentes.

D'autres points sont ajoutés aux points principaux mentionnés ci-dessus. Ce sont les suivants :

- En plus des quatre points mentionnés dans la question, Q12 est aussi concerné par la comparaison des écritures arabes.
- Q12 partage son intérêt avec Q4 dans le domaine de la variante textuelle.
- Q13 est le seul qui est intéressé par les incipit ( la première phrase de texte manuscrit) ; de même pour la comparaison entre la mise en page et les images.
- L'identification de l'auteur est aussi signalée par Q14.

### 3.1.2.6. Le domaine de la codicologie ( l'étude matérielle)

| Réponse / Domaine | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 | Q13 | Q14 | Q15 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| La composition des cahiers | X | | X | | X | X | X | X | | | | X | X | | | 8 |
| Les types de support | X | | | | X | X | | X | X | | | X | X | | | 7 |
| Les réclames | | | | | | X | X | | | | | X | | | | 3 |
| La technique de reliure | | | | | | X | X | | X | | | | | X | X | 5 |
| La mise en page | X | | | | | X | | X | X | X | | X | X | X | | 9 |
| Autres | Les notes marginales | | | | | | | Q10 | | | | | | | | |
| | La datation du papier | | | | | | | Q15 | | | | | | | | |

**Tableau n°.26 : Le domaine de la codicologie ( l'étude matérielle)**

- La mise en page d'un manuscrit ( la réglure des pages, le nombre des lignes, le paragraphe, le chapitre et les sous-chapitres) est la plus mentionnée (9) dans nos réponses.
- Huit sont intéressés par *la composition de cahier* (Cahiers de cinq ou de dix etc,)
- Ensuite, il y a *les types de support* (papier ou parchemin) (7 réponses).
- Alors que *la technique de reliure* prend la quatrième place avec 5 réponses.
- Il y a enfin *les réclames* dans 3 réponses.

D'autres éléments sont ajoutés par Q10 et Q15 ; il s'agit des *notes sur la marge* et *la datation du papier*.

Dans le paragraphe qui suit, nous montrerons en détail les aspects les plus importants dans l'étude de la mise en page d'un manuscrit.

| Réponse / Domaine | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 | Q13 | Q14 | Q15 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Le nombre de lignes par page | X | | | X | X | X | X | X | | | | | X | | X | 8 |
| La réglure | X | | | X | | X | X | X | X | | | X | X | X | | 9 |
| La structure du texte manuscrit (chapitres, sous-chapitres, etc.) | X | | | X | X | X | X | X | | X | | X | X | | | 9 |

**Tableau no27 : l'étude de la mise en page d'un manuscrit**

Le tableau ci-dessus montre que *la structure du texte manuscrit* (chapitre et sous-chapitres) de même que *la réglure* de page viennent au même niveau (neuf réponses). Le

*nombre de* lignes par *page* est aussi important dans le domaine de la mise en page mais il vient plutôt en deuxième étape dans l'intérêt des répondants.

### 3.1.2.7. L'histoire des manuscrits

| Réponse / Domaine | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 | Q13 | Q14 | Q15 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Les noms des possesseurs (personne ou institution) | **X** | **X** | **X** | **X** | X | **X** | **X** | **X** | **X** | | | | **X** | **X** | | **11** |
| Les noms des collectionneurs | **X** | **X** | | **X** | **X** | | | **X** | | | | | **X** | **X** | **X** | **8** |
| Les cachets | **X** | **X** | | **X** | | **X** | X | **X** | | | | | | **X** | **X** | **8** |

| | | |
|---|---|---|
| Autres | Les noms de lieux où les manuscrits ont été copiés | Q13 |
| | Le colophon | Q14 |
| | La dédicace | Q14 |
| | Tous les repères qui aident à la datation | Q15 |
| | La place du manuscrit dans une tradition scientifique ou philosophique | Q10 |
| | Le manuscrit et sa relation avec Jérusalem – période islamique. | Q5 |

**Tableau n°.28 : L'histoire des manuscrits**

- Pour suivre l'histoire d'un manuscrit donné, 11 répondants ont choisi comme moyen le plus efficace *le nom du possesseur (personne ou institution)*.
- L'identification historique d'un manuscrit par *les noms de collectionneurs* et *les cachets* occupe la deuxième place. (Huit réponses).

D'autres propositions sont ajoutées, concernant l'identification historique.

- Q13 par exemple cite les lieux où les manuscrits ont été réalisés comme un moyen pour tracer l'histoire d'un manuscrit (andalou, moyen-oriental etc.).
- Q14 rejoint les onze réponses mentionnées dans le tableau qui concerne *les objectifs des chercheurs lors de leur étude de manuscrits,* en proposant l'étude des colophons comme un moyen pour suivre l'histoire d'un manuscrit donné. J'ai trouvé très logique cette proposition.
- La dédicace est aussi proposée par Q14 en plus de la recommandation précédente.
- Q10 a proposé d'étudier la place d'un manuscrit dans une tradition scientifique ou philosophique.
- Q15 a été plus général et il cite *tous les repères qui aident à la datation,* c'est-à-dire tous les éléments mentionnés ci-dessus.

### *3.1.3. L'étude paléographique des manuscrits*

| Réponse / Domain | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 | Q13 | Q14 | Q15 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| La morphologie | X | | X | | X | X | | X | | | | X | | | | 6 |
| Le style | | | | | X | X | X | X | | | | X | | | X | 6 |
| La calligraphie | | | X | | | X | | X | X | | X | | | | X | 6 |
| Autres | La possibilité de dater et de situer géographiquement un manuscrit | | | | | | | | | | | | | Q10 | | |
| **Tableau n°.29 : L'étude paléographique des manuscrits** | | | | | | | | | | | | | | | | |

Presque la moitié des répondants sont intéressés par la paléographie en général.

- Les éléments de morphologie, le style, de même que la calligraphie (malgré la richesse et la beauté de la calligraphie arabe) sont tous au même niveau d'intérêt.
- Q10 est le seul qui propose un autre élément (la possibilité de dater et de situer géographiquement un manuscrit) à partir de la paléographie.

### *3.1.4. La recherche de l'information*

Dans cette partie nous voudrions pointer les difficultés rencontrées dans la recherche d'informations à partir d'un manuscrit sur un support papier. Ensuite, nous voudrions connaître le point vue des répondants pour une recherche menée à partir d'un format électronique de manuscrits après la numérisation. Ceci sera la quatrième partie de notre questionnaire.

#### 3.1.4.1. La recherche d'informations textuelles dans un manuscrit.

| **Réponse / Les moyens** | **Q1** | **Q2** | **Q3** | **Q4** | **Q5** | **Q6** | **Q7** | **Q8** | **Q9** | **Q10** | **Q11** | **Q12** | **Q13** | **Q14** | **Q15** | **Total** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Par la table des matières, s'il y en a une | **X** | **X** | | **X** | X | | | **X** | | **X** | | | **X** | | | **7** |
| En repérant les titres des chapitres, des sous chapitres, etc. | **X** | **X** | **X** | | **X** | **X** | | X | | **X** | | **X** | **X** | | | **9** |
| En feuilletant les manuscrits | **X** | **X** | | **X** | | **X** | X | **X** | **X** | **X** | **X** | **X** | **X** | **X** | **X** | **13** |
| Autres | Par les catalogues | | | | | | | | | | | | | | Q4 | |
| | Par la consultation de moyens de recherche bibliographiques majeurs, comme Brockelman, Sezgin, etc. | | | | | | | | | | | | | | Q8 | |
| | Les données biographiques et bibliographiques en plus des idées doctrinales. | | | | | | | | | | | | | | Q15 | |
| **Tableau n°.30 : La recherche d'informations textuelles dans un manuscrit** | | | | | | | | | | | | | | | | |

A partir de tableau ci-dessus, on peut faire les remarques suivantes :

- Treize des répondants trouvent les informations dont ils ont besoin en feuilletant le manuscrit. C'est un moyen peu pratique et pénible pour les chercheurs, car il prend beaucoup de temps et d'énergie. Ce moyen est très utilisé malgré toutes ses difficultés, du fait de l'absence d'autres moyens pratiques de recherche comme la table des matières par exemple.
- Neuf repèrent les titres des chapitres et des sous-chapitres dans le but de trouver les informations.
- Sept font la recherche par la table des matières. Mais dans la plupart des cas on ne trouve pas de table des matières dans les manuscrits.

Trois autres possibilités de recherche sont ajoutées par les répondants :

- Par *les catalogues* comme le cas de Q4 ;
- Q8 trouve les informations par la consultation des moyens de recherche bibliographique majeurs, comme Brockelman, Sezgin, etc,
- Selon Q15, par les données biographiques et bibliographiques en plus des idées doctrinales.

En réponse à la question « **trouvez-vous facilement l'information dont vous avez besoin ?** », nos répondants nous ont donné les résultats suivants (voir le tableau ci-dessous)

Six disent qu'ils trouvent toujours facilement les informations dont ils ont besoin. Contrairement à Q1, Q4 et Q8 qui les trouvent facilement, mais pas toujours.

| **Oui /non et pourquoi** | | **Numéro de questionnaire** | **Total** |
|---|---|---|---|
| Oui | Oui mais pas toujours (quelques fois) | Q1, Q4 | **2** |
| | Oui toujours | Q6, Q7, Q11, Q13, Q14 et Q15 | **6** |
| | Oui, relativement, mais dépend de l'expérience | Q8 | **1** |
| Non | Non, toujours | Q3, Q5, Q9 | 3 |
| | Non, c'est variable selon les cas | Q10 | 1 |

**Tableau n°. 31 : Les réponses à la question « trouvez-vous facilement l'information dont vous avez besoin**

### 3.1.4.2. Les problèmes rencontrés dans le processus de recherche manuelle d'information.

Les difficultés trouvées dans la recherche sont dues à plusieurs facteurs selon nos répondants comme les suivants :

- Mon ignorance, comme l'indique Q3 dans sa réponse ;
- Difficultés de toutes sortes : textes acéphales (sans tête) et le désordre des folios, etc. comme le dit Q10 ;
- Le manque de tables des matières et d'index dans la plupart des manuscrits.
- Selon Q5, il y a plusieurs facteurs :
    - Trop de manuscrits originaux ne sont pas disponibles, sont manquants ou dispersés à travers le monde.
    - Pour faire des comparaisons, on ne trouve pas de manuscrits sur le même sujet qui permettent de comparer l'un à l'autre.
    - Beaucoup de manuscrits en Palestine ne sont pas bien conservés;
    - Il devrait y avoir des échanges de manuscrits au niveau international;
    - Nous avons besoin d'équipement de haute technologie pour sauvegarder les manuscrits.

Notre dernier répondant, Q5, de nationalité palestinienne a évoqué le problème qui touche les chercheurs palestiniens. On peut résumer ainsi ses réponses:

- En premier lieu, le manque de copies pour le même manuscrit pour faire l'étude comparée.
- En deuxième lieu, dans la plupart des cas, on ne trouve pas le manuscrit original (soit il est perdu, soit il est ailleurs quelque part dans le monde).
- En troisième lieu, ce qui est évident et qui a été dit dans la partie « description d'un corpus de manuscrits de Jérusalem » c'est que les manuscrits sont dans une condition de conservation très mauvaise, ce qui rend la recherche très difficile.
- Pour surmonter ces difficultés, Q5 a proposé deux solutions (les deux derniers points de ses réponses).
- Le premier est l'échange de manuscrits à l'échelle mondiale.

- Le second est l'utilisation d'équipements de haute technologie pour sauvegarder les manuscrits et en assurer une bonne diffusion.

### *3.1.5. La numérisation des manuscrits*

#### 3.1.5.1. Que pensez-vous de la numérisation des manuscrits ?

Douze parmi les quinze répondants sont favorables au processus de numérisation des manuscrits arabes pour les raisons suivantes :

- Q2 pense que la numérisation a permettra un grand bond en avant dans l'étude des manuscrits.
- Pour Q3, la numérisation permettra d'éliminer une part des problèmes de lecture et aidera au classement comparatif, mais selon lui à une condition « si on parvient à instaurer une grille respectable dans un maximum de cas ».
- La réponse de Q4 est aussi conditionnelle : elle sera une véritable aide si l'accès est fait et la résolution bonne « I think it is good idea, if accessing and resolution are possible ».
- Du fait de son expérience et de la particularité des corpus de manuscrits qu'il a travaillés, Q5, reste toujours prudent. Selon lui, c'est très bien, mais on a besoin d'experts pour que la numérisation des manuscrits donne une bon résultat avec l'utilisation de l'ordinateur «Very good, still we need experts for making this digitalisation of the manuscripts on excellant one through the use of computers».
- Q7, quant a lui, estime que la numérisation est une technique d'accès intéressante à condition d'être utilisée à bon escient.
- Q11 rejoint l'idée de Q7 en disant que la numérisation « sera un moyen très efficace et peut être une révolution pour accéder à tous les manuscrits arabes du monde ».
- Q9 est le seul qui prend en considération l'intérêt des utilisateurs pour lui, la numérisation facilitera beaucoup le travail des utilisateurs « un great facilitation for users ».
- Pour Q11, la numérisation est « bonne pour la protection des originaux dont elle peut réduire les consultations ».

- Q8 partage ses idées avec plusieurs répondants ; pour lui, la numérisation facilite les taches suivantes :
    - Cela permettra un accès bien plus large aux manuscrits arabes « This well certainly provides wider access to the Arabic manuscripts ».
    - De même, la numérisation fournira un excellent substitut aux originaux, particulièrement dans les cas des trésors que constituent enluminures et illustrations, de même que pour les manuscrits dont on n'a qu'un seul exemplaire. « Also digitisation will provide an excellent surrogate of the originals, particularly the illuminated and illustrated treasure items, as well as unique manuscripts".
    - C'est d'un grand profit pour la conservation et la préparation des manuscrits ; en effet, les lecteurs peuvent étudier les originaux sans les toucher ni, par conséquent, les abîmer. « Benefit for conservation and preservation, readers can study the manuscripts without handling and damaging the original manuscripts."
- La réponse de Q10 « Je ne connais pas d'exemple» indique qu'il n'a pas d'expérience dans ce domaine. Par conséquent, il n'a pas donné de réponse spécifique.

« **Pensez-vous que l'accessibilité des manuscrits sur l'Internet facilite la recherche** ? »

Dans le même domaine de numérisation, et pour cette question, on peut classer les réponses en deux catégories :

| | **OUI/ Pourquoi ?** |
|---|---|
| Q1 | Je ne sais pas à vrai dire mais je l'espère |
| Q2,Q3,Q4,Q13,Q14 | Oui (sans commentaire) |
| Q6 | Oui mais attention ! Il y a des chercheurs qui déjà lisent trop vite les textes ! Avec les moyens modernes, ils seront tentés d'aller encore plus vite. |
| Q7 | Oui : accès plus rapide à des données essentielles ; possibilité de visualiser immédiatement les manuscrits |
| Q8 | Oui: les utilisateurs éloignés peuvent consulter les manuscrits en ligne et effectuer la majeure partie de leur recherche depuis leur domicile avant du consulter les originaux. Encore une fois, il y a moins de manipulation des précieux manuscrits. « Mais tout ceci dépend de la qualité de la numérisation. *Yes, remote users can consult the manuscripts on-line and do much of their research from home before consulting the originals. Again, less handling of precious manuscripts. But this all depends on the quality of the digitisation, etc."* |
| Q9 | On peut étudier des manuscrits à partir de son propre ordinateur ; il n'y a pas de problème pour obtenir une copie des manuscrits et pas besoin de voyager de |

| | |
|---|---|
| | bibliothèques en bibliothèque.<br>*"You can look into manuscripts from your computer and don't have troubles in getting copies of manuscripts. Or travelling from library to library".* |
| Q10 | Encore faudrait-il que de très nombreux manuscrits soient accessibles |
| Q11 | Oui, si on arrive à avoir un grand nombre de fonds |
| | NON / Pourquoi ? |
| Q5 | Pas encore, un grand nombre d'informations sont contenues dans ces manuscrits. Les installer sur Internet peut mettre leur propriétaire en danger. Peut-être que cela présentera moins de problème si c'est réalisé au niveau international.<br>"*Not yet: there are a lot of important information engulfing these manuscript. Putting it on the Internet might endanger the owners and so on. Yet if this thing take place on the international level it might be fine.*" |
| **Tableau n°.32 : Les réponses à la question « Pensez-vous que l'accessibilité des manuscrits sur l'Internet facilite la recherche ? »** | |

Dans les réponses ci-dessus, on peut distinguer quatre groupes:

- *Le premier groupe* est d'accord avec le processus de numérisation mais sans donner aucune explication (Q2, Q3,Q4, Q13, Q14) ;
- *Le deuxième groupe* a répondu par *Oui* pour montrer les facilités que peut permettra cette nouvelle technologie. On peut retenir les points suivants:
  - Un accès plus rapide à des données essentielles et la possibilité de visualiser immédiatement les manuscrits comme l'indique Q7.
  - Il en est de même pour Q10 qui rejoint Q7 dans la même position. Il ajoute ceci : « Encore faudrait-il que de très nombreux manuscrits soient accessibles ».
- *Pour le troisième groupe,* Q8, Q9 et Q11, la numérisation facilite la tâche des utilisateurs en leur donnent un accès à distance aux manuscrits comme l'indique Q11 « C'est bien, surtout pour les chercheurs éloignés des grandes bibliothèques ». Q9 aussi est d'accord avec le même principe, surtout que cela évite le déplacement des chercheurs d'une bibliothèque à l'autre. Également, la numérisation donne à l'utilisateur la possibilité de trouver facilement les copies des manuscrits : « *You can look into manuscripts from your computer and don't have troubles in getting copies of manuscripts, or travelling from library to library* ». Q8, qui est aussi d'accord avec les idées précédentes, ajoute que la numérisation aide à la conservation des manuscrits en réduisant l'utilisation directe des documents. Grâce à la numérisation, les chercheurs peuvent facilement les consulter par l'intermédiaire de l'ordinateur. Q8 dit que « Yes, remote users can consult the manuscripts on-line and do much of their research

from home before consulting the originals, Again, less handling of precious manuscripts ».

- *Le quatrième groupe* est aussi d'accord avec le principe mais leur *Oui* est conditionnel. Q8, dans sa dernière remarque, rejoint le quatrième groupe à condition que la numérisation soit de bonne qualité « *But this all depends on the quality of the digitisation, etc.* ». Q6 est plus concerné par le comportement des utilisateurs : « *Oui* mais attention ! Il y a des chercheurs qui déjà lisent trop vite les textes ! Avec les moyens modernes, ils seront tentés d'aller encore plus vite ». Mais pour Q15, le plus important est la rassemblement d'un fond de manuscrits numérisés : « *Oui*, si on arrive à avoir un grand nombre de fonds ».
- *Le cinquième groupe* est tout à fait en opposition avec les quatre premiers groupes, mais il ne s'agit que d'une seule réponse. Q5 n'est pas d'accord avec le processus de numérisation. Ses précautions concernent surtout les informations et les possesseurs de manuscrits. Les informations contenues dans les manuscrits risquent de circuler par l'intermédiaire de l'Internet, ce qui, peut-être, peut mettre en danger les possesseurs des manuscrits. Q5 serait d'accord avec les autres répondants à une seule condition : « si tous les manuscrits étaient diffusés à une échelle mondiale, autrement dit si le projet de numérisation deverait un projet mondial ». Selon lui : « *Not: yet: there are a lot of important information engulfing these manuscript. Putting it on the Internet might endanger the owners and so on. Yet if this thing take place on the International level it might be fine*".

### 3.1.5.2. L'attente d'une recherche électronique sur les manuscrits

Treize ont répondu à la question : « Qu'attendez-vous d'une recherche électronique sur les manuscrits ? ». Leurs attentes des recherches électroniques peuvent être classées comme suit :

- Faciliter la recherche : Q1, Q6, et Q7 sont d'accord sur la même idée. Pour Q1, la numérisation «facilite la recherche ». Pour Q6, la recherche électronique fait progresser le travail de recherche « Il faut voir la chose de près et au fur et à mesure de l'avancement du travail de recherche ». Pour Q7, «il faut qu'elle soit aussi complète que possible, tout en restant aisée et rapide ».

- Faciliter la comparaison des textes manuscrits : l'attente de Q8 est d'être capable de comparer les manuscrits de différentes collections, autrement dit d'assembler les manuscrits « *Being able to compare manuscripts from different collections, collate manuscripts, etc.*» . Pour Q15, « cela devrait faciliter les comparaisons ».
- Selon Q3, la recherche électronique peut « systématiser au maximum et donc « globaliser » et réunir des connaissances éparses et par conséquent : permettre de nouvelles conclusions».
- L'identification de texte : l'attente de Q4 lors de la recherche électronique est d'abord de bien lire le texte manuscrit « *To be able to read it* ». La recherche sur le vocabulaire du manuscrit est une autre attente exprimée par Q10 : « par exemple, les possibilités d'identification du texte, de recherches sur le vocabulaire, etc, comme on peut le faire sur CD-Rom pour les textes imprimés».
- Qualité d'image très élevée : l'espérance de Q2 et Q9 est d'avoir des images de manuscrits de grande qualité. Q2 aussi souhaite la possibilité de bien manipuler l'imagerie « High quality, manipulatable imagery ».
- Un catalogue correct et complet : pour faire une recherche électronique, il faut un catalogue électronique cohérent, comme le propose Q2.
- Pour Q14, il faut une vision entière des manuscrits, même le « feuilleton » ;
- Pour Q5, la numérisation et par conséquent la recherche électronique peut endommager le texte, ce qui le rendra difficile à comparer avec d'autres manuscrits dans le même domaine : « *will damage the texts, even will make it very difficult to compare it to other manuscripts which related to subject of this manuscripts* ».

Les répondants nous ont proposé d'autres éléments qu'ils souhaitent trouver par une recherche électronique :

| **Les éléments proposés** | **Numéro de questionnaire** |
|---|---|
| Noms propres | **Q1** |
| Thèmes | **Q1** |
| Titres | **Q7** |
| Chapitres | **Q7** |
| Données codicologiques (date de copie, nom du copiste etc.) | **Q7** |
| Différents types d'enluminures | **Q8 et Q14** |
| Différents types d'illustration | **Q8** |
| Colophons | **Q8** |
| Index avec le titre des chapitres | **Q9** |

| | |
|---|---|
| Les miniatures | **Q14** |
| Avoir dans la main le plus grand nombre d'éléments manuscrits d'un ou plusieurs auteurs. | **Q15** |
| **Tableau n°.33 : Les éléments proposés par les répondants qu'ils souhaitent trouver par une recherche électronique.** | |

En plus des onze éléments mentionnés dans le tableau ci-dessus, on a eu d'autres réponses qu'on peut considérer comme des moyens de recherche, tels que les noms propres, les thèmes, les titres, les titres de chapitres, les données codicologiques (la date de la copie, le nom du copiste, etc.) les indexes, les miniatures, les enluminures, les illustrations, les colophons, etc.). Q5 souhaite avoir un service de recherche gratuity: « *No need to pay money for the wanted manuscripts whatever* ». Q6 souhaite que la recherche électronique puisse aider à l'avancement dans le travail de recherche.

### *3.1.6. Propositions générales*

En réponse à la question « Avez-vous des précisions à apporter sur des éléments qui n'ont pas été cités auparavant dans le questionnaire ?», nous avons reçu les propositions suivantes :

- Q3 est intéressé par la classification électronique des manuscrits.
- Q15 est intéressé par l'exploitation scientifique du contenu.
- Alors que l'intérêt de Q5 est tout à fait différent des autres. Il souhaite que le projet de numérisation soit un moyen pour rassembler tous les manuscrits qui concernent la Palestine et en particulier Jérusalem, et les faire revenir en Terre Sainte : « *The project of bringing all the scattered manuscripts relating to Palestine, in particular to Jerusalem, back to the Holy Land* ».

### *3.1.7. Conclusion :*

Les résultats que nous avons obtenus lors du questionnaire nous ont permis de construire et de définir les métadonnées propre aux besoins de nos répondants. Les métadonnées proposées dans le tableau restent toujours à enrichir par l'étude d'autres projets de numérisation tels que MASTER, EAMMS ou DEBORA. Le tableau ci-dessous est un tableau récapitulatif qui rassemble les résultats obtenus par les réponses au questionnaire :

| ***N°.*** | ***Les Métadonnées proposées par les répondants*** |
|---|---|
| **1** | Auteur |
| | Copiste |

| | | | | |
|---|---|---|---|---|
| | Nom du possesseur | | | |
| | Nom du collectionneur | | | |
| **2** | Titre | Titre des manuscrits | | |
| | | Titre des chapitres | | |
| | | Titre des sous-chapitres | | |
| | | Le titre du manuscrit dans le colophon | | |
| | | Incipit | | |
| **3** | Colophon | Date | | |
| | | Lieu | | |
| **4** | Période étudiée | Classés par période seulement | Islamique | |
| | | | IV°-X° siècles | |
| | | | VI°-VII° siècles | |
| | | | Médiévale islamique (VII°-XV°) | |
| | | | VIII°-XVI° (700-1500) | |
| | | | X°-XX° siècles | |
| | | | XII°-XIV° siècles | |
| | | | XII°-XV° siècles | |
| | | | Toutes les périodes | |
| | | Classés par thème et période | La codicologie (IX°-XVI°) | |
| | | | La paléographie (toutes les périodes) | |
| | | | Les textes grammaticaux (IX°-XVI°) | |
| | | | Jérusalem islamique et pré-Islamique | |
| **5** | Les éléments qui aident à identifier la date des manuscrits | | Le nom des possesseurs (personne ou institutions) | |
| | | | Le nom du collectionneur | |
| | | | Le cachet | |
| | | | Le nom du lieu où le manuscrit a été copié | |
| | | | Le colophon | |
| | | | La dédicace | |
| **6** | Domaine d'intérêt (thèmes) | | L'édition de textes | |
| | | | La consultation de textes | |
| | | | L'histoire des collections | |
| | | | L'histoire de l'art | |
| | | | La paléographie | |
| | | | La codicologie | |
| **7** | Catégorie de manuscrits | | Arabo-islamique | |
| | | | Arabo-chrétien | |
| **8** | Type de manuscrit (sujet) | | Coranique | |
| | | | Autres textes religieux | |
| | | | Scientifiques | |
| | | | Littéraires | |
| | | | Documentaires | |
| | | | Méthodologie juridique musulmane | |
| | | | Philosophiques | |
| | | | Historiques | |
| | | | Grammaire | |
| | | | Mystique et philosophie arabo-islamique | |
| **9** | Codicologie L'étude matérielle de document | | La composition du cahier | |
| | | | Les types de support | |
| | | | Les réclames | |
| | | | La technique de reliure | |
| | | | La mise en page | Le nombre de lignes par page |
| | | | | La réglure |
| | | | | La structure du texte manuscrit (chapitres, sous-chapitres, etc.) |
| | | | Les notes marginales | |
| | | | La datation de papier | |
| **10** | Paléographie (l'étude de l'écriture) | | La morphologie | |
| | | | Le style | |

| | | | |
|---|---|---|---|
| | | | La calligraphie |
| **11** | Table des matières | | Oui |
| | | | Non |
| **12** | Index | | Oui |
| | | | Non |
| **13** | Thèmes | | |
| **14** | Décor des textes | Enluminures | |
| | | Illustrations | |
| | | Miniatures | |
| | | Frontispice | |
| | | Décor de reliure | |

**Tableau n°.34 : Les Métadonnées proposées par les répondants**

## 3.2. Définition des métadonnée

### 3.2.1. Introduction

Dans ce chapitre, nous définirons les métadonnées et leur grammaire (la DTD). Ces éléments nous servent au balisage et à la description des manuscrits arabes, en prenant en considération leur structure hiérarchique. Chaque élément est défini en tenant compte de ses attributs et de sa relation avec les éléments fils. Notre intention également est de comparer les métadonnées des manuscrits arabes avec celles du projet MASTER que nous avons mentionné dans la premiere partie du thèse.

Parmi les éditeurs XML qui existent dans le marché, nous avons choisi XML Spy pour définir notre DTD. Notre choix est dû au fait que cet éditeur est, jusqu'à maintenant, le plus facile et le plus avancé dans le domaine de la publication électronique de document sur XML format.

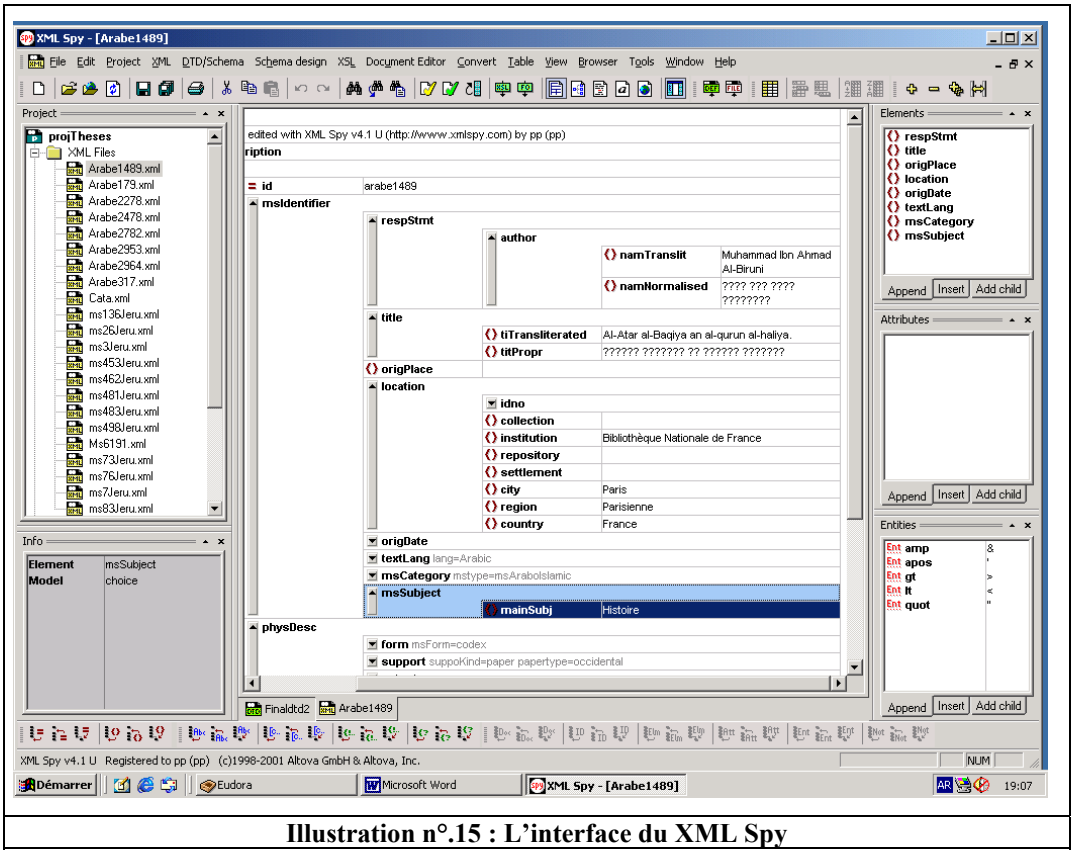



**Illustration n°.15 : L'interface du XML Spy**

Son interface consiste en trois fenêtres : la fenêtre principale se trouvant au milieu avec une fenêtre de chaque côté.

La fenêtre à droite consiste en trois petites fenêtres permettant l'insertion et l'ajout d'éléments et d'attributs. La première fenêtre en haut à droite affiche les éléments fils qui

appartiennent à un élément racine dans la fenêtre principale du milieu de l'écran. La deuxième, située au milieu, affiche les attributs en relation avec l'élément racine, alors que la troisième, située en bas, est consacrée aux entités•,

Cependant pour *les fenêtres de gauche*, la première en haut est réservée pour l'information sur le projet créé par l'utilisateur pour mettre tous les fichiers, soit en forme XML, soit en forme DTD, soit en forme de schéma, etc., tandis que la deuxième, située en bas, est consacrée à l'information générale.

Les icônes en bas de l'écran aident à la création et à l'ajout, facilement et rapidement, d'éléments, d'attributs et de commentaires.

Dans notre travail actuel, le DTD MASTER a été notre guide mais nous avons trouvé nécessaire d'y apporter quelques modifications, qui résultent des caractéristiques particulières des manuscrits arabes. Donc, on trouve des points de rencontre ainsi que des points des divergences avec le DTD MASTER. Nous avons tenté, dans notre travail, de définir le plus grand nombre possible des métadonnées trouvés dans l'ensemble des manuscrits, tout en sachant que ces métadonnées ne seraient pas appliquées à chacun de ces manuscrits.

### 3.2.1.1. Quelques remarques explicatives

Il est indispensable dans ce paragraphe d'expliquer les schémas que nous allons montrer dans ce chapitre comme exemples.

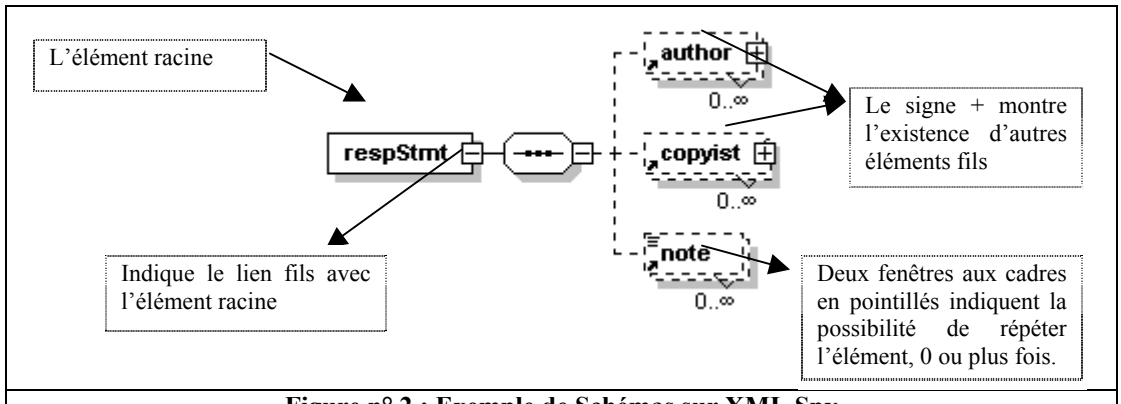



**Figure n°.2 : Exemple de Schémas sur XML Spy**

• Le signe qui représente un mot donné comme *amp* est l'entité du signe (&) ; le *apos* est l'entité de (') ; le *gt* est l'entité du (>) ; le *It* est l'entité de (<) et le *quot* est l'entité du ("), etc.

### 1. Les éléments de racine et les éléments fils

Comme il a été montré dans la figure ci-dessus, le rectangle avec le mot « respStmt » représente l'élément racine, qui peut être suivi par des éléments fils pour affiner la définition de l'élément princeps : dans ce cas de figure, l'auteur, le copiste et la note sont des éléments fils. Le petit cadre entre l'élément racine et les éléments fils signifie qu'il y a un lien entre les deux. Le signe (+) sur les deux premiers cadres des éléments fils indique que ces deux éléments ont aussi des éléments fils, alors que le cadre en pointillé est l'indication que l'élément est répétable soit 0 ou une fois, soit une ou plusieurs fois, soit 0 ou plusieurs fois.

### 2. Les attributs

L'élément racine ainsi que leurs éléments fils ont des attributs qui ont pour but d'ajouter des valeurs à l'élément lui-même, comme par exemple l'adjectif «française » ajouté à l'élément « langue » indique que la langue utilisée est la langue française, etc.

## *3.2.2. Les DTD des manuscrits arabes*

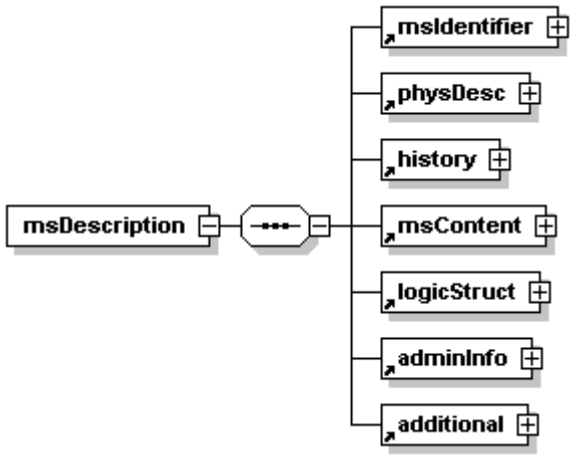
Cent soixante tries champs, dont soixant tries possèdent en moyenne deux ou trois attributs, ont été retenus, pour définir la structure des documents décrivant les manuscrits arabes.

Comme point de départ, nous avons choisi le terme « msDescription » qui sert de base à partir de laquelle sont établis tous les autres éléments fils et les sous-éléments. De même, dans le projet MASTER, le terme « msDescription » a aussi été utilisé comme élément de base.

**msDescription :** l'élément « msDescription » est divisé en six éléments principaux. Il s'agit de msIdentifier, physDesc, history, msContent, logicStruct, additional, (cf. les figures suivantes). Alors que dans le projet MASTER, les éléments msIdentifier, msHeading, msContents, physDesc, history, additional et msPart ont été choisis comme éléments fils pour l'élément msDescription.

**element msDescription**

diagram




children **msIdentifier physDesc history msContent logicStruct adminInfo additional**

attributes

| Name | Type | Use | Default | Fixed |
|---|---|---|---|---|
| status | xs:NMTOKEN | | uni | |
| type | xs:string | | | |

source

```
<xs:element name="msDescription">
 <xs:complexType>
  <xs:sequence>
   <xs:element ref="msIdentifier"/>
   <xs:element ref="physDesc"/>
   <xs:element ref="history"/>
   <xs:element ref="msContent"/>
   <xs:element ref="logicStruct"/>
   <xs:element ref="adminInfo"/>
   <xs:element ref="additional"/>
  </xs:sequence>
  <xs:attribute name="status" default="uni">
   <xs:simpleType>
    <xs:restriction base="xs:NMTOKEN">
     <xs:enumeration value="uni"/>
     <xs:enumeration value="comp"/>
     <xs:enumeration value="frag"/>
     <xs:enumeration value="def"/>
     <xs:enumeration value="unknown"/>
    </xs:restriction>
   </xs:simpleType>
  </xs:attribute>
  <xs:attribute name="type" type="xs:string"/>
 </xs:complexType>

</xs:element>
```

**Figure n°.3: Les éléments msDescription**

**1. msIdentifier** (manuscript identifier) cet élément englobe tous les éléments qui permettent l'identification du manuscrit ou d'un fragment de manuscrit.

Notre démarche de choisir les éléments fils de « msIdentifier » diverge de celle du projet MASTER. Pour ce dernier, les éléments d'identification d'un manuscrit donné sont le pays (country), la région (region), l'habitation (settlement), l'institution (institution), le dépositaire (repository), la collection (collection), alors que, pour nous, les éléments d'identification dans le MASTER sont des éléments fils de l'élément « location » qui

forme, avec les autres éléments fils, l'élément d'identification. Les neuf éléments suivants ont été choisis comme éléments d'identification parmi le DTD de manuscrit arabe.

element **msIdentifier**

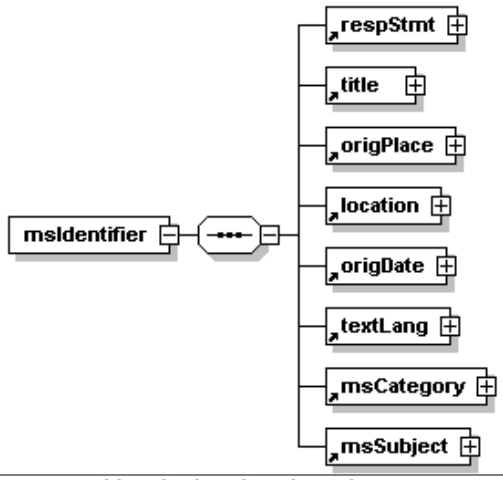
| | |
|---|---|
| diagram |  |
| children | respStmt title origPlace location origDate textLang msCategory msSubject |
| used by | element msDescription |
| source | <xs:element name="msIdentifier"><br><xs:complexType><br><xs:sequence><br><xs:element ref="respStmt"/><br><xs:element ref="title"/><br><xs:element ref="origPlace"/><br><xs:element ref="location"/><br><xs:element ref="origDate"/><br><xs:element ref="textLang"/><br><xs:element ref="msCategory"/><br><xs:element ref="msSubject"/><br></xs:sequence><br></xs:complexType><br></xs:element> |

**Figure n°.4: Les éléments msIdentifier**

**1.1. respStmt** : cet élément contient les informations sur les noms des personnes qui sont responsables de l'élaboration et de la rédaction du manuscrit comme l'auteur, le copiste, etc.

Par contre, MASTER utilise cet élément pour indiquer les noms des personnes, autres que l'auteur ou le copiste, responsables d'une partie du texte (comme l'illustration par exemple), l'élément auteur étant mis comme un élément fils de l'élément « msHeading ».

element **respStmt**

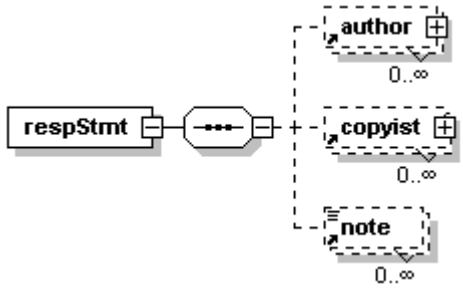
| | |
|---|---|
| diagram |  |
| children | **author copyist note** |
| used by | element **msIdentifier** |
| source | <xs:element name="respStmt"><br><xs:complexType><br><xs:sequence><br><xs:element ref="author" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="copyist" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="note" minOccurs="0" maxOccurs="unbounded"/><br></xs:sequence><br></xs:complexType><br></xs:element> |

**Figure n°. 5: Les éléments respStmt**

**1.1.1. author:** il contient le nom de l'auteur principal à l'origine du manuscrit.

Pour l'élément « author », on a trois éléments fils:

1. nomNormalised : le nom officiel reconnu par le spécialiste car, pour le nom d'un auteur arabe, il arrive de trouver le nom écrit de différentes manières, soit dans le même document, soit dans d'autres documents bibliographiques.

2. namTranslit (nom translittéré) : il fournit le nom d'un auteur arabe écrit en caractères latins. Nous avons mis cet élément pour deux raisons, d'une part comme solution en l'absence de logiciel informatique de langue arabe et, d'autre part, pour garder un nom translittéré comme aide aux utilisateurs qui ne savent pas lire l'arabe.

3. *otherNames* : dans « respStmt", l'élément otherNames contient un autre nom par lequel un auteur ou un copiste est connu (un surnom par exemple).

4. *profession* : nous avons trouvé nécessaire de mettre la profession de l'auteur comme information supplémentaire, en sachant que cet élément n'arrive pas au même niveau que l'autre mais qu'il indique seulement la profession de l'auteur du manuscrit.

5. Date de naissance et de morte.

element **author**

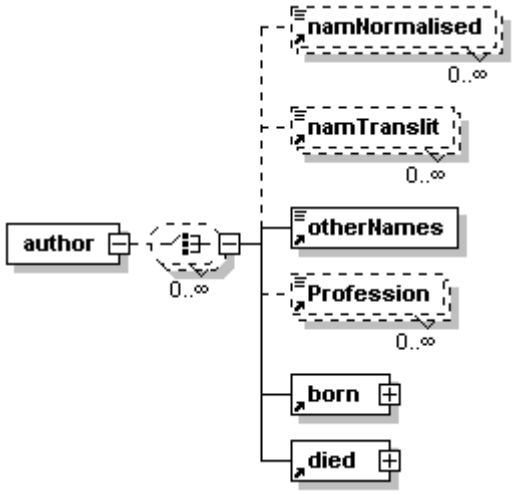
| | |
|---|---|
| diagram |  |
| children | namNormalised namTranslit otherNames Profession born died |
| used by | elements msItem respStmt |
| source | <xs:element name="author"><br><xs:complexType><br><xs:choice minOccurs="0" maxOccurs="unbounded"><br><xs:element ref="namNormalised" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="namTranslit" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="otherNames"/><br><xs:element ref="Profession" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="born"/><br><xs:element ref="died"/><br></xs:choice><br></xs:complexType><br></xs:element> |

**Figure n°.6: Les éléments author**

**1.1.2. copyist:** il contient le nom du copiste qui exécute le travail. En contrepartie, MASTER utilise le mot « scribe » comme attribut pour indiquer le nom du copiste.

Comme pour l'auteur, le nom du copiste a été classé selon les éléments suivants :

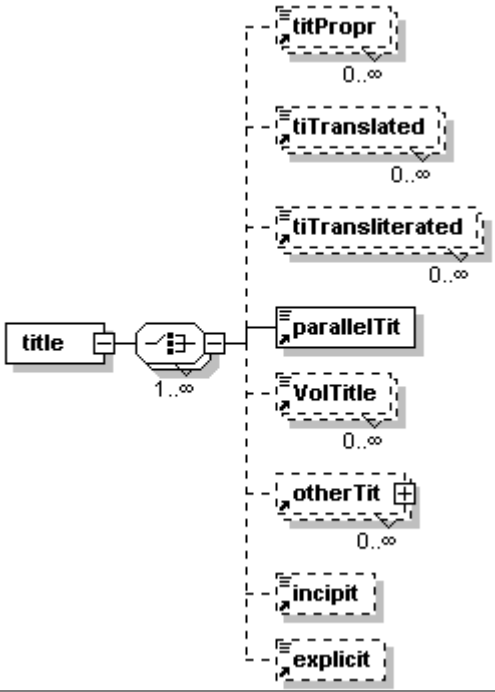
1. namNormalised : il indique le nom officiel du copiste.
2. namTranslit : il contient le nom du copiste écrit en caractères latins.
3. otherNames : il fournit les noms autres que le nom officiel du copiste.

**1.1.3. note :** cet élément contient n'importe quelle description supplémentaire qui concerne la responsabilité intellectuelle du manuscrit autre que l'auteur et le copiste, comme le peintre par exemple.

**1.2. title :** il fournit le titre du document ou d'une partie de document. Le même élément a été utilisé dans MASTER mais comme élément fils de l'élément msHeading. Nous avons trouvé nécessaire de mettre à l'intérieur de l'élément racine *titre* les éléments fils suivants :

element **title**

diagram




children titPropr tiTranslated tiTransliterated parallelTit VolTitle otherTit incipit explicit

used by elements msIdentifier msItem

source

```
<xs:element name="title">
  <xs:complexType>
    <xs:choice maxOccurs="unbounded">
      <xs:element ref="titPropr" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="tiTranslated" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="tiTransliterated" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="parallelTit"/>
      <xs:element ref="VolTitle" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="otherTit" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="incipit" minOccurs="0"/>
      <xs:element ref="explicit" minOccurs="0"/>
    </xs:choice>
  </xs:complexType>
</xs:element>
```

**Figure n°.7: Les element title**

**1.2.1. titPropr** : il s'agit du véritable titre qui a été retrouvé répété de la même façon dans plusieurs sources, soit sur la première page du manuscrit, dans le colophon, au dos de l'ouvrage, dans des catalogues de bibliothèques ou dans des livres bibliographiques et spécialisés, tels que «Brokelman » par exemple.

**1.2.2. tiTranslated** (titre traduit) : il contient le titre du manuscrit traduit dans une autre langue que l'arabe.

**1.2.3. tiTransliterated** (titre translittéré) : il fournit le titre du manuscrit en langue arabe mais écrit en caractères latins.

**1.2.4. parallalTit** (titre parallèle) : il fournit le titre parallèle du titre propre qui se trouve dans certains manuscrits écrit soit dans la même langue, soit dans une autre langue que l'originale.

**1.2.5. VolTitle** (titre du volume) : il donne le titre de chaque volume dans le cas où le manuscrit se compose de plusieurs volumes.

**1.2.6. otherTit** (autres titres) : il donne la possibilité au « catalogueur » de mettre d'autres titres qui n'ont pas été mentionnés ci-dessus.

element **otherTit**

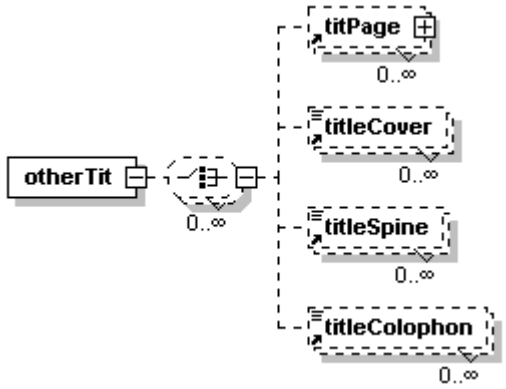
| | |
|---|---|
| diagram |  |
| children | titPage titleCover titleSpine titleColophon |
| used by | element title |
| source | <xs:element name="otherTit"><br><xs:complexType><br><xs:choice minOccurs="0" maxOccurs="unbounded"><br><xs:element ref="titPage" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="titleCover" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="titleSpine" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="titleColophon" minOccurs="0" maxOccurs="unbounded"/><br></xs:choice><br></xs:complexType><br></xs:element> |

**Figure n°.8: Les element otherTitle**

**1.2.6.1. titCover** : il contient le titre qui se trouve sur la « premières de couverture »

**1.2.6.2. titleSpine** : il fournit le titre du manuscrit trouvé écrit sur le dos du livre.

**1.2.7. incipit** : il contient la première phrase du manuscrit au cas où le titre n'existerait pas ou pour ajouter des informations supplémentaires en plus du titre.

element **incipit**

| | | | | | |
|---|---|---|---|---|---|
| diagram |  | | | | |
| type | extension of **xs:string** | | | | |
| used by | elements msItem title | | | | |
| attributes | Name | Type | Use | Default | Fixed |
| | type | xs:string | | | |
| | defective | xs:NMTOKEN | | no | |

source

```
<xs:element name="incipit">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute name="type" type="xs:string"/>
    <xs:attribute name="defective" default="no">
     <xs:simpleType>
      <xs:restriction base="xs:NMTOKEN">
       <xs:enumeration value="yes"/>
       <xs:enumeration value="no"/>
       <xs:enumeration value="unknown"/>
      </xs:restriction>
     </xs:simpleType>
    </xs:attribute>
   </xs:extension>
  </xs:simpleContent>
 </xs:complexType>
</xs:element>
```

**Figure n°.9 : L'élément incipit**

**1.2.8. explicit** : Il contient la dernière phrase du manuscrit.

element **explicit**

diagram


type: extension of **xs:string**

used by: elements **msItem** **title**

attributes

| Name | Type | Use | Default | Fixed |
|---|---|---|---|---|
| type | xs:string | | | |
| defective | xs:NMTOKEN | | no | |

source

```
<xs:element name="explicit">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute name="type" type="xs:string"/>
    <xs:attribute name="defective" default="no">
     <xs:simpleType>
      <xs:restriction base="xs:NMTOKEN">
       <xs:enumeration value="yes"/>
       <xs:enumeration value="no"/>
       <xs:enumeration value="unknown"/>
      </xs:restriction>
     </xs:simpleType>
    </xs:attribute>
   </xs:extension>
  </xs:simpleContent>
 </xs:complexType>
</xs:element>
```

**Figure n°.10 : L'élément explicit**

A l'exception de l'incipit et de l'explicit, les cinq premiers éléments fils de l'élément titre n'existent pas dans MASTER. Par contre, il utilise le mot « titre », indistinctement pour les différents titres trouvés dans le manuscrit.

**1.3. origPlace** : il s'agit du lieu d'origine du manuscrit. Le lieu d'origine se compose en trois éléments fils: le nom de la ville, le nom de la région, et le nom du pays. Dans MASTER, le mot « origPlace » englobe toutes les formes des noms de lieux utilisés pour identifier la provenance du manuscrit ou d'une partie de manuscrit.

element **origPlace**

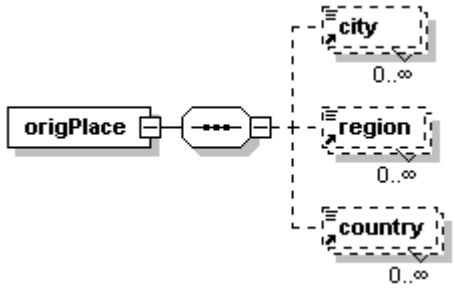
| | |
|---|---|
| diagram |  |
| children | **city region country** |
| used by | element **msIdentifier** |
| source | <xs:element name="origPlace"><br><xs:complexType><br><xs:sequence><br><xs:element ref="city" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="region" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="country" minOccurs="0" maxOccurs="unbounded"/><br></xs:sequence><br></xs:complexType><br></xs:element> |

**Figure n°. 11: Les éléments origPlace**

**1.3.1. city** (Ville) : il contient le nom d'une unité géopolitique plus petite que la région. Il indique la ville d'origine où le manuscrit a été achevé.

**1.3.2. region** (région) : il contient le nom d'une unité géopolitique plus large que celle de la ville et plus petite qu'un pays. Il montre le nom de la région dans laquelle la ville est située.

**1.3.3. country** (pays): il contient le nom d'un endroit géopolitique plus grand que la région. Il fournit le nom du pays d'origine du manuscrit.

**1.4. Localisation :** il contient les éléments nécessaires pour localiser un manuscrit donné dans une ou plusieurs bibliothèques. Dans cette partie, nous avons utilisé une autre démarche que celle de MASTER, en faisant la distinction entre le lieu d'origine du manuscrit et l'endroit de sa présence actuelle.

element **location**

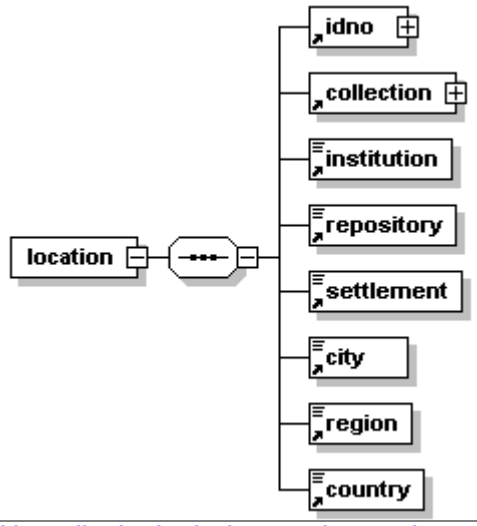
| | |
|---|---|
| Diagram |  |
| Children | idno collection institution repository settlement city region country |
| used by | element mSIdentifier |
| Source | <xs:element name="location"><br><xs:complexType><br><xs:sequence><br><xs:element ref="idno"/><br><xs:element ref="collection"/><br><xs:element ref="institution"/><br><xs:element ref="repository"/><br><xs:element ref="settlement"/><br><xs:element ref="city"/><br><xs:element ref="region"/><br><xs:element ref="country"/><br></xs:sequence><br></xs:complexType><br></xs:element> |

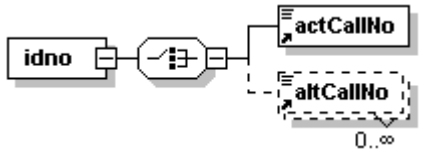
**Figure n°.12: Les éléments location**

**1.4.1. idno** (cote): il s'agit de la série des abréviations et des chiffres utilisés pour identifier un manuscrit ou un livre dans une ou plusieurs bibliothèques. Le « idno » consiste en éléments fils tels que :

**1.4.1.1.actCallNo** :cote actuelle du manuscrit dans une bibliothèque.

**1.4.1.2. altCallNo** : cote alternative pour d'éventuelles copies de ce manuscrit, se trouvant dans d'autres bibliothèques.

element **idno**

| | |
|---|---|
| Diagram |  |
| Children | actCallNo altCallNo |
| used by | element location |
| source | <xs:element name="idno"><br><xs:complexType><br><xs:choice><br><xs:element ref="actCallNo"/> |

```
      <xs:element ref="altCallNo" minOccurs="0" maxOccurs="unbounded"/>
    </xs:choice>
  </xs:complexType>
</xs:element>
```

**Figure n°.13: Les éléments idno**

**1.4.2. collection** : il contient le nom de la collection dans une bibliothèque ou dans un autre lieu où se trouve le manuscrit.

Dans le schéma ci-dessous, le « p » désigne un paragraphe, c'est à dire que le catalogueur a la possibilité d'écrire ce qu'il veut. La double fenêtre avec le nombre (1) autour du « p » signifie qu'il est possible de répéter le paragraphe au moins une ou plusieurs fois.

**1.4.3. institution**: il contient le nom de l'institution, que ce soit bibliothèque ou université dans lequel le manuscrit existe.

**1.4.4. repository** : il permet de localiser le manuscrit dans la partie de la bibliothèque ou de l'institution, où il se trouve.

**1.4.5. settlement** : il contient le nom d'un lieu plus petit qu'une ville, tel qu'un village par exemple.

Les trois derniers éléments fils (« city », « region » et « country ») sont déjà définis dans l'élément « origPlace ».

**1.5. origDate :** il contient n'importe quelle date utilisée pour identifier la date d'origine d'un manuscrit ou d'une partie de manuscrit.

element **origDate**

diagram


children **Date**

used by element **msIdentifier**

source

```
<xs:element name="origDate">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Date"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
```

**Figure n°.14 : Les éléments origDate**

**1.5.1. Date** : pour la date, il y a un élément fils « p » dans lequel on peut ajouter la date sous n'importe quelle forme.

Autre possibilité, pour faciliter le tâche des catalogueurs, nous avons mis tous les attributs des dates trouvées pendant notre étude des manuscrits, et ce plus particulièrement pour les manuscrits arabo-chrétiens.

| attributes | Name | Type | Use | Default | Fixed |
|---|---|---|---|---|---|
| | Day | xs:string | | | |
| | Month | xs:string | | | |
| | JCEra | xs:string | | | |
| | Hegira | xs:string | | | |
| | diffDates | xs:string | | | |
| | AdamEra-5508BC | xs:string | | | |
| | AlexandEra-356BC | xs:string | | | |
| | MartyrEra-283AC | xs:string | | | |
| | notBefore | xs:string | | | |
| | notAfter | xs:string | | | |
| | evidence | xs:NMTOKEN | | | |

source

```
<xs:element name="Date">
 <xs:complexType>
  <xs:sequence>
   <xs:element ref="p"/>
  </xs:sequence>
  <xs:attribute name="Day" type="xs:string"/>
  <xs:attribute name="Month" type="xs:string"/>
  <xs:attribute name="JCEra" type="xs:string"/>
  <xs:attribute name="Hegira" type="xs:string"/>
  <xs:attribute name="diffDates" type="xs:string"/>
  <xs:attribute name="AdamEra-5508BC" type="xs:string"/>
  <xs:attribute name="AlexandEra-356BC" type="xs:string"/>
  <xs:attribute name="MartyrEra-283AC" type="xs:string"/>
  <xs:attribute name="notBefore" type="xs:string"/>
  <xs:attribute name="notAfter" type="xs:string"/>
  <xs:attribute name="evidence">
   <xs:simpleType>
    <xs:restriction base="xs:NMTOKEN">
     <xs:enumeration value="internal"/>
     <xs:enumeration value="external"/>
     <xs:enumeration value="conjecture"/>
    </xs:restriction>
   </xs:simpleType>
  </xs:attribute>
 </xs:complexType>
</xs:element>
```

**Tableau n°.35: Les elements date**

Le tableau suivant explique chaque datation.

| Name | Type |
|---|---|
| Day | Il contient le jour d'achèvement du manuscrit, s'il existe |
| Month | Il contient le mois au cours duquel le travail s'est terminé |
| JCEra | Date selon l'ère de Jésus Christ |
| Hegira | Date selon l'ère hégire |
| diffDates | Date autre que les précédentes |
| AdamEra-5508BC | Date selon l'ère d'Adam (5508 avant Jésus Christ) |
| AlexandEra-356BC | Date selon l'ère d'Alexandre (356 avant Jésus Christ) |
| MartyrEra-283AC | Date selon l'ère du Martyr (283 après Jésus Christ) |
| notBefore | datation estimée pour un manuscrit non daté (pas anrérieure à telle période) |
| notAfter | datation estimée pour un manuscrit non daté (pas postérieure à telle période) |
| evidence | Il précise le degré d'évidence pour une date fournie par autre attribut. |

**Tableau n°.36 : Explication pour les éléments date**

*Evidence*: pour le dernier élément de la liste, « evidence », il existe trois possibilités d'attributs : une source interne, une source externe ou une conjecture.

Les trois derniers éléments dans le tableau de datation sont les mêmes que ceux du MASTER, avec un nouvel élément qui s'appelle « certainty », ce dernier spécifiant le degré de certitude quant à la date fournie par d'autres attributs.

Il est indispensable de mentionner dans cette partie que l'on doit utiliser une forme normalisée de date, selon le standard international connu (ex : jour, mois, année).

**1.6. textLang :** il décrit la langue officielle ou une combinaison de deux ou trois langues utilisées pour écrire le texte du manuscrit. Même dans MASTER, l'élément textLang indique la langue officielle du texte.

element **textLang**

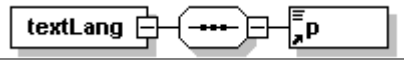
| | |
|---|---|
| diagram |   |
| children | p |
| used by | elements Language msIdentifier |
| attributes | Name: lang; Type: xs:NMTOKEN; Use: ; Default: ; Fixed: |
| source | <xs:element name="textLang"><br><xs:complexType><br><xs:sequence><br><xs:element ref="p"/><br></xs:sequence><br><xs:attribute name="lang"><br><xs:simpleType><br><xs:restriction base="xs:NMTOKEN"><br><xs:enumeration value="Arabic"/><br><xs:enumeration value="Arabic-Coptic"/><br><xs:enumeration value="Arabic-Grece"/><br><xs:enumeration value="Arabic-Syriac"/><br><xs:enumeration value="Arabic-Coptic-Syriac"/><br><xs:enumeration value="Copt"/><br><xs:enumeration value="French"/><br><xs:enumeration value="Greek"/><br><xs:enumeration value="Latine"/><br><xs:enumeration value="Persian"/><br><xs:enumeration value="Syriac"/><br><xs:enumeration value="Turkish"/><br></xs:restriction><br></xs:simpleType><br></xs:attribute><br></xs:complexType><br></xs:element> |

**Figure n°.15: Les élément et les attributs du textLang**

Notre proposition des langues suivantes provient de notre étude de manuscrits, notamment de manuscrits arabo-chrétiens où certains textes religieux sont écrits en deux ou trois langues, tels que :

Arabic-Coptic, Arabic-Grece, Arabic-Syriac, Arabic-Coptic-Syriac et quelquefois Arabic-French, arabic-Latine, Arabic-Persian, Arabic-Turkish) .

**1.7. LangUsage :** cet élément définit une combinaison particulière de deux langues (telle que "langue espagnole écrite en caractères arabe" ou aljamiado-morisque). Dans MASTER, l'élément langUsage indique le même phénomène.

**1.8. msCategory :** Il indique le catégorie de manuscrits soit arabo-islamiques (msAraboIslamic), soit arabo-chrétiens (msAraboChristian). Cet élément manque dans MASTER.

element **msCategory**

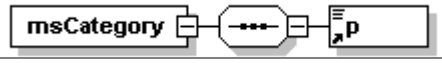
| | |
|---|---|
| diagram |   |
| children | p |
| used by | element msIdentifier |
| attributes | Name: mstype; Type: xs:NMTOKEN; Use: ; Default: ; Fixed: |
| source | <xs:element name="msCategory"><br><xs:complexType><br><xs:sequence><br><xs:element ref="p"/><br></xs:sequence><br><xs:attribute name="mstype"><br><xs:simpleType><br><xs:restriction base="xs:NMTOKEN"><br><xs:enumeration value="msAraboIslamic"/><br><xs:enumeration value="msAraboChristian"/><br></xs:restriction><br></xs:simpleType><br></xs:attribute><br></xs:complexType><br></xs:element> |

**Figure n°.16: Les element msCategory**

**1.9. msSubject:** il s'agit d'une catégorie qui n'existe pas dans MASTER et qui indique le sujet ou les sujets du manuscrit, msSubject est reparti en trois sous-éléments : sujet principal « mainSubj », autres sujets « otherSubj », et mots clés « keywords ».

element **msSubject**

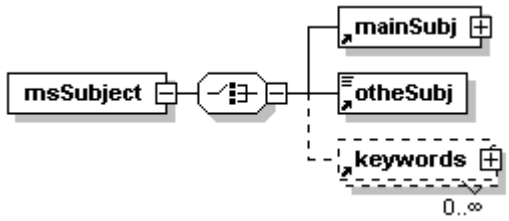
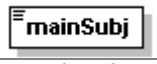
| | |
|---|---|
| diagram |   |
| children | mainSubj otheSubj keywords |
| used by | element msIdentifier |
| source | <xs:element name="msSubject"><br><xs:complexType><br><xs:choice><br><xs:element ref="mainSubj"/><br><xs:element ref="otheSubj"/><br><xs:element ref="keywords" minOccurs="0" maxOccurs="unbounded"/><br></xs:choice><br></xs:complexType><br></xs:element> |

**Figure n°.17: Les éléments msSubject**

**1.9.1. mainSubj**: il contient un ou plusieurs sujets principaux du manuscrit.

element **mainSubj**

| | |
|---|---|
| diagram |  |
| type | extension of **xs:string** |
| used by | element **msSubject** |

| attributes | Name | Type | Use | Default | Fixed |
|---|---|---|---|---|---|
| | type | xs:NMTOKEN | | | |
| | p | xs:string | | | |

source

```
<xs:element name="mainSubj">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute name="type">
     <xs:simpleType>
      <xs:restriction base="xs:NMTOKEN">
       <xs:enumeration value="Algebra"/>
       <xs:enumeration value="araLangGram"/>
       <xs:enumeration value="Arithemetic"/>
       <xs:enumeration value="Astronomy"/>
       <xs:enumeration value="Botany"/>
       <xs:enumeration value="chrisTheology"/>
       <xs:enumeration value="Documentary"/>
       <xs:enumeration value="Ethics"/>
       <xs:enumeration value="Fiqh"/>
       <xs:enumeration value="Genealogy"/>
       <xs:enumeration value="Geography"/>
       <xs:enumeration value="Geometry"/>
       <xs:enumeration value="Hadith"/>
       <xs:enumeration value="History"/>
       <xs:enumeration value="islAraPhilos"/>
       <xs:enumeration value="IslamTheology"/>
       <xs:enumeration value="Juridical"/>
       <xs:enumeration value="Koran"/>
       <xs:enumeration value="Language"/>
       <xs:enumeration value="langLiter"/>
       <xs:enumeration value="Literature"/>
       <xs:enumeration value="Medicine"/>
       <xs:enumeration value="metaphysics"/>
       <xs:enumeration value="Mystic"/>
       <xs:enumeration value="Pharmacy"/>
       <xs:enumeration value="Philosophy"/>
       <xs:enumeration value="PoliticalScience"/>
       <xs:enumeration value="Science"/>
       <xs:enumeration value="Tafsir"/>
       <xs:enumeration value="Travels"/>
       <xs:enumeration value="Zoology"/>
      </xs:restriction>
     </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="p" type="xs:string"/>
   </xs:extension>
  </xs:simpleContent>
 </xs:complexType>
</xs:element>
```

**Figure n°.18: Les attributs mainSubject**

**1.9.2. otherSubj** : il indique les sujets autres que le sujet ou les sujets principaux.

Guidée par notre étude de manuscrits arabes et inspirée par l'enquête menée auprès de spécialistes de manuscrits arabes, nous allons proposer les sujets suivants afin de faciliter la classification de ces ouvrages :

Coran (Koran), hadith (Hadith) interprétation du Coran (tafsir), jurisprudence (Fiqh), texte religieux chrétien (chrisReligTex), science (science), littérature (literature), documentaire (documentary), juridique (Juridical), philosophie (philosophy), histoire (history), grammaire de langue arabe (araLangGram), mystique (mystic), islamique philosophie (islAraPhilos).

element **otheSubj**

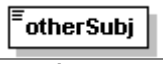
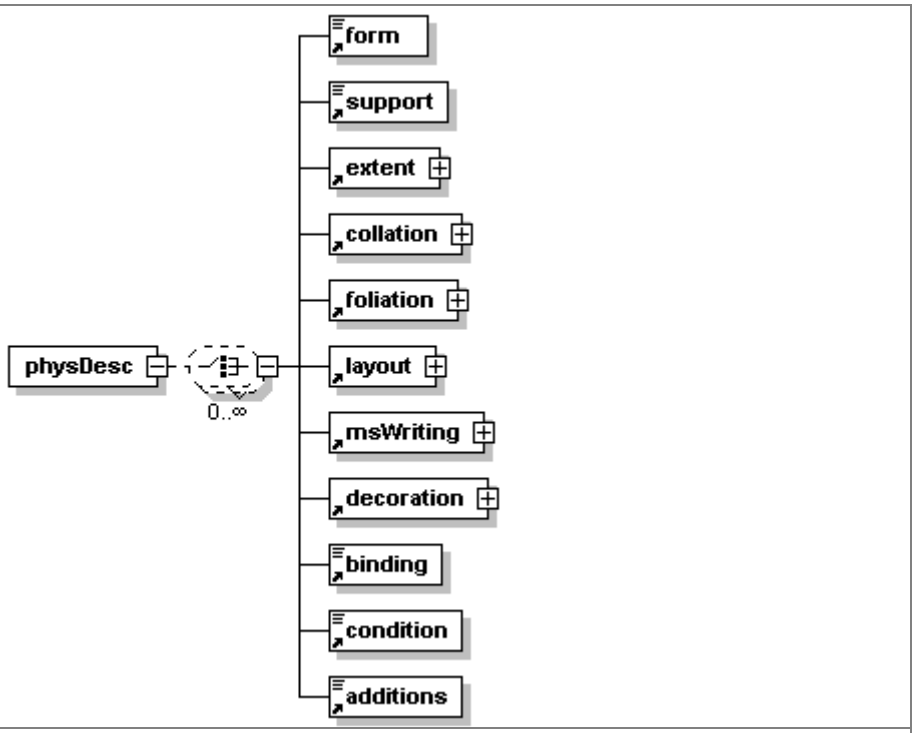
| | |
|---|---|
| diagram |   |
| type | **xs:string** |
| used by | element **msSubject** |
| source | <xs:element name="otherSubj" type="xs:string"/> |

**Figure n°.19: L'élément otherSubj**

**1.9.3. keywords** : il fournit un ou plusieurs mots clés autres que le sujet pour affiner le sujet du document.

**2. physDesc:** Il contient des informations sur la description physique d'un manuscrit ou d'une partie d'un manuscrit, comme la forme, la collation, la composition du cahier, le nombre de folios, la réclame, le cachet, etc. (cf. la figure suivante)

element **physDesc**

| | |
|---|---|
| diagram |   |
| children | **form support extent collation foliation layout msWriting decoration binding condition additions** |
| used by | element **msDescription** |

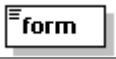
source

```
<xs:element name="physDesc">
 <xs:complexType>
  <xs:choice minOccurs="0" maxOccurs="unbounded">
   <xs:element ref="form"/>
   <xs:element ref="support"/>
   <xs:element ref="extent"/>
   <xs:element ref="collation"/>
   <xs:element ref="foliation"/>
   <xs:element ref="layout"/>
   <xs:element ref="msWriting"/>
   <xs:element ref="decoration"/>
   <xs:element ref="binding"/>
   <xs:element ref="condition"/>
   <xs:element ref="additions"/>
  </xs:choice>
 </xs:complexType>
</xs:element>
```

**Figure n°.20: Les éléments et les attributs du *physDesc***

**2.1. Form** : il décrit la forme dans laquelle le manuscrit a été écrit, soit sous la forme de codex, de rouleau, soit sous la forme de charte.

element **form**

diagram 


type extension of **xs:string**

used by element **physDesc**

attributes

| Name | Type | Use | Default | Fixed |
|---|---|---|---|---|
| msForm | xs:NMTOKEN | | | |

source

```
<xs:element name="form">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute name="msForm">
     <xs:simpleType>
      <xs:restriction base="xs:NMTOKEN">
       <xs:enumeration value="codex"/>
       <xs:enumeration value="roller"/>
       <xs:enumeration value="chart"/>
      </xs:restriction>
     </xs:simpleType>
    </xs:attribute>
   </xs:extension>
  </xs:simpleContent>
 </xs:complexType>
</xs:element>
```

**Figure n°.21: L'élément et les attributes *form***

**2.2. support** : le support se repartit en trois attributs importants.

La première est le «suppKind» : il s'agit de la variété du support utilisé (papier, parchemin ou papyrus).

Le deuxième «paperType» indique, en cas de support papier, le type de papier utilisé (arabique ou occidental).

Le troisième «other» donne la possibilité au catalogueur de mettre une autre information qu'il trouve nécessaire et qui n'est pas indiquée auparavant.

En cas d'utilisation de papier «arabique», il existe des attributs qui aident à l'identification de ce genre de papier (Sulimani, Talhi, Nohi, Faraouni, Jaafari, Tahiri); un autre attribut est aussi présent pour d'autres types que les précédents «otherType».

Par contre, si le type de papier est occidental, on distingue les attributs suivants : avec filigrane (WaterMark) et sans filigrane. Si c'est un papier avec filigrane, il existe un champ avec WaterMarkType pour mettre le type de filigrane utilisé pour la fabrication du papier.

element **support**

| | |
|---|---|
| diagram |   |
| type | extension of **xs:string** |
| used by | element **physDesc** |

| attributes | Name | Type | Use | Default | Fixed |
|---|---|---|---|---|---|
| | suppoKind | xs:NMTOKEN | | | |
| | papertype | xs:NMTOKEN | | | |
| | others | xs:string | | | |

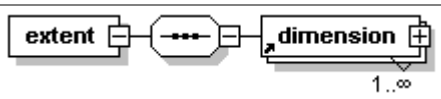
source

```
<xs:element name="support">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute name="suppoKind">
     <xs:simpleType>
      <xs:restriction base="xs:NMTOKEN">
       <xs:enumeration value="paper"/>
       <xs:enumeration value="parcheman"/>
       <xs:enumeration value="papyrus"/>
      </xs:restriction>
     </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="papertype">
     <xs:simpleType>
      <xs:restriction base="xs:NMTOKEN">
       <xs:enumeration value="Arabic"/>
       <xs:enumeration value="occidentale"/>
      </xs:restriction>
     </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="others" type="xs:string"/>
   </xs:extension>
  </xs:simpleContent>
 </xs:complexType>
</xs:element>
```

**Figure n°.22: Les attributs *support***

**2.3. extent**. Il décrit la taille approximative du manuscrit, spécifiée dans n'importe quelle unité adéquate, en cm ou en mm, etc.

element **extent**

| | |
|---|---|
| diagram |   |
| children | **dimension** |
| used by | element **physDesc** |
| source | <xs:element name="extent"><br><xs:complexType> |

```
    <xs:sequence>
      <xs:element ref="dimension" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
```

**Figure n°.23 : Les éléments *extent***

**2.3.1. dimension** : l'élément dimension est un élément fils de l'élément « extent » qui contient les informations sur la dimension du folio (leaves), de l'espace réglé (ruled), l'espace percé (pricked) et l'espace écrit (writtensurface) ainsi que la dimension des miniatures, la dimension de la reliure (binding) et la dimension de l'étui (box). Les sous-éléments : hauteur (height), largeur (width) et profondeur (depth) sont des outils dans l'élément « dimension » pour mesurer les différentes parties du manuscrit mentionnée ci-dessus.

element **dimension**

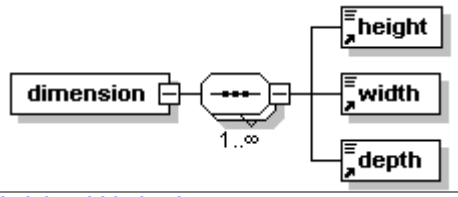
| | |
|---|---|
| Diagram |  |
| Children | **height width depth** |
| used by | element **extent** |
| Attributes | Name: type; Type: xs:NMTOKEN; Use: ; Default: ; Fixed: |
| Source | <xs:element name="dimension"><br><xs:complexType><br><xs:sequence maxOccurs="unbounded"><br><xs:element ref="height"/><br><xs:element ref="width"/><br><xs:element ref="depth"/><br></xs:sequence><br><xs:attribute name="type"><br><xs:simpleType><br><xs:restriction base="xs:NMTOKEN"><br><xs:enumeration value="leaves"/><br><xs:enumeration value="ruled"/><br><xs:enumeration value="pricked"/><br><xs:enumeration value="writtensurface"/><br><xs:enumeration value="miniatures"/><br><xs:enumeration value="binding"/><br><xs:enumeration value="box"/><br></xs:restriction><br></xs:simpleType><br></xs:attribute><br></xs:complexType><br></xs:element> |

**Figure n°.24: Les éléments et les attributs *dimansion***

**2.4. collation** il contient le nombre de folios composant un manuscrit. Il est répartit en trois éléments fils: formula, cahierComp, other. Dans <p> le catalogueur peut mettre de façon libre le nombre de folios trouvés.

element **collation**

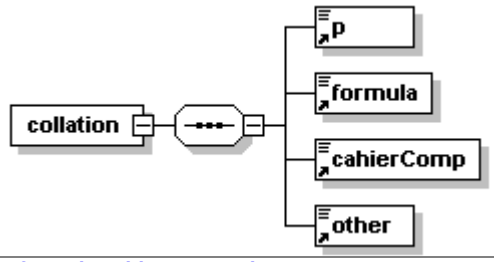
| | |
|---|---|
| Diagram |  |
| Children | p formula cahierComp other |
| used by | elements msItem<br>physDesc |
| source | <xs:element name="collation"><br><xs:complexType><br><xs:sequence><br><xs:element ref="p"/><br><xs:element ref="formula"/><br><xs:element ref="cahierComp"/><br><xs:element ref="other"/><br></xs:sequence><br></xs:complexType><br></xs:element> |

**Figure n°.25 : Les éléments *collation***

**2.4.1. formula** : il décrit des informations particulières qui peuvent être trouvées dans le manuscrit, telles des pages écrites dans un style différent et qui sont répertoriées de telle page à telle page. Dans MASTER, l'élément extent est utilisé dans le même sens mais se situe comme élément fils de l'élément « collation ».

**2.4.2. cahierComp** : le sous-élément cahierComp est le troisième parmi les éléments fils d'extent ; il contient l'information sur la composition du cahier, surtout son élément fils noBifolia. Nous avons ajouté les attributs (ternion, quaternion, quinion, senion) pour aider le catalogueur à choisir le type de composition de chaque cahier composant le manuscrit. Il s'agit de trois, quatre, cinq ou six bi-folios etc.

element **cahierComp**

| | |
|---|---|
| diagram |  |
| type | extension of **xs:string** |
| used by | element collation |

| attributes | Name | Type | Use | Default | Fixed |
|---|---|---|---|---|---|
| | noBifolia | xs:NMTOKEN | | | |
| | other | xs:string | | | |

source

```
<xs:element name="cahierComp">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute name="noBifolia">
     <xs:simpleType>
      <xs:restriction base="xs:NMTOKEN">
       <xs:enumeration value="ternion"/>
       <xs:enumeration value="quaternion"/>
       <xs:enumeration value="quinion"/>
       <xs:enumeration value="senion"/>
      </xs:restriction>
```

```
        </xs:simpleType>
      </xs:attribute>
      <xs:attribute name="other" type="xs:string"/>
    </xs:extension>
  </xs:simpleContent>
 </xs:complexType>
</xs:element>
```

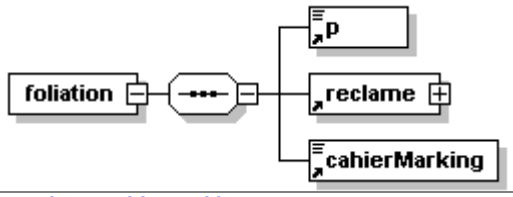
**Figure n°.26 : Les éléments et les attributs cahierComp**

**2.4.3. other** : l'élément « other » contient des informations sur un autre type de cahier que n'a pas été mentionné auparavant. Cet élément n'existe pas dans MASTER car cet élément ne concerne que les manuscrits arabes.

**2.5. foliation** : élément décrivant une ou plusieurs formes de pagination appliquée dans un manuscrit, comme la pagination de folios ou de cahier, etc. Il est réparti en deux sous-éléments : le réclame qui sert comme une sorte de pagination et le cahierMarking qui sert de type de marquage de cahier.

element **foliation**

diagram




children p reclame cahierMarking

used by element physDesc

source

```
<xs:element name="foliation">
 <xs:complexType>
  <xs:sequence>
   <xs:element ref="p"/>
   <xs:element ref="reclame"/>
   <xs:element ref="cahierMarking"/>
  </xs:sequence>
 </xs:complexType>
</xs:element>
```

**Figure n°.27 : Les éléments *foliation***

**2.5.1. reclame** : il contient des informations sur la réclame et sa composition (les trois ou quatre derniers caractères du mot ou le dernier mot entier de la page précédente

**2.5.2. cahierMarking** : il fournit des informations sur le type de marquage trouvé dans le manuscrit. Nous avons proposé, dans le sous-élément markType, les attributs suivants :

*allAraLet* : il indique que toutes les lettres de marquage sont en langue arabe. L'attribut *allAraSyriLet* montre que le marquage de certains manuscrits est fait en deux langues : arabique et syriaque. Alors que *coptNos* expose le marquage en chiffres coptes. Et *araNos* présente le marquage en chiffres arabes. Cependant le sous-élément *other* donne

la possibilité d'ajouter d'autres informations qui ne sont pas mentionnées parmi les attributs proposés auparavant.

element **cahierMarking**

| | |
|---|---|
| diagram | cahierMarking |
| type | extension of **xs:string** |
| used by | elements **foliation**<br>**physDesc** |

| attributes | Name | Type | Use | Default | Fixed |
|---|---|---|---|---|---|
| | markType | xs:NMTOKEN | | | |
| | other | xs:string | | | |

source

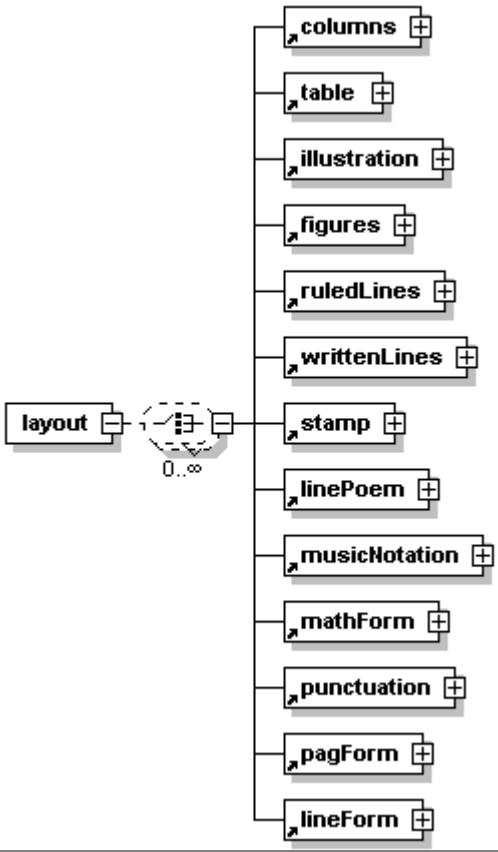
```
<xs:element name="cahierMarking">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute name="markType">
     <xs:simpleType>
      <xs:restriction base="xs:NMTOKEN">
       <xs:enumeration value="allAraLet"/>
       <xs:enumeration value="allAraSyriLet"/>
       <xs:enumeration value="coptNos"/>
       <xs:enumeration value="araNos"/>
      </xs:restriction>
     </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="other" type="xs:string"/>
   </xs:extension>
  </xs:simpleContent>
 </xs:complexType>
</xs:element>
```

**Figure n°.28: Les éléments *cahierMarking***

**2.6. layout** : l'élément disposition de page ou la mise en page (layout) est utilisée pour décrire la manière dont le texte et l'illustration sont disposés sur les pages du manuscrit. Dans l'élément « pagPresentation » nous avons mis les éléments qui peuvent être trouvés dans un manuscrit comme la colonne (columns), le tableau (table), l'illustration (illustration), les figures (figures) le réglage des lignes (ruledLines) et le nombre des lignes d'écriture (writtenLines), surtout dans l'élément fils nomLigne. LignePoem est un autre sous-élément pour indiquer l'existence de poème dans le texte, à quelle page et à quelle ligne.

element **layout**

diagram




children **columns table illustration figures ruledLines writtenLines stamp linePoem musicNotation mathForm punctuation pagForm lineForm**

used by element **physDesc**

source

```
<xs:element name="layout">
  <xs:complexType>
    <xs:choice minOccurs="0" maxOccurs="unbounded">
      <xs:element ref="columns"/>
      <xs:element ref="table"/>
      <xs:element ref="illustration"/>
      <xs:element ref="figures"/>
      <xs:element ref="ruledLines"/>
      <xs:element ref="writtenLines"/>
      <xs:element ref="stamp"/>
      <xs:element ref="linePoem"/>
      <xs:element ref="musicNotation"/>
      <xs:element ref="mathForm"/>
      <xs:element ref="punctuation"/>
      <xs:element ref="pagForm"/>
      <xs:element ref="lineForm"/>
    </xs:choice>
  </xs:complexType>
</xs:element>
```

**Figure n°. 29: Les éléments *layout***

**2.6.9. musicNotation** : il contient des informations concernent les notes musicales trouvées dans le texte.

**2.6.10. mathForm** : il contient ce qui concerne des informations dans le texte qui ne sont pas d'écriture normale comme une formule mathématique par exemple.

**2.7. msWriting** : il contient la description des différentes écritures utilisées pour écrire un manuscrit. L'élément msWriting est réparti en deux sous-éléments : handDesc et « p ».

element **msWriting**

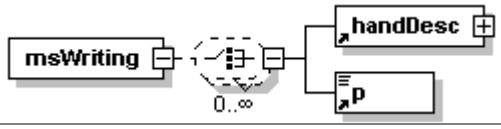
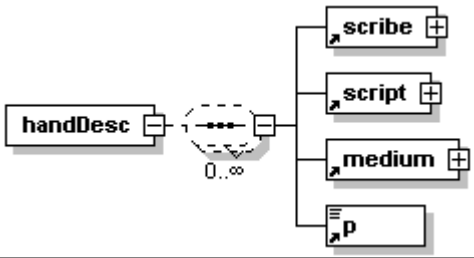
| | |
|---|---|
| diagram |   |
| children | handDesc p |
| used by | element physDesc |
| attributes | Name: hands; Type: xs:NMTOKEN; Use: ; Default: ; Fixed: |
| source | <xs:element name="msWriting"><br><xs:complexType><br><xs:choice minOccurs="0" maxOccurs="unbounded"><br><xs:element ref="handDesc"/><br><xs:element ref="p"/><br></xs:choice><br><xs:attribute name="hands"><br><xs:simpleType><br><xs:restriction base="xs:NMTOKEN"><br><xs:enumeration value="firstHand"/><br><xs:enumeration value="secondHand"/><br><xs:enumeration value="thirdHand"/><br><xs:enumeration value="secFol"/><br></xs:restriction><br></xs:simpleType><br></xs:attribute><br></xs:complexType><br></xs:element> |

**Figure n°.30: Les éléments *msWriting***

**2.7.1. handDesc** : l'élément fils handDesc décrit tout se qui concerne l'écriture du texte.

element **handDesc**

| | |
|---|---|
| diagram |   |
| children | scribe script medium p |
| used by | elements msWriting physDesc |
| attributes | Name: scope; Type: xs:NMTOKEN; Use: ; Default: ; Fixed: |
| source | <xs:element name="handDesc"><br><xs:complexType><br><xs:sequence minOccurs="0" maxOccurs="unbounded"><br><xs:element ref="scribe"/><br><xs:element ref="script"/><br><xs:element ref="medium"/><br><xs:element ref="p"/><br></xs:sequence><br><xs:attribute name="scope"><br><xs:simpleType><br><xs:restriction base="xs:NMTOKEN"><br><xs:enumeration value="sole"/> |

```
        <xs:enumeration value="major"/>
        <xs:enumeration value="minor"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:attribute>
 </xs:complexType>
</xs:element>
```

**Figure n°.31: Les éléments *handDesc***

**2.7.1.1. scribe** : il contient le nom(s) de(s) personne(s) autres que le copiste principal qui a participé à la rédaction du manuscrit

**2.7.1.2. : script** : il décrit le style d'écriture principal et les autres styles utilisés pour l'écriture du texte manuscrit. Dans writStyle nous avons proposé les styles d'écriture à partir de notre étude mais nous avons gardé la possibilité d'ajouter d'autres styles dans le sous-élément otherStyle.

Les styles proposés sont les suivants : Coufi, Diwani, Farisi, Higazi, Houroufal-Taaj, Ijaza, Kufi, Kufi-Occidental-Tunisie, Kufi-Oriental-Iraq-Iran, Magribi, Magribi-Andalou, Muhaqqaq, Muhaqqaq-Arabe, Muhaqqaq-Turquie, Nashki, Nashki-Arabe, Nashki-Egypt, Nashki-Inde, Nashki-micrographie-Egypt, Nashki-Persan, Nashki-Syrie, Nashki-Tulut-Iraq, Nastaliq-Persan, Orientale-Egypt, Rouqa, Taghra, Tulut, Tulut-Muhaqqaq, Tulut-Iran, Tulut-Muhaqqaq-Nashki-Turquie.

element **script**

| | |
|---|---|
| diagram |   |
| children | **image** |
| used by | elements **handDesc msItem** |
| attributes | Name: writStyle, otherStyle; Type: xs:NMTOKEN, xs:string; Use; Default; Fixed |
| source | <xs:element name="script"><br><xs:complexType><br><xs:sequence minOccurs="0" maxOccurs="unbounded"><br><xs:element ref="image"/><br></xs:sequence><br><xs:attribute name="writStyle"><br><xs:simpleType><br><xs:restriction base="xs:NMTOKEN"><br><xs:enumeration value="Diwani"/><br><xs:enumeration value="Farisi"/><br><xs:enumeration value="Higazi"/><br><xs:enumeration value="Houroufal-Taaj"/><br><xs:enumeration value="Ijaza"/><br><xs:enumeration value="Kufi"/><br><xs:enumeration value="Kufi-Occidental-Tunisie"/><br><xs:enumeration value="Kufi-Oriental-Iraq-Iran"/><br><xs:enumeration value="Magribi"/><br><xs:enumeration value="Magribi-Andalou"/><br><xs:enumeration value="Muhaqqaq"/><br><xs:enumeration value="Muhaqqaq-Arabe"/><br><xs:enumeration value="Muhaqqaq-Turquie"/> |

```
        <xs:enumeration value="Naskhi"/>
        <xs:enumeration value="Naskhi-Arabe"/>
        <xs:enumeration value="Naskhi-Egypt"/>
        <xs:enumeration value="Naskhi-Inde"/>
        <xs:enumeration value="Naskhi-micrographie-Egypt"/>
        <xs:enumeration value="Naskhi-Persan"/>
        <xs:enumeration value="Naskhi-Syrie"/>
        <xs:enumeration value="Naskhi-Tulut-Iraq"/>
        <xs:enumeration value="Nastaliq-Persan"/>
        <xs:enumeration value="Orientale-Egypt"/>
        <xs:enumeration value="Rouqa"/>
        <xs:enumeration value="Taghra"/>
        <xs:enumeration value="Tulut"/>
        <xs:enumeration value="Tulut-Muhaqqaq"/>
        <xs:enumeration value="Tulut-Iran"/>
        <xs:enumeration value="Tulut-Muhaqqaq-Nash-Turquie"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:attribute>
  <xs:attribute name="otherStyle" type="xs:string"/>
 </xs:complexType>
</xs:element>
```

**Figure n°.32: Les éléments *script***

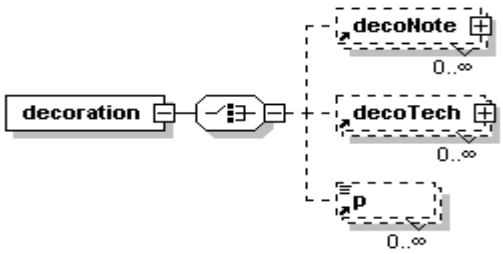
**2.7.1.3. medium**: le médium d'écriture, c'est à dire la teinte ou le type d'encre utilisé. Il donne la possibilité de mettre les différentes couleurs d'encre trouvées, soit pour le texte intégral, soit pour une ou des parties du texte comme le titre de chapitre et de sous chapitre qui se trouvent écrits en couleur autre que celle du reste du texte.

Cependant, scope est un attribut à l'intérieur de handDesc qui spécifie le taux de participation d'un copiste dans l'écriture des manuscrits (sole, major ou minore).

**2.8. decoration:** il contient une description du décor trouvé dans le texte. Le décor est réparti en deux éléments fils principaux :

element **decoration**

diagram




children **decoNote decoTech p**

used by element **physDesc**

source

```
<xs:element name="decoration">
 <xs:complexType>
  <xs:choice>
   <xs:element ref="decoNote" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="decoTech" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="p" minOccurs="0" maxOccurs="unbounded"/>
  </xs:choice>
 </xs:complexType>
</xs:element>
```

**Figure n°.33 : Les elements *decoration***

**2.8.1 decoNote**: il s'agit des notes qui décrivent les composants du décor. Cet élément est similaire de celui du MASTER, sauf en ce qui concerne ses attributs, car dans MASTER,

il définit une liste d'attributs tels que type, sub-type, technique, qualité, figurative - alors que pour faciliter le tâche des catalogueurs, nous avons trouvé utile de mettre « p » comme zone libre pour d'autres descriptions.

element **decoNote**

| | |
|---|---|
| diagram |   |
| children | p |
| used by | element decoration |

| attributes | Name | Type | Use | Default | Fixed |
|---|---|---|---|---|---|
| | type | xs:string | | | |
| | subtype | xs:string | | | |
| | quality | xs:string | | | |
| | figurative | xs:NMTOKEN | | na | |
| | illustrative | xs:NMTOKEN | | u | |

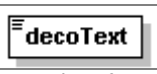
source

```
<xs:element name="decoNote">
 <xs:complexType>
  <xs:sequence minOccurs="0" maxOccurs="unbounded">
   <xs:element ref="p"/>
  </xs:sequence>
  <xs:attribute name="type" type="xs:string"/>
  <xs:attribute name="subtype" type="xs:string"/>
  <xs:attribute name="quality" type="xs:string"/>
  <xs:attribute name="figurative" default="na">
   <xs:simpleType>
    <xs:restriction base="xs:NMTOKEN">
     <xs:enumeration value="yes"/>
     <xs:enumeration value="no"/>
     <xs:enumeration value="na"/>
    </xs:restriction>
   </xs:simpleType>
  </xs:attribute>
  <xs:attribute name="illustrative" default="u">
   <xs:simpleType>
    <xs:restriction base="xs:NMTOKEN">
     <xs:enumeration value="y"/>
     <xs:enumeration value="n"/>
     <xs:enumeration value="u"/>
    </xs:restriction>
   </xs:simpleType>
  </xs:attribute>
 </xs:complexType>
</xs:element>
```

**Figure n°.34: les éléments et les attributs *decNote***

**2.8.2. decoTech** : la technique de décor. Il contient des descriptions sur la caractéristique de technique du décor qui existent dans les manuscrits arabes. Nous avons divisé cet élément en trois parties : la technique de texte, la technique de décor dans le texte coranique et le décor de reliure.

element **decoText**

| | |
|---|---|
| diagram |   |
| type | extension of **xs:string** |
| used by | element decoTech |

| attributes | Name | Type | Use | Default | Fixed |
|---|---|---|---|---|---|
| | decType | xs:NMTOKEN | | | |

| | |
|---|---|
| | decPlace xs:NMTOKEN |
| source | <xs:element name="decoText"><br><xs:complexType><br><xs:simpleContent><br><xs:extension base="xs:string"><br><xs:attribute name="decType"><br><xs:simpleType><br><xs:restriction base="xs:NMTOKEN"><br><xs:enumeration value="Shamsas"/><br><xs:enumeration value="miniatures"/><br><xs:enumeration value="illustration"/><br><xs:enumeration value="drawings"/><br><xs:enumeration value="arabisque"/><br><xs:enumeration value="DecoMargin"/><br></xs:restriction><br></xs:simpleType><br></xs:attribute><br><xs:attribute name="decPlace"><br><xs:simpleType><br><xs:restriction base="xs:NMTOKEN"><br><xs:enumeration value="DecoFullPage"/><br><xs:enumeration value="decoSection"/><br><xs:enumeration value="decoColophon"/><br><xs:enumeration value="decoTextDivision"/><br><xs:enumeration value="decoPagFram"/><br><xs:enumeration value="DecoMargin"/><br><xs:enumeration value="decoLining"/><br></xs:restriction><br></xs:simpleType><br></xs:attribute><br></xs:extension><br></xs:simpleContent><br></xs:complexType><br></xs:element> |

**Figure n°.35: Les éléments et les attributes *decText***

**2.10.2.1. decoText** : il décrit la forme de décor trouvé dans le texte.

Dans le decType qui est un élément fils de decoText, nous avons mis la liste des attributs suivants qui représentent le décor trouvé dans le texte comme (Shamsas), un style de décor connu sous ce nom, la miniature, les illustrations, les dessins (drawings), les arabesques et le décor sur la marge (DecoMargin), alors que le deuxième élément fils decoPlace indique la place du décor à l'intérieur de texte. Il s'agit du décor de pages entières (DecoFullPage) ou du décorde certaines parties du manuscrit (decoSection), du décor du colophon (decoColophon), du décor de séparation ou de division de textes (decoTextDivision), du décor de cadre des pages (decoPagFram), du décor de frontispice (frontspice), du décor sur la marge (DecoMargin) et du décor de doublure (decoLining).

element **decoText**

| | | | | | |
|---|---|---|---|---|---|
| diagram |  | | | | |
| type | extension of **xs:string** | | | | |
| used by | element **decoTech** | | | | |
| attributes | Name | Type | Use | Default | Fixed |
| | decType | xs:NMTOKEN | | | |
| | decPlace | xs:NMTOKEN | | | |
| source | <xs:element name="decoText"> | | | | |

```
<xs:complexType>
 <xs:simpleContent>
  <xs:extension base="xs:string">
   <xs:attribute name="decType">
    <xs:simpleType>
     <xs:restriction base="xs:NMTOKEN">
      <xs:enumeration value="Shamsas"/>
      <xs:enumeration value="miniatures"/>
      <xs:enumeration value="illustration"/>
      <xs:enumeration value="drawings"/>
      <xs:enumeration value="arabisque"/>
      <xs:enumeration value="DecoMargin"/>
     </xs:restriction>
    </xs:simpleType>
   </xs:attribute>
   <xs:attribute name="decPlace">
    <xs:simpleType>
     <xs:restriction base="xs:NMTOKEN">
      <xs:enumeration value="DecoFullPage"/>
      <xs:enumeration value="decoSection"/>
      <xs:enumeration value="decoColophon"/>
      <xs:enumeration value="decoTextDivision"/>
      <xs:enumeration value="decoPagFram"/>
      <xs:enumeration value="DecoMargin"/>
      <xs:enumeration value="decoLining"/>
     </xs:restriction>
    </xs:simpleType>
   </xs:attribute>
  </xs:extension>
 </xs:simpleContent>
</xs:complexType>
</xs:element>
```

**Figure n°.36: Les éléments *decText***

**2.10.2.2. decoTeKoran** : il décrit le genre de décor trouvé dans le texte coranique, surtout le décor de la séparation des sourates, le frontispice et le commencement des sourates, etc. L'élément fils type fournit les attributs qu'il s'agisse du décor de première page du manuscrit (frontispice), du décor du titre de chaque sourate (decoTiSourate), des décors qui séparent les versets en différents niveaux après chaque verset (af-1-Verse), après cinq versets (af-5-Verse), après dix versets (af-10-Verse). L'attribut (divRamadan) concerne le décor des divisions du Coran, destinées à la lecture pendant le mois de Ramadan.

**2.10.2.3. decBinding** : il contient la description du style de décor utilisé dans la fabrication de la reliure.

BindingDesc est un élément fils de decBinding. Il fournit les informations sur la technique de reliure.

**2.11. Binding** : l'élément reliure (binding) contient l'élément fils concernant le style de reliure (bindTypes) avec les attributs suivants : reliure française (bindFrench), reliure maghrébine (bindMaghriban), reliure orientale (bindOriental), reliure égyptienne

(bindEgyptian), reliure ottomane (bindOthoman) et la dernière étant l'attribut autre (others) qui concerne d'autres reliures que les précédentes.

**2.12. condition** : il fournit des informations sur la condition physique du document. En ce qui concerne la condition physique « conPhysique », il existe trois attributs : bon (good), moyen (medium) et mauvais (bad).

element **condition**

| | |
|---|---|
| diagram | condition |
| type | extension of **xs:string** |
| used by | element **physDesc** |
| attributes | Name: conPhysique; Type: xs:NMTOKEN; Use: ; Default: ; Fixed: |
| source | <xs:element name="condition"><br><xs:complexType><br><xs:simpleContent><br><xs:extension base="xs:string"><br><xs:attribute name="conPhysique"><br><xs:simpleType><br><xs:restriction base="xs:NMTOKEN"><br><xs:enumeration value="good"/><br><xs:enumeration value="medium"/><br><xs:enumeration value="bad"/><br></xs:restriction><br></xs:simpleType><br></xs:attribute><br></xs:extension><br></xs:simpleContent><br></xs:complexType><br></xs:element> |

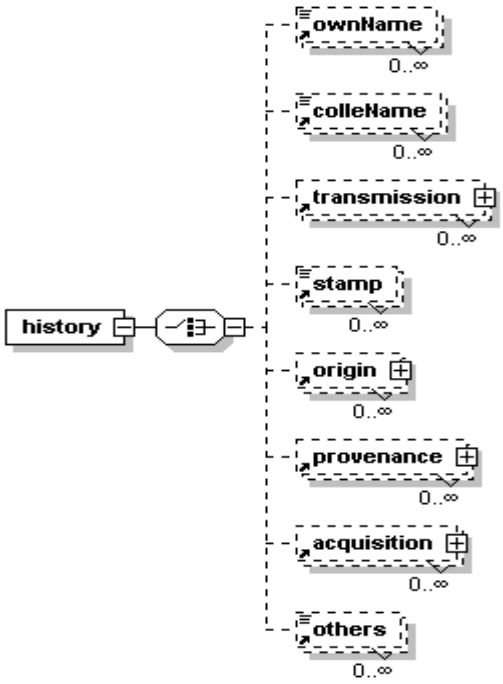
**Figure n°.37 : Les éléments et les attributs *condition***

**2.13. additions** : pour fournir d'autres éléments afin de décrire l'état du document et qui n'ont pas été mentionnés ci-dessus.

**3. history :** il regroupe les éléments qui décrivent le histoire entière du manuscrit. Il contient les éléments fils suivants :

element **history**

diagram




children **ownName colleName transmission stamp origin provenance acquisition others**

used by element **msDescription**

attributes

| Name | Type | Use | Default | Fixed |
|---|---|---|---|---|
| Status | xs:NMTOKEN | | | |
| notBefore | xs:string | | | |
| notAfter | xs:string | | | |
| evidence | xs:NMTOKEN | | | |

source

```
<xs:element name="history">
 <xs:complexType>
  <xs:choice>
   <xs:element ref="ownName" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="colleName" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="transmission" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="stamp" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="origin" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="provenance" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="acquisition" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="others" minOccurs="0" maxOccurs="unbounded"/>
  </xs:choice>
  <xs:attribute name="Status">
   <xs:simpleType>
    <xs:restriction base="xs:NMTOKEN">
     <xs:enumeration value="Autographes"/>
     <xs:enumeration value="Apographes"/>
     <xs:enumeration value="Unique"/>
     <xs:enumeration value="waqf"/>
    </xs:restriction>
   </xs:simpleType>
  </xs:attribute>
  <xs:attribute name="notBefore" type="xs:string"/>
  <xs:attribute name="notAfter" type="xs:string"/>
  <xs:attribute name="evidence">
   <xs:simpleType>
    <xs:restriction base="xs:NMTOKEN">
     <xs:enumeration value="internal"/>
     <xs:enumeration value="external"/>
     <xs:enumeration value="conjecture"/>
    </xs:restriction>
   </xs:simpleType>
  </xs:attribute>
 </xs:complexType>
```

| | |
|---|---|
| | </xs:element> |

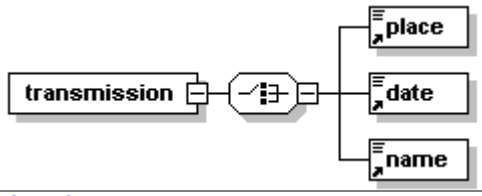
**Figure n°. 38: Les éléments *history***

**3.1. ownName** (le nom du possesseur) : il contient des informations sur le(s) nom(s) de(s) possesseur(s).

**3.2. colleName** (le nom du collecteur) : il contient des informations sur le(s) nom(s) de(s) collecteur(s).

**3.3. transmission** : les informations mentionnées au cours de la transmission. Il contient les noms des personnes mentionnées au cours de la transmission, surtout le(s) nom(s) de(s) personne(s) qui écoute (nt) « sama », la personne qui a lu le manuscrit «qirah » et le(s) nom(s) de(s) personne(s) qui donne (nt) le diplôme « ijaza » à la personne qui a lu le manuscrit. Il contient également le nom du lieu et de la date de la cérémonie.

element **transmission**

| | |
|---|---|
| diagram |  |
| children | **place** **date** **name** |
| used by | element **history** |

| attributes | Name | Type | Use | Default | Fixed |
|---|---|---|---|---|---|
| | sama | xs:NMTOKEN | | | |
| | qirah | xs:NMTOKEN | | | |
| | ijaza | xs:NMTOKEN | | | |

source

```
<xs:element name="transmission">
 <xs:complexType>
  <xs:choice>
   <xs:element ref="place"/>
   <xs:element ref="date"/>
   <xs:element ref="name"/>
  </xs:choice>
  <xs:attribute name="sama">
   <xs:simpleType>
    <xs:restriction base="xs:NMTOKEN">
     <xs:enumeration value="nOfTimes"/>
     <xs:enumeration value="perListening"/>
     <xs:enumeration value="place"/>
    </xs:restriction>
   </xs:simpleType>
  </xs:attribute>
  <xs:attribute name="qirah">
   <xs:simpleType>
    <xs:restriction base="xs:NMTOKEN">
     <xs:enumeration value="NOfTimes"/>
     <xs:enumeration value="perReading"/>
     <xs:enumeration value="place"/>
    </xs:restriction>
   </xs:simpleType>
  </xs:attribute>
  <xs:attribute name="ijaza">
   <xs:simpleType>
    <xs:restriction base="xs:NMTOKEN">
     <xs:enumeration value="Donor"/>
```

```
        <xs:enumeration value="place"/>
        <xs:enumeration value="date"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:attribute>
 </xs:complexType>
</xs:element>
```

**Figure n°.39: Les éléments *transmission***

**3.4. stamp** (le cachet) : il fournit la possibilité d'indiquer l'existence du cachet et dans quelle page il se situe. Le cachet est un élément important pour qu'un spécialiste dans l'histoire des manuscrits puisse suivre l'histoire d'un manuscrit donné.

element **stamp**

| | |
|---|---|
| diagram |  |
| type | extension of **xs:string** |
| used by | elements **history** **layout** |

| attributes | Name | Type | Use | Default | Fixed |
|---|---|---|---|---|---|
| | exist | xs:NMTOKEN | | | |
| | pagNomb | xs:string | | | |

source

```
<xs:element name="stamp">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute name="exist">
     <xs:simpleType>
      <xs:restriction base="xs:NMTOKEN">
       <xs:enumeration value="yes"/>
       <xs:enumeration value="no"/>
      </xs:restriction>
     </xs:simpleType>
    </xs:attribute>
    <xs:attribute name="pagNomb" type="xs:string"/>
   </xs:extension>
  </xs:simpleContent>
 </xs:complexType>
</xs:element>
```

**Figure n°.40: les éléments et les attributes *stamp***

**3.5. origin** (l'origine) : élément qui permet de décrire l'origine d'un manuscrit ou d'une partie de manuscrit.

**3.6. Others** (autres) : élément qui fournit d'autres indications qui peuvent servir à identifier l'histoire du manuscrit. Dans MASTER, l'élément history inclut les éléments fils suivants : <origin>, <provenance> et <acquisition>.

Mais comme dans MASTER, nous avons trouvé nécessaire de mettre les attributs suivants comme un moyen d'aide à définir une date approximative pour le manuscrit non daté.

Les attributs sont :

| Name | Type |
|---|---|
| Status | xs:NMTOKEN |

| | |
|---|---|
| notBefore | xs:string |
| notAfter | xs:string |
| evidence | xs:NMTOKEN |

**a**) Status : c'est un attribut qui contient des informations sur la copie du manuscrit, du point de vu originalité, comme :

**a.a)** Autographes : il est utilisé pour indiquer que la copie du manuscrit est écrite directement par l'auteur lui-même.

**a.b)** Apographes: il indique que la copie du manuscrit actuelle est écrite directement à partir de la copie originale.

**a.c)** Unique: il indique que la copie du manuscrit est la seule copie qui existe dans le monde.

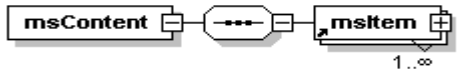
**a.d)** Waqf: il indique que la copie fait partie du waqf ( héritage familial ou religieux).

**b)** certainty : il désigne le niveau de confiance associé à la datation indiquée par certains attributs dans le manuscrit comme un niveau très élevé (high), moyen (medium) ou bas (low).

**c)** evidence : il indique le genre d'évidence ou de témoignage à la datation d'un manuscrit, une évidence intérieure (internal), extérieure (external) ou hypothétique (conjecture).

**4. msContent (***manuscript content*) : il décrit tous les éléments qui aident à identifier le contenu d'un manuscrit donné.

element **msContent**

| | |
|---|---|
| diagram |   |
| children | **msItem** |
| used by | element **msDescription** |

| attributes | Name | Type | Use | Default | Fixed |
|---|---|---|---|---|---|
| | defective | xs:NMTOKEN | | no | |

source

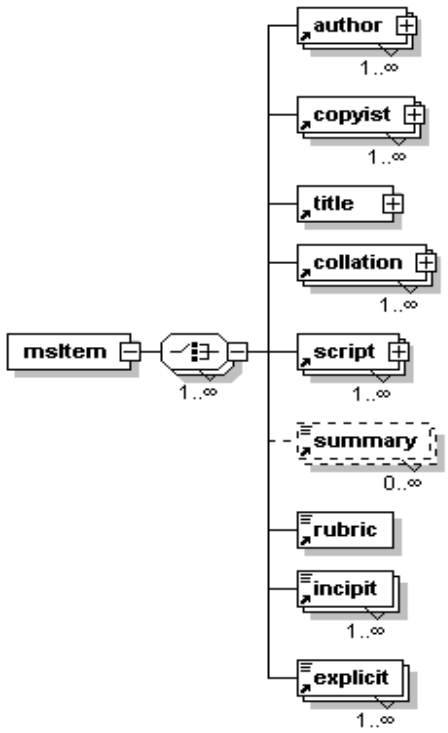
```
<xs:element name="msContent">
 <xs:complexType>
  <xs:sequence>
   <xs:element ref="msItem" maxOccurs="unbounded"/>
  </xs:sequence>
  <xs:attribute name="defective" default="no">
   <xs:simpleType>
    <xs:restriction base="xs:NMTOKEN">
     <xs:enumeration value="yes"/>
     <xs:enumeration value="no"/>
     <xs:enumeration value="unk"/>
    </xs:restriction>
   </xs:simpleType>
  </xs:attribute>
 </xs:complexType>
</xs:element>
```

**Figure n°.41: Les elements et les attributes *msContent***

**4.1. msItem** : il s'agit d'une unité descriptive dans laquelle on trouve des informations sur la composition du manuscrit en volumes et l'information qui concerne chaque volume comme: l'auteur (author), le copiste (copyist), le titre du volume (title), la collation (collation), le style d'écriture (writStyle), le script (script) le résumé (summary), la rubrique (rubric), l'incipit (incipit) et l'explicit (explicit).

element **msItem**

diagram




children **author copyist title collation script summary rubric incipit explicit**

used by element **msContent**

source

```
<xs:element name="msItem">
  <xs:complexType>
    <xs:choice maxOccurs="unbounded">
      <xs:element ref="author" maxOccurs="unbounded"/>
      <xs:element ref="copyist" maxOccurs="unbounded"/>
      <xs:element ref="title"/>
      <xs:element ref="collation" maxOccurs="unbounded"/>
      <xs:element ref="script" maxOccurs="unbounded"/>
      <xs:element ref="summary" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="rubric"/>
      <xs:element ref="incipit" maxOccurs="unbounded"/>
      <xs:element ref="explicit" maxOccurs="unbounded"/>
    </xs:choice>
  </xs:complexType>
</xs:element>
```

**Figure n°.42: Les elements *msItem***

**5. logicStruct** : il s'agit de la structure logique du document, sa composition en page de titre, de table de matières, de parties, de chapitre, etc.

element **logicStruct**

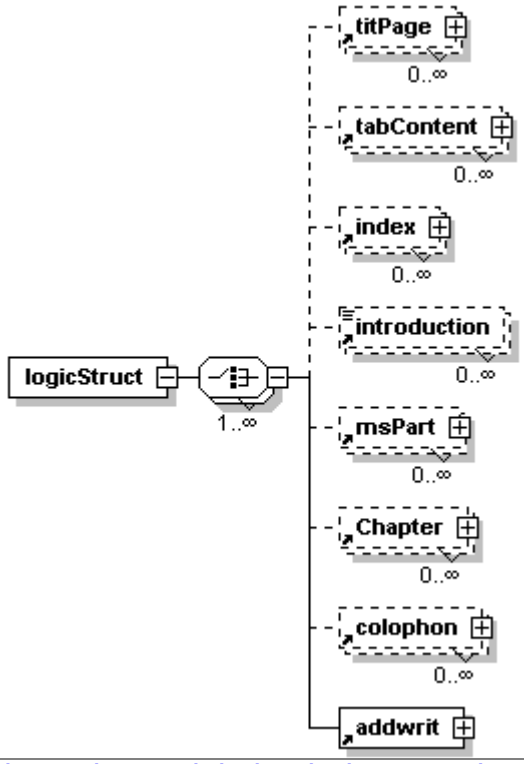
| | |
|---|---|
| diagram |  |
| children | titPage tabContent index introduction msPart Chapter colophon addwrit |
| used by | element msDescription |
| source | <xs:element name="logicStruct"><br><xs:complexType><br><xs:choice maxOccurs="unbounded"><br><xs:element ref="titPage" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="tabContent" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="index" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="introduction" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="msPart" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="Chapter" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="colophon" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="addwrit"/><br></xs:choice><br></xs:complexType><br></xs:element> |

**Figure n°.43 : Les éléments *logicStruct***

**5.1.titlePage** : il contient des informations qui indiquent l'existence de page de titre dans le manuscrit. Les attributs suivants sont ajoutés pour identifier le type de titre de page s'il existe dans le texte soit écrit dans une page séparée (separate), soit mélangé avec le corpus du texte (notSeparate)

**5.2. tabContent** : il contient des informations sur la table des matières. Le sous-élément <p> est là pour donner au catalogueur de mettre le nombre de pages et d'autres informations concernant la table des matières.

element **tabContent**

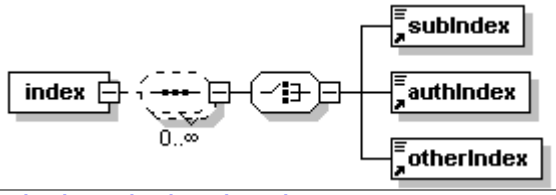
| | |
|---|---|
| diagram |   |
| children | p |
| used by | element logicStruct |
| attributes | Name: existe, tabConType<br>Type: xs:NMTOKEN, xs:NMTOKEN<br>Use:<br>Default: no<br>Fixed: |
| source | <xs:element name="tabContent"><br><xs:complexType><br><xs:sequence minOccurs="0" maxOccurs="unbounded"><br><xs:element ref="p"/><br></xs:sequence><br><xs:attribute name="existe" default="no"><br><xs:simpleType><br><xs:restriction base="xs:NMTOKEN"><br><xs:enumeration value="yes"/><br><xs:enumeration value="no"/><br><xs:enumeration value="unDet"/><br></xs:restriction><br></xs:simpleType><br></xs:attribute><br><xs:attribute name="tabConType"><br><xs:simpleType><br><xs:restriction base="xs:NMTOKEN"><br><xs:enumeration value="subjTable"/><br><xs:enumeration value="SoraTable"/><br><xs:enumeration value="chapTitTable"/><br><xs:enumeration value="otherTable"/><br></xs:restriction><br></xs:simpleType><br></xs:attribute><br></xs:complexType><br></xs:element> |

**Figure n°.44: Les éléments *tabContent***

Les attributs suivants sont ajoutés pour décrire le type de table des matières trouvée dans le texte, surtout le sous-élément tabConType, soit une table classée par sujet (subjTable), par titre de sourate (SoraTable), par titre de chapitre (chapTitTable), soit par d'autres titres (otherTable) qui ne seraient pas mentionnés ci-dessus.

5.3. index : il contient des informations sur l'existence d'un ou plusieurs index dans le manuscrit.

element **index**

| | |
|---|---|
| diagram |   |
| children | subIndex authIndex otherIndex |
| used by | element logicStruct |
| attributes | Name: exist<br>Type: xs:NMTOKEN<br>Use:<br>Default:<br>Fixed: |
| source | <xs:element name="index"><br><xs:complexType> |

```
    <xs:sequence minOccurs="0" maxOccurs="unbounded">
      <xs:choice>
        <xs:element ref="subIndex"/>
        <xs:element ref="authIndex"/>
        <xs:element ref="otherIndex"/>
      </xs:choice>
    </xs:sequence>
    <xs:attribute name="exist">
      <xs:simpleType>
        <xs:restriction base="xs:NMTOKEN">
          <xs:enumeration value="yes"/>
          <xs:enumeration value="no"/>
          <xs:enumeration value="pagNomb"/>
          <xs:enumeration value="undit"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
  </xs:complexType>
</xs:element>
```

**Figure n°.45: Les éléments et les attributs *index***

Il est divisé en trois éléments fils :

**5.3.1.subIndex** (index par sujet) : il contient des informations sur l'existence d'un index par sujet.

**5.3.2.authIndex** (index par auteur) : il fournit des informations sur l'existence d'un index par auteur.

**5.3.3.otherIndex** : un autre genre d'index qui peut être trouvé dans d'autres manuscrits et qui n'ait pas été mentionné jusqu'ici.

**5.4. msPart** : il contient des informations sur les parties composant un manuscrit.

element **msPart**

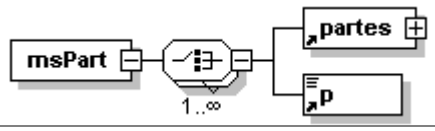
| | |
|---|---|
| diagram |   |
| children | partes p |
| used by | element logicStruct |
| source | <xs:element name="msPart"><br><xs:complexType><br><xs:choice maxOccurs="unbounded"><br><xs:element ref="partes"/><br><xs:element ref="p"/><br></xs:choice><br></xs:complexType><br></xs:element> |

**Figure n°.46: Les elements *msPart***

L'élément msPart a été divisé entre trois parties :

**5.4.1. Partes** : Procédure utilisée par l'auteur et le copiste dans leur classement des textes des manuscrit arabes. Avec les attributs suivants, nous avons tenté de définir les différentes parties composantes comme le (djuz), le (kytab), le (bab), le (fasl), le (matlab) et le (masalah). L'équivalent de ces classements en langue française est difficile à fournir

en raison de la différence dans la logique de classement. Mais pour cette raison, nous avons trouvé nécessaire de mettre le classement par chapitre pour le manuscrit qui ne suit pas le classement précédent.

element **partes**

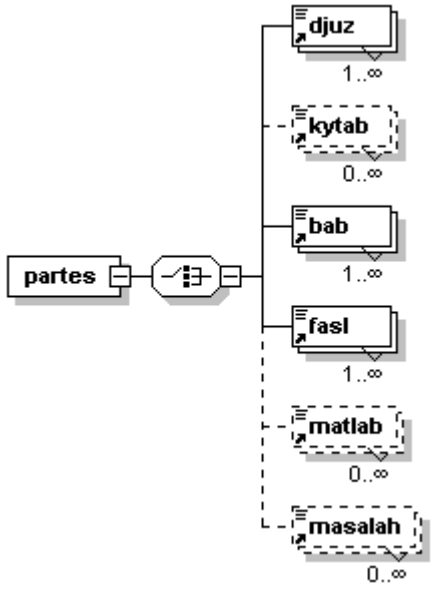
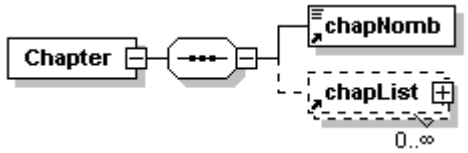
| | |
|---|---|
| diagram |  |
| children | djuz kytab bab fasl matlab masalah |
| used by | element msPart |
| source | <xs:element name="partes"><br><xs:complexType><br><xs:choice><br><xs:element ref="djuz" maxOccurs="unbounded"/><br><xs:element ref="kytab" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="bab" maxOccurs="unbounded"/><br><xs:element ref="fasl" maxOccurs="unbounded"/><br><xs:element ref="matlab" minOccurs="0" maxOccurs="unbounded"/><br><xs:element ref="masalah" minOccurs="0" maxOccurs="unbounded"/><br></xs:choice><br></xs:complexType><br></xs:element> |

**Figure n°.47 : Les éléments *partes***

**5.4.2. chapter**: il contient des informations sur la division par chapitre ; il est divisé en deux éléments fils :

element **Chapter**

| | |
|---|---|
| diagram |  |
| children | chapNomb chapList |
| used by | element logicStruct |
| source | <xs:element name="Chapter"><br><xs:complexType><br><xs:sequence><br><xs:element ref="chapNomb"/><br><xs:element ref="chapList" minOccurs="0" maxOccurs="unbounded"/> |

```
    </xs:sequence>
  </xs:complexType>
</xs:element>
```

**Figure n°.48: Les éléments *chapter***

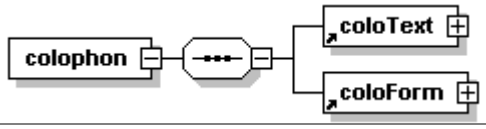
**5.4.2.1. chapNomb** : il a pour but de fournir les données sur le nombre de chapitres.

**5.4.2.2. chapList** : l'élément fils liste de chapitre est un moyen de lister les titre de chapitre, si possible.

**5.5. colophon** : il fournit les informations sur le colophon de manuscrit. Il est divisé en deux éléments fils : la première par le texte du colophon (coloText) pour citer les informations trouvées à l'intérieur du colophon et la deuxième, la forme de colophon « coloForm » qui décrit la forme dans laquelle le colophon a été présenté.

element **colophon**

diagram




children **coloText coloForm**

used by element **logicStruct**

attributes

| Name | Type | Use | Default | Fixed |
|---|---|---|---|---|
| presence | xs:NMTOKEN | | | |

source

```
<xs:element name="colophon">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="coloText"/>
      <xs:element ref="coloForm"/>
    </xs:sequence>
    <xs:attribute name="presence">
      <xs:simpleType>
        <xs:restriction base="xs:NMTOKEN">
          <xs:enumeration value="yes"/>
          <xs:enumeration value="no"/>
          <xs:enumeration value="unDet"/>
          <xs:enumeration value="torn"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
  </xs:complexType>
</xs:element>
```

**Figure n°.49: Les éléments *colophon***

Le <P> dans coloText est une zone libre pour citer l'information trouvée dans le colophon.

Cependant, le type dans « coloForm » indique les attributs des différents types de forme trouvés pour décrire le colophon.

Dans les éléments fils de l'élément type, nous avons mis les attributs de la forme du colophon trouvés dans notre étude de manuscrits : pour la première forme, il s'agit d'un triangle pointé en haut (upTriangle), au contraire de la première forme, le deuxième est

une forme de triangle pointé en bas (dowTriangle), la troisième prenant la forme de double triangle (doubTriangle). Alors que (others) est réservé pour d'autres formes de colophon qui n'ait pas été mentionné auparavant.

**6. Additional** : il inclut des informations additionnelles qui concernent la situation actuelle du manuscrit dans la bibliothèque. Il contient les éléments fils suivants:

element **additional**

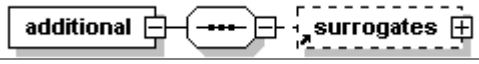
| | |
|---|---|
| diagram |   |
| children | surrogates |
| used by | element msDescription |
| source | <xs:element name="additional"><br><xs:complexType><br><xs:sequence><br><xs:element ref="surrogates" minOccurs="0"/><br></xs:sequence><br></xs:complexType><br></xs:element> |

**Figure n°.50: les éléments *additional***

**6.1. adminInfo** : il fournit des informations sur la situation administrative du manuscrit à l'intérieur de la bibliothèque.

element **adminInfo**

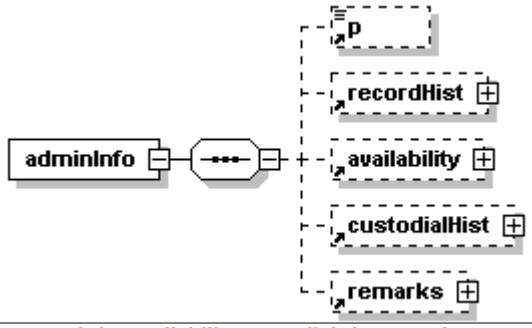
| | |
|---|---|
| diagram |   |
| children | p recordHist availability custodialHist remarks |
| used by | element msDescription |
| source | <xs:element name="adminInfo"><br><xs:complexType><br><xs:sequence><br><xs:element ref="p" minOccurs="0"/><br><xs:element ref="recordHist" minOccurs="0"/><br><xs:element ref="availability" minOccurs="0"/><br><xs:element ref="custodialHist" minOccurs="0"/><br><xs:element ref="remarks" minOccurs="0"/><br></xs:sequence><br></xs:complexType><br></xs:element> |

**Figure n°.51: les elements *adminInfo***

Il est divisé en cinq sous-éléments : un sous-élément <p> pour mettre n'importe quelle information nécessaire.

**6.1.1. Le « recordHist »** fournit des informations sur la source du manuscrit et sa copie d'origine.

« recordHist » est divisé en deux sous-éléments : source et change. Source avec l'élément <p> qui fournit des informations sur l'origine du manuscrit, alors que l'élément « change » fournit toutes les informations concernant le changement qui a permis d'arriver à la situation actuelle du manuscrit.

**6.1.2. Availability** : il fournit des informations sur la disponibilité du manuscrit dans la bibliothèque et sur l'éventuelle restriction de son utilisation, autrement dit sur le règlement d'utilisation de ce manuscrit dans la bibliothèque.

**6.1.3. custodialHist** : il fournit des informations sur l'histoire d'acquisition du manuscrit par la bibliothèque, soit par achat, soit par donation, etc.

element **custodialHist**

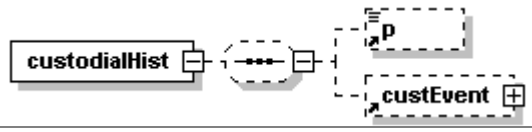
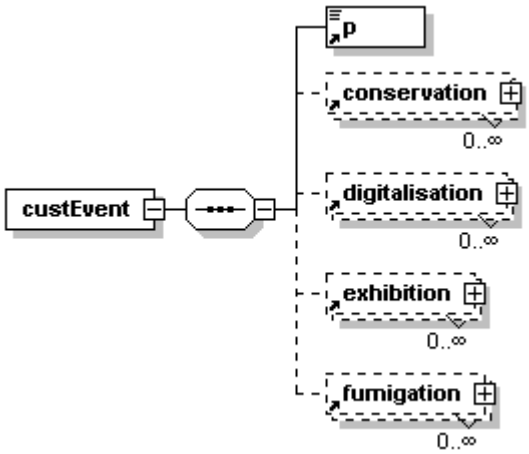
| | |
|---|---|
| diagram |   |
| children | p custEvent |
| used by | element adminInfo |
| source | <xs:element name="custodialHist"><br><xs:complexType><br><xs:sequence minOccurs="0"><br><xs:element ref="p" minOccurs="0"/><br><xs:element ref="custEvent" minOccurs="0"/><br></xs:sequence><br></xs:complexType><br></xs:element> |

**Figure n°.52: Les elements *custodialHist***

**6.1.3.1. custEvent** : il décrit les « traitements » qui ont été appliqués au document lors de son acquisition par la bibliothèque, comme la conservation, la présentation dans une exposition, ou la numérisation (digitisation) et la fumigation. Les même termes et la même structure ont été utilisés dans MASTER, à l'exception de la photographie car nous n'avons pas trouvé nécessaire de la mettre dans cette catégorie. Par contre, nous avons trouvé nécessaire d'ajouter la fumigation, élément qui n'existe pas dans MASTER.

element **custEvent**

diagram




children **p conservation digitalisation exhibition fumigation**

used by element **custodialHist**

attributes

| Name | Type | Use | Default | Fixed |
|---|---|---|---|---|
| notBefore | xs:string | | | |
| notAfter | xs:string | | | |
| certainty | xs:NMTOKEN | | | |
| evidence | xs:NMTOKEN | | external | |
| type | xs:string | | | |

source

```
<xs:element name="custEvent">
 <xs:complexType>
  <xs:sequence>
   <xs:element ref="p"/>
   <xs:element ref="conservation" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="digitalisation" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="exhibition" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="fumigation" minOccurs="0" maxOccurs="unbounded"/>
  </xs:sequence>
  <xs:attribute name="notBefore" type="xs:string"/>
  <xs:attribute name="notAfter" type="xs:string"/>
  <xs:attribute name="certainty">
   <xs:simpleType>
    <xs:restriction base="xs:NMTOKEN">
     <xs:enumeration value="high"/>
     <xs:enumeration value="medium"/>
     <xs:enumeration value="low"/>
    </xs:restriction>
   </xs:simpleType>
  </xs:attribute>
  <xs:attribute name="evidence" default="external">
   <xs:simpleType>
    <xs:restriction base="xs:NMTOKEN">
     <xs:enumeration value="internal"/>
     <xs:enumeration value="external"/>
     <xs:enumeration value="conjecture"/>
    </xs:restriction>
   </xs:simpleType>
  </xs:attribute>
  <xs:attribute name="type" type="xs:string"/>
 </xs:complexType>
</xs:element>
```

**Figure n°.53: Les éléments *custEvent***

**6.1.3.1.1. Conservation** : il fournit les informations sur l'état de conservation des manuscrits ; l'élément <p> est un espace libre pour le catalogueur, lui permettant d'ajouter des informations si nécessaire.

**6.1.3.1.2. digitisation** : il contient l'information sur la situation du manuscrit, du point de vue numérisation et date de numérisation ; de même, l'élément <p> a été ajouté pour le catalogueur, lui permettant de noter des informations  si nécessaire.

**6.1.3.1.3. exhibition** : il fournit les données sur les évènements passés concernant le manuscrit tels que sa participation à des expositions, etc.

**6.1.3.1.4. fumigation** : il donne des informations sur la dernière date de fumigation.

**6.1.4.** L'élément fils **« remarks »** contient n'importe quelle remarque que le catalogueur trouve nécessaire de faire, afin de permettre de décrire un évènement qui n'ait pas été défini ailleurs.

**6.2. surrogates** : il fournit des informations sur la copie numérisée, la photocopie, le microfilm ou une copie publiée de ce même manuscrit et qui existerait au sein de la bibliothèque ou à l'extérieur.

element **surrogates**

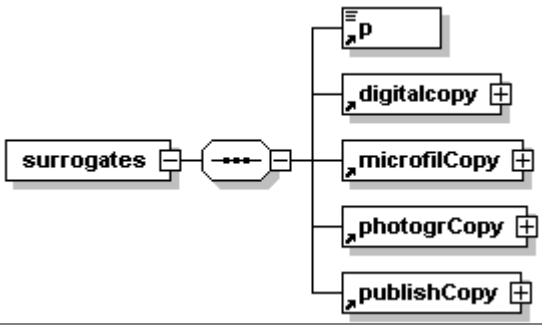
| | |
|---|---|
| diagram |  |
| children | p digitalcopy microfilCopy photogrCopy publishCopy |
| used by | element additional |
| source | <xs:element name="surrogates"><br><xs:complexType><br><xs:sequence><br><xs:element ref="p"/><br><xs:element ref="digitalcopy"/><br><xs:element ref="microfilCopy"/><br><xs:element ref="photogrCopy"/><br><xs:element ref="publishCopy"/><br></xs:sequence><br></xs:complexType><br></xs:element> |

**Figure n°.54: Les éléments *surrogates***

**6.2.1. digitalCopy**: il contient des informations sur la copie numérisée de ce même manuscrit.

**6.2.2. microfilCopy**: il contient des informations sur la copie microfilm de ce même manuscrit.

element **microfilCopy**

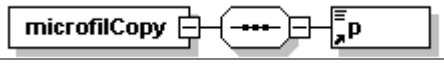
| | |
|---|---|
| diagram |   |
| children | p |
| used by | element surrogates |
| attributes | Name Type Use Default Fixed<br>lieu xs:string<br>date xs:string<br>institution xs:string<br>other xs:string |
| source | <xs:element name="microfilCopy"><br><xs:complexType><br><xs:sequence><br><xs:element ref="p"/><br></xs:sequence><br><xs:attribute name="lieu" type="xs:string"/><br><xs:attribute name="date" type="xs:string"/><br><xs:attribute name="institution" type="xs:string"/><br><xs:attribute name="other" type="xs:string"/><br></xs:complexType><br></xs:element> |

**Figure n°.55: les éléments et les attributs *microfilCopy***

**6.2.3. photogrCopy**: il contient des informations sur la copie photographique de ce même manuscrit.

**6.2.4. publishCopy**: il contient des informations sur la copie publiée de ce même manuscrit.

### *3.2.3. Conclusion :*

La description des métadonnées et sa grammaire se veut une aide à l'analyse d'un manuscrit ou d'une partie de manuscrit, et une aide utile pour le catalogueur afin d'encoder ces manuscrits sous forme électronique numérisée. Il y a peu de manuscrits qui contiennent tous les éléments et sous-éléments mentionnés dans cette partie ; cependant, nous avons trouvé nécessaire de définir le plus large « éventail » possible d'éléments qui permettent de faire une analyse exhaustive des caractéristiques des manuscrits arabes.

## 3.3. L’extraction automatique des métadonnées par analyse d’image

(*Cette partie résulte d'un travil en commun avec Frank LeBourgeois du laboratoire LIRIS-RFV de l'INSA* de Lyon*)

### 3.3.1. Présentation

Nous avons pour objectif d'étudier la faisabilité du traitement des images par ordinateur afin d'extraire automatiquement les métadonnées et les caracteristiques des manuscrits arabes. Notre corpus présente trois difficultés majeures pour le traitement automatique des images :

- L'écriture arabe : il existe très peu de travaux de recherche sur la reconnaissance automatique des documents arabes. Seuls, quelques travaux sont récemment apparus sur la lecture automatique des documents arabes imprimés qui n'ont pas encore permis d'améliorer les rares systèmes de lecture optique commerciaux. Les performances actuelles des quelques OCR commercialisés sur l'Arabe imprimé sont très inférieures à celles des OCR sur les textes latins.
- L'écriture manuscrite arabe : s'il existe quelques travaux de recherche sur les textes arabes imprimés, en revanche, il n'y a quasiment pas eu d'études sur l'analyse des textes manuscrits en arabe. Le même constat existe sur les textes manuscrits anciens d'Europe. Cela s'explique à la fois par l'émergence de ce domaine et par les difficultés qu'il soulève. De plus les manuscrits arabes présentent des difficultés qui sont différentes de celles que l'on rencontre sur les manuscrits latins et qui rendent impossible les adaptations des autres travaux sur les manuscrits anciens d'Europe.
- La médiocre qualité des images : une grande partie des images du corpus proviennent de microfilms numérisés. Nous savons actuellement que ce support n'est pas adapté à une numérisation de qualité. En effet, il n'est pas possible de numériser des microfilms en niveaux de gris car le procédé photographique du microfilmage enlève toutes les nuances de niveaux de gris pour ne laisser apparaître que du blanc ou du noir afin de pouvoir réduire considérablement la taille de l'image. Les images numériques issues de microfilms sont donc des images binaires qui ne peuvent pratiquement plus être corrigées. Les taches

* Institut National de Science Appliquer

mélangées au texte ne peuvent plus être enlevées et les dégradées des peintures et des ornements ont définitivement été perdues. Dans ces conditions, l'information perdue ne peut pas être retrouvée et les textes effacés ne peuvent pas plus être segmentés.

Les métadonnées que nous cherchons à extraire ne nécessitent pas la reconnaissance des textes car les annotations, les titres et les illustrations sont parfaitement visibles et reconnaissables sans recours au contenu des textes. Nous avons donc demandé au laboratoire LIRIS-RFV de l'INSA de Lyon de développer un logiciel d'analyse d'images capable de reconnaître certaines de nos métadonnées. Le délai très court imposé à cette étude n'a pas permis de réaliser un logiciel abouti mais seulement un démonstrateur avec lequel nous avons pu mesurer les performances réelles sur notre corpus pour conclure sur la faisabilité du traitement automatique des manuscrits anciens par analyse d'images.

### *3.3.2. Construction d'une chaîne d'analyse d'image*

L'analyse des images de documents est un processus complexe qui ne peut pas toujours s'effectuer séquentiellement car les opérations de segmentation et de reconnaissance sont étroitement liées. Les ordinateurs actuels basés sur le traitement séquentiel des données ne sont donc pas adaptés à l'analyse d'image. Pour pallier ce problème, on cherche à découper le processus d'analyse d'images en étapes plus ou moins séquentielles, plus adaptées à l'architecture de nos ordinateurs. Le choix du découpage du processus va déterminer les limites fonctionnelles d'un système d'analyse d'images. La chaîne traditionnelle de traitement consiste à simplifier progressivement l'image pour segmenter les formes puis à soumettre ces derniers à des algorithmes de reconnaissance. Dans un premier temps, on procède à une suite d'étapes appelée ***segmentation*** qui consiste à convertir l'image couleur en image à niveaux de gris puis en image binaire pour extraire les différents objets de l'image. Dans un deuxième temps, on effectue une phase de ***reconnaissance*** qui analyse et mesure les différents objets segmentés pour les classer suivant leurs formes.

#### 3.3.2.1. La segmentation des images

La segmentation des images consiste à trouver tous les objets porteurs d'une information dans l'image. Puisque cette phase précède celle de la reconnaissance et que nous avons séparé ces deux étapes qui sont pourtant étroitement liées, la segmentation va donc s'effectuer sans l'aide de la reconnaissance des formes. Les images de notre corpus sont très variées car elles peuvent être en couleurs, en niveaux de gris ou binaires quand elles sont obtenues à partir de microfilms. Pour traiter la grande variété des images et pour réutiliser au maximum les algorithmes adaptés à un certain type d'image, un module de pré-traitement a été réalisé. Il permet de restaurer et de convertir les images couleurs ou en niveaux de gris en images binaires. La segmentation proprement dite est effectuée à partir de l'image binaire. La phase de restauration des images couleurs et en niveaux de gris et de conversion en image binaire est donc importante pour les performances globales du système. Cela explique pourquoi, en présence d'images déjà binaires de mauvaise qualité, la restauration des images est impossible.

##### 3.3.2.1.1. Le pré-traitement et la restauration des images

Le pré-traitement consiste à simplifier progressivement l'image et à restaurer l'information contenue dans les couleurs ou les nuances de gris pour obtenir une image binaire où tous les objets importants apparaissent. Les objets qui nous intéressent dans les textes sont constitués de traits. En terme d'analyse d'image, il faut donc chercher tous les traits possibles quelles que soient leurs couleurs ou leurs nuances de gris. Cette étape est trop complexe pour être appliquée directement sur les images couleurs, il faut procéder dans un premier temps à une conversion de l'image couleur en image à niveaux de gris en minimisant les pertes d'informations concernant les traits des objets que l'on désire conserver. Dans un deuxième temps, l'image à niveaux de gris est transformée en image binaire où chaque pixel n'est représenté que par deux valeurs possibles, (0) pour les pixels du fond et (1) pour les pixels des traits des objets.

##### 3.3.2.1.2. La conversion d'images couleurs en images à niveaux de gris

Dans une image couleur, chaque pixel de l'image est représenté par un triplet de valeurs (R,V,B) qui mesure l'intensité dans chacun de canaux Rouge, Vert, Bleu. Chaque canal

mesurant une valeur entre 0 et 255, nous avons $256^3$ couleurs possibles soit plus de 16 millions de couleurs. Trois méthodes différentes sont applicables dans le démonstrateur suivant le niveau de difficulté de séparation entre le fond et la forme des objets que l'on désire segmenter. L'utilisateur choisi la méthode la plus adaptée après une phase de test sur quelques images. Ces méthodes sont choisies, pour leurs performances, leurs robustesses et leurs généricités qui leurs permettent de traiter automatiquement une grande variété d'images. En contre parties, ces méthodes de conversion demandent beaucoup de temps de calcul par image mais elles ne nécessitent presque pas de paramètres et fonctionnent automatiquement sans l'assistance de l'utilisateur.

**a) Conversion en luminance** :

quand tous les objets conformant image couleur sont reconnaissables en fonction des intensités communes à tous les canaux R,V, B, alors on peut calculer une image à niveaux de gris à partir de la moyenne des canaux R,V, B. Cette méthode, très simple, réduit considérablement les temps de calcul et reste adaptée à quelques ouvrages pour lesquels la couleur n'est pas une information importante pour la segmentation des textes.

**Image originale couleur RVB** **Analyse de la luminance**

**Illustration n°.16 :Conversion en luminance**

**b) Analyse statistique des couleurs :**

Guand les textes ou les illustrations sont réalisés avec des couleurs différentes, il faut donc analyser chaque information d'intensité Rouge, Vert et Bleu séparément. Pour cela, on utilise les outils statistiques d'analyse de la variance pour trouver la combinaison optimale ($u_1,u_2,u_3$) des canaux R,V, B pour minimiser la perte d'information.

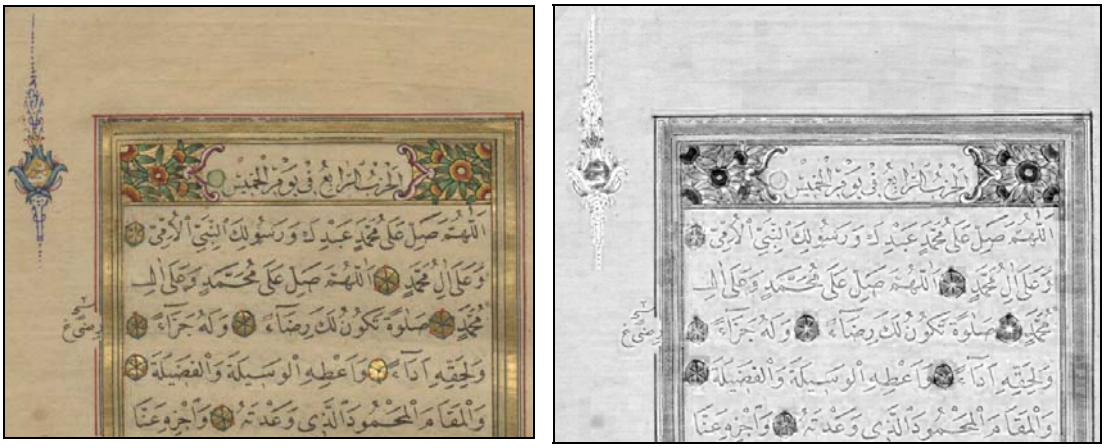


*Image originale couleur RVB* *Analyse statistique des couleurs*

**Illustration n°.17 : Analyse statistique des couleurs**

### c) Classification automatique des couleurs :

Quand les différents objets n'ont pas une couleur homogène et que l'information de couleur doit être analysée de façon plus subtile, il convient de réaliser une classification automatique des couleurs en K classes distinctes. Cette opération est réalisée en appliquant une classification automatique à tous les pixels de l'image dans l'espace tridimensionnel des couleurs. L'algorithme universel des K-means permet de classer automatiquement tous les pixels dans K classes. L'utilisateur doit donner a priori le nombre de classes qu'il souhaite obtenir. Cette méthode performante permet de séparer le recto du verso sur les images couleurs ou d'isoler les différentes couleurs utilisées dans un document. Cependant cette méthode d'analyse est très coûteuse en temps de calcul à cause du grand nombre de pixels à classer et peut prendre plusieurs minutes sur des images de grande dimension.

*Image originale couleur RVB* *Classification des couleurs en K=5 classes*

**Illustration n°.18 : Classification automatique des couleurs**

### 3.3.2.1.3. La conversion des images à niveaux de gris en images binaires

La seconde étape consiste à convertir l'image en niveaux de gris en image binaire adaptée à la segmentation des objets. Cette étape critique appelée aussi ***binarisation*** va conditionner les performances de la segmentation des objets lors des étapes suivantes. Nous avons encore choisi trois méthodes robustes et réputées performantes sur une grande variété de documents :

- **La binarisation automatique globale** : chaque pixel est comparé à un seuil optimal calculé automatiquement par le critère statistique de Fisher. Ce critère garantit de trouver, dans la distribution statistique des nuances de gris représentée sous la forme d'un *histogramme*, exactement 2 classes de nuances de niveaux de gris séparées par un seuil. Tous les pixels dont le niveau est inférieur à ce seuil sont classés 0 (noir) et les autres classés 1 (blanc). Le seuil est identique pour tous les pixels de l'image. La méthode de seuillage globale est adaptée aux documents très contrastés pour lesquels tous les objets ont un niveau de gris suffisamment différent de celui du support papier. Il ne convient pas comme le montre la figure suivante sur des images où les traits ont des nuances faiblement contrastées.

*Image originale en niveaux de gris* *Seuillage automatique global*

**Illustration n°. 19 : La binarisation automatique globale**

- **La binarisation adaptative** : Pour les documents où certains traits ont des nuances de gris proches de celui du support papier, il convient d'appliquer une méthode de seuillage adaptatif qui va calculer pour chaque pixel un seuil localement adapté. Parmi toutes les méthodes adaptatives locales, nous avons pris la méthode la plus fiable appelée Niblack, du nom de son auteur. Un abaissement du niveau de seuil dans les zones de l'image faiblement contrastées permet

d'augmenter la sensibilité de la détection des traits dans les zones de niveaux de gris homogènes.

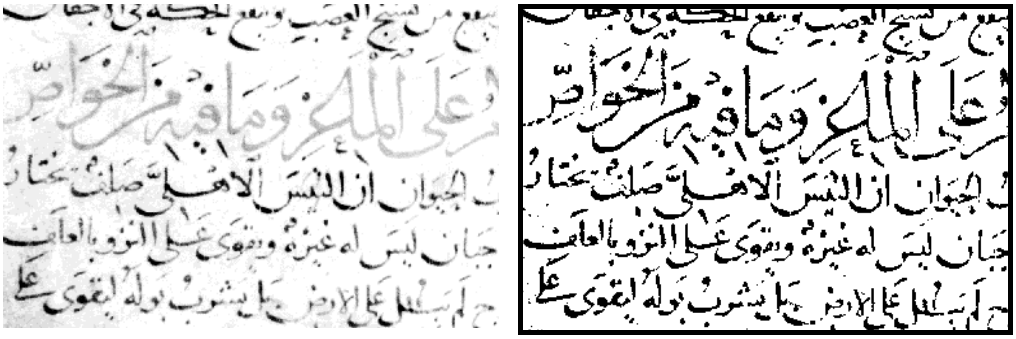


Image originale en niveaux de gris Seuillage adaptatif

**Illustration n°.20 : La binarisation adaptative**

- **La binarisation par classification** : Comme pour les images couleurs, la classification automatique peut réaliser une binarisation des images à niveaux de gris. Dans un premier temps, on cherche automatiquement K classes parmi toutes les nuances de gris trouvées dans l'image, la valeur de K étant saisie par l'utilisateur. Dans un deuxième temps, on attribue les pixels de la première moitié des classes à la valeur 0 et la seconde moitié des classes à la valeur 1. Sans autre information locale, cette méthode peut s'apparenter à une binarisation automatique globale, car la classification détermine globalement les classes sur tous les pixels de l'image. Les résultats obtenus sont cependant très différents de la première méthode car elle ne tient pas compte du nombre de pixels dans chacune des classes. La classification trouvée favorise donc plus les traits statistiquement moins présents dans l'image que les nuances de gris correspondant au support papier. L'image binaire obtenue montre un épaississement des traits des textes et une tendance à faire apparaître les taches.

*Classification en K=4 classes* *Seuillage par classification*

**Illustration n°.21 : La binarisation par classification**

### 3.3.2.2. La segmentation des objets

#### 3.3.2.2.1. Le choix de la méthode

Nous avons choisi d'effectuer une analyse ***ascendante*** de l'image en partant de l'information élémentaire qu'est le pixel pour obtenir une information plus interprétée comme les objets pour enfin aboutir à des informations encore plus évoluées comme celle de la zone principale de texte. Cette approche ascendante, aujourd'hui classique, est en opposition avec l'approche ***descendante*** qui consiste à partir des connaissances *a priori* sur le contenu des images et de segmenter les différents objets à partir de ces connaissances. Le choix entre une méthode ascendante et une méthode descendante s'effectue en fonction de la possibilité ou de la non possibilité de modéliser le contenu des images. Ainsi pour certains textes imprimés, on privilégiera une méthode descendante pour localiser les caractères, les mots, les lignes et les paragraphes car les règles typographiques actuelles sont suffisamment rigides pour pouvoir réaliser un modèle généraliste de segmentation. A l'inverse, les textes anciens et en particulier les textes manuscrits montrent une plus grande variabilité dans leurs formes et leurs structures. C'est pour cette raison que nous avons privilégié une méthode ascendante plus souple et qui ne nécessite pas d'étude préalable très longue sur une grande quantité de textes pour créer un modèle de segmentation robuste de tous les textes manuscrits arabes.

#### 3.3.2.2.2. La définition d'un objet par les connexités de l'image

L'image binaire obtenue par les étapes précédentes est constituée principalement de traits et de points. Cependant les blocs de texte sont constitués d'objets intermédiaires qui sont les éléments connexes.

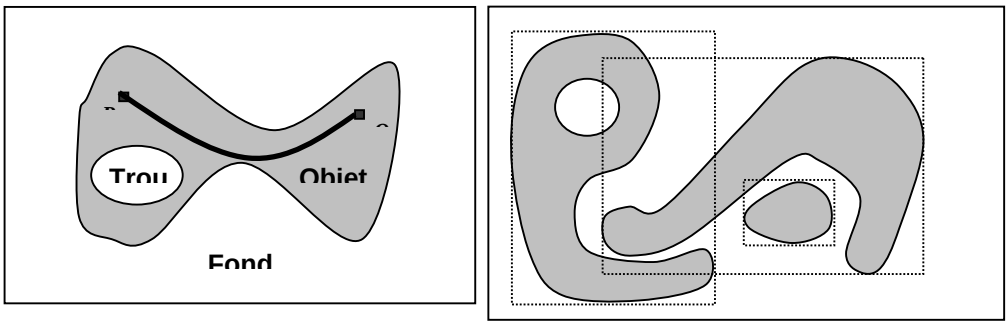
On définit une connexité comme étant un sous-ensemble de points de l'image de même valeur 0 ou 1, entre lesquels il existe toujours un chemin qui relie deux points P et Q quelconques.

Dans une image binaire, on classifie les connexités dans 3 familles distinctes :

- Les connexités de valeur 0 (noir) représente par convention les objets portant une information

- Les connexités de valeur 1 (blanc), adjacent aux bords de l'image, représentent le fond de l'image c'est à dire le support papier
- Les connexités de valeur 1 (blanc) qui ne font pas parties du fond, représentent les trous inscrits dans les objets.

On définit le rectangle circonscrit à une connexité, le plus petit rectangle qui contient cette connexité. Cette notion de rectangle circonscrit, fréquemment utilisé, est ambigu. Un rectangle circonscrit à une connexité peut contenir plus d'une connexité ou peut avoir une intersection non vide avec d'autres rectangles circonscrits à d'autres connexités. Notre choix de construire les objets autour des connexités est adapté à l'écriture arabe car celle-ci est formée de blocs connexes espacés.




*Définition d'une connexité, des objets du fond, des trous et du rectangle circonscrit*

*Ambiguïté dans la représentation des connexités par des rectangles*

**Figure n°.56 : La définition d'un objet par les connexités de l'image**

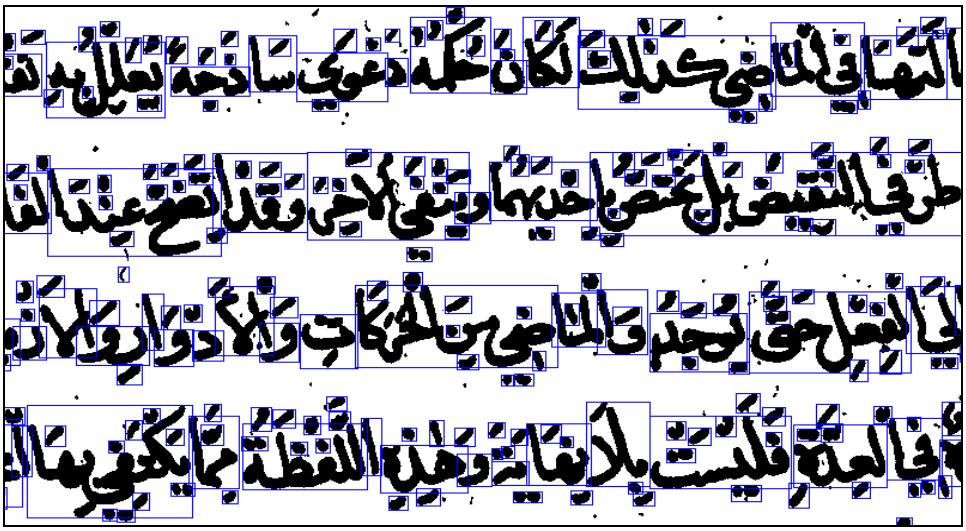


**Illustration n°. 22 : Segmentation des connexitiés : rectangles cironscrits aux connexitiés**

Pour éviter de capturer des objets trop petits comme les points isolés ou les petites taches, nous éliminons toutes les connexités trop petites dont la largeur ou la hauteur est inférieure à taille donnée. Ainsi les traits peuvent s'éliminer s'ils sont horizontaux et verticaux et s'ils possèdent une épaisseur suffisamment faible. Cependant les cadres continus forment de grandes connexités qui ne peuvent pas être supprimés. Le choix d'utiliser la notion de connexité pour définir les objets va donc poser des problèmes sur le traitement des illustrations ou les textes qui touchent les cadres ainsi que les tableaux. La première figure montre le dessin de la roue touchant le cadre, il fait donc partie de l'objet «cadre». De même les textes qui touchent le cadre feront parti intégrante du cadre. Sur la figure suivante, le cadre est continu sur la parti haute de l'image et forme une grande connexité englobant l'illustration jusqu'au milieu de la page. A cause de la mauvaise qualité de l'image, le cadre est représenté en bas de l'image par des traits pointillés qui ont été supprimés dans l'analyse. Sur cette image, nous n'obtenons ni un objet « cadre » complet, ni un objet « image » séparé du cadre. Enfin sur la dernière image, le cachet étant imprimé sur une partie du titre et du cadre, nous obtenons un objet commun regroupant tous ces éléments.

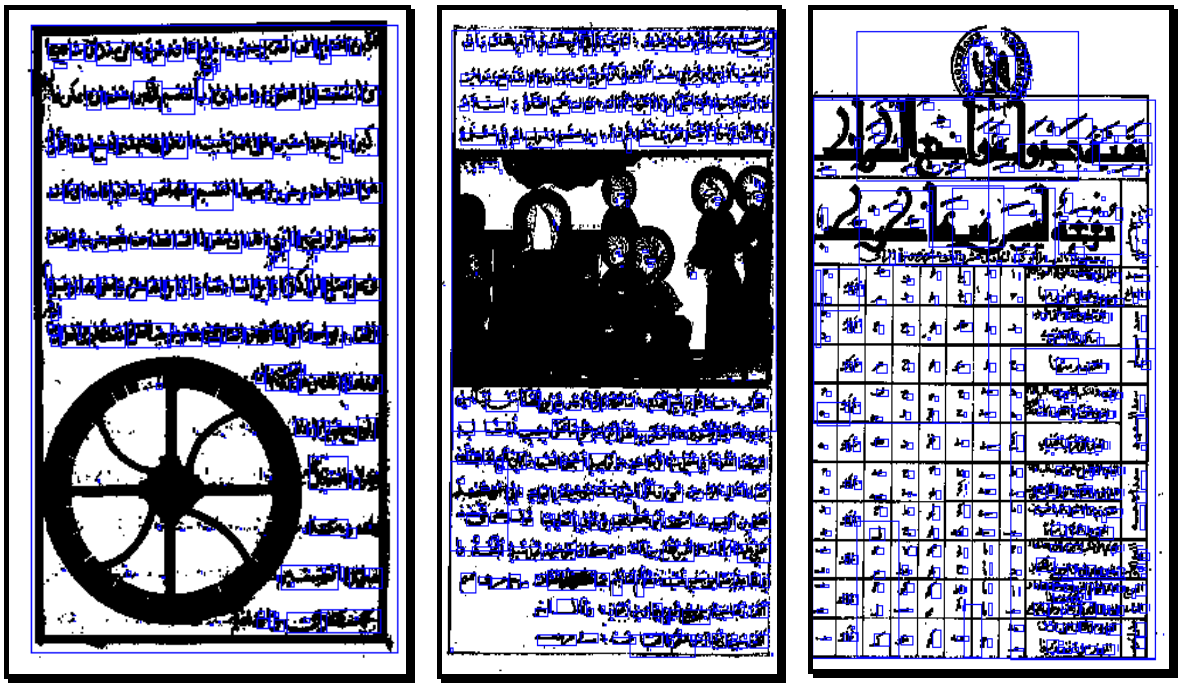


**Illustration n°. 23 : Problèmes rencontrés par la segmentation des objets à partir de la définition de connexité**

La segmentation de l'image en objets connexes est riche en information, mais elle est sensible à la qualité des images et nécessite que tous les objets soient espacés les uns des autres. Il existe des méthodes de traitement des images qui réduisent les points de contact entre les objets, mais appliquées à l'ensemble de l'image elles casseraient les blocs de texte en plusieurs objets indépendants.

#### 3.3.2.2.3. Le traitement des cadres et des illustrations

L'autre solution consiste simplement à utiliser les outils de traitements morphologiques pour mesurer les tailles de tous les objets et de séparer tous les objets dont la taille est supérieure à une valeur, c'est à dire les bordures, les taches, les illustrations et les cadres. Une fois séparée, on traite ces éléments graphiques pour distinguer les bordures dont l'épaisseur est faible des illustrations qui sont des grands objets.

La morphologie permet d'affecter à toute connexité une valeur correspondant à la taille maximale, celle-ci pouvant être mesurée dans n'importe quelle direction. Par exemple sur l'image suivante Arabe 2478 (R18271) image n°0120, on a appliqué la morphologie pour affecter à tous les objets une valeur correspondant à la hauteur et à la largeur maximale. Si les objets de grande dimension sont dans les deux cas bien détectés, en revanche, la mesure de la largeur maximale fait apparaître certains longs mots du texte arabe pour de grands objets et qu'il ne faut pas séparer.

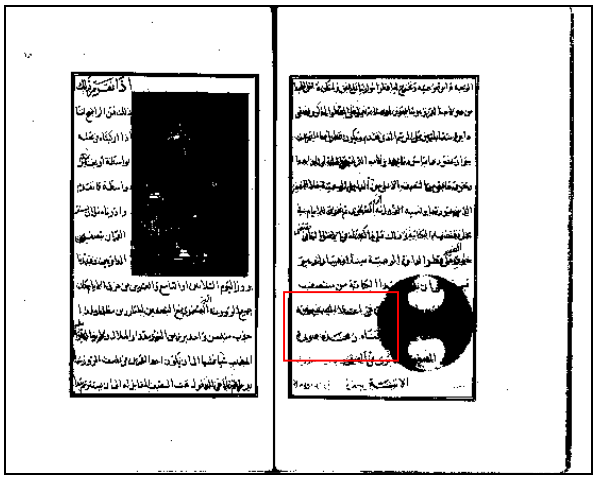
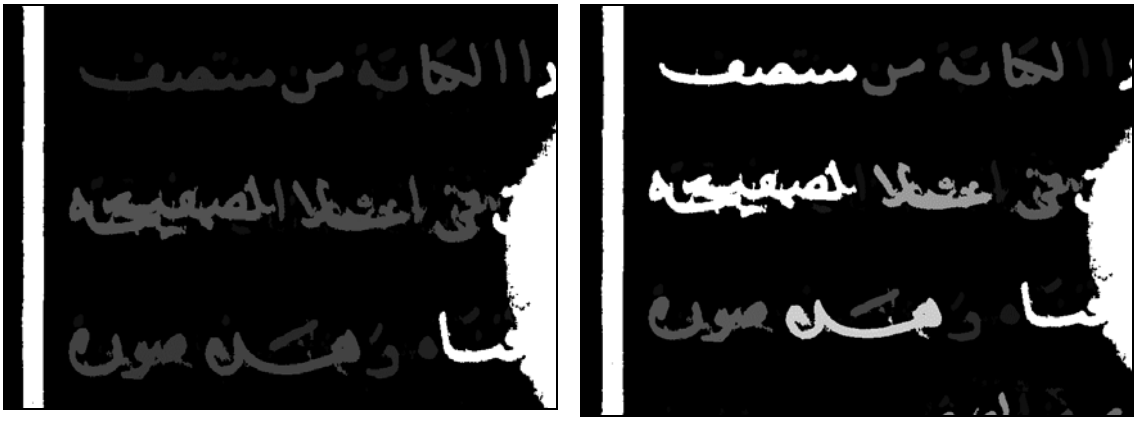


**Illustration n°.24 : Zone d'étude sur R18271 (Arabe2478) :image 0120**



*Mesure de la hauteur maximale* *Mesure de la largeur maximale*

**Illustration n°.25 : mesure de la hauteur et de la largeur maximale**

Par conséquent, nous avons choisi de mesurer tous les objets par la hauteur maximale, de façon à ce que le texte ne soit jamais séparé et classé comme un grand objet. La figure suivante montre que tous les objets de grande taille sont affectés d'une valeur maximale, alors que le texte possède une hauteur maximale négligeable. Cependant les petits objets qui touchent un grand objet comme le texte qui touche le cadre ou les illustrations feront partie intégrante de l'objet graphique.

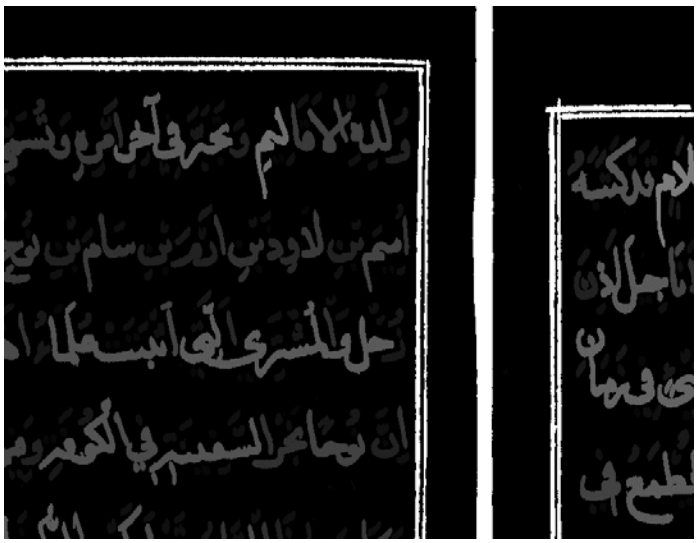


**Illustration n°.26 : Résultat de la mesure des objets par la hauteur maximale**

On sépare donc les objets de grande taille (illustrations, bordures du livre, cadres etc.) dans une autre image séparée de celle du texte.

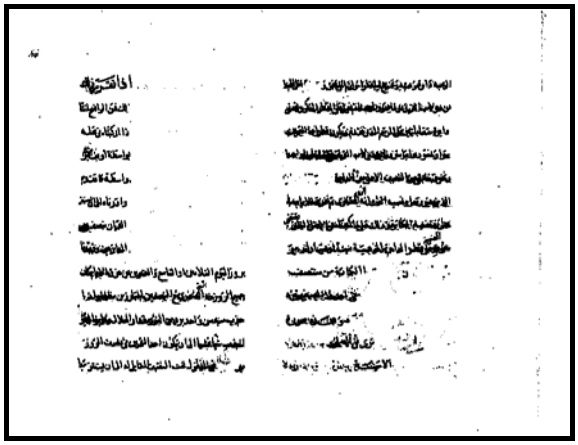


*Image A : Objets de faible hauteur (Texte)*

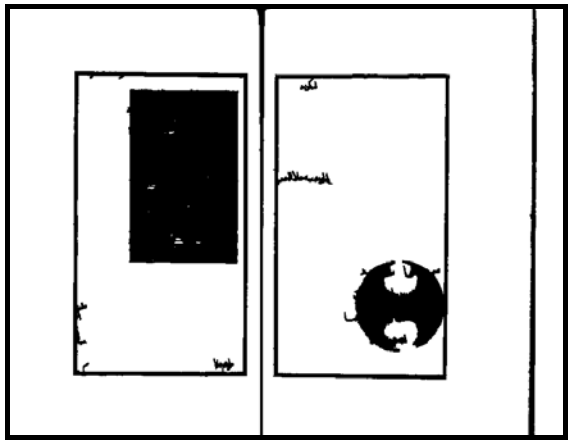


*Image B : Objets de grande hauteur (cadre, bordures du livre, illustrations*

**Illustration n°.27 : La séparation des objets de grande taille**

L'image des objets de grande taille va être traitée encore par morphologie pour retirer tous les objets de faible épaisseur comme les bordures et la reliure du livre ainsi que les cadres. Il ne restera de cette opération que les illustrations.

*Image C : Filtrage par l'épaisseur des objets graphiques de l'image B*

*Image D : Union de l'image de l'image C et A*

**Illustration n°.28 : La séparation des objets de grande taille**

Le texte ne sera pas affecté par ce traitement car la suppression des objets de faible épaisseur ne s'effectue que sur l'image contenant que les grands objets.

### 3.3.2.2.4. Détection automatique de la zone principale

***Problématique* :**

La zone principale délimite la région d'intérêt où se situent les textes. Cette zone principale peut être simple sur les images contenant une seule page (figure 2a) ou double

sur les images d'un livre ouvert (figure 2b). La détection de la zone principale de texte permet de définir les textes hors champs comme les annotations.

***Figure 2a*** *Zone principale simple* ***Figure 2b*** *Zone principale double*

**Illustration n°.29 : Détection automatique de la zone principale du texte**

Sur certains ouvrages, la zone de texte est formalisée par un cadre explicite (figure 3a) alors que pour d'autres ouvrages, elle ne peut se voir implicitement que par la justification des textes sur les bords des pages (figure 3b).

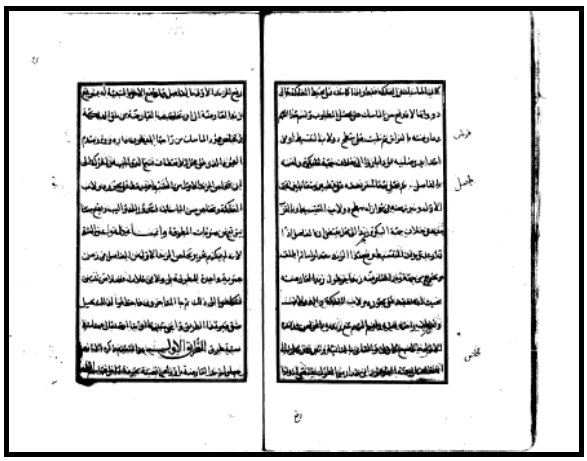


***Figure*** *3a zone principale encadrée explicitement* ***Figure 3b*** *zone principale définie implicitement par la justification du texte*

**Illustratin n°. 30 : Exemples de zone principale du texte**

Quand la zone principale est délimitée par un cadre, ce dernier n'apparaît pas toujours comme une ligne continue à cause de la mauvaise qualité de l'image ou d'une mauvaise binarisation (figure 4).

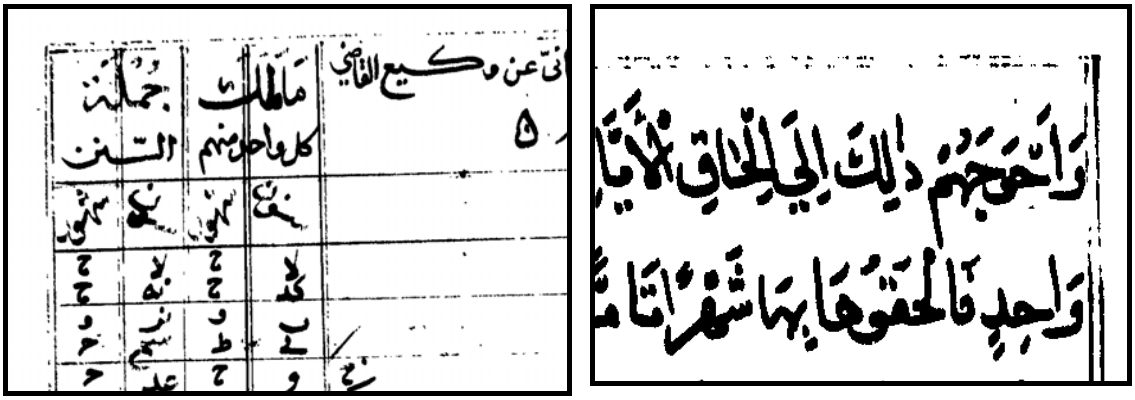


***Figure 4*** *Cadres partiellement visibles à cause de la mauvaise qualité des images.*

**Illustration n°.31 : Cadres partiellement visibles à cause de la mauvaise qualité des images**

La localisation des cadres peut s'effectuer facilement par la localisation des alignements horizontaux et verticaux de pixels noirs. Cependant, l'interprétation de tous les alignements par un programme n'est pas toujours simple. En effet, les bords du livre et de la reliure produisent des alignements qui peuvent être interprétés comme des cadres potentiels et qu'il faudra ignorer (figure 5a). De même, certaines illustrations peuvent contenir des cadres plus petits qui mettront en échec une interprétation automatique (figure 5b).

*Figure 5a présence des bords du livre et de la reliure*

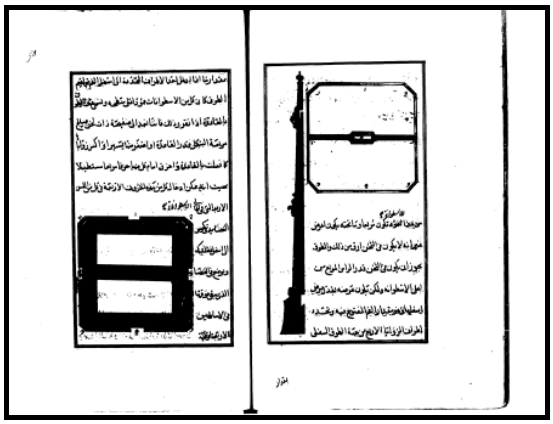


*Figure 5b présence de cadres dans les illustrations*

**Illustration n°. 32 : Presence des bords du livre, de la reliure et de cadre dans les illustrations**

Lorsque la zone principale n'est pas explicitement définie par un cadre, sa détection par analyse d'image peut être rendue difficile quand les textes ne sont pas justifiés (figure 6a) ou bien en présence de larges zones d'illustrations ou de tableaux (figure 6b).

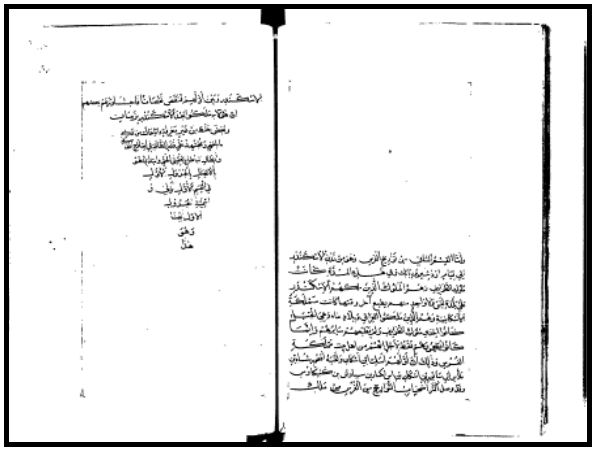


*Figure 6a Textes non justifiés et présence de larges zones sans texte*

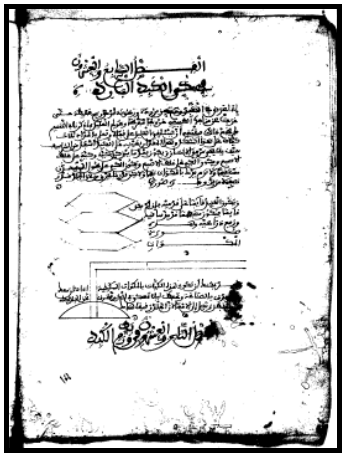


*Figure 6b Présence d'illustrations qui génèrent des espaces sans texte*

**Illustration n°.33 : Exemples de texte non justifié et la présence d'illustrations**

***Proposition***

La présence de cadres explicites étant peu fréquente sur le corpus, la méthode de localisation de la zone principale qui a été retenue est la détection de textes justifiés sachant que cette méthode peut tout de même échouer sur les documents présentant de larges zones vides ou bien avec des textes non justifiés.

Comme la reconnaissance est effectuée après la segmentation physique de la page, le logiciel d'analyse d'images n'a pas encore la connaissance de l'identité de tous les objets présents et donc des objets qui correspondent à la définition d'un texte ou non. Pour localiser les zones potentielles de texte, il faut procéder grossièrement à une pré-classification des objets en texte/non texte pour ensuite localiser les alignements des objets textuels seuls. A ce niveau la pré-classification des objets étant hasardeuse, on a choisi d'attribuer à chaque objet une probabilité *P(x)* d'être ou non du texte en fonction des tailles des objets. A priori, les objets textuels étant plus nombreux et de taille homogène, ils peuvent être statistiquement détectés en analysant la taille moyenne de tous les objets situés dans l'image.

Comme la taille moyenne des objets est proche de la taille moyenne d'un bloc de texte, on peut définir, pour chaque objet *x*, une valeur de probabilité *P(x)* entre 0 et 1 à partir de l'écart de la taille de *x* à la taille moyenne de tous les objets.

$$P(x)= 1 - \frac{|Taille(x)-tailleMoyen|}{\underset{\text{tout objet } x}{Max}\,|Taille(x)-tailleMoyen|}$$

En projetant horizontalement et verticalement à la position de chaque objet *x* la valeur *P(x)* on construit deux histogrammes *Xprofile* et *Yprofile* :

*Xprofile* = projection verticale de *P(x)* sur la largeur de *x*

*Yprofile* = projection horizontale de *P(x)* sur la hauteur de *x*

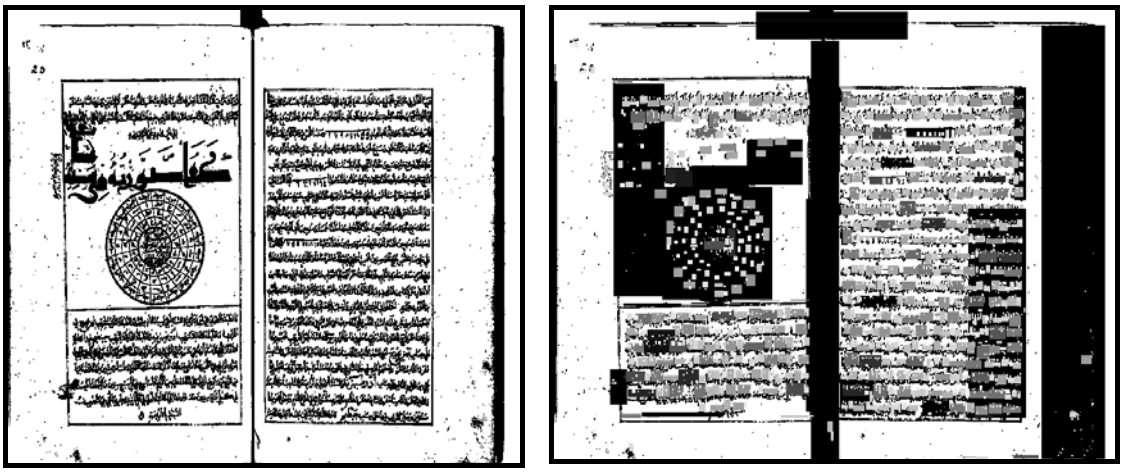
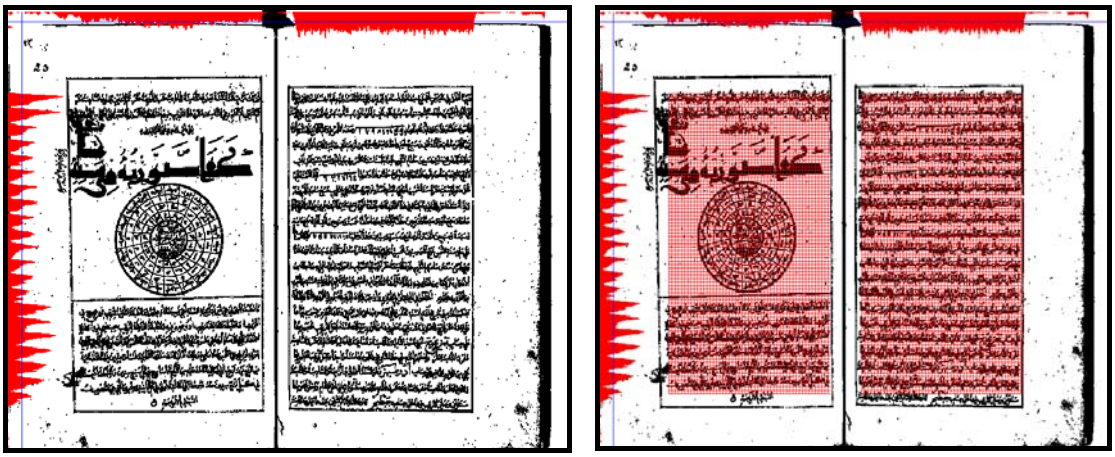


*Image originale* *Affichage de P(x) sur chaque objets*

**Illustration n°. 34 : l'affichage de P(X) sur chaque objets**



*Xprofile et Yprofile et seuil de détection* *Localisation des corps de texte*

**Illustration n°.35 : Localisation des corps de texte**

Les projections *Xprofile* et *Yprofile*, affichées sur le haut et la gauche de l'image, montrent des valeurs élevées aux coordonnées où se situent les objets qui ont une taille proche de la taille moyenne des objets, c'est à dire précisément là où se situe le texte. Un seuil, affiché sous la forme d'une ligne qui coupe les "profiles", est automatiquement

calculé pour déterminer les bords des zones qui contiennent le plus de texte. On balaye les deux "profiles", en partant des coordonnées extrêmes de l'image, jusqu'à atteindre une valeur de "profile" supérieur au seuil calculé, sur les 4 coins de l'image. On effectue un traitement supplémentaire pour détecter une double zone principale en cherchant au milieu de la zone trouvée, des valeurs de *Xprofile* inférieur au seuil. Si il existe une telle une zone, alors on balaye *Xprofile* vers la gauche puis vers la droite jusqu'à ce que l'on détecte des valeurs supérieures au seuil, indiquant la présence du bord du texte au centre du livre.

L'avantage d'un calcul automatique du seuil, c'est qu'en présence d'une faible quantité de texte, le seuil s'abaisse automatiquement. En l'absence d'objets de taille moyenne en grand nombre, ce seuil peut s'abaisser au point de détecter les bords du livre. Il reste encore des problèmes de détection ; dans certains cas extrêmes comme sur les pages présentant des annotations trop nombreuses peuvent faire dévier la détection jusqu'à les englober. De même, la présence de grandes zones d'illustrations peut diminuer localement la quantité de blocs de texte et donc fausser le calcul des projections et donc la détection.

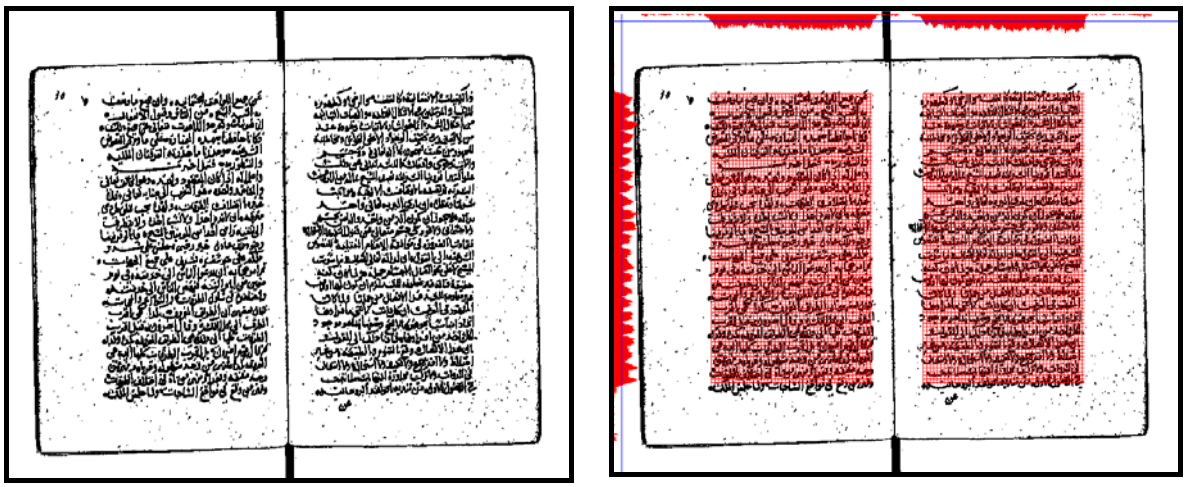
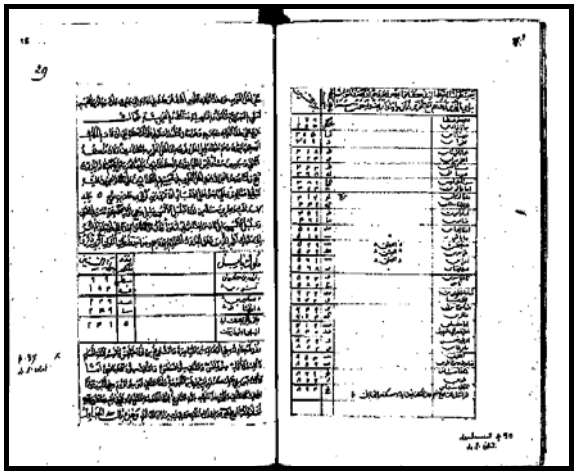
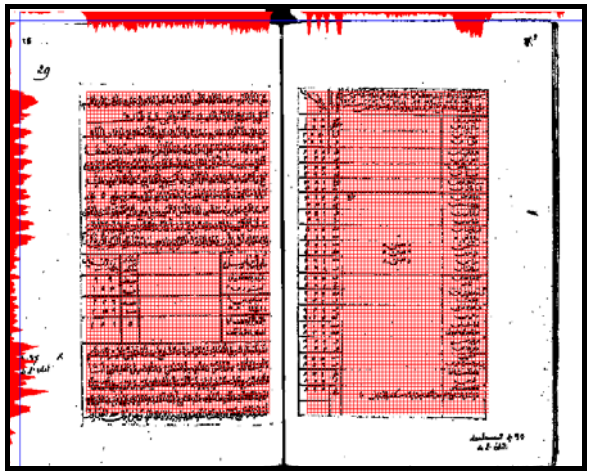


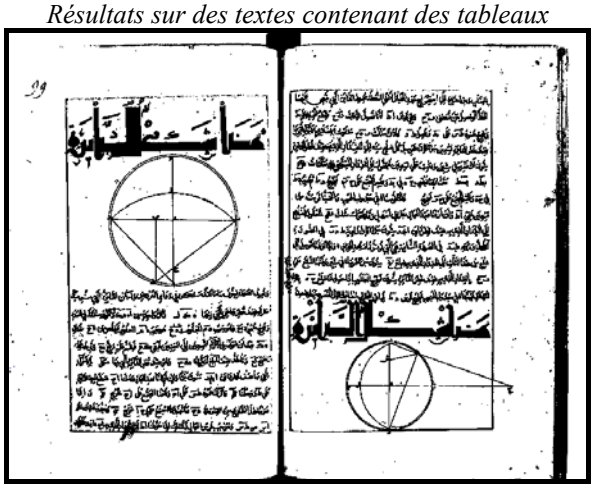
**Illustration n°.36 : Résultats sur des textes inclinés**



*Résultats sur des textes contenant des tableaux*



*Résultats sur des textes contenant des illustrations*



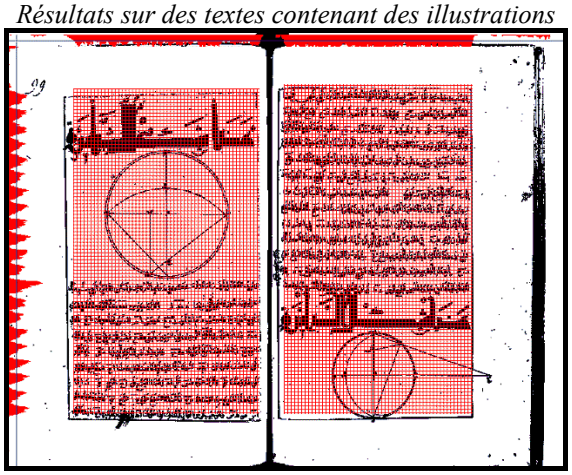


**Illustration n°.37 : Résultats sur des textes contenant des tableaux et des illustration**

Les objets vont être séparés en deux groupes :

- Les objets situés dans les zones principales
- Les objets situés en dehors des zones principales

Tous les objets seront soumis à la reconnaissance des formes car il est important de savoir la nature (texte/graphique/ornement) de tous les objets y compris ceux qui sont situés en dehors des zones principales.

### 3.3.2.3. La reconnaissance des formes

Le processus de reconnaissance des formes s'effectue en deux étapes :

- L'analyse et la mesure des objets : cette phase correspond à la nécessaire *caractérisation* des objets par différentes mesures aussi bien géométriques, de forme ou sur la couleur quand cette information existe. Le choix des descripteurs

qui caractérisent les objets est essentiel pour les performances. Si les descripteurs ne sont pas adaptés aux objets et aux classes souhaitées, l'étage de reconnaissance ne pourra pas réaliser une classification performante.

- La classification des objets : Cette étape importante permet de classer les objets et déterminer si ils sont des textes, des titres, des annotations ou des graphiques. L'utilisateur donne le nombre de classes qu'il souhaite obtenir et leurs libellés.

#### 3.3.2.4. La caractérisation des objets

Nous laissons à l'utilisateur le choix entre 14 descripteurs dont 4 pour la couleur, 4 sur la forme des objets, et 6 sur la géométrie et le dimensionnement.

***Les mesures sur la couleur des objets :***

- La luminance : la valeur moyenne Y des canaux R,V, B.
- la chrominance : le couple (U,V) du système de couleur YUV de la norme standard de codage des images couleurs. Le système de couleur YUV étant qu'une rotation du système de couleur RVB pour faire coïncider l'un de ses axes sur l'axe principal de la luminance Y. Le système de couleur YUV permet de décorréler la luminance Y des informations sur la chrominance (U,V) qui détermine la couleur des objets.
- La saturation : La saturation est mesurée par la distance d'une couleur dans l'espace tridimensionnel RVB par rapport à l'axe de la luminance Y. Plus le pixel est proche de l'axe de la luminance plus sa couleur est proche d'une nuance de gris traduisant une saturation presque nulle. Inversement, plus la valeur d'un pixel est éloignée de l'axe de la luminance et plus sa couleur est saturée.
- La teinte : La teinte est l'angle que fait une couleur avec l'axe de la luminance Y.

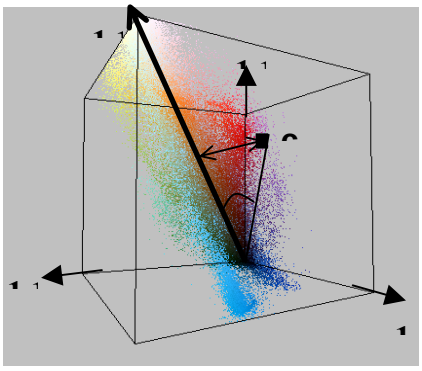
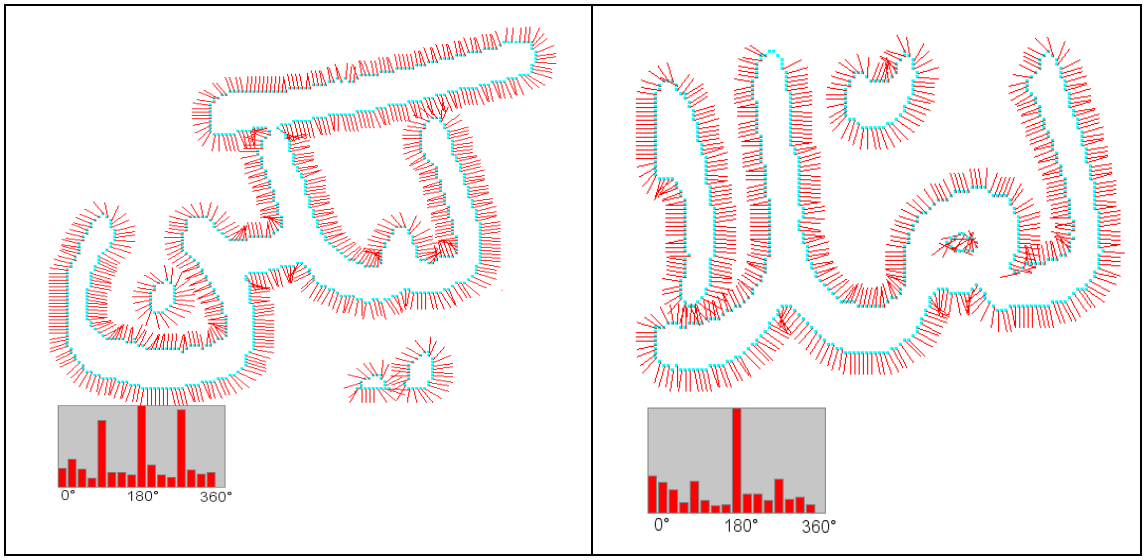


**Figure n°57. Mesure de la saturation S et la teinte T d'une couleur c dans l'espace RVB**

La mesure de la saturation et de la teinte sont calculées à partir des informations de chrominance. Il n'est donc pas utile de conserver à la fois les informations sur la chrominance et les informations de teinte et de saturation. Le choix entre l'une et l'autre des représentation de la couleur se justifie par l'importance de la saturation des couleurs comme mesure intéressante pour caractériser les objets d'une classe particulière. Si la saturation n'apporte pas d'information supplémentaire, on choisira alors par défaut l'information de chrominance.

***Les mesures sur la forme des objets :***

- La densité : cette mesure correspond à la surface relative qu'occupe un objet dans le rectangle englobant. Il est calculé par le rapport entre le nombre de pixels noirs appartenant à l'objet et le nombre de pixel blancs appartenant à l'arrière-plan.
- La structure : la variation des épaisseurs des traits
- La courbure : la distribution des niveaux de courbures dans 16 directions.



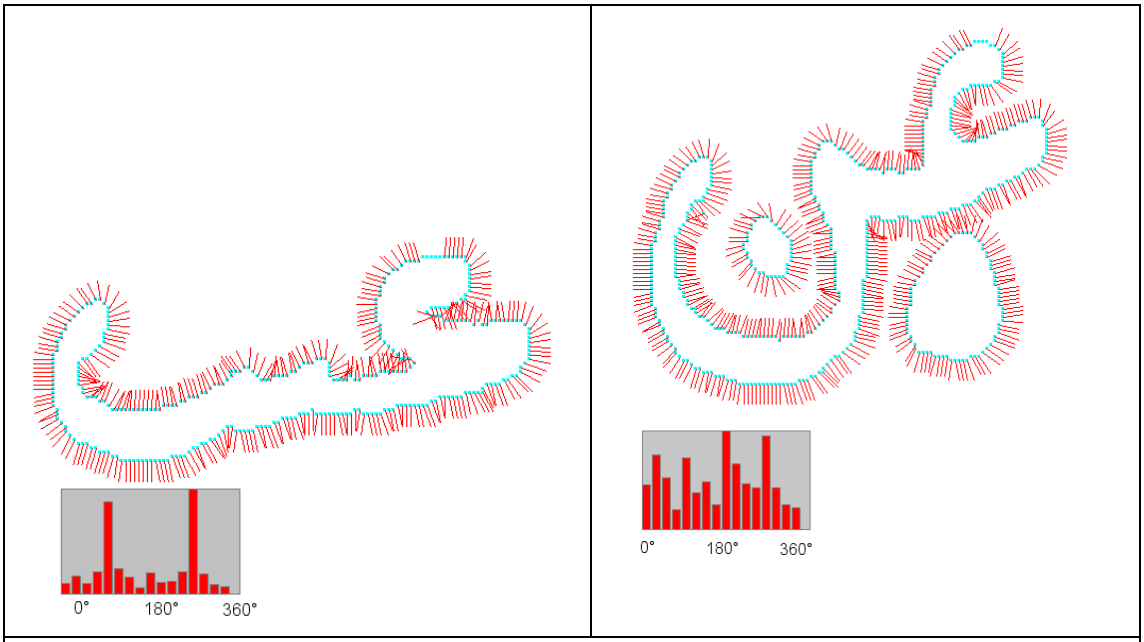



**Illustration n°.38 : La distribution des niveaux de courbures dans 16 directions**

- Les Projections : la projection horizontale et verticale de l'objet sur une échelle de 5 valeurs respectivement.

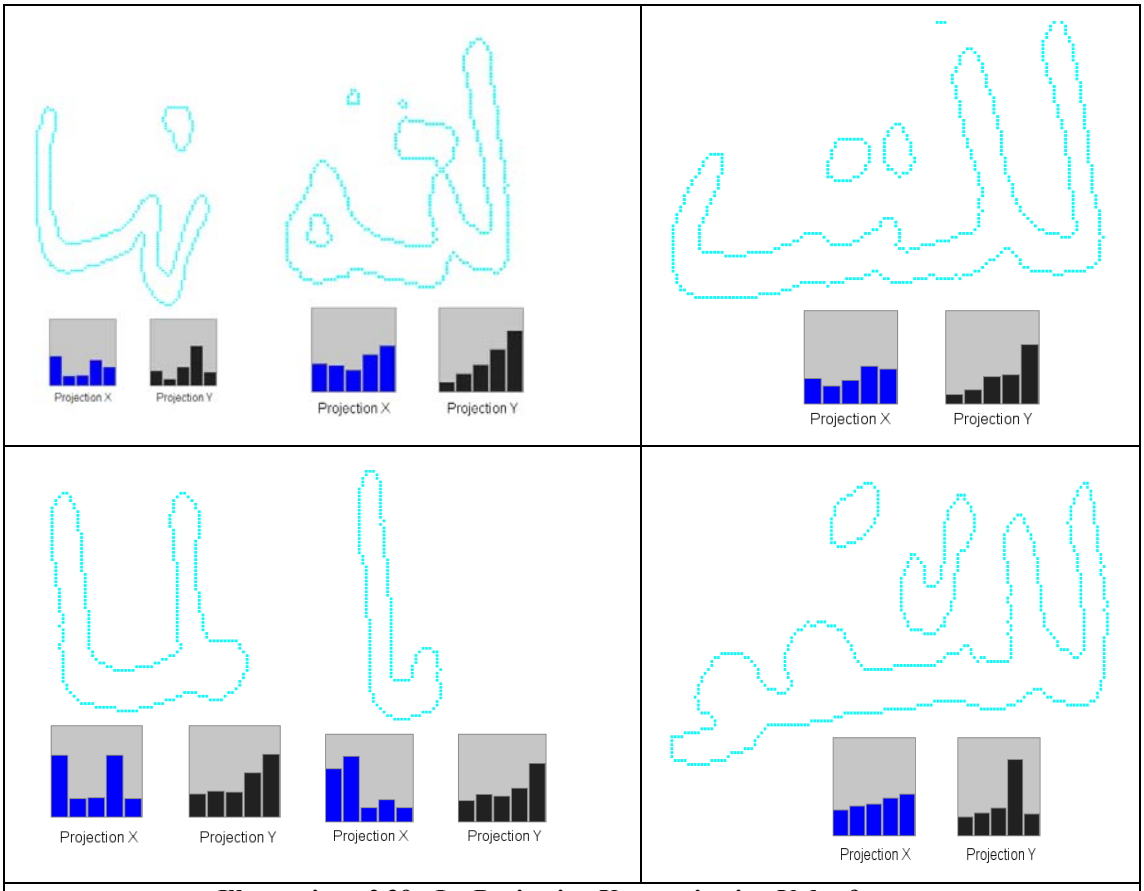


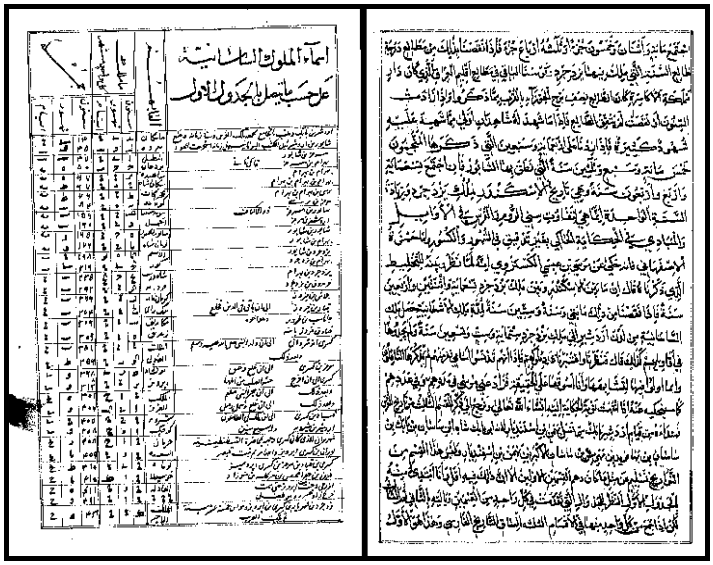

**Illustration n°.39 : La Projection X et projection Y des formes**

***Les mesures sur la géométrie des objets :***

- La hauteur de l'objet
- La largeur de l'objet
- L'épaisseur : l'épaisseur moyenne des traits
- La longueur : la longueur de tous les traits de l'objet.
- La position relative dans l'image en abscisse
- La position relative dans l'image en ordonnée

#### 3.3.2.5. La prise en compte des relations spatiales

Il existe des métadonnées qui ne peuvent pas être reconnues à partir de la seule forme des objets analysée indépendamment les uns des autres. Par exemple, pour le manuscrit *Arabe 1489* (R28062), l'alignement entre les objets et la répartition spatiale régulière des objets doivent être pris en compte dans la reconnaissance des textes situés dans les tableaux. En effet, la forme des textes est identique à l'intérieur comme à l'extérieur des tableaux et la présence des bordures des tableaux n'est pas toujours apparente dans l'image. Dans un autre exemple sur le manuscrit *Arabe 179* (R60914), le texte est identique en taille et en épaisseur à celui des titres et aucune des primitives décrites précédemment ne peut permettre la reconnaissance des titres à partir de la seule forme des textes. Les titres sont reconnaissables seulement à partir de l'indentation du texte par rapport à la bordure de la page et à la distance avec le texte supérieur et inférieur. Ces deux exemples illustrent combien l'alignement, la justification et les distance entre les blocs de texte sont importants pour l'extraction de certaines des métadonnées.



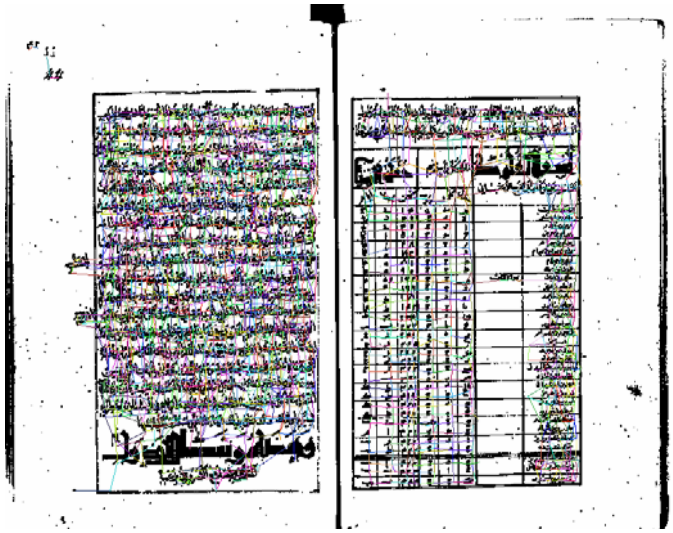
*Exemple où la régularité de la répartition spatiale entre les objets est nécessaire pour différencier les textes dans les tableaux du texte principal (Arabe 1489 (*R28062) image « 0275 »)

*Exemple où la distance entre les objets doit être prise en compte pour la reconnaissance des titres (interligne et indentation) (Arabe 179 (R60914) image «008 »).*

**Illustration n o.40 : Exemple : La prise en compte des relations spatiales**

Nous avons donc rajouté des primitives qui traduisent les relations spatiales entre un objet et ses voisins. Pour exprimer les notions de régularités, de distances entre objets et d'alignements nous devons procéder par étape :

- Chercher pour chaque objet les 4 voisins les plus proches dans les directions principales *(nord, sud, est, ouest)*. S'il n'existe pas de voisins proche d'une distance inférieure à un seuil, alors le champ reste vide.
- Calculer les caractéristiques spatiales entre chaque objet et ses voisins qui sont : les distances, les alignements verticaux et horizontaux.



*Recherche des voisins limitée par une distance maximale fixe.*

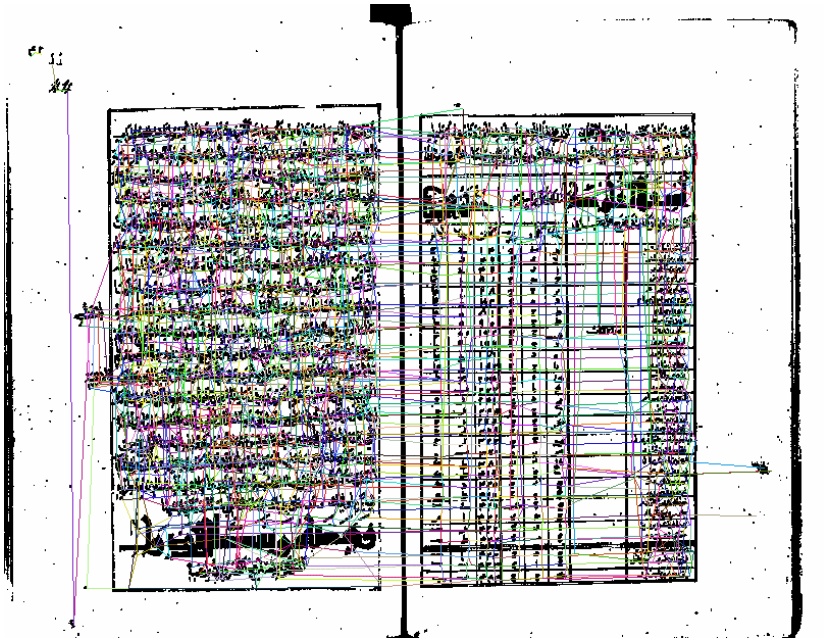


**Illustration n°. 41 : Recherche de tous les voisins sans limitation de distance**

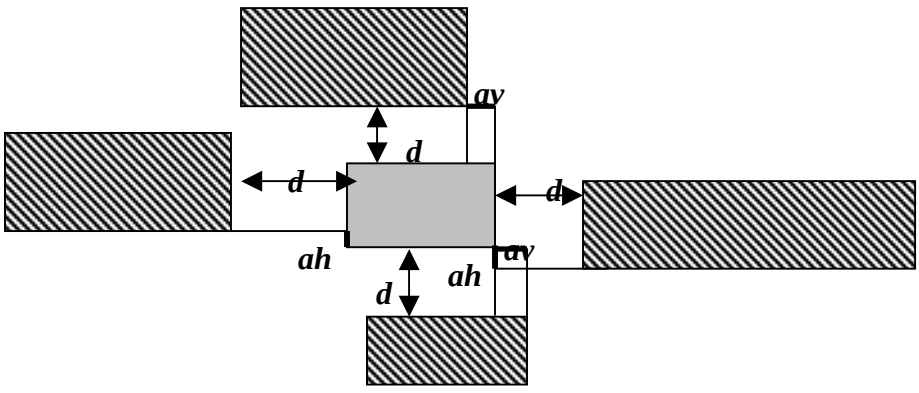



**Figure n°. 58** *: Caractéristiques spatiales entre objets voisins : mesures d'alignement et de distance entre un objet et ses 4 voisins : 2 mesures d'alignements horizontaux ah, 2 mesures d'alignements verticaux av et 4 distances d.*

Les alignements verticaux *av* avec les objets situés au-dessus et en dessous, sont mesurés à partir des bords situés à droite des objets afin de tenir compte du sens naturelle de lecture en arabe. La longueur des mots étant variable, seule la justification à droite est intéressante. Les alignements horizontaux *ah* expriment les variations de l'alignement horizontal des objets voisins par rapport à la ligne de base car les blocs de texte ont une hauteur variable. S'il n'y a pas de voisin proche dans une direction, les primitives associées avec ce voisin sont mises à zéro. La distance *d* entre les objets voisins est prise à partir des bords des objets les plus proches pour limiter les effets de la variation en longueur et en hauteur des mots.

#### 3.3.2.5.1. La reconnaissance

Cette reconnaissance a pour objectif d'identifier des classes d'objets (titres, illustrations, cadres, textes, etc.). Ce processus, qui est en fait une classification, peut s'effectuer de deux manières différentes :

- *La classification supervisée par l'utilisateur* : Elle consiste, pendant une phase *d'apprentissage*, à rentrer un certain nombre d'observations qui permettront la prise de décision par la machine pour le classement de nouveaux objets. Cette approche permet de diriger complètement le système de reconnaissance grâce au choix des observations que le système doit apprendre. Plus le nombre d'observations par classe sera élevé, plus la classification sera juste mais plus la phase d'apprentissage sera longue et fastidieuse.

  Pour une étude de faisabilité, nous avons utilisé une méthode très simple comme le K-PPV (K Plus Proches Voisins) qui attribue la classe majoritaire parmi les K observations les plus proches de l'objet à reconnaître. Le nombre K d'observations dépend du nombre de classes et du nombre d'observations par classes lors de l'apprentissage. Un classifieur 1-PPV, qui tient compte que de l'observation la plus proche, est sensible aux cas particuliers et donne une classification trop dépendante de la justesse de la base d'apprentissage. Inversement un classifieur K-PPV avec un nombre K élevé est indépendant des cas particuliers et gagne en généralité et en justesse. Cependant pour augmenter le

nombre K, il faut augmenter le nombre d'observations et donc le temps d'apprentissage.

- *La classification automatique non supervisée* : C'est un algorithme qui, à partir des caractéristiques des objets et du nombre de classes désirées, va effectuer tout seul une classification automatique sans l'assistance de l'utilisateur. La fastidieuse phase d'apprentissage est évitée mais cette approche ne donne pas de résultats reproductibles et conformes aux souhaits de l'usager en raison de l'absence d'observations pour guider la classification. Cette méthode est cependant intéressante pour tester la pertinence des caractéristiques et évaluer les performances futures d'une classification supervisée en fonction du nombre de classes souhaitées.

  La méthode retenue est celle des K-MEANS qui consiste à classer itérativement les objets par rapport à K centres de classes pris au hasard puis à recalculer ces centres en effectuant la moyenne des caractéristiques des objets de chaque classe. La classification s'arrête quand les centres restent immobiles. La classification des observations en K classes est alors optimale.

#### 3.3.2.5.2. L'apprentissage

L'apprentissage consiste à désigner les observations pertinentes et à indiquer leurs classes respectives. C'est une étape critique dont va dépendre la qualité de la reconnaissance. Le choix des observations et leur nombre par classe sont deux facteurs importants. Plus le nombre d'observations est élevé, plus la reconnaissance sera sûre. Naturellement, l'apprentissage doit être réalisé pour chaque ouvrage, car toutes les pages d'un même ouvrage gardent une présentation homogène et affichent des métadonnées communes qui peuvent être reconnues

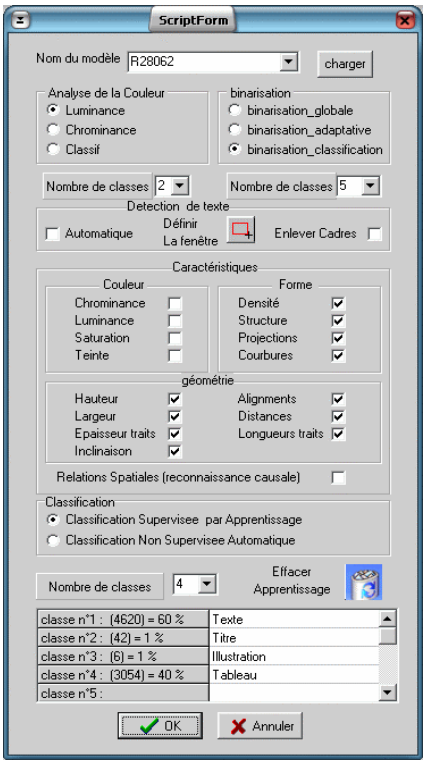
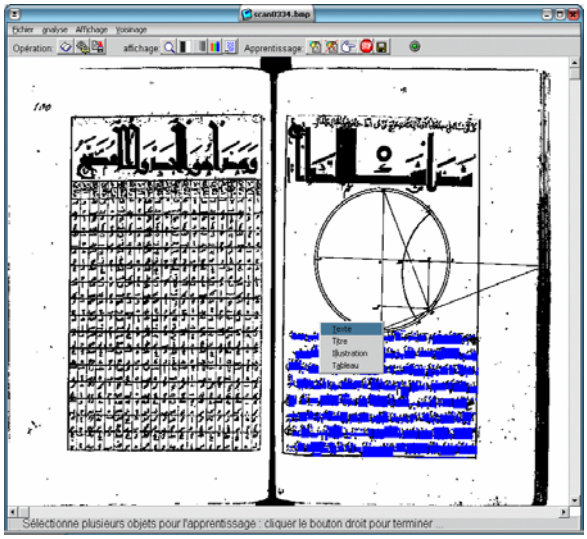



**Illustration n°. 42 : Interface du script**

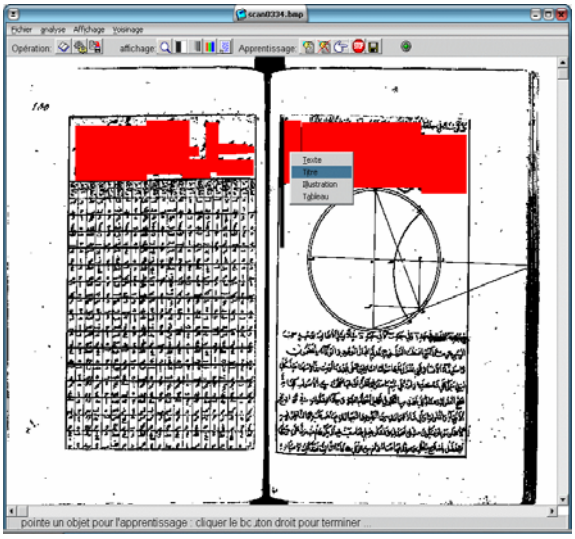
automatiquement. A l'inverse, deux ouvrages différents ont rarement la même présentation ni les mêmes métadonnées. Cette constatation nous a poussé à définir un script pour chaque ouvrage qui permet de conserver tous les paramètres nécessaires au traitement de ce dernier.

Les informations conservées par ce script concernent à la fois les méthodes de segmentation et leurs paramètres ainsi que les caractéristiques que l'utilisateur va choisir, pour définir les objets, le choix de la méthode de reconnaissance, le nombre de classes et l'apprentissage des observations. L'interface indique aussi le nombre d'observations acquises lors de l'apprentissage et permet aussi de nommer chaque classe des métadonnées.

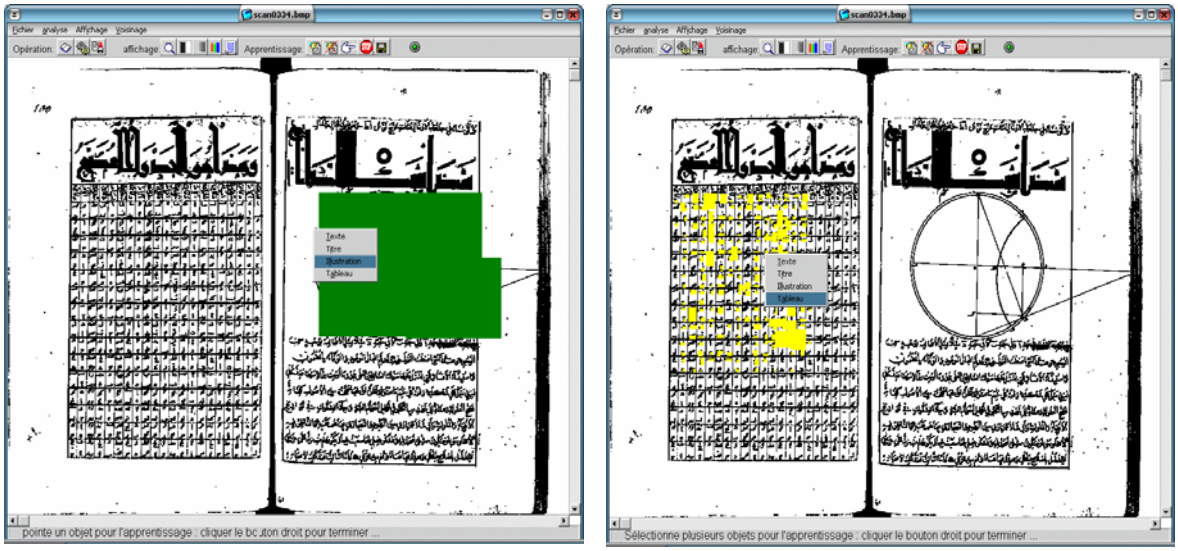
Très conviviale, l'interface utilisateur simplifie la phase d'apprentissage. Un certain nombre d'outils permettent de sélectionner (ou déselectionner), en pointant avec la souris, un ou plusieurs objets. Un menu contextuel apparaît pour déterminer la classe des objets sélectionnés.




*Interface client : Saisie de la classe n°1 (Texte)*




*Interface client : Saisie de la classe n°2 (Titre)*



*Interface client : Saisie de la classe n°3 (Illustration)* *Interface client :Saisie de la classe n°4 (Tableau)*

**Illustration n°.43 : Interface client avec les saisie des différentes classes**

L'utilisateur peut à tout moment vérifier la progression de l'apprentissage en relançant le processus de reconnaissance sur la même page avec les nouvelles observations saisies et vérifier l'amélioration de la reconnaissance. Si la reconnaissance se dégrade au fur et à mesure de l'apprentissage, alors il faut remettre en cause le choix des caractéristiques ou celui des métadonnées. Nous avons par exemple constaté des difficultés à différencier les illustrations des cadres illuminés qui ont les mêmes caractéristiques. Dans l'impossibilité de différencier certaines métadonnées, nous préférons les regrouper ensemble. En cas d'échec de séparation entre certaines métadonnées il faut alors réfléchir sur la formalisation de nouvelles caractéristiques physiques qui permettraient de les différencier de façon fiable et répétitive sur l'ensemble d'un ouvrage. C'est ce travail difficile qui nécessite du temps et des connaissances approfondies des manuscrits anciens.

### 3.3.2.6. Résultats

Cinq manuscrits ont été choisis pour valider le processus de reconnaissance car ils avaient une certaine richesse dans leurs métadonnées. Il est juste question ici d'évaluer la faisabilité de l'analyse d'image comme aide à l'extraction de métadonnées dans les images numérisées de manuscrits anciens en langue arabe. Ce ne sont donc que des premiers résultats qui devront être confirmés par la suite sur un nombre plus conséquent de manuscrits.

### 3.3.2.6.1. Résultats sur MS6191

Les images de l'ouvrage MS6191 ont été obtenues par la numérisation directe de l'ouvrage original en couleur. Nous serions dans les conditions optimales de qualité en terme de numérisation, si les images n'avaient pas été réduites en terme de résolution et comprimées avec la compression JPEG avec une perte d'information visible qui gène l'analyse d'image. Cependant l'information couleur est suffisamment importante pour pallier la perte de résolution et les déformations engendrées par la compression JPEG.

#### 3.4.2.6.1.1. Reconnaissance non supervisée sans apprentissage

En mode de reconnaissance non supervisé. Nous avons testé le pouvoir discriminant des informations extraites dans les images. Nous avons demandé au système de classer automatiquement de façon optimale tous les objets en 8 classes en tenant compte de toutes les primitives possibles. Dans la première classe, nous avons retrouvé les ponctuations ; dans les classes n°2 et n°5, on retrouve les signes diacritiques ; les autres classes représentent des mots triés suivant leur forme, leur longueur, leur épaisseur etc.

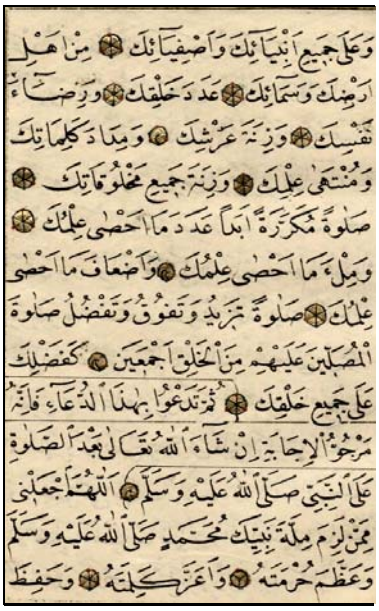
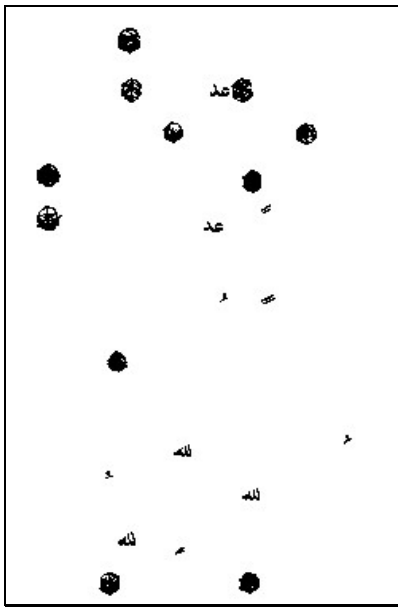
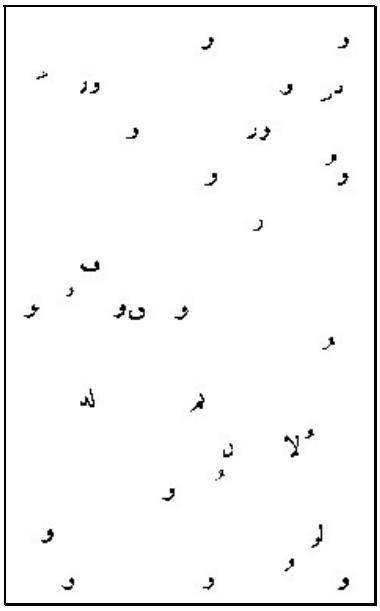


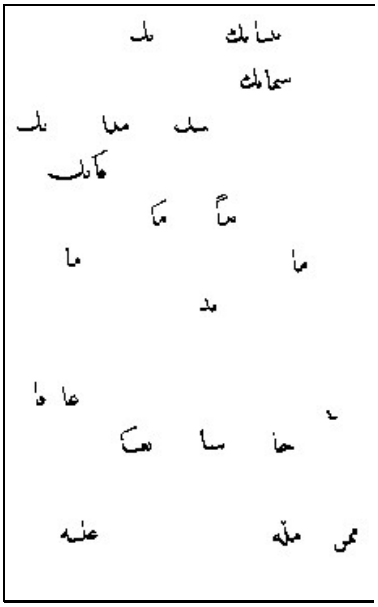
*Image originale*



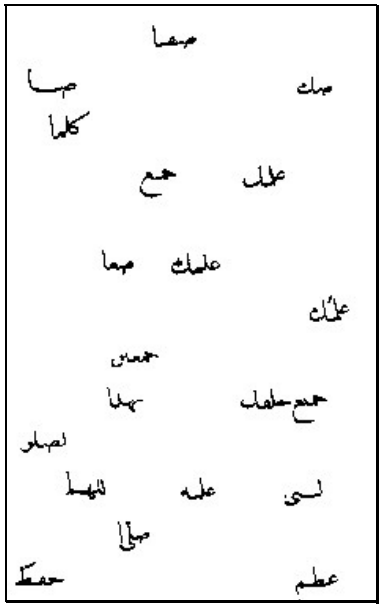
Classe 1



*Classe 2*

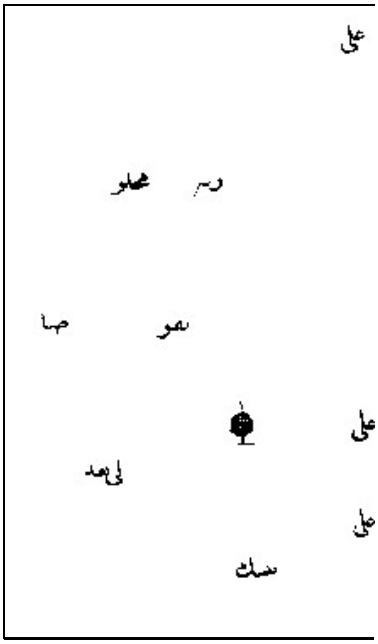


*Classe 3*

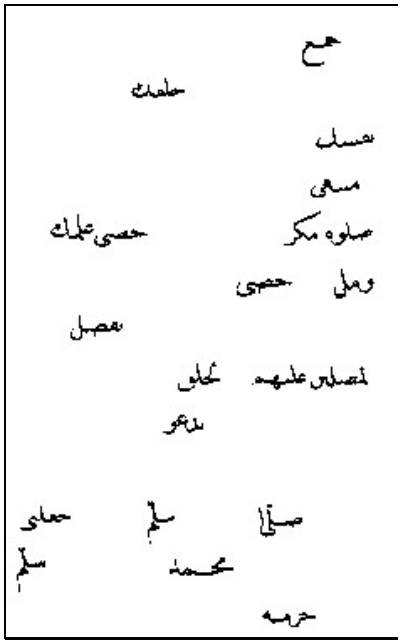


Classe 4

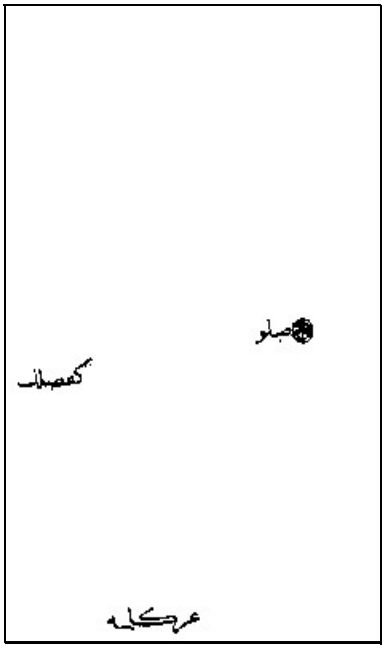
*Classe 5*



*Classe 6*



Classe 7



*Classe 8*

**Illustration n°. 44 : Résultats de reconnaissance non-supervisée sur le manuscrits MS6191**

On s'aperçoit déjà que la connexion entre objets de classes différentes va mettre en échec la reconnaissance des formes. Par exemple, les ponctuations rattachées au texte sont classées dans les classes 5, 6 et 8. Le système a tenu compte de l'importance du nombre d'objets dans chaque classe pour classer automatiquement les objets entre eux sans l'intervention extérieure de l'utilisateur. Le mode de reconnaissance non supervisé est intéressant pour tester le pouvoir discriminant des primitives mais ne peut pas être utilisé

en tant que tel pour l'extraction des métadonnées. C'est à l'usager de désigner les objets qui doivent être reconnus comme appartenant à une classe donnée lors d'un apprentissage. Le mode de reconnaissance supervisé est donc le mode le plus intéressant pour l'extraction automatique des méta-données.

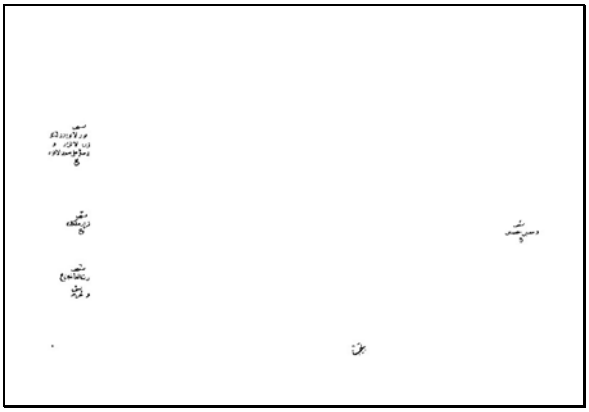
### 3.4.2.6.1.2. Reconnaissance supervisée par apprentissage

En mode de reconnaissance supervisé, nous avons fait l'apprentissage sur les 5 premières pages en quelques minutes avant de lancer la reconnaissance sur les 70 pages restantes. Le travail a été fini en moins de 40 minutes pour toutes les pages en tenant compte de la séparation des couleurs, de la binarisation, de l'extraction des objets et des relations entre eux, de les mesures des primitives et la reconnaissance. Le temps de calcul est d'environ 35 secondes par page sur un PC 1,9GHZ. Nous avons modélisé 4 méta-données à reconnaître :

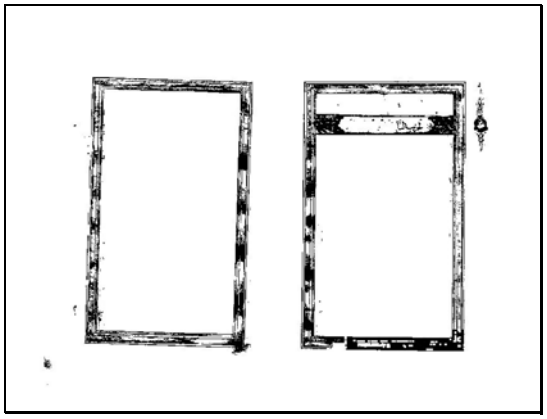
- Le texte principal
- Les annotations
- Les dessins et cadres illuminés
- Les ponctuations

*a) Image originale*

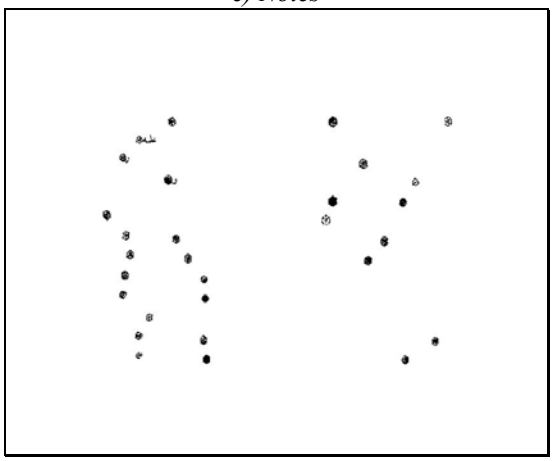
*b) Texte Principal*



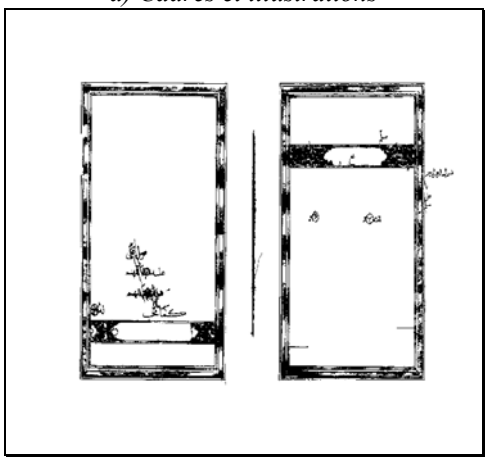
*c) Notes*



*d) Cadres et illustrations*



*e) Ponctuation*



*f) Erreurs de reconnaissance : Texte connecté aux cadres et ponctuation connectée au texte*

**Illustration n°. 45 : Résultat de la reconnaissance supervisée sur l'ouvrage MS6191**

Les ponctuations qui touchent le texte ont été systématiquement classés comme du texte, car le système a probablement retenu lors de l'apprentissage la forme géométrique circulaire des ponctuations. Cette forme n'apparaît pas quand la ponctuation touche le texte. La connexion de texte au cadre illuminé constitue une autre erreur assez fréquente. C'est la faible résolution qui empêche à l'analyse d'image de trouver un espace vide entre le texte et le cadre ou les éléments de ponctuation. Seule l'augmentation de la résolution permettrait de pallier ces difficultés. Les taux de reconnaissance sont très satisfaisants et permettent d'exploiter directement les résultats pour l'enrichissement de la base de donnée sans correction.

#### 3.3.2.6.2. Résultats sur le manuscrit *Arabe 2782* (R12051)

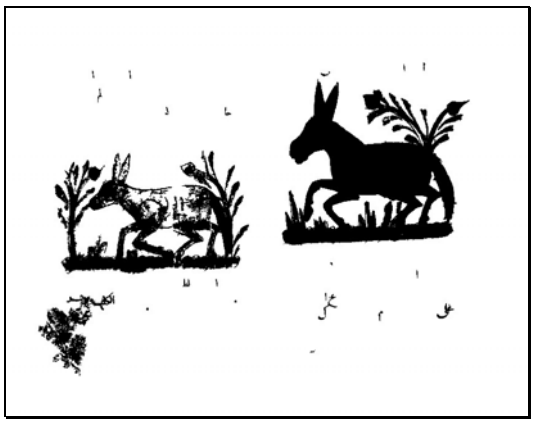
Le manuscrit *Arabe 2782* (R12051) a été numérisé en niveaux de gris sur un microfilm de très mauvaise qualité. De plus la qualité de la numérisation n'est pas régulière en

terme de contraste et de luminosité d'une page à l'autre traduisant un réajustement manuel de l'opérateur entre chaque prise d'image. La mauvaise qualité du manuscrit rentre aussi en ligne de compte avec la présence de nombreuses taches qui occultent le contenu des pages. Mais c'est l'irrégularité de la luminosité et du contraste qui va le plus perturber l'analyse d'image car les méthodes de segmentation sélectionnées et leurs paramètres ne marchent pas pour toutes les images présentant des contrastes différents. La segmentation des caractères qui étaient écrits en rouge (qui apparaissent donc en gris clair) va échouer sur toutes les images surexposées et ils ne pourront donc être reconnus plus tard. Et cette même nuance de gris apparaît sur le texte courant, quelques pages plus loin, à cause d'une surexposition de l'image lors de la numérisation. Une normalisation des images en terme de luminosité et de contraste est alors nécessaire pour rendre toutes les images comparables. Cependant, cet outil n'a pas été encore développé dans le prototype et donc les résultats sont très partiels. Les métadonnées sélectionnées sont :

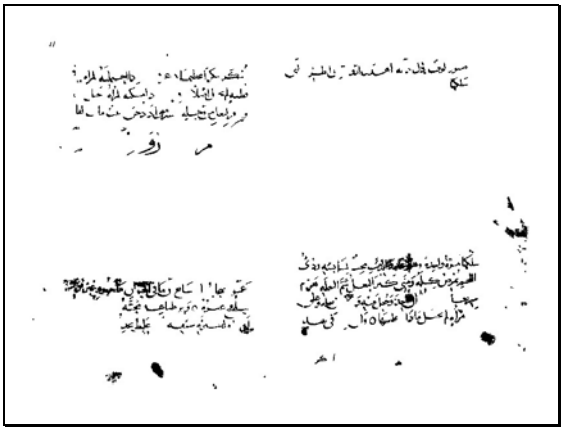
- Texte Noir
- Texte Rouge (apparaissant gris clair)
- Illustrations

Pour les images suffisamment contrastées, les résultats de la reconnaissance est satisfaisante.

*Image originale*



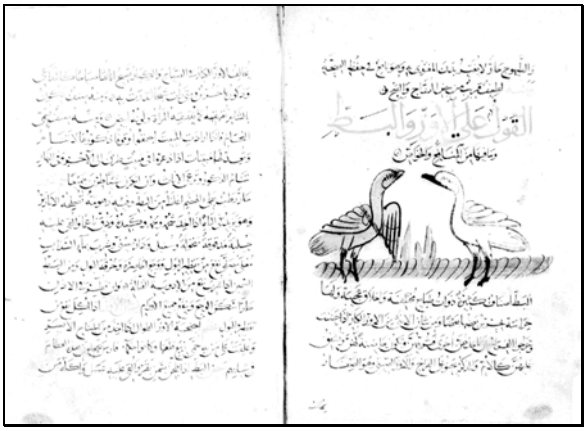
*Illustrations*



*Texte noir* *Texte rouge (gris claire)*

**Illustration n°.46 : Résultats de la reconnaissance supervisée sur le manuscrit *Arabe 2782* (R12051)**

En revanche pour les pages surexposées ou présentant des tâches et des zones d'ombres, le résultat de l'analyse est insuffisant. Le taux de reconnaissance dépend donc de la quantité d'image surexposée ou tâchée.



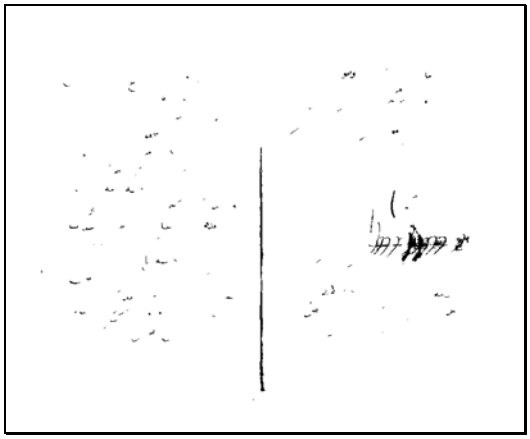


*Image originale surexposée* *Erreurs de reconnaissance du texte rouge*

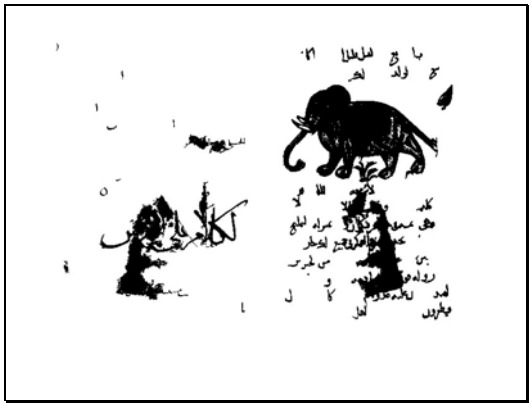


*Image originale tachée* *Erreurs de reconnaissance des illustrations*

**Illustration n°. 47 : Exemples des erreurs de reconnaissance**

### 3.3.2.6.3. Résultats sur *Arabe 2478* (R18271)

Les images du manuscrit R18271 sont d'assez bonne qualité bien qu'ayant été obtenu par la numérisation des microfilms. La résolution est suffisante pour séparer tous les objets sauf certaines illustrations du bord du cadre. Nous avons défini 3 métadonnées :

- Texte
- Illustrations
- Annotations

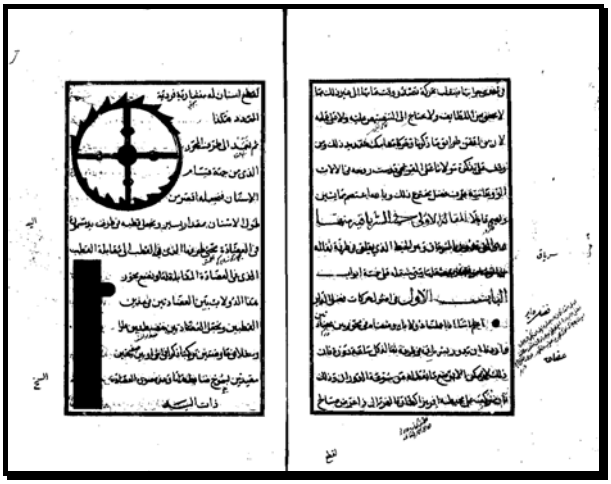
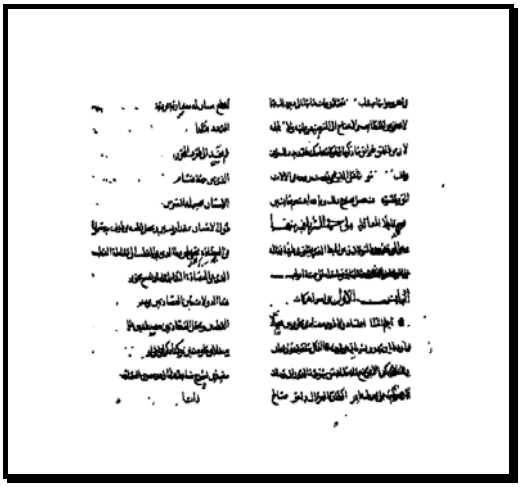
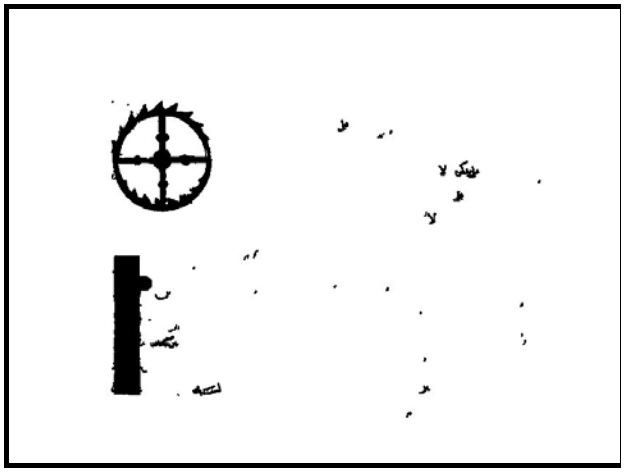


*Image Originale Noir et blanc*



*Texte seul*



*Illustrations*

*Annotations*

**Illustration n°. 48 : Résultats de la reconnaissance supervisiée sur Arabe 2478 (R18271)**

La présence du cadre explicite rend la reconnaissance des annotations certaine. Les seules erreurs observées sont toutes causées par une mauvaise séparation physique entre des objets de classe différente comme le texte connecté aux cadres ou aux illustrations.

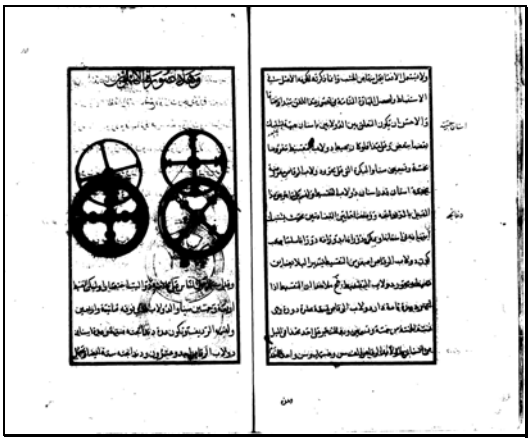
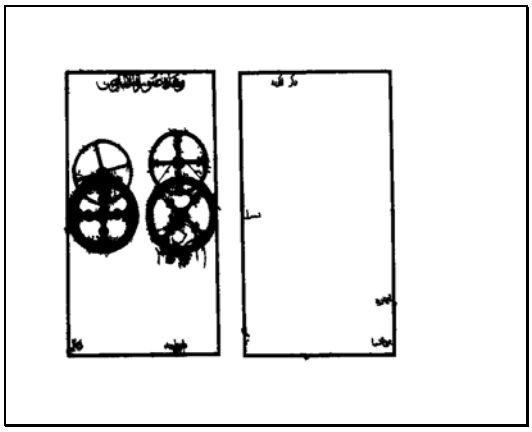


*Image originale*



*Erreur dans la reconnaissance des illustrations : connexion Texte/cadre/dessins*

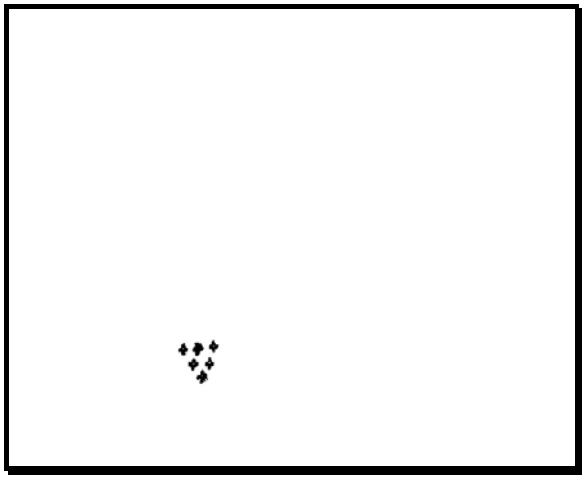
**Illustration n°.49 : Erreur dans la reconnaissance des illustrations : connexion Texte/cadre/dessins**

### 3.3.2.6.4. Résultats sur *Arabe 179* (R60914)

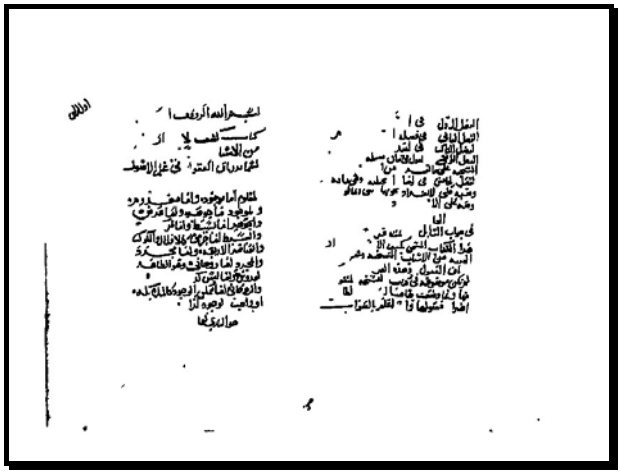
La qualité de la numérisation de cet ouvrage est suffisante en terme de résolution et de qualité d'images pour l'extraction automatique des métadonnées. Cependant cet ouvrage est assez pauvre en métadonnées et nous avons donc utilisé l'analyse d'images pour affiner les métadonnées comme la séparation des styles d'écritures et la présence de prolongement des mots. Nous avons donc défini cinq classes de métadonnées :

- Les décorations florales
- Le texte courant
- Les prolongements de textes
- Les autres styles de texte
- Texte de titre (reconnaissable par l'indentation)

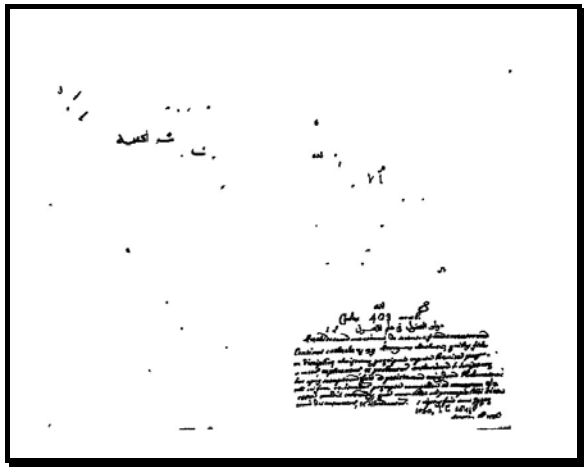
*Image originale*



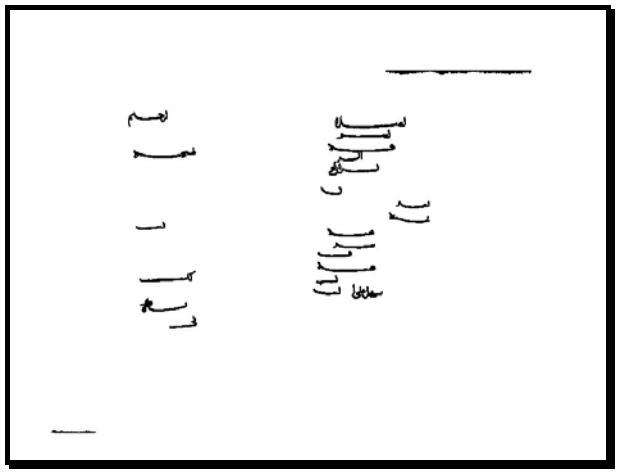
*Décorations florales*



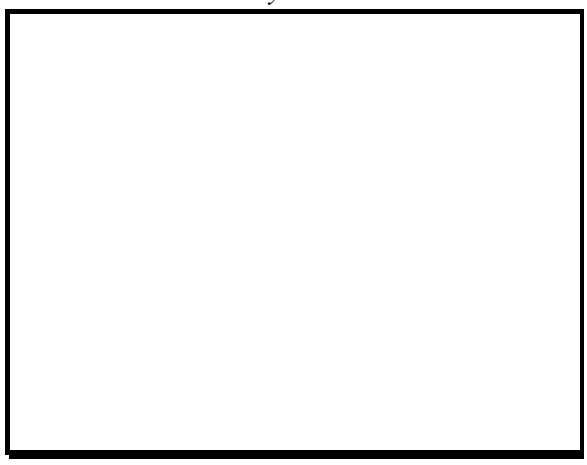
*Texte Courant*



*Autre style de texte*



*Prolongement de Texte*



*Texte de titre (erreur de reconnaissance)*

**Illustration n°. 50 : Résultats de la reconnaissance supervisée sur Arabe 179 (R60914)**

La reconnaissance des titres par l'indentation s'est avérée impossible car la mesure d'alignements verticaux avec les blocs voisins *av* n'est pas différente de celle du texte

normal. Pour trouver l'indentation du texte de titre, il faut détecter la justification à droite du texte ou du bord de la page et d'ajouter une mesure d'indentation par rapport à celle-ci. Le problème vient de la conception même du logiciel qui ne considère que les objets noirs, les zones vides blanches n'étant pas pris en compte. Ce sera l'objet de développements futurs.

#### 3.3.2.6.5. Résultats sur Arabe 1489 (R28062)

C'est l'ouvrage le plus riche en terme de métadonnées et qui aussi présente des mises en page d'une très grande complexité. La qualité des images est assez médiocre car elle est encore issue de la numérisation de microfilms. L'image binarisée présente des pertes d'informations et affiche des objets coupés ou collés ainsi qu'un grand nombre de taches. Nous rappelons qu'une image binaire ne peut pas être restaurée par traitement d'images car l'information perdue lors de la binarisation ne peut plus être retrouvée. La bordure des tableaux n'est pas utilisable car elle est représentée trop souvent par des lignes discontinues jusqu'à des pointillés. Le manque de temps sur le développement informatique du logiciel ne nous a pas permis d'extraire toutes les métadonnées que l'on souhaitait et notamment les textes en zigzag. Nous nous sommes limités à des métadonnées qui étaient susceptibles d'être reconnues avec une taux suffisamment élevé de reconnaissance pour exploiter les résultats en terme d'indexation :

- Texte
- Titres
- Tableaux
- Illustrations

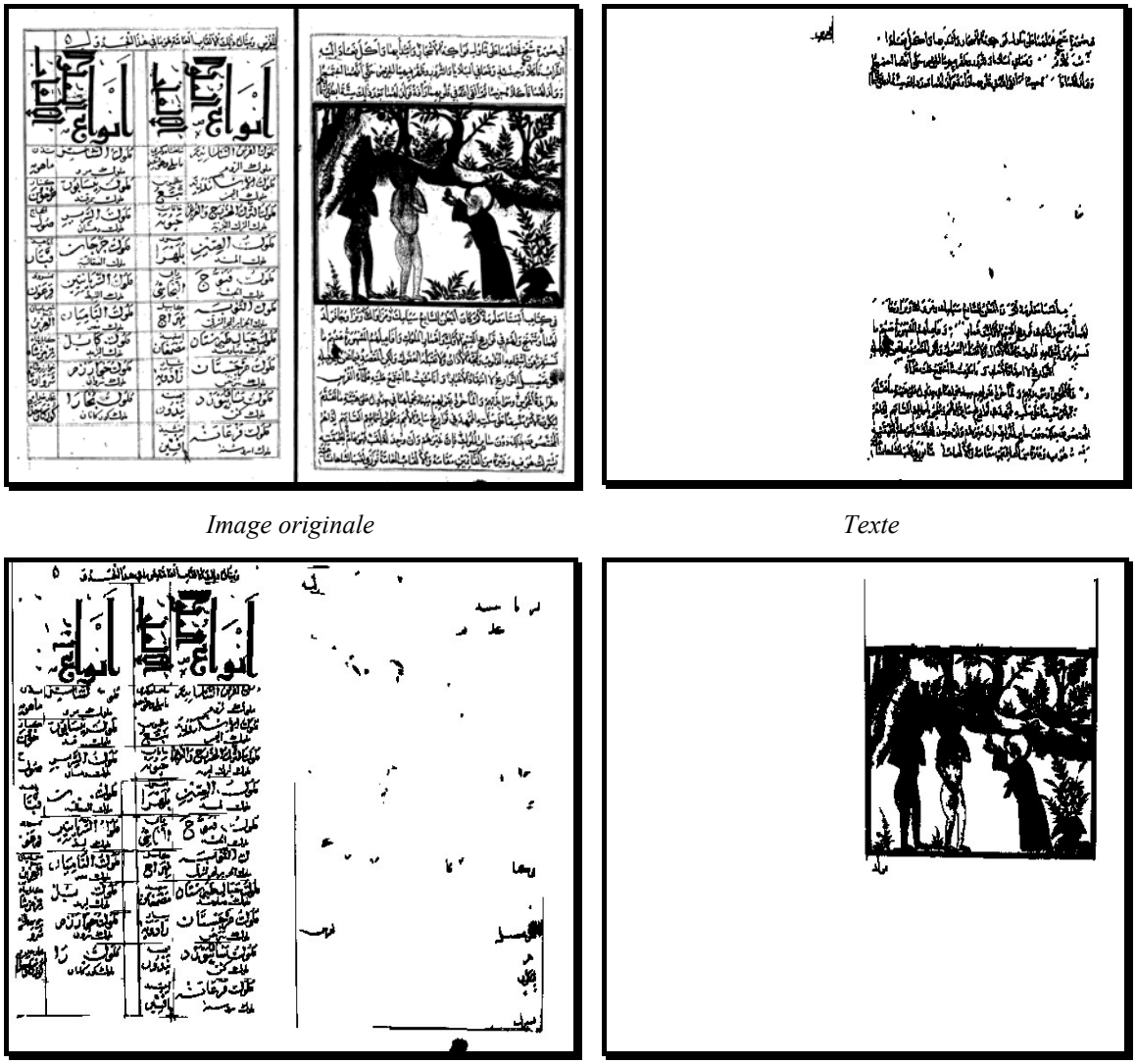


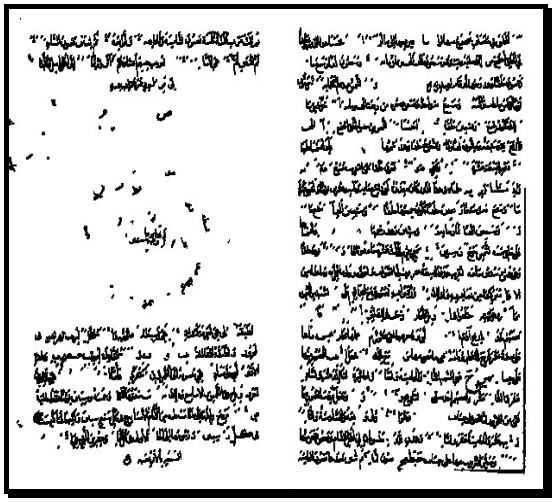
*Image originale* *Texte*

*Tableau* *Illustrations*

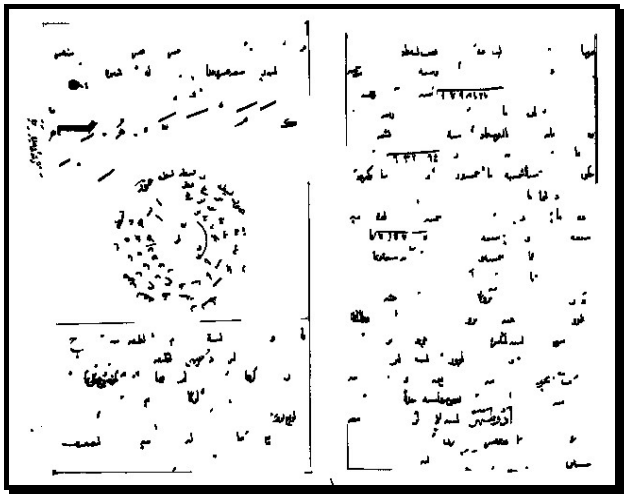
**Illustration n°.51 : Résultats de la reconnaissance supervisée sur Arabe 1489 (R28062)**

Cependant la complexité des métadonnées de ce manuscrit rend sa description très difficile. Par exemple certains chiffres surlignés apparaissent dans la classe des tableaux à cause de l'alignement et de l'équidistance entre les caractères. De même les diacritiques des titres et le texte dans certains graphiques de forme circulaire sont alignés et apparaissent également dans la classe des tableaux. Et la partie verticale du titre n'a pas été reconnue, car, d'une part, celui-ci touche le cadre et, d'autre part, cette orientation particulière n'a pas été vue lors de l'apprentissage.

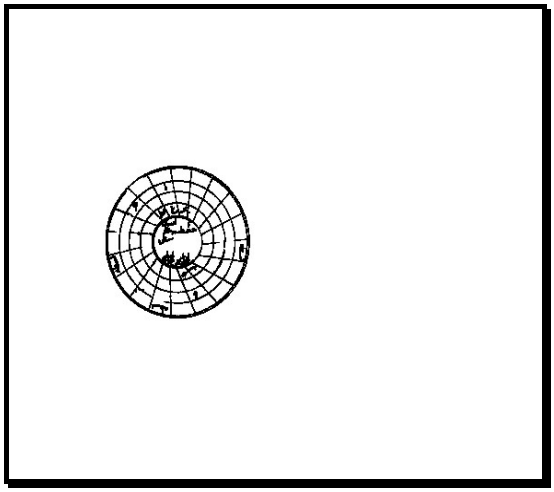
*Image originale*



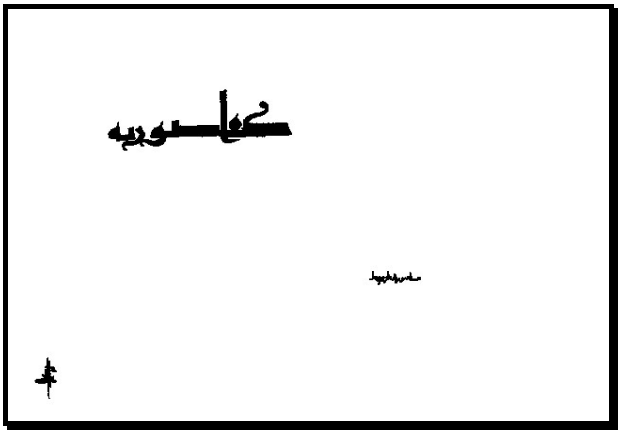
*Texte*



*Tableau*



*Illustrations*

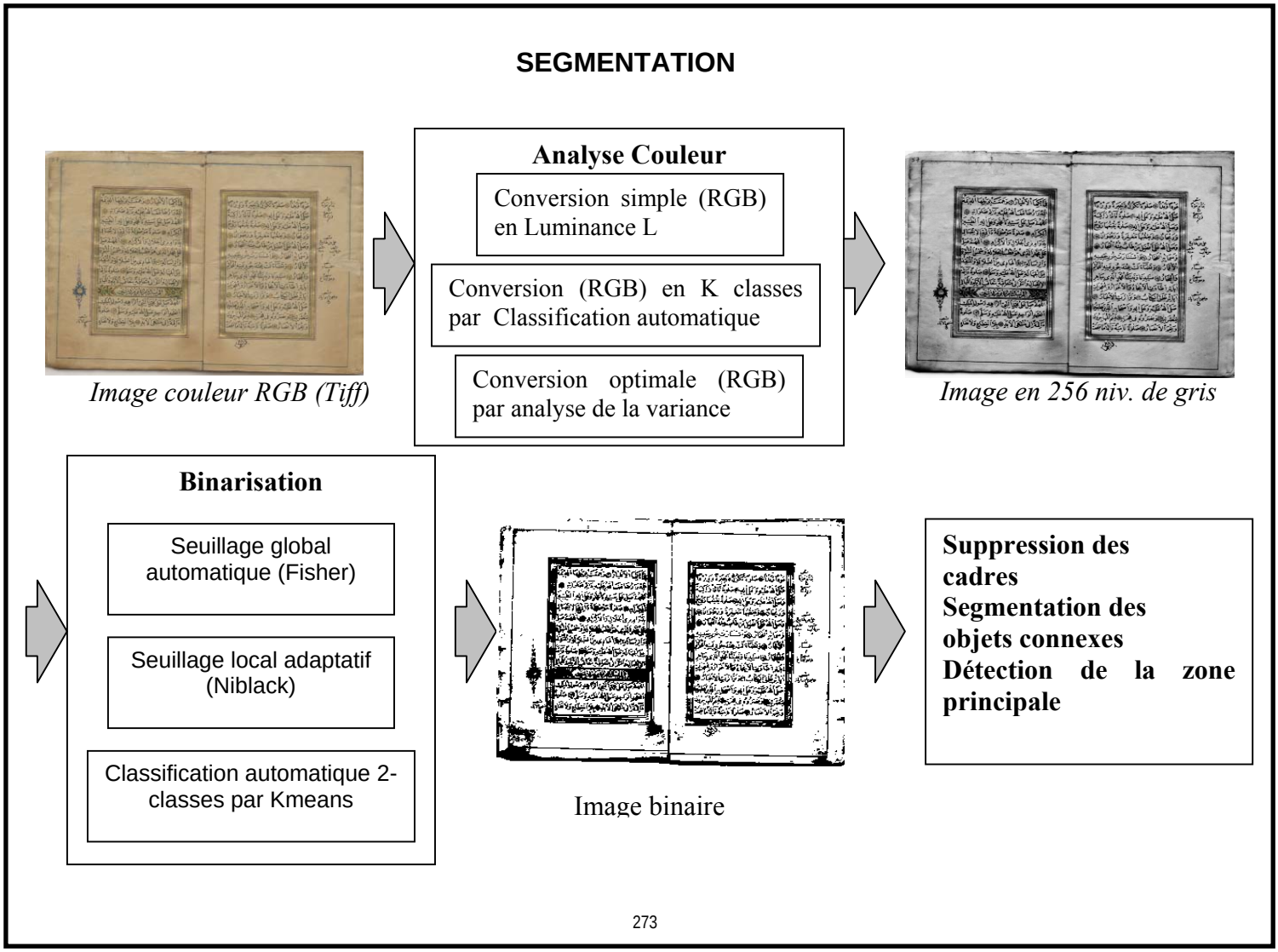


*Titres*

**Illustration n°. 52 Exemples des erreurs de la reconnaissance sur Arabe 1489 (R28062)**

### 3.3.2.7. Conclusion sur l'analyse d'images

La reconnaissance des métadonnées dans les images est très dépendante de la qualité de celle-ci et de leur richesse d'information en terme de résolution et du nombre de couleurs. Des taux très élevés de reconnaissance ont pu être mesurés sur des manuscrites couleurs malgré la faible résolution comme le manuscrit MS6191. En revanche, nous obtenons des taux très bas de reconnaissance sur des images dégradées issus de la numérisation de microfilms et qui présentent des tâches rendant impossible la séparation des objets (voir *Arabe 2953* (R3414)). La numérisation en niveaux de gris de microfilms apporte certes plus d'information mais si l'état du microfilm présente de défaut de régularités d'éclairage et des tâches sombres (voir *Arabe 2782* (R12051)), alors l'analyse d'image ne réussit pas à extraire correctement les métadonnées demandées. Enfin dans le cas où le manuscrit est numérisé en noir et blanc mais avec une résolution suffisante pour pouvoir séparer les objets et qui ne présentent pas de tâches, alors l'analyse d'image donne des résultats exploitables en terme d'indexation.



Objets localisés

**Extraction des caractéristiques sur les objets**

**Couleur**

*Chrominance (U,V)*
*Teinte (T)*
*Saturation (S)*
*Luminance (L)*

**Géométrie**

*Hauteur*
*Largeur*
*Longueur*
Epaisseur
Position

**Forme**

*Densité*
*Structure*
*Orientation*
*Courbure*

Reconnaissance

Classification automatique non supervisée (Kmeans)

Reconnaissance par classification supervisée (K-PPV)

Edition des métadonnées en XML

**Figure n°.59 Schéma du logiciel de reconnaissance des méta-données par *analyse d'images***