# مجموعات الكتابات اليدوية العربية على الويب:
# دراسة فى المعالجة والتنظيم والاسترجاع

**د. سيد ربيع سيد إبراهيم**
مدرس نظم استرجاع المعلومات
كلية الآداب - جامعة بنى سويف

**مستخلص**

تستهدف هذه الدراسة تناول مواد الكتابات اليدوية العربية على الويب كمواد معلومات قديمة حديثة، فهذه النصوص المخوطة بخط اليد تمثل وافدا جديدا بين مواد معلومات الويب، وتتطلع الويب وأدوات بحثها إلى توصيف معايير وآليات التعامل مع الكتابات اليدوية معالجة وتكشيفا واسترجاعا. وانطلقت الدراسة من أسس اختلاف الكتابات اليدوية كونها نصوص في شكل صور رقمية **OCR**، تختلف في طبيعتها بين النص منفردا والصورة منفردة. وقد قدمت الدراسة توصيفا للبنية الرقمية المقترحة لملفات مواد الكتابات اليدوية، وأساليب التكشيف المبنية على المحتوى والنص للتعامل مع التحليل الموضوعي لهذه المواد. وقد خرجت الدراسة بتوصيف مقترح لبنية وإجراءات الإدراك والمعالجة وأساليب البحث في أدوات بحث الويب التي تتوافق وطبيعة الكتابات اليدوية، كأول دراسة تتناول مواد الكتابات اليدوية في بيئة الويب.

**الكلمات المفتاحية :** الكتابات اليدوية، المخطوطات، تكشيف المخطوطات، إدراك النصوص اليدوية، القراء الآلية للحروف **OCR**، نظم استرجاع المعلومات، الويب.

**مقدمة الدراسة :**

**٠ / ١ تمهيد**

مثلت الويب منذ بدايتها مستودعا كونيا للمعلومات والمعرفة البشرية بكافة أشكالها وانماط حمل المعلومات المتنوعة، غير أن الويب كمستودع للمعلومات لم تكن منوطة بتنظيم المعلومات الرقمية وذلك لأن الهدف من الويب كان نشر المعلومات واتاحتها دون التطرق الى أساليب التنظيم وآليات المعالجة. ومن ثم ترتب على تنوع وتعدد الملفات

وبنية ووسائط حمل المعلومات الرقمية على الويب أن تحول التعامل مع المعلومات الرقمية من عمليات التنظيم الى آليات البحث والاسترجاع الملائمة للأنماط المختلفة التي تزداد تنوعا وتحديثا يوما بعد يوم في مستودع الويب. وكما شملت الويب النصوص والمواد المصورة المسموعة، أضيف اليها أنماطا مستحدثة من المعلومات الرقمية مثل الرسوم الحيوية animations والمواد ذات الأصل التقليدي كالمخطوطات والكتابات اليدوية.

ولقد شكل المحتوى العربي من المعلومات والمواقع على الويب مشكلة تقنية في عمليات البحث والاسترجاع لأدوات بحث الويب ؛ حيث تميزت اللغة العربية بخصائص لغوية عجزت أدوات البحث عن مجاراتها والتعامل معها مما انعكس بالسلب على استرجاع وتداول المعلومات العربية على الويب حتى كانت الدراسات التي تناولت خصائص البحث باللغة العربية للمحتوى العربي من المعلومات الرقمية. وما نتج عن ذلك تمثل في انتاج آليات بحث واسترجاع تتوافق وخصائص اللغة العربية معتمدة بالأساس على استخدام تقنيات التحليل الصرفي والمعاجم الآلية لمحركات البحث، وكان تتويج ذلك بتوافر قدرات البحث عن المعلومات العربية مثل البحث بالترادف والبحث بجذور الكلمات العربية والبحث بمعاني ومقابلات المفردات العربية، حتى خرجت الى فضاء الويب محركات بحث كاملة تدعم البحث باللغة العربية بشكل معمق.

وقد عمل تطور التقنيات والبرمجيات المستخدمة لتحويل المعلومات التقليدية الى الشكل الرقمي على إستقبال الويب لمجموعات من المعلومات الرقمية العربية التي قد سبقت الويب في الوقت من حيث نشاتها واتاحتها مثل المخطوطات والكتابات العربية التراثية التي تعود الى أزمان الترجمة والتعريب ثم ازدهار الانتاج الفكري العربي لقرون الاسلام الأولى. هذا فضلا عن توافر برمجيات الكتابة اليدوية مباشرة على الوسيط الرقمي في عصرنا الحديث كالكتابة على الحاسبات اللوحية والهواتف النقالة التي تمثل أدوات نشر للمعلومات الرقمية على الويب. ولقد مثل ذلك كله ظاهرة انتشار الكتابات اليدوية العربية على الويب سواء تلك التراثية أو الحالية التي تستدعي أدوات بحث الويب وأدوات استرجاعها الى الاتجاه نحو آليات ومعايير للبحث والمعالجة تلائم تنظيم وبحث واسترجاع الكتابات اليدوية العربية.

**٠ / ٢ مشكلة البحث**

تتمثل مشكلة البحث في تزايد مجموعات المعلومات من الكتابات اليدوية العربية على الويب سواء تلك التي تم تحويلها الى الشكل الرقمي بهدف الاتاحة والوصول والاستخدام الأكثر انتشار من خلال شبكة الويب مع الاحتفاظ بالنسخ الأصلية المتحفية التي تمثل قيمة تاريخية، او تلك المعلومات الرقمية المكتوبة يدويا وتم نشرها واتاحتها على الويب بشكلها المخطوط يدويا. وانطلاقا من وظيفة الويب الأساسية في توفير أدوات وآليات بحث لمختلف أشكال المعلومات التي تحملها، فإن مجموعات المعلومات العربية الرقمية يدوية الكتابة لا تتوافر لها أدوات وآليات البحث التي تتوافق وطبيعة نصوصها وكلماتها وحروفها التي تختلف وتتنوع بشكل يصعب على أدوات بحث الويب بحثها واسترجاعها بآلياتها الحالية. وتتعدد أطراف المشكلة في ايجاد انساق محددة لفهم تراكيب وبنية النصوص اليدوية العربية، أيضا آليات التعامل مع الحروف الرقمية يدوية الكتابة هل من خلال كونها حروفا ذات دلالة أم بكونها صورا بمضاهاتها؟ فضلا عن توفير أساليب للبحث تستطيع الوصول الى الاستدعاء والتحقيق اللازمين في بحث مواد المعلومات يدوية الكتابة. ومن ثم فإن البحث يقوم على فرضية رئيسة هي (إن بيئة وأدوات بحث الويب لا تتوافق آلياتها وتقنياتها وأداؤها مع مواد الكتابات اليدوية العربية الرقمية كأحدث أشكال المعلومات المتاحة على الويب)

**٠ / ٣ تساؤلات الدراسة**

يطرح البحث تساؤلات هي :-

١. ما التقنيات اللازمة للتعامل مع معالجة وتكشيف الكتابات العربية اليدوية على الويب؟
٢. ما آليات ومعايير البحث اللازمة لأدوات بحث الويب لاسترجاع الكتابات العربية اليدوية؟
٣. ما خصائص قواعد البيانات ومستودعات الكتابات العربية اليدوية على الويب؟
٤. ما القدرات التقنية والبرمجية المطلوبة للتعامل مع مفردات النصوص العربية اليدوية على الويب؟

**٠ / ٤ أهداف الدراسة**

تهدف هذه الدراسة الى تحقيق ما يلي :-

١) فهم آليات وتقنيات تناول الكتابات اليدوية على الويب كمواد معلومات رقمية.

٢) تحديد معايير وآليات البحث المطلوبة لدى أدوات البحث للتعامل مع النصوص العربية اليدوية.
٣) رسم صورة مبدئية عن بنية وتكوين مستودعات مجموعات المعلومات النصية اليدوية على الويب.
٤) وضع رؤى مقترحة حول معالجة وتنظيم وتشكيف مجموعات النصوص اليدوية الرقمية.

**٠ / ٥ أهمية الدراسة**

تكتسب هذه الدراسة اهميتها من المنطلقات التالية :-

١. ان النجاح في فتح أفق المعالجة والبحث امام النصوص العربية اليدوية يمثل بداية انطلاق نحو فتح واستكشاف كنوز المخطوطات والكتب التراثية العربية عدد المخطوطات فقط فيها يصل الى ثلاثة ملايين مخطوطة تقريبا ؛ حيث ستتحول هذه النصوص الى مجموعات من المعلومات الرقمية قابلة للبحث والتكشيف على مستوى النصوص والكلمات والبحث في آفاقها الموضوعية الموسوعية.
٢. ستكتسب الويب نوعا جديدا من أدوات البحث قادرا على التعامل مع بحث واسترجاع النصوص اليدوية الرقمية، مما يجعل انتشار الكتابات اليدوية أسرع بقدر السرعة التي يكون عليها بحث واسترجاع النصوص المكتوبة يدويا.
٣. يمكن بناء نظم استرجاع متخصصة للنصوص اليدوية الرقمية بحيث تتحول مستودعات المخطوطات التراثية ومشروعات رقمنة التراث العربي الى مكتبات رقمية متخصصة في المخطوطات والكتابات العربية اليدوية.

**٠ / ٦ منهج الدراسة**

تتبنى الدراسة استخدام منهج البحث **الوصفي التحليلي** وذلك لتتبع مجموعات الكتابات اليدوية العربية على الويب وتحري آليات المعالجة الحالية التي تتبناها أدوات بحث الويب في التعامل مع هذه المعلومات، واستقراء الانتاج الفكري والدراسات المرتبطة بهذه الظاهرة. بالاضافة الى استخدام منهج البحث **التجريبي** وذلك لملاحظة وتتبع مواقع الويب المختصة ببحث واسترجاع المعلومات النصية اليدوية.

أما **أدوات البحث** فتتمثل في أداة البحث الوثائقي في جمع الانتاج الفكري المرتبط بظاهرة البحث، واستخدام أداة الابحار والمعايشة مع الويب الى جانب استخدام أداة الملاحظة في الجوانب التطبيقية للبحث.

أما **عينة الدراسة** فقد عمل الباحث على اختيار عينة غرضية من مواقع بحث المخطوطات العربية والأجنبية لاستقراء واقع تنظيم واسترجاع مواد الكتابات اليدوية في بيئة الويب، وقد تمثلت العينة في الجدول رقم (١).

**جدول رقم (١) عينة الدراسة المستخدمة**

| قواعد البيانات العربية | إدارة النظام | الدولة |
|---|---|---|
| http://wamcp.bibalex.org/ | **مكتبة الاسكندرية** | **مصر** |
| http://www.manuscripts.idsc.gov.eg/Manuscript/About.aspx | **مجلس الوزراء** | **مصر** |
| http://www.almajidcenter.org/ar/index.php | **مركز جمعه الماجد** | **الإمارات** |
| قواعد البيانات الأجنبية | | |
| http://www.fihrist.org.uk | **المكتبة البريطانية** | **انجلترا** |
| http://www.westafricanmanuscripts.org/default.php | **جامعة الينوي** | **موريتانيا** |
| http://ddc.aub.edu.lb/projects/jafet/manuscripts/ | **الجامعة الأمريكية** | **بيروت** |

٠ / ٧ **حدود الدراسة :** تتحدد الدراسة بالحدود التالية :-

- **الحدود الموضوعية** : تتناول الدراسة موضوع المعالجة والبحث لمواد المعلومات العربية المكتوبة يدويا من وثائق ومخطوطات وغيرها من أشكال الكتابات اليدوية العربية.
- **الحدود النوعية** : تركز الدراسة على مواد الكتابات العربية وتعامل أدوات بحث الويب من حيث توفير آليات ومعايير البحث الملائمة.
- **الحدود المكانية** : ترتبط حدود الدراسة بمجموعات الكتابات العربية اليدوية على الويب في البيئة الرقمية، سواء تلك التي تنتمي للعالم العربي أو التابعة لمؤسسات دولية.
- **الحدود اللغوية** : تركز الدراسة على موضوع الكتابات العربية اليدوية دون اللغات الأخرى، واستخدام أدبيات الانتاج الفكري باللغتين العربية والانجليزية التي تتناول الموضوع بالدراسة.

٠ / ٨ **الدراسات السابقة**

اتفقت أدبيات الانتاج الفكري العربي والأجنبي إجمالا على عدم تناول مواد الكتابات اليدوية على الويب من حيث التنظيم والاسترجاع كونها مواد معلوماتية لها طبيعة تنظيم

وتحليل واسترجاع مميزة عن غيرها. وقد اختلف كل من الانتاج الفكري العربي والأجنبي في زاوية تناول المخطوطات والكتب التراثية الرقمية ؛ حيث ركز الانتاج الفكري باللغة الأجنبية الذي جاء كاملا من علوم الحاسب على النواحي التقنية وعمليات الإدراك والتحليل لصور OCR حتى أنه استخدم مصطلح التكشيف Indexing للتعبير عن ناتج عملية الإدراك recognition، وقد توقفت الدراسات باللغة العربية عند تناول رقمنة المخطوطات وعمليات الفهرسة والضبط الببليوجرافي، وقد تناول البعض منها تحليلا لمواقع المخطوطات من حيث التصميم والتقنيات المستخدمة. أما دراستنا الحالية فتعد الأولى في تناول مواد الكتابات اليدوية العربية من مخطوطات ووثائق تاريخية وكتب تراثية كمواد معلومات رقمية في بيئة الويب كتبت بخط اليد ويتم تنظيمها ومعالجتها وتكشيفها واسترجاعها باستخدام أدوات ومحركات بحث الويب. وقد استخدم الباحث قاعدة بيانات الرسائل الجامعية UMI proquest التي أفرزت الدراسات التالية :-

**الدراسة الأولى :**

Parves, Mohamed Tanvir. Arabic Handwritten text Recognition using structure and syntactic pattern attributes, King Fahd university, phd, 2010. http://search.proquest.com/pqdtglobal/docview/876013772/fulltextPDF?source=fedsrch&accountid=37552

تناولت هذه الدراسة تقنيات وبرمجيات تحليل بنية صور نصوص الكتابات اليدوية الرقمية من حيث تحليل بنية الكلمات والحروف المصورة للتعرف عليها وإدراكها، اعتمدت الدراسة في محورها على القراءة الآلية بنماذج ماركوف غير المرئية التي تستطيع تحليل صور وأبعاد الكلمات والحروف. ولم تخرج الدراسة عن إطارها التقني دون التطرق الى عمليات البحث والاسترجاع على محركات البحث.

**الدراسة الثانية :**

Cao, Huaigu. Indexing and retrieval of low quality handwritten documents, University of New York at Buffalo, PHD, UMI ProQuest LLC, 2008, cited at http://search.proquest.com/pqdtglobal/docview/305521651/fulltextPDF?accountid=37552

تناولت الدراسة خصائص وطبيعة الصور الرقمية لمجموعات المخطوطات المعتمدة على تقنية **OCR**، وقد تناولت هذه الدراسة مصطلحات الاسترجاع والتكشيف غير أن تناول

هذه المصطلحات لم يكن وفقا لمفاهيم نظم المعلومات، وإنما بشكل برمجي وتقني جعلها تنعت عملية الإدراك للكلمات والحروف بمصطلح التكشيف، ومن ثم أتى مصطلح التكشيف **Indexing** والإدراك **Recognition** في شكل المترادفات.

**الدراسات العربية :**

أما الدراسات العربية مثل ( مولاي امحمد. المحتوى الرقمي العربي المخطوط على شبكة الانترنت : دراسة تقييمية ) فقد ركزت هذه الدراسة على استخدام تقنية **temesis** لتحليل مواقع المخطوطات العربية والأجنبية التي تضم مجموعات من المخطوطات المرقمنة، وذلك لقياس مجموعة من خصائص مواقع الانترنت كالرؤية والتصميم والتقنية والخدمات والمحتوى المقدم في هذا الموقع. وقد اهتمت الدراسة بالجوانب التقييمية التحليلية لمواقع المخطوطات عن دراسة المعالجة والتنظيم لمواد المخطوطات ذاتها. ودراسة أخرى مثل (حافظي زهير، أ. مزلاح رشيد. فهرسة ورقمنة مخطوطات مكتبة جامعة الأمير عبد القادر للعلوم الإسلامية ووضعها ضمن شبكة الإنترنت) حيث ركزت هذه الدراسة على عمليات الرقمنة والضبط الفني للمخطوطات المرقمنة وبناء مجموعاتها في البيئة الرقمية، دون التطرق الى تحليلها أو الحديث عن اساليب البحث والتكشيف.

**٠ / ٩ مصطلحات اجرائية**

- **الكتابات اليدوية** : تمثل الكتابات اليدوية مجموعة مصادر المعلومات المكتوبة بخط اليد سواء تلك الكتب التراثية والمخطوطات أو مجموعات النشرات والمعلومات المكتوبة بخط اليد مباشرة باستخدم الحاسبات اللوحية الحديثة.
- **معالجة الكتابات اليدوية** : عمليات المعالجة للكتابات اليدوية الرقمية هي توفير قدرات المسح الضوئي والتعرف على أشكال الكلمات والحروف وهي الممهدة عمليات التكشيف والتحليل الموضوعي لنصوص الكتابات اليدوية.
- **إدراك الكتابات العربية : Handwritten Recognition** هي عملية تحليل وقراءة آلية لصور الكلمات والحروف ومقاطع الكلمات الواردة في نصوص الكتابات اليدوية وفقا لتقنية **OCR**.

## المبحث الأول : الكتابات اليدوية العربية على الويب

شكل المحتوى العربي بصفة عامة والنصوص الرقمية العربية بشكل خاص صعوبة كبيرة أمام أدوات بحث الويب من حيث عمليات التنظيم والمعالجة والبحث والاسترجاع ؛ حيث ترتبط بمواد المعلومات النصية العربية العديد من التحديات المتمثلة في الخصائص

الصرفية والتراكيب في اللغة العربية، الى جانب طرائق الكتابة والأشكال المتعددة من الصياغات الفصحى والعامية، فضلا عن إشكاليات الدلالة والترادف والمشتركات اللفظية واختلاف المعنى للكلمة الواحدة. وقد انتقلت هذه المشكلات الى أدوات بحث الويب التي تعاملت معها عبر العديد من التقنيات والبرمجيات منتجة أشكال البحث باللغة الطبيعية وغيرها من آليات البحث، لكن ما شكل الصعوبة الأكبر هو انتقال تلك المشكلات الى النصوص العربية المخطوطة بالأيدي خلاف تلك التي تم ادخالها عن طريق لوحات الكتابة على الحاسب الآلي. فرقمنة الكلمات العربية منذ بدايتها تجعل نظم استرجاع المعلومات أكثر ادراكا للحروف وتراكيب الكلمات العربية أما انتقال النصوص العربية المكتوبة يدويا الى الشكل الرقمي يجعل ادراك معاني الكلمات غير ممكن في ظل انتقالها على هيئة صور للكلمات وليس حروف رقمية الأصل.

تختص النصوص العربية يديوية الكتابة بالعديد من الخصائص التي تمثل معها عملية بحثها واسترجاعها في بيئة الويب تحديا كبيرا ؛ فالكتابات العربية المخطوطة تتكون من ثمانية وعشرين حرفا، تكتب بأشكال مختلفة، خلاف عمليات التشكيل والاعجام التي تتميز بها العديد من النصوص، هذا فضلا عن مجموعات الوثائق والكتابات والمخطوطات والكتب التراثية التي لم تشهد حروفها وضع النقاط ؛ حيث لم توضع النقاط على الحروف العربية الا في وقت لاحق لعمليات الكتابة والنسخ العربية تحددت تقريبيا بالقرن السابع الميلادي. كما تختص نصوص اللغة العربية بعلامات مساعدة كالشدة وعلامات الضبط النحوي وغيرها بما يعود على نظام القراءة الآلي بمزيد من صعوبات الادراك لمواد المعلومات. [١]

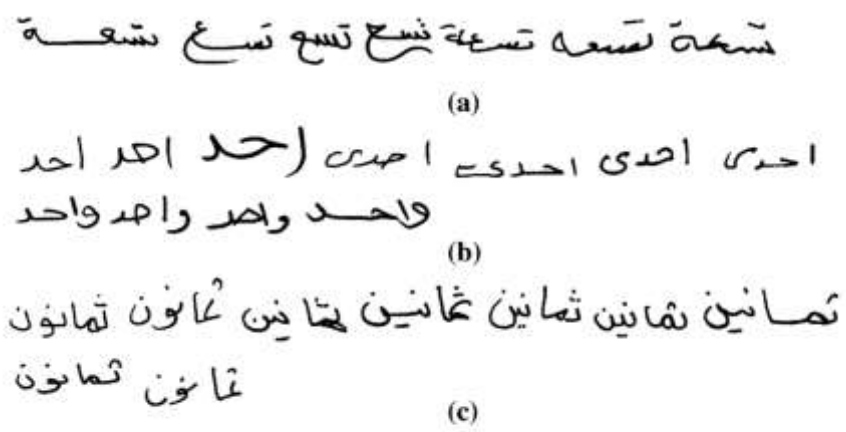


Fig. 1. Examples of Arabic words written in different fonts and styles: (a) nine (b) one or part eleven, and (c) eighty.

**شكل رقم (١) يوضح الاشكال المتعددة لكتابة الكلمات العربية**[٢]

وقد أورد **zaher al aghbari and salama brook** بعضا من خصائص اللغة العربية في الكتابة عامة والكتابات اليدوية خاصة التي تمثل تحديا أمام نظم إدراك وتحليل نصوص الكتابات اليدوية على الويب، تتمثل في :-[٣]

١) تتكون اللغة العربية من ثمان وعشرين حرفا، تتعد أشكال كتابة الحرف الواحد بين شكلين الى أربعة أشكال، فضلا عن الحروف المتشابهة رسما والمختلفة في النقاط التابعة لها.

٢) تكتب كلمات اللغة العربية بحروف متشابكة من اليمين الى اليسار، مع وجوب اتباع الفواصل بين الكلمات والحروف.

٣) تتكون أكثر كلمات اللغة العربية من مقطعين أو أكثر تمثل صعوبة أمام الحاسب أثناء التعرف على الكلمات والمفردات في عملية الإدراك.

٤) تكتب حروف اللغة العربية بالكتابة في رسم يميل الى أسفل أو يرتفع إلى أعلى السطر مع تميز بعض الحروف بالتفاف في الخط أثناء كتاباتها ( ط، ض).

٥) تتبع الحروف في اللغة العربية توابع ثانوية مثل النقاط التي تأتي أسفل أو أعلى الحروف، والهمزة التي تأتي مع بعضها الآخر (ك، ؤ).

**١ / ١ أشكال الكتابات اليدوية الرقمية**

تتنوع الكتابات الرقمية ذات الشكل اليدوي على الويب ؛ حيث تستمد الكتابات اليدوية هذا التنوع اعتمادا على أصل النشأة لهذه المواد، فمنها ما هو تقليدي الأصل مثل الكتب التراثية والمخطوطات والوثائق التاريخية التي تدخل الى بيئة الرقمنة اعتمادا على أجهزة النسخ الضوئي **optical character recogntion** المعروفة باسم **OCR** ؛ حيث يتم من خلالها استيعاب النصوص والمتون التقليدية أو المطبوعة في البيئة الرقمية وإدراك ما تحمله هذه المواد من كلمات وحروف. ومن مواد المعلومات أيضا ما يعود الى أصل رقمي النشأة وهو ما يتجلى في استخدام الحاسبات اللوحية **Tablet PC** وأجهزة الهواتف النقالة ذات أنظمة التشغيل الذكية مثل **Andorid** المنتجة بواسطة شركة **google**.[٤] وتعمل هذه التقنيات الحديثة على دعم حروف اللغة العربية منذ ادخالها وامكانية التعامل مع الكلمات المكتوبة فور ادخالها بالمعالجة والبحث والاسترجاع. ومن ثم فإن النصوص العربية اليدوية في البيئة الرقمية تتنوع بين شكلين أساسيين هما :-

**١ / ١ / ١ الكلمات والحروف المصورة**

إن الكلمة المصورة هي الوحدة الاصغر في قواعد بيانات النصوص والكتابات العربية اليدوية الرقمية ؛ حيث تعمل تقنيات المسح الضوئي على التقاط صور شاملة لصفحات ووثائق الكتابات العربية ومن ثم يبدأ نظام قاعدة البيانات على استقراء الكلمات والحروف باستخدام تقنيات معالجة الصور الرقمية. وتتحول حروف اللغة العربية في هذه الحالة الى مقاطع مصورة لصورة أكبر هي الكلمة المكتوبة التي بدورها تكون مقطعا في صورة أكبر للوثيقة التاريخية أو صفحة المخطوط العربي. ويتعامل نظام إدارة المجموعات النصية هنا بآليات القراءة والاسترجاع للصور الرقمية الثابتة ؛ فتباعد الحروف وارتباطها ومقاطع الكلمة الواحدة والمسافات الفاصلة بينها وبين الكلمات الأخرى لا تعدو كونها مسافات رقمية تقاس بنقاط كثافة الصور الرقمية (البيكسل) **pixels**. وتنبني على هذه الآلية العديد من أساليب الحسابات الرقمية التي تتحكم في التعامل مع الكلمات الرقمية المصورة ؛ فيمكن للنظام على سبيل المثال تحديد قياس ثابت وشكل تصويري محدد لحرف الألف ( أ ) الذي يعبر في أغلب حالات اللغة العربية على بداية كلمة جديدة، بما يعني حساب المسافات والقياسات البادئة والمنتهية لهذه الكلمات ثم تحديد نقاط الكثافة للكلمة وتحديد الشكل أو الأشكال التي وردت عليها، ومن ثم اشتقاقها واختزانها في ملف الكلمات المشتقة من النص لتجهيزها لعمليات المضاهاة والبحث فيما بعد. وقد نفذ فريق من الباحثين هذه الآلية في تصميم نظام للتعامل مع الوثائق العربية المكتوبة يدويا سمي

بـــ **CEDARABIC system** اعتمد في بناء قاعدة بياناته على عشرة وثائق يدوية الكتابة لعشرة كتاب مختلفين حتى يتم التعامل مع الخطوط المتباينة للكلمات العربية وتكونت كل وثيقة من كلمات تراوح عددها بين ١٥٠ الى ٢٠٠ كلمة بما يعادل اجمالا ٢٠٠٠ كلمة يمكن التعامل معها في قاعدة بيانات النظام.[٥]

**١ / ١ / ٢ الكلمات والحروف النصية**

يكمن الفارق في العمل بين الكلمات والحروف المصورة وهذه النصية في اعتماد أدوات المعالجة في حالة الكلمات والحروف النصية المكتوبة يدويا على قاعدة بيانات مسبقة لمجموعات الأشكال المختلفة من حروف اللغة العربية، التي تستطيع التقارب مع الكلمات حالية الادخال بواسطة يد المستخدم ومن ثم إدراك الكلمات والحروف في الوقت الحقيقي لعملية الكتابة **real time recognition**. ويمكن القول بشكل آخر أن عملية المضاهاة وانتاج وقواعد بيانات الكلمات والحروف المكتوبة يدويا في حالة النصوص اليدوية ذات الاصل المطبوع او التقليدي تكون لاحقة لعملية القراءة والادراك، أما في حالة الكلمات النصية فإنها تكون سابقة لعملية الادراك ولعملية الكتابة والنشأة بشكل رئيسي. وعلى ذلك فإن قواعد بيانات نصوص الكتابات اليدوية ذات الأصل الرقمي تكون أكثر جاهزية وفاعلية في التعامل مع أدوات بحث واسترجاع الكتابات اليدوية عنها في النصوص المصورة، ذلك لما لها من إدراك وتعريف سابقين، حيث يمكن تمييز وتحديد الكلمات والحروف موضع البحث والاسترجاع.[٦]

**١ / ٢ رقمنة الكتابات اليدوية ( OCR )**

تمثل تقنية القراءة الضوئية للحروف **optical character recognition (OCR)** الباب الرئيس للتعامل مع مواد الكتابات اليدوية ليس من خلال الويب وفقط، وإنما من خلال البيئة الرقمية ككل. فتحويل النص الى قدرة القراءة الآلية وإدراك الحروف والكلمات ينبني عليه كل ما يلي من إجراءات معالجة للكتابات اليدوية. وتعرف تقنية **OCR** على أنها عملية نقل الحروف والنصوص التقليدية الى بيئة الحاسب الآلي من خلال أدوات تقنية كالكاميرات الرقمية وأجهزة الماسح الضوئي **scanners** لرقمنة هذه النصوص والتعامل معها حفظا وتنظيما وبحثا ككونها صورة رقمية. وتقوم تقنية **OCR** على آلية تصوير النصوص التقليدية وتحويلها الى صور رقمية تقرأ داخل الحاسب الآلي بأكواد مخصصة كما في شفرات **ASCII text**. وتواجه تقنية المسح الضوئي نوعين مختلفين من أنواع الكلمات والخطوط التي تتعامل معها، فالكتابات التقليدية المطبوعة هي أكثر يسرا في

الادراك والمعالجة عن تلك الكتابات اليدوية ؛ حيث تختلف مواد الكتابات اليدوية في أنواع الخطوط من كاتب إلى آخر، فضلا عن الكتابات ذات الحروف المتصلة، التي تمثل تعقيدا أكبر أمام التعرف على كل حرف بشكل منفصل. وحري بالذكر أن معالجة **OCR** للنصوص التقليدية لا يتوقف عند كونها أداة رقمنة وحسب، بل تمتد الى مستويات القراءة والإدراك وفهم النصوص المصورة رقميا، فهي بذلك تتعدى حاجز الرقمنة الى المعالجة والبحث للكتابات اليدوية.[٧]

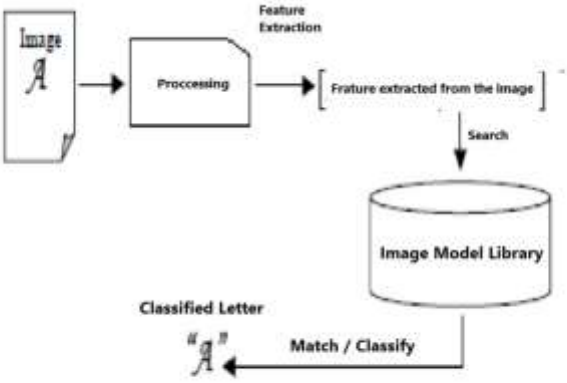
تمتلك نظم **OCR** مجموعة من قدرات التشغيل التي توفر سرعة إدراك وقراءة النصوص المصورة ؛ حيث يمكن لهذه النظم إدراك أنواع الخطوط وخاصة في الكتابات اليدوية، أيضا حجم كتابة الكلمات وتنسيقات الفقرات المكونة للنصوص، بالاضافة الى الجداول والرسوم التوضيحية والصور المرفقة ضمن النصوص. وتعمل هذه التقنية بمعدل قراءة وإدراك صفحة مطبوعة من حجم **A4** في مدة دقيقة واحدة، ويمكن لهذه التقنية ان تعمل على أنواع مخصصة من الكتابات والنصوص، بل ويمكن لهذه التقنية إدراك الأنواع المخصصة من الكلمات والأرقام كتلك التي توجد في طوابع البريد واللقطات المصغرة. وتعتمد تقنية **OCR** على آلية انتخاب للكلمة او المصطلح المقروء آليا حتى يمكن تحديد الحرف أو الرقم بشكل مؤكد. وقد ذكرا **Sebastian Stoliński, Wojciech Bieniecki** أن رقمنة وإدراك النصوص المطبوعة تتم في أربعة خطوات هي :-[٨]

**الخطوة الأولى** : معالجة صفحات الكتابات اليدوية عن طريق إزالة التحشيات والكتابات الزائدة وضبط خصائص السطوع والتعارض في ألوان الصورة المنتجة.

**الخطوة الثانية** : هي إدراك الصفحة من خلال عنصرين أولهما : تحليل الصفحة والتعرف على عناصرها الداخلية مثل الارقام والحروف والجدوال، ثانيهما : إدراك الحروف والتعرف عليها من خلال تحديد حروف وكلمات فرضية للمقطع أو الحرف المصور.

**الخطوة الثالثة** : ضبط علاقات الحروف والمقاطع ببعضها والربط بين الحروف والنقاط وعلامات الاعجام وغيرها.

**الخطوة الرابعة** : البحث داخل الحروف المعرفة ومحاولة استخدامها لإدراك الحروف التالفة.




II. Phases of General Character Recognition System

**شكل رقم (٢) يوضح الخطوات المتبعة لرقمنة وإدراك الكتابات اليدوية وفقا لتقنية OCR** [٩]

يكمن الهدف الرئيس لتقنية **OCR** في مرحلة ما بعد المعالجة في التأكيد على سلامة وصحة الحروف والكلمات التي تم إدراكها والتعرف عليها، وذلك من خلال قوائم الحروف والكلمات المسجلة مسبقا في قاعدة البيانات تبعا للغة التي يتم معالجتها. وتستند قواعد بيانات **OCR** هنا على استخدام الاساليب الاحصائية في التعرف على الكلمات والمقاطع بجانب المعالجم اللغوية ومحللات اللغة، ويمثل نموذج **hidden Markov model (HMM)** اهم آليات العمل في تقنيات إدراك وتحليل الحروف الضوئية. وعلى ذلك فإن التكامل بين نظام إدارة قواعد بيانات الكتابات اليدوية **DBMS** وبين نظام معالجة الكلمات الضوئية أمر ضروري، وذلك حتى تعمل قاعدة البيانات بالكفاءة المطلوبة في عمليات البحث والاسترجاع للكتابات اليدوية وفقا لحاجة المستفيد الموضوعية. ويرى الباحث أن ثمة مبادئ لا بد أن تتوافر في نظم معالجة **OCR** تتمثل في أولا المرونة التي يجب أن يتحلى بها النظام في التعامل مع الكتابات اليدوية وخاصة العربية منها، ثانيا الدقة في عمليات التحليل والادراك الاحصائية التي تنبني بالأساس على استخدام الصيغ الاحتمالية في التحديد النهائي للحروف والكلمات. ثالثا التكامل في الاستعانة بأدوات لغوية تتميز بالتعمق والدقة في عرض مفردات اللغة.[١٠]

**١ / ٣ البنية الرقمية للكتابات اليدوية (الكلمة المصورة)**

تتعامل الويب وأدوات بحثها مع مواد ووسائط حمل المعلومات معالجة وتنظيما واسترجاعا وفقا لطبيعة البنية والتكوين اللذين تتواجد بهما المعلومات الرقمية. ويسري هذا المبدأ على مواد الكتابات اليدوية الرقمية، فبحسب بنية الكتابات اليدوية تبدأ أدوات بحث الويب في تجهيز وتصميم البرامج وآليات المعالجة والاسترجاع المتوافقة وبنية الكتابات اليدوية الرقمية، وخاصة العربية منها لما لها من خصائص لغوية مميزة. وبالنظر لواقع البنية والبنيان الرقمي للكتابات اليدوية على الويب، فإنها تجمع في صفاتها الوراثية المعلوماتية بين وسيطي النص والصورة الرقميين، فالنصوص الرقمية تتعامل معها أدوات البحث معالجة واسترجاعا وفقا لبنية صفحات ومحارف **HTML**، أما وسيط الصورة الرقمية فإن أدوات بحث الويب تسلك سبلا اخرى مثل تحليل المحتوى **content based** من خلال تحليل البنية واللون والشكل **texture, color and shape**. وعلى ذلك فإن بنية الكتابات اليدوية تمثل مشكلة كبيرة أمام أدوات بحث الويب في كونها نص مصور أو صورة نصية تحتاج الى دراسة آليات الإدراك والمعالجة لحروف وكلمات لا تقاس بالمعاجم اللغوية أو التحليل الصرفي إنما تقاس بمدى كثافة الصورة **image resolution** ووحداث قياس الاشكال المصورة **pixels**. ويرى الباحث أن نظم استرجاع الكتابات اليدوية على الويب تحتاج الى معايير معالجة وآليات تنظيم واستراتيجيات بحث مزيجة بين ما تستخدم للنصوص والصور، كما أن هذه النظم ستحتاج بشدة الى أدوات عمل لغوية كالمعاجم اللغوية الرقمية والمحللات الصرفية، بالاضافة إلى برمجيات معالجة وقراءة الصور الرقمية.

تتعامل أدوات بحث الويب مع الصور الرقمية من خلال النصوص المصاحبة للصورة المتمثلة في عناصر الميتاداتا **metadata** والنصوص الواردة في متن صفحة الويب التي وردت بها الصورة وتأخذ منها امتدادها الرقمي **URL**. وذلك حيث لا تتوافر لدى محركات بحث الصور الرقمية حتى الآن تقنيات التعامل مع ملفات الصور الرقمية وتحليل عناصرها الرسومية المصورة، غير أن هذا التفاعل لا يتوافق وتحليل صور الكتابات اليدوية ؛ حيث لن يقف حد المعالجة والبحث عند النصوص المصاحبة أو حقول الميتاداتا، إنما يتعدى ذلك الى تحليل أشكال الحروف والكلمات الواردة في الصورة الرقمية، حتى يتم بحث نصوص الكتابات اليدوية ذاتها، وهو المستهدف من رقمنة الكتابات اليدوية على الويب. وما يمكن أن يحدث تأثيرا إضافيا هو موقع صور الكتابات اليدوية على الويب، كونها توجد داخل قواعد بيانات رقمية متاحة على الويب، أو من خلال التواجد داخل

متون صفحات ومواقع الويب ؛ حيث تحتلف بنية الملف الرقمي الذي يحمل الصورة في قاعدة البيانات المهيكلة عنها في ملفات الصور على صفحات **HTML**. فتحمل الصورة الرقمية في قاعدة البيانات مجموعة من العناصر الببليوجرافية للميتاداتا مثل **Dublin Core** التي تمثل وصفا فنيا معياريا لمواد الصور الرقمية، وفي الوقت ذاته تفتقر ملفات الصور داخل صفحات الويب الى هذه المعايير. ويؤكد ذلك كله أن الصور الرقمية التي تحمل نصوص الكتابات اليدوية سوف تتباين في ملاءمتها للمعالجة الاسترجاع داخل محركات البحث. [١١]

إن الاهتمام بما ستكون عليه البنية الرقمية لصور الكتابات اليدوية الرقمية على الويب المرقمنة بالمسح الضوئي هو أول خطوات التجهيز لما قبل المعالجة والتفاعل مع محركات بحث الويب. فالبنية الرقمية المزودة بملف رقمي يشتمل على عناصر وصف فني معيارية يمكن محركات البحث من دقة التنظيم والبحث للكتابات اليدوية الرقمية. هذا بالإضافة الى استحداث عناصر وصف للمتن والنص تمثل مرآة نصية للكلمات والحروف الواردة في صورة الوثيقة أو المخطوط، ويقترح الباحث إضافة حقول أخرى تتوافق وصور الكتابات الرقمية مثل حقل **content** أو المحتوى لنقل النص الوارد في صورة الوثيقة يحث يمكن لمحركات البحث استخدامه في البحث النصي ومعالجته وفقا للمعاجم والمحللات الصرفية. ويتم إضافة هذه الحقول الى تلك التي خلصت اليها أطروحة الباحث للماجستير عن الصور الرقمية، ومن ثم تكون العناصر الموافقة صور الكتابات اليدوية كالتالي :- [١٢]

**جدول رقم (٢) عناصر المعالجة والضبط المقترحة لمواد الكتابات اليدوية على الويب**

| العنصر | المستهدف | المعيار |
|---|---|---|
| **العنوان :** | عنوان المادة | **DC** |
| **(المؤلف) :** | المؤلف | **DC** |
| **الموضوع :** | رؤوس موضوعات | **DC** |
| **تاريخ التقاط الصورة :** | تاريخ OCR | **DC** |
| **تاريخ صنع الصورة الرقمية :** | النشر على الويب | **DC** |
| **تاريخ وضع التسجيلة :** | بنية قواعد البيانات | **DC** |
| **الناشر :** | اسم قاعدة البيانات أو موقع الويب | **DC** |
| **شكل ملف الصورة :** | برمجيات القراءة والعرض | **DC** |

| | | |
|---|---|---|
| **حجم ملف الصورة :** | الحجم بوحدات القايس بالميجابيت | **DC** |
| **حقوق نشر الصورة :** | الهيئة المسئولة عن النشر | **DC** |
| **الناسخ أو الكاتب** | كتب التراث المنسوخة او الوثائق | **مقترح** |
| **النوع** | (مخطوط او وثيقة) | **مقترح** |
| **النص الكامل** | النص الكامل للصفحة أو الوثيقة | **مقترح** |
| **الهيئة المسئولة** | الهيئة التراثية أو المتحف المسجلة لدية المادة | **مقترح** |
| **المصدر على الويب URL** | مسار الوصول الرقمي على الويب | **مقترح** |

## المبحث الثاني : معالجة وتكشيف الكتابات اليدوية على الويب

### ٢ / ١ بيئة المعالجة الرقمية للكتابات اليدوية

فرضت البيئة الرقمية للويب بتقنياتها المختلفة مجموعة من آليات معالجة مواد المعلومات الرقمية تمثل بدورها فارقا جليا بين طبيعة المعالجة داخل أدوات بحث الويب**on-line**، وبين معالجة نفس مواد المعلومات في قواعد البيانات التي تعمل خارج أو داخل النطاق الشبكي **off-line**. حيث يمكن تشبيه الفارق هنا بلفارق بين آلية العمل داخل النظم المغلقة والعمل داخل النظم المفتوحة ؛ فتمثل قواعد بيانات مواد المعلومات الرقمية خارج نطاق الويب بيئة النظم المغلقة، فلا يرتبط نظام إدارة قاعدة البيانات بغير مجموعات المعلومات المختزنة داخل قاعدة البيانات، فعليه تتم عمليات المعالجة وعليها أيضا تتم عمليات البحث والاسترجاع من جانب مستخدمي قاعدة البيانات دون اتصال او ارتباط شبكي على الخط المباشر. أما انتقال نظام إدارة قاعدة البيانات الي تقديم البحث والاسترجاع من خلال الربط الشبكي فإنه بذلك يتخطى حاجز النظم المغلقة الى حدود النظم المفتوحة التي تتبادل المعلومات في اتجاهين مستمرين هما استقبال الاستفسارات البحثية وارسال النتائج المسترجعة داخل شبكة المعلومات (الخط المباشر).

ويرى الباحث أن اختلاف نمط العمل على مجموعات الكتابات اليدوية بين قواعد البيانات خارج الربط الشبكي وقواعد البيانات على الويب يمثل بيئتين مختلفتين تماما للعمل على معالجة وبحث الكتابات اليدوية الرقمية. فثمة متطلبات وآليات للمعالجة ينبغي توافرها لاستكمال عمليات المعالجة والبحث المتكاملة للكتابات اليدوية إذا ما أرادت قواعد بيانات الكتابات اليدوية تقديم خدماتها على الخط المباشر من خلال مواقع الويب، وهو في غالب الأمر سيتمثل في مواقع المكتبات الرقمية التراثية ومراكز حفظ التراث والمخطوطات. وبناء على ما تقدم فإن معالجة الكتابات اليدوية تتم في بيئتين مختلفتين

تحتاج كل منهما الى آليات ومعايير مختلفة، وذلك كما يلي :-

**٢ / ١ / ١ المعالجة في بيئة قواعد البيانات** off-line

تقف معالجة الكتابات اليدوية في قواعد البيانات عند حدود الوصف والإدراك للكلمات والحروف داخل الوثائق المكتوبة يدويا، ولا تتعدى ذلك الى عمليات المعالجة المعلوماتية التي تشمل عمليات التحليل الموضوعي وبناء آليات البحث والوصف الفني المعياري لمواد المعلومات داخل قاعدة البيانات. فالنتيجة النهائية لقاعدة البيانات هي كلمات بحثية تتطابق وصور الكلمات والحروف موضع المعالجة. وتعمل قواعد البيانات وفقا لنظرية النظام المغلق في معالجة مجموعات الكتابات اليدوية على عناصر داخلية متكاملة وأهم هذه العناصر هو مستودع من الحروف والكلمات في شكل صور ثابتة ؛ حيث ينبني على استخدام هذا المستودع محاولة إدراك الكلمات والحروف داخل الكتابات اليدوية. وتتوقف عمليات الضبط الببليوجرافي في قواعد البيانات عن أرقام الترميز **IDs** التي تعطى لكل وحدة أو كلمة داخل قاعدة البيانات حتى يتم استرجاعها فيما بعد كما في شكل (٣). ولا يتطلب العمل في قواعد البيانات خارج الويب التحلي بدعائم الربط مع محركات البحث مثل محارف الترميز**meta tages** أو حقول الوصف العام لمواد المعلومات التي من خلالها تستطيع محركات البحث تدقيق عمليات التكشيف والتحليل. وقد ذكر **Yefeng Zheng, Yi Li, David Doermann** في دراستهم عن تحليل وادراك الكتابات اليدوية آلية العمل التي تسير عليها قواعد البيانات، وهي :-[١٣]

**أولا** : يتم رصد الكلمات ومقاطع الكلمات داخل الصفحة مع ترميزها برموز محددة **IDs**.

**ثانيا** : يتم الوقوف على الشكل النهائي للمقاطع او الكلمات مع عرضها على مستودع الكلمات المصورة حتى يتم تحديد شكل المقطع أو الحرف.

**ثالثا** : يتم حساب أو وزن كل مقطع بالنسبة للسطر الذي وجد به حتى يرمز له بكونه مقطع او كلمة كاملة يتم التعرف عليه.

**رابعا** : المقاطع التي لا تمثل كلمة او حرف محدد يتم ترتيبها وفقا لمقاييسها ثم وضعها في أسطر لإعادة معالجتها مرة اخرى.

**خامسا** : يتم رصد الصورة الأقرب للمقطع التي تدل على هوية المقطع إذا ما تجاوزت درجة التقارب نسبة ٥٠%.

**سادسا** : يتم اعتبار المقاطع التي تم التعرف عليها مقاطع معرفة أو محددة **processed** .

**سابعا** : يتم تجميع المقاطع الاقرب كتابة أو الاكثر تلازما واعتبارها كلمة واحدة لمقاطع متعددة.

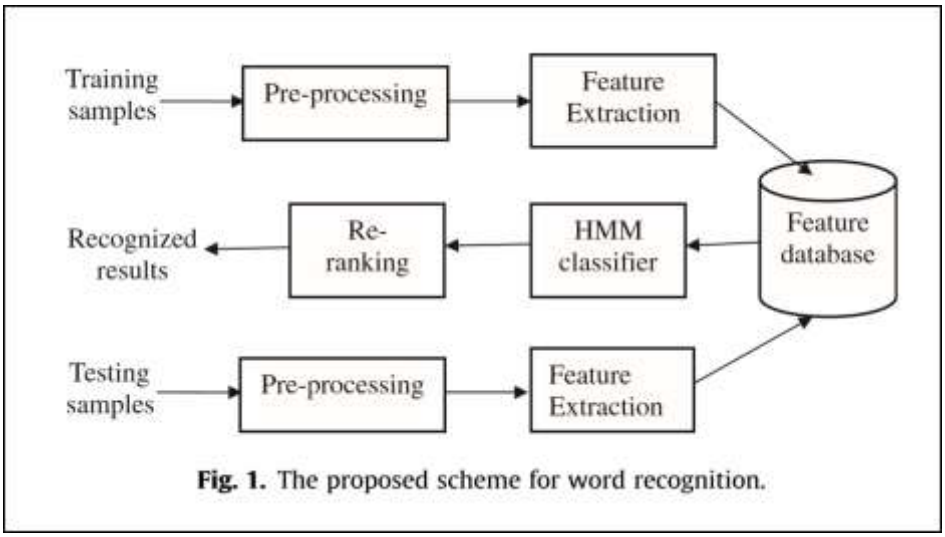



**Fig. 1.** The proposed scheme for word recognition.

**شكل رقم (٣) معالجة الكتابات اليدوية في بيئة قواعد البيانات خارج الخط المباشر**[١٤]

٢ / ١ / ١ المعالجة في بيئة الويب **on-line**

تضم بيئة الويب شتاتا ضخما متنوعا من مواد المعلومات الرقمية تختلف فيما بين النصوص والصور والمواد السمعية تقليدية الأصل ورقمي النشأة، وتتبع الويب درب ما يضاف اليها من أنماط جديدة من المعلومات الرقمية بالمعالجة والبحث والاسترجاع من خلال أدوات بحثها التي تعمل على مدار الثانية تنظيما واسترجاع لفضاء الويب. ولعل الكتابات اليدوية الرقمية أحدث أشكال مواد المعلومات التي تنظر لها الويب بعين التنظيم بغية إحكام التوافق بينها وبين محركات البحث حتى تيسر لمستخدمي الويب الوصول الى تلك الكتابات وتحقيق مبدأي الاتاحة والاستخدام. وإذا كانت قواعد البيانات خارج الويب تمثل النظم المغلقة، فإن الويب بفضائها وأدوات بحثها تمثل بحرا لا شطآن له من مواد المعلومات الرقمية تبحر داخلها سفن محركات البحث لجمع واصطياد كل ما تجده من مواد معلومات لإيصاله الى المستفيدين. ومن ثم فإن تواجد مواد الكتابات اليدوية الرقمية في فضاء الويب سواء أكانت في جوف قواعد البيانات أو على مواقع وصفحات **html**، يجعل محركات البحث منوطة بالوصول الى كلمات وحروف الوثائق والمخطوطات الرقمية ومعالجتها لدعم الوصول والاتاحة أمام مستخدمي الويب.

إن الويب تعتمد على أدوات بحث ذات مكونات محددة متكاملة الأداء والعمل تهدف جميعها الى تحقيق غاية أدوات البحث في ضم ومعالجة واتاحة مواد المعلومات الرقمية على الويب، ومن ثم فإن أول ما يمكن التفكير به هو جعل الكتابات اليدوية العربية من مخطوطات ووثائق بصبغة رقمية تتوافق وطبيعة عمل برامج أدوات البحث ؛ فكيف يمكن لبرامج الزاحف أو العنكبوت **spider or crawler** أن تصل الى الكتابات اليدوية الرقمية وتستطيع نسخها واضافتها الى قاعدة بيانات محرك البحث دون أن تكون هذه المواد بصبغة رقمية تسمح لبرامج الزاحف بالتعامل معها، أيضا لا يمكن لبرامج المكشف **indexer** أداء مهامها داخل محرك البحث بتكشيف وتحليل محتوى مواد معلومات الويب دون أن تكون مواد الكتابات العربية اليدوية على الشكل الذي يمكّن برنامج المكشف من التعرف على الكلمات الدالة أو الكشفية في الوثيقة أو المخطوط. وعلى ذلك فإن مواد الكتابات العربية اليدوية تحتاج الى التوافق وبنية **HTML** التي تسبح محركات البحث داخل محارفها وتعتمدها في اتمام عملها.

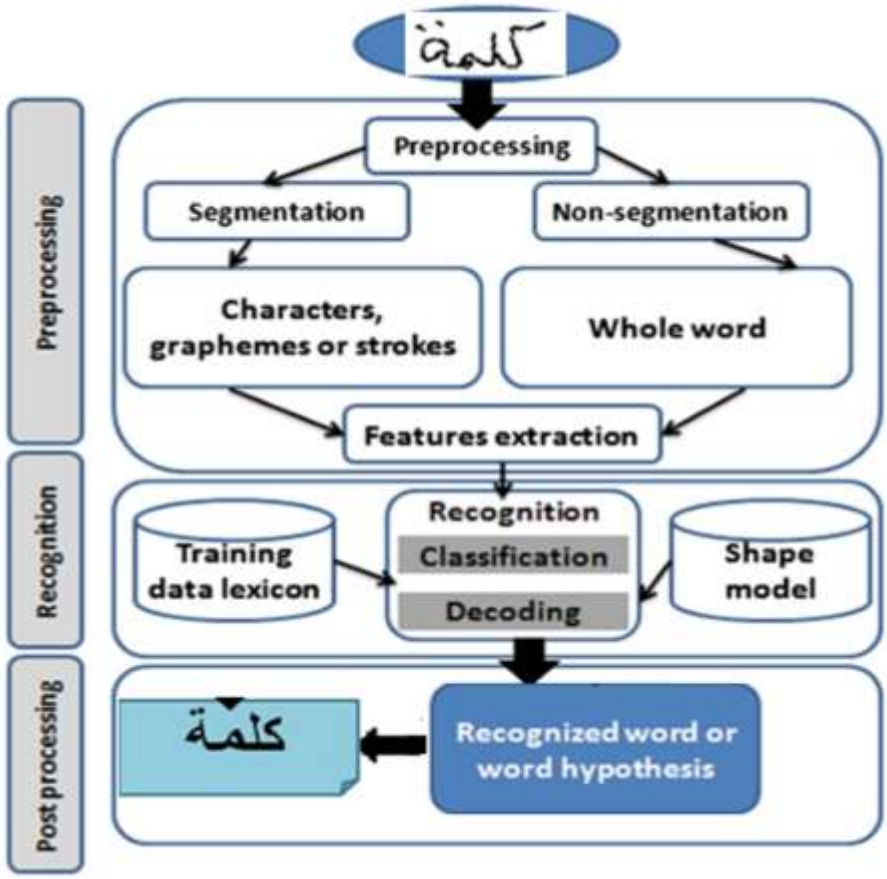
يبقى في هذا الصدد الحديث عن تحد مهم يكاد يقطع الطريق على محركات البحث في التعامل المباشر مع مواد الكتابات اليدوية العربية، ألا وهو أن المخطوطات والوثائق التي تحمل النصوص اليدوية العربية تتواجد في بيئة الويب في شكل صور ثابتة تحمل أشكالا من الكلمات والحروف لا تكاد تعيها محركات البحث أو تستطيع إدراكها سوى كونها صورة رقمية **digital image**. وقد تمحورت دراسات محركات البحث في التعامل مع الصور الرقمية من خلال النصوص المصاحبة التي تتواجد في موقع او صفحة الويب التي تحوي الصورة، وبناء على ذلك فإن محركات بحث الويب تحتاج الى آلية تحليل الكلمات داخل الصورة وإدراكها فيما يعرف بآلية البحث باستخدام المحتوى **content based image retrieval** مضاف إليها تقنيات تحليل وإدراك الكلمات المصورة **word image recognition**.

## ٢ / ٢ معالجة وإدراك الكتابات اليدوية الرقمية

إن ثمة حقيقة مؤكدة في واقع التعامل مع الكتابات اليدوية على الويب من خلال أدوات بحث الويب، ألا وهي أن مواد الكتابات العربية على الويب عبارة عن مواد نصية في وسائط معلومات مصورة. واعتمادا على هذه الحقيقة المعلوماتية يمكن البناء لكل ما سيأتي فيما بعد في التعامل مع مواد الكتابات اليدوية كمواد معلومات يتم معالجتها

وتنظيمها وبحثها واسترجاعها من خلال محركات البحث أو باقي أدوات بحث الويب. لقد تناول الباحث في دراسة له من قبل التعامل مع الصور الثابتة وكيفية معالجتها وآليات توافق محركات البحث معها تنظيما واسترجاعا، غير أن الفارق الجوهري في حالة الكتابات اليدوية على الويب يكمن في أن المعالجة والتنظيم يمتد الى داخل بنية الصورة والمتمثلة في النصوص الكتابية المخطوطة بالأيدي، التي تستمد الكتابات اليدوية أهميتها والحرص على تنيظمها من النصوص والمتون التي هي غاية رقمنة الكتابات اليدوية بالاساس.

لقد عملت الويب على تعديل العديد من المفاهيم والمصطلحات المرتبطة بنظم استرجاع المعلومات ؛ حيث سيمر البحث هنا بمرحلتين متتاليتن كانتا قد انضمتا فيما سبق في التعامل مع جل مواد المعلومات الرقمية على الويب، ثم يرى الباحث هنا حتمية فصلهما وهما مرحلتي المعالجة والتكشيف لمواد الكتابات اليدوية على الويب. فمرحلة المعالجة هنا تركز على كيفية قراءة وإدراك أشكال الكلمات المصورة داخل مواد الكتابات اليدوية والعمل على تحويلها الى نصوص مفهومة من قبل نظم وأدوات البحث على الويب حتى يمكن التعامل معها تنظيما واسترجاعا. أما العنصر التالي بعد ذلك ترتيبا وتفعيلا هو تكشيف الكتابات اليدوية على الويب والتعرف على الكلمات الكشفية الدالة على الموضوعات والأفكار التي تزخر بها الكتابات اليدوية مثل المخطوطات والوثائق التاريخية. ويمكن الحديث هنا عن المعالجة في إطار المراحل التي تأخذها مواد الكتابات اليدوية منذ كونها مواد معلومات تقليدية أو مطبوعة إلى كونها مواد رقمية على الويب. وتتمثل في ثلاث مراحل هي ما قبل المعالجة والمعالجة ذاتها ثم مرحلة ما بعد المعالجة من إجراءات، وذلك كما في شكل (٤).




**شكل رقم (٤) مراحل وإجراءات المعالجة والإدراك لكلمات الكتابات اليدوية على الويب[١٥]**

يوضح شكل (٤) البنية العامة لعملية المعالجة والادراك لكلمات الكتابات اليدوية على الويب بمراحلها الثلاث والإجراءات المتداعية داخل كل مرحلة، وذلك على النحو التالي :-

٢ / ٢ / ١ مرحلة ما قبل المعالجة **Pre- Processing**

تسبق مرحلة المعالجة والادراك للكتابات اليدوية بعض الاجراءات التجهيزية لمواد الكتابات اليدوية تشكل معا مرحلة ما قبل المعالجة، وذلك انطلاقا من أن إدراك الكلمات المصورة داخل الكتابات اليدوية يتم بشكل رقمي خالص مما يستوجب معه ضبط هذه

النصوص حتى تتسم عملية الادراك بالدقة الكافية لمعرفة الكلمات والحروف والمكتوبة المكونة لمتن الكتابات اليدوية. وتتلخص الاجراءات التجهيزية فيما يلي :-

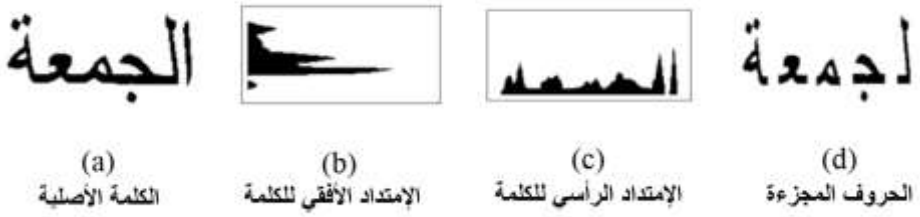
**٢ / ٢ / ١ / ١ تطبيع الكتابات اليدوية الرقمية** normalization

تتسم عملية المسح الضوئي عامة بتباين درجات جودة الناتج النهائي من الصور الضوئية الرقمية، وهذا على اختلاف طبيعة مواد المعلومات التي يتم رقمنتها من خلال الماسح الضوئي. وتزداد درجات الجودة في حالة المعلومات المصورة عنها في المعلومات النصية، خاصة تلك النصوص ذات الكتابات اليدوية التي تخضع جودتها لجودة خط الكاتب، وعدم تأثر الوثيقة الأصلية بعوامل التلف والطمس. وعلى ذلك فإن ثمة عملية يطلق عليها التطبيع لا بد أن تمر خلالها مواد الكتابات اليدوية الرقمية بعد رقمنتها وقبل إخضاعها لبرامج الإدراك والقراءة. وتهتم عملية التطبيع برفع درجة جودة صور الكلمات والحروف في الكتابات اليدوية من خلال حذف الزوائد والصور التشويشية حتى يمكن تأمين دقة قياس الارتفاع والانخفاض والامتداد لحروف ومقاطع وكلمات الكتابات اليدوية، ومن ثم تحديدها بواسطة الحاسب والتعرف عليها. وتنتهي عملية التطبيع بوضع خط وهمي يمر بمنتصف الكلمات والحروف عرضه يساوي نقطة واحدة في كثافة الصورة **one-pixel-wide** حتى تمثل مرتكزا لقياس أبعاد الحروف والكلمات.[١٦]

**٢ / ٢ / ١ / ٢ التقسيم والتجزئة** Segmentation

تحتوي صفحات الوثائق والمخطوطات على العديد من أشكال المواد المصورة والنصية، تختلف بين الصور والجدوال والعلامات والكلمات. وحيث أن التركيز هنا يكون على معالجة وإدراك الكلمات والحروف، فإنه يتم تجزئة وتقسيم صفحة الوثيقة أو المخطوط الى أسطر منفصلة يستقل كل منها بمعالجة محددة، ثم يتوالي بعد ذلك تقسيم وتجزئة السطر الواحد الى مجموعة الكلمات والمقاطع داخل السطر، ومنها يتم تقسيم كل مقطع وكلمة الى الحروف المكونة لها. وذلك كله يأتي في منظومة التعرف على الوحدة الاصغر في النص ثم بناء التراكيب والتتابع اللازمين لتكوين شكل الكلمة النهائية اعتمادا على المعاجم اللغوية. وتزداد صعوبة تقسيم الأسطر وتمييزها من الكتابات المطبوعة الى الكتابات اليدوية، وذلك لاتسام الكتابات اليدوية بالعديد من الخصائص مثل امكانية تداخل الخطوط وفقا لجودة الكتابة، أو تعمد متابعة الاسطر وتحرير المساحات الفارغة في الوثيقة، أو التحشيات الدخيلة على النص من تعليقات وغيرها.[١٧]

وتنتقل عملية التجزئة من الأسطر الى تجزئة الكلمات لحروف ومقاطع يطلق عليها **graphemes**، وتعتمد نظم الإدراك على آليات تستخدم نقاط الضبط الرقمية المحددة للاتجاهات الأفقية والرأسية للكلمات **horizontal and vertical projections**. كما تستخدم المعاملات الرياضية والتصويرية **mathematical morphological techniques** للتعرف الحروف المكونة للكلمة من خلال استخدام نقاط قياس الكثافة والرسم شكل تقريبي للمسارات التي يمر بها الحرف في صورة الكلمة ككل، ومن ثم التعرف على الشكل النهائى لهذا الحرف، كما فى شكل رقم (٥).




**شكل رقم (٥) تجزئة الكلمات الى حروف باستخدام الامتدادات التصويرية للكلمة**

**٢ / ٢ / ١ / ٣ استخراج الملامح** Feature Extraction

يمثل إجراء استخراج أو اشتقاق ملامح الكلمات والحروف خطا فاصلا بين مرحلتي ماقبل المعالجة والإدراك للكتابات اليدوية، فاستخراج الملامح الشكلية التي تصف اتجاهات وأبعاد الحروف يعمل على تمييز شكلي محدد لحروف اللغة في ذاكرة نظام إدراك الكتابات اليدوية. ويترتب على عملية التحديد المميز لكل حرف امكانية التواصل مع المعاجم اللغوية بشكل محدد للحرف يسهل استخراج الحرف الموافق للملاح الشكلية المشتقة. وتتم عملية اشتقاق الملامح اعتمادا على ترحيل شكل الحرف من الوثيقة الى شاشات النظام لتصفيته الرقمية حتى يختفي الشكل الحقيقي للحرف وتظهر نقاط رسومية ممثلة لبنية الحرف الرقمية من اتجاهات وأبعاد وزوايا يمر بها الحرف في كتابته. وتراعي نظم إدراك الكتابات اليدوية خصائص حروف اللغة العربية التي تتميز عن غيرها من اللغات بإمكانية كتابة الحرف الواحد على أكثر من جزء، فبعض حروف اللغة العربية مثل ( ك، ن، ت، ي) تكتب على مقطعين احدهما يمثل الجزء الرئيس من الحرف والآخر يمثل الجزء الثانوي التابع للحرف مثل النقاط في حروف (ن، ت، ض)، وحرف الهمزة مع حرف الكاف، وبالاعتماد على الحسابات الرياضية والرسومية تستطيع تحديد

الاجزاء الثانوية التابعة لكل حرف وضمه أو إلغائه عند تحديد الملامح الشكلية النهائية لكل حرف.[١٨]

تستهدف عملية اشتقاق ملامح الحروف والكلمات انتاج مجموعة من المعلومات التي تصف شكل الحرف بشكل محدد يميزه النظام بها عن غيره من الحروف. وهذه المعلومات حول وصف الحروف تأتي اعتمادا على مدخلين أساسيي في وصف حروف الكتابات اليدوية ؛ هما أولا الملامح البنيويةStructural features وتمثلها طبيعة البنية الرقمية التي يوجد بها الحرف المتمثلة في شكل البداية والنهاية للحرف والزوايا والاتجاهات التي تمر بها كتابة الحرف. ثانيا الملامح الإحصائية **Statistical features** المتمثلة في نقاط كثافة الصورة التي يمر بها الحرف، وقياس المسافات الرقمية بين اجزاء الحرف. ومن خلال تدقيق المعلومات الاحصائية والبنيوية لحروف الكتابات اليدوية يمكن لنظام الإدراك التمييز والفصل بين حروف اللغة العربية التي تتشابه في الجزء الرئيس لها وتختلف في اتجاهات الأجزاء التابعة لها مثل ما هو كائن في حرفي ( ن، ب ) أيضا امكانية التمييز بين الحروف ذات الرسم الواحد مع اختلاف عدد النقاط التابعة لها مثل حروف ( ب، ت، ث ) وحروف ( س، ش ). وبالتوصل الى هذه الملامح يبدأ نظام الإدراك في الانتقال الى مرحلة المعالجة لاستخراج الحروف النصية والكلمات من المعجم اللغوي والتعامل مع النصوص. [١٩]

**٢ / ٢ / ٢ مرحلة المعالجة (الإدراك)** Processing (Recognition)

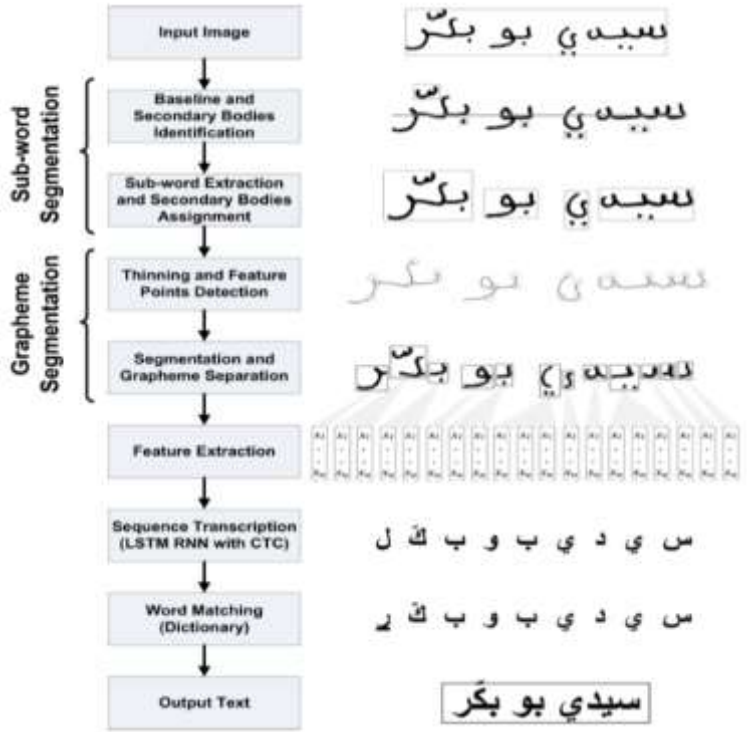
تستهدف مرحلة المعالجة والإدراك استخدام مجموعة الملامح الشكلية للحروف في عملية المضاهاة مع المعاجم الآلية داخل النظام للتعرف النهائي على نص الكلمة والحرف، ومن ثم استخدام الحروف والكلمات لبناء النص النهائي المقابل لصور الكتابات اليدوية واختزانها في قواعد البيانات النهائية لإجراء عمليات البحث والاسترجاع من قبل مستخدمي النظام. وتتم عملية المعالجة وإدراك الكتابات اليدوية وفقا لخطوات ممنهجة تعتمد في أدائها على أدوات عمل كنموذج الأشكال والمعاجم الآلية.

**٢ / ٢ / ٢ / ١ الإدراك** Recognition

تمثل خطوة الإدراك واسطة العقد لكل مراحل وخطوات المعالجة لمواد الكتابات اليدوية الرقمية ؛ ففي هذه العملية يبزغ النص النهائي المتولد عن مضاهاة أشكال الحروف المصورة مع المعاجم الآلية. وسيحل النص المتولد عن هذه العملية محل الكتابات اليدوية الأصيلة في باقي عمليات البحث والاسترجاع داخل نظام استرجاع الكتابات اليدوية على

الويب، ومن خلال النص المنتج يمكن إجراء المعالجات اللغوية والصرفية على مجموعات المخطوطات والوثائق التراثية التي لم يكن لأحد أن يفكر في تطويعها بحثا واسترجاعا في بيئة الويب. حيث سيمثل النص المتولد عن إدراك الكتابات اليدوية الوجه المطابق لمادة المخطوط أو الوثيقة الذي يمكّن أدوات بحث الويب من سهولة التعامل مع توفير النموذج المصور النهائي للمخطوطات والوثائق للعرض النهائي.

تبنت العديد من الدراسات آلية محددة لإجراء عمليات المعالجة والادراك للكتابات اليدوية، يطلق على هذه الآلية نماذج ماركوف غير المرئية **Hidden Markov Models (HMM)**. وترتكن هذه الآلية إلى استخدام التحليل الشكلي لحروف الكتابات اليدوية واستخدام الملامح الشكلية الناتجة في الخطوات السابقة في المضاهاة المقننة بين صور الحروف ونظائرها الأكثر تشابها في المعاجم اللغوية. وقد وصفت هذه الآلية بغير المرئية نظرا لأن الصور الرقمية للحروف موضع القياس تكون عبارة عن نقاط توزيع لبنية الحرف، دون ظهور الحرف ذاته. وتتبع آلية **HMM** أسلوبا منظما يقضي بمقاربة حروف الكلمة الواحدة الناتجة من التحليل الشكلي وفقا للترتيب والتتابع الذي جاءت به في الكتابات اليدوية مع قواعد بيانات المعاجم الآلية، وهي بذلك تحافظ على ترتيب الحروف للكلمة وترتيب مقاطع الكلمة الواحدة، بحيث ينتج الشكل النهائي للكلمات والعبارات الواردة في الكتابات اليدوية، كما في شكل رقم (٦). ويأتي الاصرار من جانب الباحثين على استخدام هذه الآلية لما اثبتته من نجاح في التعامل مع إدراك الكلمات الشفوية على مواد المعلومات الصوتية.[٢٠]




**شكل رقم (٦) طبيعة المعالجة والادراك لإحدى عبارات الكتابات اليدوية باستخدام آلية HMM**
[٢١]

### ٢ / ٢ / ٢ / ٢ التصنيف والتوزيع classification

تتفاعل الملامح الشكلية لحروف الكتابات الرقمية مع وحدات التحليل اللغوي في قاعدة البيانات للوقوف على النصوص النهائية للكتابات اليدوية، ويتم ذلك في ظل ما تزخر به اللغة العربية من أشكال متعددة للحروف والكلمات المكونة للنصوص العربية، وخاصة الكتابات اليدوية. وعلى أساس التجانس بين الكلمات أو الحروف المحددة والعديد من الكلمات والحروف داخل المعاجم اللغوية، فإن آلية الإدراك في نظام بحث الكتابات اليدوية تبدأ بعمل توزيع وتصنيف لكل ما خرج من ملامح شكلية للحروف والمقاطع وكل ما يقابلها من كلمات وردت متشابهة في عملية المضاهاة. إن كلمة واحدة في اللغة العربية يقابلها العديد من الأشكال والصور التي قد تأتي بها في مواد المخطوطات والوثائق، وعلى الجانب الآخر فإن صورة كلمة ما ودت في إحدى مواد الكتابات اليدوية يقابلها

العديد من الكلمات الفرضية التي قد تكون دالة على الملامح الشكلية للكلمة موضع التحليل.

**٢ / ٢ / ٢ / ٣ فك الترميز** decoding

تتكون قاعدة بيانات نظام إدراك الكتابات اليدوية من قواعد بيانات فرعية تختص كل منها بمهام محددة وتسجيل مجموعة متكاملة مع غيرها من بيانات النظام الأخرى. ويختزن النظام مجموعات الملامح الشكلية والصور الرقمية لكلمات وحروف الكتبات اليدوية، بجانب مجموعات الكلمات المسجلة في قاعدة بيانات المعجم اللغوي، وإن ثمة ربط تقني وأرقام تحديد **IDs** تستخدم لإكمال عمليات البحث والاسترجاع لا بد من تكوينها داخل النظام ككل. إن عملية التجزئة والتقسيم **segmentation** الواردة في مرحلة ما قبل المعالجة والإدراك قد شملت أسطر وكلمات الكتابات اليدوية الواردة في صفحات متعددة للمخطوطات والوثائق، وقد نتج عنها سلسلة متتابعة من الحروف والمقاطع المتتالية غير محددة ببداية أو نهاية لكلمات مُعرفة. مثلت هذه السلسلة المتتابعة من صور الحروف الرقمية ترميزا وشفرة معقدة تحتاج الى تدقيق متعمق في قاعدة بيانات المعاجم الآلية التي تستطيع وحدها فك رموز سلسلة الحروف المتتابعة، من خلال تكوين كلمات لها دلالة لغوية وشكل كتابي محدد، تبدأ بحرف محدد وتنتهي عند حرف بذاته. ومن ثم تتحول سلسلة الحروف المشفرة الى مجموعات من الكلمات المعرفة ذات دلالة محددة بما يعكس فك تشفير ورموز حروف الكتابات اليدوية الرقمية داخل النظام. وهو ما توضحه الخطوتان الأخيرة وما قبلها في شكل رقم ( ). وينتهي النظام في مرحلة المعالجة بتكوين قاعدة بيانات فرعية للكلمات النصية المؤكد ورودها في متن الكتابات اليدوية، مع الكلمات الفرضية التي توافق بنسبة كبيرة صور الكلمات والحروف.

**٢ / ٢ / ٢ / ٤ أدوات نظام إدراك الكتابات اليدوية**

تنتهي مرحلة الإدراك للكتابات اليدوية بانتهاء عملية فك الترميز وتكوين الكلمات النصية النهائية المعبرة عن صور الحروف والكلمات، غير أن الباحث لا يمكنه إغفال الحديث عن أداتين أساسيتين لا تعتمد عليهما عملية الإدراك وحسب، وإنما يعتمد عليهما نظام البحث والاسترجاع للكتابات اليدوية ككل. وهاتان الأداتان هما قاعدة نماذج الأشكال **shape model** والمعاجم اللغوية **lexicon**، وعلى الرغم من اختلاف ما تحمله كل منهما من معلومات، غير أن نجاح نظام إدراك الكتابات اليدوية يعتمد على درجة التوافق في التفاعل بينهما. فقاعدة نماذج الأشكال تحوي مجموعات الصور الرقمية لحروف وكلمات

الكتابات اليدوية، وتحوي المعاجم الآلية مجموعات الكلمات النصية التي تقاس على أساسها الملامح المقتبسة وتحديد الكلمات الدالة عليها تلك الملامح. ونظرا لما تفعله هذه الأدوات من تأثير على عملية إدراك الكتابات اليدوية، فإن الاهتمام بمعاجم آلية وفيرة الكلمات يساعد على تحقيق الدقة والكفاءة المأمولين في بحث واسترجاع المستفيدين لمواد الكتابات اليدوية.[٢٢]

**٢ / ٢ / ٣ مرحلة ما بعد المعالجة** Post Processing

لا تنتهتي مراحل وخطوات المعالجة والإدراك للكتابات اليدوية العربية عند خطوة الإدراك والتعرف على الكلمات المصورة وتحويلها الى كلمات نصية مستخرجة من المعاجم الآلية، إنما تتم المراحل بمرحلة ما بعد المعالجة والادراك. ويتم في هذه المرحلة عمل مختلف الإجراءات التنظيمية والتنقيحية للنصوص المستخرجة من عملية الإدراك ؛ فمواد المعلومات العربية التراثية من مخطوطات ووثائق وغيرها تتلاءم واستخدام اللغة الطبيعية في الكتابة أكثر من اللغة الفصحى المقيدة المحددة بقواعد التراكيب والصرف في اللغة العربية. وفي مجال نظم استرجاع المعلومات الرقمية، فإن معالجة نص على طبيعته وعلى الشكل المكتوب به دون تغيير أو تعديل، يستلزم ذلك استخدام آليات وأساليب معالجة اللغة الطبيعية **Natural Language Processing (NLP)** في التنظيم والاسترجاع. فضلا عن أمر مهم يتطلب معه استخدام اللغة الطبيعية، ألا وهو القيمة التراثية التاريخية التي تتميز بها هذه النصوص والدراسات التاريخية الاجتماعية التي تستدعي دراسة متون ونصوص المواد التراثية الوثائقية كما كانت عليه من كتابة وصياغة وأسلوب ومفردات مستخدمة تحاكي الوقت والعصر الذي كتبت فيه. وتقوم المعاجم الآلية في هذه المرحلة على ضبط جودة وسلامة النصوص المستخرجة من حيث الشكل الاملائي والصرفي للنص النهائي، ويتأتي ذلك من تزويد قواعد المعاجم الآلية داخل النظام بمجموعات الأشكال الشائعة من الأخطاء اللغوية والكتابية المرتبطة باستخدام تقنية **OCR** في رقمنة الكتابات اليدوية، ومن ثم يجري المعجم مسحا شاملا للنص لاستخراج الأخطاء ومعالجة الكلمات الواردة بها.

إن عمليات التقسيم والتجزئة التي تمر بها الكتابات اليدوية أثناء المعالجة من الصفحات الى الأسطر إلى الكلمات ثم الحروف المكونة لكل لكلمة، يجعل من السهل جدا حدوث أخطاء متنوعة أثناء عمليات التقسيم وتفكيك الكلمات ثم ما يتلوها من عمليات إدراك وتجميع بواسطة المعاجم الآلية، وكل ذلك من شأنه إحداث أخطاء لغوية وتركيبية بين

مقاطع الكلمات وحروفها. وتقوم المعاجم الآلية بتبيان وتوضيح طبيعة الحروف المفقودة أو التي وضعت في أماكن خاطئة أو علامات الترقيم وضبط النص "?" و "!" التي تصاحب الحروف في متن الكتابات اليدوية، بما يساعد على إعادة استقامة النص مرة أخرى.[٢٣]

**٢ / ٤ تكشيف الكتابات العربية اليدوية**

تتجه الدراسة بالحديث عن تكشيف الكتابات اليدوية من الوجهة التقنية المعلوماتية الى الوجهة المعلوماتية الخالصة، فعملية التكشيف إحدى عمليات التنظيم والتحليل الموضوعي التي تقوم عليها عمليات البحث والاسترجاع لمختلف مواد المعلومات على الويب. بل ويمكن القول أن مختلف أدوات بحث الويب تطلق مصطلح التكشيف تحديدا على عمليات التحليل الموضوعي المستخدمة مع مجموعاتها الرقمية، فالغاية المنشودة لدى أدوات البحث هي توفير مجموعات من الكلمات الكشفية الدالة على المحتوى تنظيما، واستخدام تلك المجموعات بحثا واسترجاعا من جانب مستخدمي الويب. ومن ثم اختار الباحث عملية التكشيف خاصة دون سواها من عمليات التحليل الموضوعي كالتصنيف والاستخلاصن لأنها الأكثر ملاءمة في التعامل مع مواد الكتابات اليدوية العربية الرقمية، فالكلمة الدالة هنا والمصطلح الكشفي سيكون كل كلمة وردت في متن المخطوط أو الوثيقة، أيضا الكلمة المفتاحية المتوقع طرحها من جانب المستفيد هي كل كلمة قد وردت في نفس المخطوط أو الوثيقة. وذلك انطلاقا من مبدأ أن الأهمية المعلوماتية والتاريخية التراثية للكتابات العربية المتضمنة في النص تستوجب إجراء التكشيف على مستوى الكلمات والنص ككل، كما هو متبع في حال تكشيف النصوص القانونية أو المقدسة.

تتحدد نوعية التكشيف المتبع في تحليل محتوى مواد المعلومات الرقمية وفقا لمجموعة من المحددات المستمدة من نوعية مادة المعلومات ذاتها ؛ فمواد المعلومات النصية تستخدم معها نوعية التكشيف الاشتقاقي **Extracted Indexing** أما باقي وسائط حمل المعلومات فستخدم التكشيف بالتعيين **Assigned Indexing** والفارق بينهما يأت وفقا للمصدر المأخوذة عنه المداخل الكشفية، فإذا ما كانت الكلمات الدالة من بين كلمات النص المكشف كان التكشيف اشتقاقيا كما في حالة الكتابات اليدية العربية، أما باقي المواد فتكون المداخل مختارة من المكانز وعليه فيكون التكشيف بالتعيين، كما سيلي ذكره. ويتحديد نوعية التكشيف المبني على الاشتقاق، فإنه يمكن القول أننا حددنا اللغة الطبيعية في التكشيف **Natural Language Processing (NLP)** دون اللغة المقيدة، وذلك لأن

الكتابات اليدوية سيتم تكشيف النص الكامل لها وفقا لما وردت عليه من مؤلفيها. كما تتحكم مواد الكتابات اليدوية اختيار أسلوب محدد من أسلوبي تكشيف المواد الرقمية، فأولهما التكشيف المبني على المحتوى والمتوافق ومواد المعلومات المصورة والمسموعة، وثانيهما أسلوب التكشيف المبني على النص للمواد النصية، لأن مواد الكتابات اليدوية هي نصية مصورة، فإن كلا الأسلوبين يتم استخدامهما مع تكشيف الكتابات اليدوية العربية.[٢٤]

إن ثمة اعتبارين يجب توصيفهما واعتبارهما أساسين لانطلاق عملية التكشيف للكتابات اليدوية العربية ؛ **أولهما** : أن الكتابات اليدوية العربية على الويب توجد على حالتين ؛ مختزنة في قواعد بيانات ذات بنية نظم إدارة قواعد البيانات **DBMS**، وتوجد في صفحات ومواقع الويب كصور مضافة لصفحات ذات بنية **HTML**. وهذا الاعتبار من شأنه التأثير على آليات التكشيف وأدوات التكشيف المتبعة في الحالتين، بل وقدرة أدوات بحث الويب على القيام بعملية التكشيف كما يجب. **ثانيهما** : أن الكتابات اليدوية العربية عبارة عن كلمات ونصوص تتواجد في شكل صور رقمية، فالملف المستخدم لتداولها على الويب هو ملف الصور الرقمية، أما المطلوب منها في التداول فهو النص المتضمن في الصورة الرقمية المنتجة بواسطة **OCR**. وهذا الاعتبار من شأنه التأثير على عمليات آليات وأدوات التكشيف المتبعة ؛ حيث يختلف تكشيف النص والكلمات عن تكشيف الصور. وانطلاقا من الحقيقة المذكورة سلفا بأن الكتابات اليدوية الرقمية هي نصوص في وسيط مصور، فإن خصائص وأساليب التكشيف المذكورة لاحقا سوف تصيغ عملية تكشيف متكاملة تتوافق وطبيعة مواد الكتابات اليدوية الرقمية.

يبقى في التمهيد لتكشيف الكتابات اليدوية العربية الحديث عن المستهدف من تكشيف هذه المواد، فالغاية المراد الوصول إليها ستحدد أساليب وآليات العمل الواجب اتباعها، بل وستحدد طبيعة التكشيف وركائزه المتبعة. ويرى الباحث أن غاية تكشيف الكتابات اليدوية على الويب هي توفير قدرات البحث والاسترجاع لنصوص الكتابات اليدوية العربية وفقا للشكل الذي كتبت به وخرجت عليه، دون المساس بصيغة وبنية النص المكتوب، مع عرض صور رقمية مطابقة لأصل المخطوط أو الوثيقة موضع التكشيف. وعلى ذلك فإن المراد استرجاعه ملف قاعدة بيانات أو صفحة **HTML** تحتوي على صورة رقمية للوثيقة أو صفحة المخطوط مع ميتاداتا كشفية للنص الكامل والبيانات الببليوجرافية عن المادة المسترجعة. ووفقا لهذه الغاية وهذا المراد تستكمل الدراسة باقي عناصرها.

### ٢ / ٤ / ١ خصائص تكشيف الكتابات اليدوية

تتسم عملية تكشيف الكتابات اليدوية بالعديد من الخصائص التي تتحكم في الشكل والمنتج النهائي لعملية البحث والاسترجاع على قواعد بيانات ومحركات بحث الويب. وتمثل هذه الخصائص أسس التشغيل والمعالجة والبحث، فضلا عن سياسات وبنية ادوات بحث الويب. وأهم هذه الخصائص :-

**أولا** : تتوقف عملية التكشيف بالأساس على مكان إجراء تلك العملية في دورة تداول الكتابات اليدوية على الويب، فقد تضطلع قواعد بيانات الكتابات اليدوية بإجراء التكشيف واضافة حقل **content** الى عناصر الوصف الفني (الميتاداتا) تحمل النص الكامل الوارد في متن الكتابات اليدوية. ومن ثم لا تحتاج محركات البحث الى إجراء عملية التكشيف واستخدام هذا العنصر في البحث والاسترجاع. أما في حالة تزويد محركات البحث واضافة ملفات نصوص الكتابات اليدوية مباشرة من الويب، فإن ذلك يستلزم تزويدها بآليات تحليل الصور المبني على المحتوى **content based image retrieval**.

**ثانيا** : ستعتمد عملية تكشيف الكتابات اليدوية الرقمية على تحليل صور المسح الضوئي لنصوص الكتابات اليدوية، مما يعني إدخال متطلبات آلية لمحركات بحث الويب تختص بتحليل بنية الكلمات المصورة، مثل نموذج ماركوف غير المرئي **HMM** المخصص لإدراك صور الحروف والكلمات، واستخدام المعاجم الآلية لتحديد هوية الكلمات والعبارات مع التدقيق التركيبي والصرفي لكلمات اللغة العربية.

**ثالثا** : تتجاوز عملية تكشيف الكتابات اليدوية المفاهيم المعروفة عن تكشيف النصوص الرقمية ؛ حيث تنطوي مفاهيم تكشيف النصوص على استخراج الكلمات الدالة والمعبرة عن موضوعات وأفكار النص المكشف، ويتم ذلك بواسطة محركات بحث الويب أثناء التحليل وقبل عملية البحث، أما تكشيف نصوص الكتابات اليدوية داخل صور المسح الضوئي فإنها تتمثل في اقتباس كل الكلمات ومقاطع الكلمات الواردة في صورة النص وإدراكها واستخراجها في نص منفصل، واستخدام كل الكلمات الناتجة عن ذلك ككلمات كشفية تستخدم للبحث والاسترجاع. بما يعني اجمالا أن تكشيف الكتابات اليدوية إنما هو اشتقاق كل كلمات النص وتعيينها ككلمات كشفية.

**رابعا** : تعد عملية التكشيف المحدد الرئيس لطبيعة البنية التي ستكون عليها أدوات بحث الكتابات اليدوية، ذلك لأن محركات البحث في بنيتها الطبيعية تشمل عملية التكشيف مع عمليتي الاضافة والبحث، فإذا ما تركت عملية تكشيف الكتابات اليدوية لمحركات البحث، فسوف تكون وفقا لبنيتها المطابقة لمفهوم عمل محركات البحث. وأما إذا كانت

عملية التكشيف تتم داخل قواعد بيانات مخصصة للكتابات اليدوية وتتاح هذه القواعد في بيئة الويب، فإن دور محرك البحث ثو يقف عند كونه واجهة لبحث الكتابات اليدوية دون التدخل في عمليات التكشيف والمعالجة والإدراك للكتابات اليدوية الرقمية على الويب. وستكون بنية عمل محركات البحث هنا مطابقة مع مفاهيم محركات بحث قواعد بيانات الويب غير المرئية **Deep web search engines** التي تناولها الباحث في أطروحته للدكتوراه، حيث تتبنى محركات البحث أحد خيارين في التعامل مع مجموعات الكتابات اليدوية في قواعد البيانات ؛ فإما أن يكون محرك البحث أداة لإرسال كلمات البحث واستقبال النتائج المستدعاة، أو أن يبني نموذج بحث عام يمكنه إجراء البحث مباشرة في مجموعات قواعد بيانات الكتابات اليدوية واسترجاع صورها ونصوصها.[٢٥]

**٢ / ٤ / ٢ أساليب تكشيف الكتابات اليدوية**

لقد استخدمت أدبيات الانتاج الفكري حول الكتابات اليدوية على الويب مصطلح التكشيف **indexing** كمفهوم مرادف لمصطلح الإدراك **recognition**، ويأتي ذلك من تبني وجهة نظر واحدة في التعامل مع الكتابات اليدوية كونها صور رقمية يتمثل تكشيفها في تحليل محتواها سواء أكان عناصر رسومية أم كلمات مصورة. ومن ثم فإدراك الكلمات المصور في داخل صور الكتابات اليدوية يمثل تكشيفا تقنيا أو برمجيا لدى باحثي علوم الحاسب، أما وجهة النظر المعلوماتية فتمتد إلى تحليل محتوى الكلمات النصية المشتقة من صور الكتابات اليدوية والخروج بالمصطلحات الأكثر دلالة لاستخدامها كمداخل كشفية في عملية التنظيم والاسترجاع. لذا ومن المنطقي هنا عند التعرض لأساليب التكشيف المستخدمة مع الكتابات اليدوية، فإن الأقرب الى التناول والاستخدام هو التكشيف المبني على المحتوى **content based retrieval**، ذلك لأن مادة الكتابات اليدوية ستكون في شكل صورة رقمية، الى جانب أن المراد من التكشيف هو اشتقاق النص الكامل، وليس فقط الخروج بالمداخل والمصطلحات الكشفية. وقد ساعد على أن يكون التكشيف على مستوى النص الكامل، ما وصلت إليه أدوات بحث الويب من إمكانات تقنية يمكنها البحث على مستوى النصوص الكاملة، واستخدام كافة ما ورد في البحث من كلمات ومقاطع كمداخل كشفية، وهو ما يشير إليه ظهور آليات البحث باللغة الطبيعية.

**٢ / ٤ / ٢ / ١ التكشيف المبني على المحتوى**

ينطوي التكشيف المبني على المحتوى على استخراج مجموعات الرسوم المضمنة في الصورة سواء أكانت أشكال أم كلمات، واستخدامها في عمليات تنظيم واسترجاع محتوى

الصور الرقمية وفقا لأساليب تنظيم وبحث تقنية تقوم على برمجيات تحليل اللون والبنية والشكل للصور الرقمية. وتقف صور الكتابات اليدوية في مكانة وسطى بين الصور الرقمية التي تحوي أشكال مصورة، وبين النصوص الرقمية على الويب، ولذلك تبدأ عملية تكشيف صور الكتابات اليدوية العربية وتنتهي مع عملية إدراك واشتقاق الكلمات المصورة الواردة في صور الكتابات اليدوية. ويعتمد تكشيف الصور المبني على المحتوى على ثابت مهم هو اشتقاق العناصر المصورة وإدراكها ثم استخدامها بنفس بنيتها الرقمية في إجراء عمليات البحث والاسترجاع، مما يعني بدوره أن استخدام تكشيف الكتابات اليدوية المبني على المحتوى سوف ينتج عنه صور لكلمات نصية تستخدم كأشكال مصورة أثناء عمليات البحث. ولكن ما تسعى إليه هذه الدراسة هو استخدام معاملات إدراك الكتابات اليدوية ومقومات المعاجم الآلية في تحويل الكلمات المصورة المشتقة من صور الوثائق والمخطوطات وتحويلها الى نصوص من كلمات وحروف يمكن للمستفيد البحث داخلها بكلمات مفتاحية وليس بصور أو أشكال مصورة. وقد ظهر مفهوم **Content Based Images Retrieval (CBIR)** على يد **Kato** عام ١٩٩٢ عندما بدأ العمل في اتجاه استرجاع الصور للصورة ويكون برنامج نظام الاسترجاع ذا خصائص تمنحه القدرة على تحليل مكونات الصورة الأساسية وهي اللون والشكل **shape** والبنية **texture**.[٢٦]

**٢ / ٤ / ٢ / ٢ التكشيف المبني على النص**

يهدف التكشيف المبني على النص إلى استخدام النص الكامل لمادة المعلومات في استخراج المصطلح أو المصطلخات الدالة على فكرة وموضوعات محتوى مادة المعلومات النصية. وقد عكست عملية معالجة وإدراك نصوص الكتابات اليدوية أنه لا مكان لمفهوم التكشيف المبني على النص في هذه العملية، وذلك لأن صور الكتابات اليدوية بتقنية **OCR** تستخدم برمجيات وآليات حاسوبية تعمل على إدراك مختلف الأسطر والكلمات والحروف المكونة لهذه الأسطر، وتحويلها الى نصوص معرّفة لدى أدوات بحث الويب تستخدم في البحث والاسترجاع. بما يعني أن النص الكامل للكتابات اليدوية الرقمية هو بالأصل نص مصور لا يمكن معه استخدام آليات التكشيف المتبعة لدى أدوات البحث، فضلا عن ذلك فإن أدوات بحث الويب تستخدم كل كلمات النصوص الكاملة كمصطلحات كشفية ومداخل بحث لاسترجاع مواد المعلومات النصية، وذلك لتحقيق ميزة تنافسية مع باقي أدوات البحث التي تمتلك قدرات متنامية في إختزان وبحث واسترجاع مواد المعلومات. ومن ثم لا تتمكن أدوات بحث الويب بالأساس من استعراض نصوص

الكتابات اليدوية، ما لم تكن تملك أدوات الإدراك والتحليل **Recognition** لصور الكتابات اليدوية بتقنية **OCR**، وستقتصر عملية التكشيف المبنية على النص على تحليل عناصر الوصف الفني (الميتاداتا) المصاحبة لملف صور الكتابات اليدوية. وعند استخدام تحليل النصوص المصاحبة لصور الكتابات اليدوية في قواعد البيانات أو ضمن نصوص صفحات الويب، فإن ثمة إجراءات محددة تمر بها عملية التكشيف المبني على النص، وهي :-[٢٧]

١) إعداد قائمة الاستبعاد للكلمات كثيرة التواجد غير الدالة كحروف العطف وظرف الزمان والمكان.
٢) يتم تجريد الكلمات من السوابق واللواحق الواردة معها.
٣) حساب عدد مرات التكرار لكل الكلمات وترتيبها تربيا تنازليا بحسب عدد مرات التكرار، ثم اختيار الكلمات التي تتصدر الترتيب.
٤) يتم اعتبار الكلمات التي تجاوز عدد مرات تكرارها عدد محدد من المرات، هي أكثر كلمات دلالة لمحتوى الوثيقة داخل نظام الاسترجاع، وإعطائها وزنا نسبيا أكبر أثناء بحث واسترجاع النصوص.

**٢ / ٤ / ٣ مقومات تكشيف الكتابات العربية**

تبنى عملية تكشيف الكتابات اليدوية على عناصر عمل متكاملة تمثل مقومات التكشيف التي لا يمكن لنظام بحث واسترجاع الكتابات اليدوية العمل دونها، وتختلف هذه المقومات بين التقنية والبرمجية واللغوية، ذلك لأن الكتابات اليدوية تجمع بين البيئة الرقمية واللغة العربية. كما أن عملية التكشيف ذاتها لهذه المواد تتم بأسلوبين أحدهما آلي تقني العمل، والآخر نصي لغوي البنية. وتتمثل مقومات عملية تكشيف الكتابات اليدوية في التالي :-

**أولا : تقنيات إدراك النصوص text recognition** : وأهم أشكالها نموذج ماركوف غير المرئي **HMM** ؛ حيث تعمل هذه التقنيات بمصاحبة تقنية **OCR** لرقمنة الكتابات اليدوية وتحليل محتوى صفحات موادها وتقسيم الاسطر الى كلمات ومنها الى حروف، بما ينتج عنه اشتقاق النص الكامل في شكل نصي بعد الشكل المصور، وهو ما يمثل تكشيفا على مستوى النص الكامل للمخطوط أو الوثيقة.

**ثانيا : المعاجم اللغوية الآلية Lexicon** : فاستخدام هذه الأدوات يأتي في مرحلتين ؛ أولاهما عند تحليل وإدراك النصوص المصورة واشتقاق الكلمات، وثانيهما عند إجراء عمليات البحث بمشتقات أو معاني أو تصريفات الكلمة الواحدة. ويتركز الحديث هنا حول

الاعتماد المقترح لاستخدام المعاجم الآلية في مرحلة التنظيم ( التكشيف) التي تتصل باشتقاق الواصفات الموضوعية الكتابات اليدوية. وتعد المعاجم الآلية العمود الفقري لتنظيم واسترجاع الكتابات اليدوية العربية حيث يتوفر بذلك إجراءات التنظيم في تحليل الأشكال المختلفة للكلمة الواحدة، ومن ثم يمكن التدقيق في حساب تكرار الكلمة إذا ما اعتمد محرك البحث على اشتقاق الكلمات المفتاحية بحسب تكرارها في النص. أما إجراءات الاستدعاء فهي تتمثل في تحليل كلمات البحث المستخدمة من جانب المستفيد لاستدعاء صفحات الويب، مما يساعد على أداء خدمات الاسترجاع في اللغة العربية مثل استدعاء المقابلات أو استدعاء الوحدات التي تحتوي على معنى كلمة البحث أو جذر كلمة البحث بمختلف السوابق واللواحق، وهو ما يتناوله الحديث في التعرف على التحليل الصرفي للغة العربية[٢٨]

**ثالثا استخدام اللغة الطبيعية : Natural Language Processing** : تمثل اللغة الطبيعية العمود الفقري لإدارة معالجة وبحث مواد الكتابات اليدوية، فاللغة الطبيعية هي لغة المؤلف حين يكتب المادة المعلوماتية، وهي لغة التكشيف عندما تعتمد نظم الاسترجاع على النص في اشتقاق كلماته الدالة ومداخله الكشفيه، وهي أيضا لغة المستفيد حينما يستخدم كلمات مفتاحية غير مأخوذة عن مكنز او قائمة رؤوس موضوعات مقننة. وعلى ذلك فإن آليات اللغة الطبيعية تمثل جناحي التكشيف والبحث لمواد الكتابات اليدوية على الويب، فضلا عن كونها الأنسب لتوجهات أدوات البحث في التكشيف على مستوى النص الكامل، وتحليل وإدراك نصوص الكتابات اليدوية كما وجدت عليه بصياغتها ومفرداتها. ولم تقتصر دراسات التطوير للغة داخل نظم استرجاع المعلومات على التقدم في معالجة اللغة الطبيعية، وإنما امتد الأمر إلى تطوير قدرات نظم استرجاع المعلومات على معالجة اللغات الحية المستخدمة في كتابة الوثائق والإنتاج الفكري على مستوى العالم. ومن هذه الدراسات ما تم عن طريق**naserdine semmar** وزملائه في تطوير قدرات نظم استرجاع المعلومات للعمل على تنظيم واسترجاع المعلومات المسجلة بلغات مختلفة. من خلال توفير برامج للتحليل اللغوي والإحصائي داخل نظام الاسترجاع يأتي دورها في العمل عند إضافة الوثائق ومعالجتها. ويعمل هذا النظام اعتمادا على مجموعة من برامج تحليل اللغة سواء المستخدمة في استفسار البحث أم تلك المستخدمة في كتابة الوثائق. وعلى ذلك فإن ثمة ترابط قد أحدثته هذه الدراسة بالجمع بين آليات المعالجة والبحث وبرامج تحليل اللغة، ومن البرامج المستخدمة في هذا النظام ما يلي:-[٢٩]

١. برامج التحليل اللغوي : **linguistic analyzer** وهي التي تقوم بالمعالجة اللغوية لكل من استفسار البحث والكتابات المكشفة.
٢. برامج التحليل الإحصائي : **statistic analyzer** وهي تقوم فقط بتكشيف مواد المعلومات.
٣. برامج تهذيب البحث : **reformulator** وتختص بتوسيع سؤال البحث وإجراء البحث داخل النتائج المسترجعة.
٤. برامج الربط أو المضاهاة : **comprator** وهي تقوم بتحديد التشابه الدلالي بين سؤال البحث والوثائق المكشفة.

## المبحث الثالث : البحث والاسترجاع للكتابات اليدوية

تكتمل عملية إدارة مواد الكتابات اليدوية على الويب بالوصول الى توفير آليات البحث والاسترجاع لنصوص هذه المواد، وتتأتى آليات البحث والاسترجاع الموافقة لكل مادة معلومات على الويب من خلال طبيعة بنية ومعالجة هذه المادة، وحيث أن بنية الكتابات اليدوية الرقمية هي النصوص المصورة كما ذكر من قبل، فإن ثمة أساليب بحث تكمن في استرجاع الكتابات اليدوية باستخدام صور الكلمات**image-to- image matching** واسترجاعها من خلال الكلمات المفتاحية النصية **word matching** . وسيتناول هذا المبحث كل ما يتعلق بعملية البحث من أساليب وبنية وواجهة البحث.

### ٣ / ١ أساليب البحث والاسترجاع للكتابات العربية

تباينت الدراسات التي تحدثت عن استرجاع الكتابات اليدوية على الويب حول جدوى استخدام البجث النصي والبحث بالصورة في عمليات الاسترجاع، فمن الآراء ما أكد على فاعلية استخدام الكلمات المصورة نظرا لطبيعة وبنية الكتابات اليدوية الرقمية، وتأكيدا على توافر أجهزة الكتابة اليدوية الرقمية مثل الحاسبات اللوحية والهواتف النقالة. أما تبني استخدام البحث بالكلمات النصية فيرتكن إلى طبيعة وجود آليات الإدراك والمعالجة لنصوص الكتابات اليدوية وتحويلها إلى كلمات وحروف يمكن تطويعها للبحث بالكلمات المفتاحية. وعلى ضوء تحديد أسلوب البحث المستخدم تتحدد آلية المضاهاة والمطابقة **matching** المستخدمة داخل النظام ؛ حيث ستتم عملية المضاهاة مباشرة مع مواد الكتابات اليدوية وصور الكلمات حال الاعتماد على البحث بصورة الكلمة، أما في حالة استخدام البحث بالكلمات المفتاحية فسيتم الاعتماد على مخرجات عملية الادراك دون الرجوع الى صورة المخطوط او الوثيقة.[٣٠]

### ٣ / ١ / ١ البحث المبني على الكلمة المصورة word image search

تتدخل تقنيات وبرمجيات استرجاع المعلومات لتحدث تطويرا مستمرا في آليات عمل نظم استرجاع المعلومات، فمواد الكتابات اليدوية التي يتم رقمنتها بواسطة **OCR** ترجع الى أزمنة قديمة لم تكن الكتابة اليدوية الرقمية على الحاسب قد ظهرت. وقد أتى الوقت الحالي ليشهد أجهزة الكتابة الرقمية اليدوية بواسطة أقلام الكترونية يستطيع المستخدم إدخال نص يدوي الكتابة دون تدخل لوحة المفاتيح ويمكن لنظم الاسترجاع معالجة هذه النصوص وإدراكها مباشرة فور كتابتها يدويا **Real Time Recognition**. وبهذه الأجهزة المستحدثة تستطيع نظم استرجاع الكتابات اليدوية من مخطوطات ووثائق استخدام نماذج كلمات البحث المكتوبة يدويا لمضاهاتها مباشرة مع نصوص الكتابات اليدوية في شكل صور واسترجاع ما يتطابق أو يقترب منها بدرجات محددة. ويطلق على هذا الأسلوب البحث المبني على الكلمات المصورة أو صورة الكلمة **Word Image**، وذلك كما يعبر عنه شكل رقم (٧). ولقد اعتمد هذا الأسلوب على الدور التقني لبرمجيات تحليل الصور الرقمية وآليات تحديد أوجه وقياسات أشكال الكلمات في نصوص الكتابات اليدوية، كتقنية **Hidden Markouv Model HMM** التي تستطيع التعرف على كلمات اللغة العربية وتقسيمها إلى مقاطع وحروف كنماذج مصورة ثم استخدام المعاجم الآلية لتحويل صور الحروف والكلمات وفقا لتتابع الحروف الى كلمات نصية ومن ثم الوصول الى النص الكامل لمتن الوثيقة أو المخطو، وتسمى هذه الآلية باكتشاف الكلمة **word spotting**.[٣١]


**Figure 1.4.** An illustration of document retrieval using word image as query.

**شكل رقم (٧) أسلوب استرجاع الكتابات اليدوية المبني على صورة الكلمة والنتائج المسترجعة**
٣٢

يوضح شكل رقم (٧) استرجاع كلمات الكتابات اليدوية من خلال صور الكلمات البحثية المدخلة يدويا من جانب مستخدم النظام، ثم تطبيق المضاهاة وفقا لأشكال ونماذج الكلمات المتطابقة بنسب مختلة مع كلمة صورة الكلمة المدخلة، كما يربط النظام في عملية الاسترجاع بين الصور المسترجعة والوثيقة أو المخطوطة المرتبطة المنتمية لها الكلمة المسترجعة، حتى يمكن الوصول الى النص الكامل للمادة.

**٣ / ١ / ٢ البحث المبني على الكلمات النصية key word seraching**

يتوافق البحث بالكلمات النصية مع التوجه العام لمستخدمي الويب سواء أكان بحثهم عن مواد معلوماتية مصورة أم مسموعة أم نصية، فمستخدم الويب لا يحتاج سوى كتابة كلمات نصية مدخلة بلوحة المفاتيح في فراغ البحث **search box** حتى يعبر عن الحاجة الموضوعية لديه في صور كلمات مفتاحية أو استفسار بحثي. وقد أدركت نظم استرجاع المعلومات هذا السلوك فعملت على تحويل البحث عن الوسائط المختلفة من صوت وصورة بكلمات بحثية تتمثل في بحث الميتاداتا المكونة لملف مادة المعلومات أو النصوص المصاحبة الواردة في صفحات الويب، كأن يتم البحث عن الصور والصوت بكلمات نصية أو عبارات بحثية. وعند استخدام الكلمات المفتاحية للبحث عن نصوص الكتابات اليدوية من وثائق ومخطوطات، فإن النظام لن يعود الى مواد معلومات الكتابات اليدوية ذاتها، إنما سيتم إجراء عمليات المضاهاة والمطابقة مع النصوص المستخرجة بتقنيات الإدراك من صور الكتابات اليدوية **OCR**. وإن كان ذلك سيحمل نقصا في عدم البحث في الوثيقة أو المخطوطة ذاتها نظرا لوجودها في شكل صور رقمية، فإنه سيحمل امكانات التحليل الصرفي واللغوي من خلال المرور على المعاجم الآلية المستخدمة في تحليل نصوص الوثائق والمخطوطات في نظم استرجاع الكتابات اليدوية العربية. وفي هذه الحالة فإنه من الضروري التأكيد على معيارية بناء ملفات الكتابات اليدوية وما تحمله من عناصر وصف ميتاداتا وقدرة ضمها لعنصر **Content** المقترح في هذه الدراسة الذي سيحمل النص الكامل المستخرج من آليات الإدراك لنصوص الكتابات اليدوية.

وعلى هذا فإن نظام استرجاع الصور الرقمية يعمل على مضاهاة كلمات البحث مع حقول ملف مادة المخطوط أو الوثيقة كاملة إضافة إلى النص الكامل لصفحة الويب المصدرية مع حقول اللغة المعيارية **HTML**، ليخرج نظام استرجاع الصور بملف أو ملفات

الكتابات اليدوية التي يتوافر في أي من حقولها كلمات البحث المدخلة إلى النظام. ولعله يمكن القول أن البحث بالكلمات المفتاحية قد يعطي في كثير من عمليات البحث نتائج غير دقيقة خلاف التصفح بالتقسيم الموضوعي؛ ذلك لاحتياج أسلوب الكلمات المفتاحية إلى آليات بحث دقيقة يستخدمها غير المتمرسين من مستخدمي الويب. ويبرز الدور المهم لأسلوب الكلمات المفتاحية في بحث المخطوطات والوثائق إذا ما تحلى نظام الاسترجاع بحلية اللغة الطبيعية (لغة النصوص العربية الكاملة) كلغة تنظيم لنصوص ووحدات الكتابات اليدوية. ويحتاج أسلوب الكلمات المفتاحية في بحث النصوص الكاملة بلغتها الطبيعية إلى مجموعة من آليات البحث القادرة على صياغة مختلف استراتيجيات البحث بما يحقق الدرجة المرجوة من التحقيق في استرجاع المخطوطات والوثائق. وتتكامل آليات بحث اللغة العربية مع آليات البحث العامة لإجراء عمليات البحث في النظام.[٣٣]

**٣ / ٢ آليات ومعايير البحث**

إن المحتوى العربي من مواد الكتابات اليدوية العربية من مخطوطات ووثائق سوف ينعكس بطبيعة الأمر على شكل نظام الاسترجاع المستهدف بناؤه، فكما سبق وأن انعكست بنية الكتابات العربية على أدوات النظام في استخدام المعاجم الآلية، وانعكاس اللغة العربية على أساليب التكشيف والبحث، فإن طبيعة محتوى الكتابات ذات اللغة العربية وطبيعتها التراثية سوف تنعكس على طبيعة آليات ومعايير البحث المستخدمة لبحث واسترجاع نصوص هذه المواد، هذا فضلا عن بنيتها التقنية كونها صور رقمية لنصوص **OCR**. ويضاف الى هذا كله طبيعة المعالجة الزمنية التي كتبت فيها هذه المخطوطات والوثائق بما ينعكس على المفردات اللغوية وأساليب الكتابة المستخدمة، مما يعني إجمالا أن المعاجم الآلية ستكون لب عمل نظام استرجاع الكتابات اليدوية على الويب، واستخدامها لأداء العديد من الوظائف الرئيسة في معالجة نصوص المواد الوثائقية ليس فقط في البحث وإنما في التحليل والإدراك والتكشيف.

لقد عملت التقنية العربية على تذليل مشكلات المفردات العربية من اشتراك لفظي وترادف وغيرها واستخدام المحللات الصرفية وتقنيات التحليل اللغوي في توفير قدرات البحث بالمعنى والمقابل والبحث بجذور الكلمات ومشتقاتها. وقد تميزت شركة صخر **Sakhr** في مجال المعالجة اللغوية لبحث واسترجاع المحتوى العربي على الويب. ومن خلال العديد من الدراسات استخلص الباحث مجموعة من آليات ومعايير البحث التي يمكن استخدامها مع بحث واسترجاع الكتابات اليدوية العربية على الويب على النحو التالي :-[٣٤]

**جدول رقم (٣) أساليب بحث الكتابات اليدوية العربية على الويب**

| نوعية معيار البحث | أسلوب واتجاهات البحث | آليات ومعاملات البحث | قدرات البحث |
|---|---|---|---|
| **أولا معيار البحث المصور** | الصور المعدة مسبقا | ١) صورة الكلمة | نموذج مصور متطابق |
| | التصفح بين الوحدات | ٢) تصفح صور الكلمات | التصفح الارشادي |
| | أجهزة الكتابة اليدوية الرقمية | ٣) الكتابة اليدوية الرقمية | أجهزة الإدخال الرقمي |
| **ثانيا معيار البحث النصي** | **آليات البحث بالكلمة الواحدة** | ▪ البحث بالتطابق | بنفس شكل الكلمة |
| | | ▪ البحث غير المطابق | (البحث باللواصق) |
| | | ▪ البحث بجذر الكلمة | (حل مشكلة الاشتقاق) |
| | | ▪ البحث بالتشكيل للمشترك اللفظي | استخدام قدرات الاعجام |
| | | ▪ البحث الحر | الامتدادات المختلفة |
| | | ▪ البحث بحساسية الحالة | وفقا لرسم الكلمة |
| | | ▪ البحث بالترادف | لحل مشكلة الترادف |
| | | ▪ البحث بالمتضادات | باستخدام المعاجم الآلية |
| | | ▪ البحث بالمعاني | باستخدام المعاجم الآلية |
| | | ▪ إمكانية التعرف على الأخطاء العربية | باستخدام المعاجم الآلية |
| | **آليات البحث بكلمات متعددة** | ❖ البحث في الحقول | عناصر الميتاداتا |
| | | ❖ البحث بالتقارب | مسح النص الكامل |
| | | ❖ استخدام المنطق البوليني | |
| | | ❖ طرح سؤال باللغة الطبيعية | |
| | | ❖ استخدام علامات التطابق " " | |
| | | ❖ البحث في الحقول | |
| | | ❖ البحث بالتقارب | |

## ٣ / ٣ بنية نظم استرجاع الكتابات اليدوية على الويب

تحتاج الكتابات اليدوية العربية على الويب تكامل مجموعة من البرامج أو النظم الفرعية حتى يصاغ النظام المتكامل لاسترجاع وبحث المخطوطات والوثائق، وقد مر البحث حتى الآن بعناصر شتى يختص كل منها بأداء مهام محددة في عملية معالجة واسترجاع الكتابات العربية على اختلاف أشكالها. ويركز الباحث هنا على محاولة توصيف نموذج مقترح لبنية نظم تنظيم وبحث الكتابات اليدوية يتالف من مجموعة من النظم الفرعية، مع توضيح اختصاصات كل نظام في البنية العامة ككل. وقد أورد الدراسات التي تناولت الكتابات اليدوية عامة والعربية خاصة على الويب مقترحات لنظم بحث الكتابات اليدوية على الويب، غير أنها قد اتسمت بمحدودية العمل، واختار الباحث منها أقربها في شكل رقم (٨) للبنية المقترحة في هذا البحث مع إضافة وتوسيع التوصيف العام لأدوات بحث الويب.

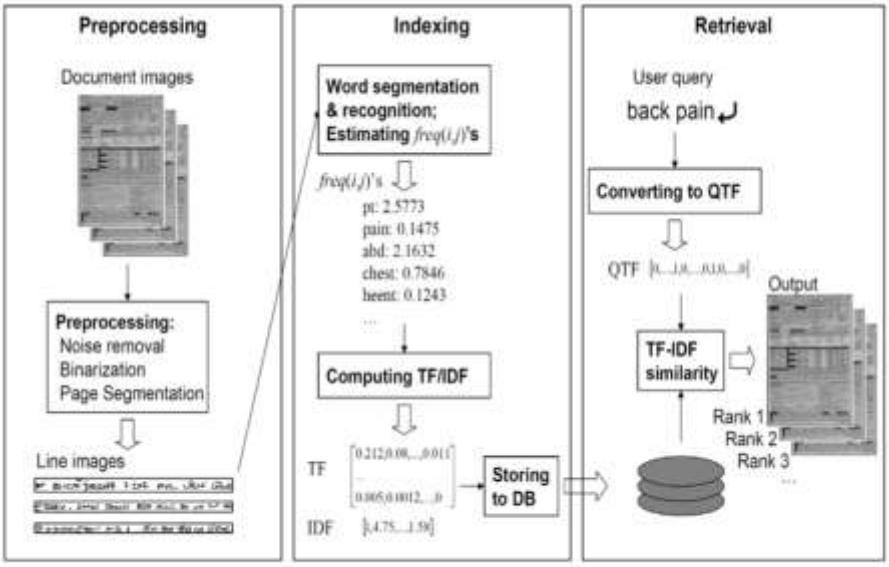



Figure 4.3. Flowchart of the search engine.

**شكل رقم (٨) محرك بحث الكتابات اليدوية المبني على المحتوى**[٣٥]

يوضح شكل رقم (٨) شكل البنية العامة التي يكون عليها نظام بحث الكتابات اليدوية، مضاف إليها ما تقترحه الدراسة، وتنحصر النظم الفرعية المقترحة فيما يلي :-

**٣ / ٣ / ١ نظام الوصول والإضافة**

تتناثر مواد الكتابات اليدوية في فضاء الويب غير محددة بحدود موقع أو قاعدة بيانات، وتتشابه مختلف الكتابات اليدوية على اختلاف لغاتها في بنيتها الرقمية التي تستمدها إما من صفحة ويب **HTML** أو من معيارية قاعدة بيانات مختصة بهذه المواد. وتعد عملية الوصول والاضافة لمادة المعلومات هي المكون الأول والإجراء المبدأي داخل نظم استرجاع المعلومات؛ وهي تقابل عملية التزويد في المكتبات. ونظرا لاختلاف أنواع النظم العاملة مع الكتابات اليدوية الرقمية بين محركات البحث وقواعد البيانات، فإن الوصول الى مواد الكتابات اليدوية الرقمية يختلف بين المحركات وقواعد البيانات ؛ حيث تعتمد محركات البحث على برامج الزاحف والعنكبوت التي تختص بعمليات الاضافة والتزويد لملفات المعلومات، غير أنها تتواصل بشكل أساسي مع مواقع وصفحات **HTML** دون قدرة الولوج الى داخل قواعد بيانات الويب، ومن ثم فإن بنية ملفات الكتابات اليدوية التي تضيفها محركات البحث لا تخضع لمعيار واحد أو بنية رقمية ذات ضبط مقنن. وهو ما عالجة البحث في الحديث عن أهمية بنية رقمية معيارية موحدة لملفات الكتابات اليدوية. أما قواعد البيانات فإن بناء مجموعاتها وإضافة ملفات المخطوطات والوثائق بها يتم غالبا من خلال الضبط المقنن ومعايير بناء الملفات الرقمية بواسطة العامل البشري، بما يعطي تقنين أكبر لشكل وعناصر البيانات المصاحبة لملف المخطوطة او الوثيقة في قاعدة البيانات. وفي الحالتين يعمل نظام الوصول على تكوين مجموعات الكتابات اليدوية داخل نظام البحث وتحديد ملامح مقننة لكل ملف يتم إضافته الى نظام البحث وفقا لاشتراطات وأسلوب معالجة وتنظيم محددين.

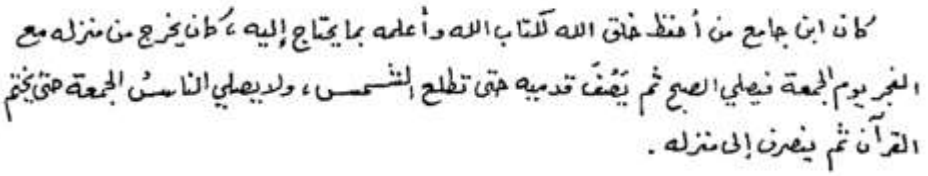


Fig. 9. A sample of Arabic manuscript.

**شكل رقم (٩) أحد ملفات الكتابات اليدوية على الويب في شكل مخطوطة** [٣٦]

وبحسب نوعية نظام البحث بين محرك وقاعدة بيانات، يحتاج نظام الوصول والإضافة المتمثل في برنامج العنكبوت لتوصيف العديد من المحددات مثل كيفية التعرف على مواد

الكتابات اليدوية، وكيفية قراءة الملف، وكيفية نسخه وضمه الى باقي المجموعات، فضلا عن المعلومات التعريفية التي تسهل عمليات المعالجة والتنظيم فيما بعد.

**٣ / ٣ / ٢ نظام المعالجة والإدراك**

يختص نظام المعالجة والإدراك بعملية التنظيم القبلية لعملية التكشيف والمعالجة المعلوماتية الفعلية داخل نظام الاسترجاع المتكامل، وتتم عملية الإدراك بشكل تقني خالص ينطوي على استخراج واشتقاق نصوص الكتابات اليدوية من هيئتها المصورة. وعلى الرغم من أن نظام الإدراك هنا هو نظام فرعي لبحث الكتابات اليدوية، إلا أنه ينطوي بذاته على ثلاثة مراحل لعملية الإدراك هي ما قبل المعالجة ثم الإدراك ثم ما بعد الإدراك، وتمر كل مرحلة بعدة إجراءات تبدأ في أول مراحلها بتهيئة الكتابات اليدوية الرقمية للإدراك ثم تنتهي آخر إجراءاتها بالرجوع الى المعاجم الآلية لتحديد المقاطع والحروف والكلمات غير المتعارف عليها التي يصنفها نظام الإدراك على أنها غير معالجة **unprocessed**. ويفرد شكل رقم ( ) عدة خطوات واضحة تتعلق بتجزئة الصفحات إلى أسطر ثم إلى كلمات ومنها إلى مقاطع وحروف حتى يتم تحليلها والتعرف عليها باستخدام تقنيات وبرمجيات مثل نموذج ماروكوف غير المرئي **HMM** وتقنيات الخط البيني **Baseline** التي تستطيع إدراك وقراءة نصوص **OCR**. وفي هذه المعالجة يتم حساب **term frequency (TF)** وربطه بمجموعة الوثائق التي جاء بها ثم حساب توارده في مجموعات الكتابات الرقمية في كل قاعدة البيانات وفقا لمعاملات تردد الوثائق **document frequency (IDF)**، وكما يتم استخدام هذه التقنية في إدراك النصوص، فإنه يتم استخدامها أيضا أثناء تحليل صور الكلمات المفتاحية المدخلة الى النظام أثناء عملية البحث، ومن ثم تحليلها ومضاهاتها مع ما يطابقها من كلمات ونصوص في مجموعات الكتابات اليدوية. وعلى الرغم من أن بعض دراسات علوم الحاسب قد نعتت هذه المرحلة بمرحلة التكشيف، إلا أنها تختلف تماما عن عملية التكشيف من وجهة النظر المعلومات التي تنطوي على استخراج الكلمات الدالة المعبرة عن محتوى الكتابات اليدوية.[٣٧]

**٣ / ٣ / ٣ نظام التكشيف والتحليل**

أفردت الدراسة العنصر الرابع في المبحث الثاني للحديث عن عمليات التكشيف التي تتم داخل نظام بحث واسترجاع الكتابات اليدوية على الويب بشكل مقصل وبمختلف زواياه ؛ حيث تناول الباحث أسلوبي التكشيف المبني على صورة الكلمات والحروف، والتكشيف المبني على تحليل الكلمات النصية المستخرجة من عملية إدراك صور الكتابات اليدوية،

إضافة إلى مقومات عملية التكشيف وخصائص تكشيف الكتابات اليدوية على الويب. وكل ما ورد في هذا العنصر يمثل توصيفا للنظام الفرعي القائم على التكشيف والتحليل الموضوعي للكتابات اليدوية.

### ٣ / ٣ / ٤ نظام الإدارة والتفعيل

يتخطى مفهوم نظام الإدارة هنا حدود العمل التقني في معالجة وبحث الكتابات اليدوية ليصل إلى حد إدارة مجموعات المعلومات من حيث أداء النظام وتفاعله مع البيئة الخارجية من قواعد بيانات ومستخدمي الويب. فتمتلك الكتابات اليدوية من مخطوطات ووثائق قيمة تاريخية تراثية وقانونية تختلف في طبيعتها عن مواد المعلومات الرقمية الأخرى، ومن ثم فالوصول إليها ورقمنتها وإتاحتها على الويب يحتاج بدوره إلى سياسات أداء مقننة، كتلك المستخدمة معها من الناحية المعلوماتية فيما يتعلق بتكشيفها وبحثها. وسيكون هذا النظام عبارة عن سياسة عمل شاملة تضبط عمل النظام ككل مما يترتب عليه المتابعة والتقييم التي تمثل أحد أسس بناء نظم استرجاع المعلومات، وستضح جليا استخدام هذه السياسة وأهميتها إذا ما وضع في الاعتبار الأشكال المختلفة التي ستكون عليها بنية النظام، فإذا ما كانت البنية متمثلة في محرك بحث يتصل بقواعد البيانات المتاحف والهيئات الوثائقية فثمة اعتبارات وتقانين لا بد من تحديدها، فضلا عن التوافق التقني اللازم لتواصل محرك البحث مع نماذج بحث قواعد البيانات وتبادل ملفات المعلومات على اختلاف بنية النظامين.

### ٣ / ٣ / ٥ نظام البحث والاسترجاع

تناولت الدراسة البحث والاسترجاع بشيء من التفصيل في المبحث الثالث كاملا لتوضيح الأساليب والآليات وتقنيات البحث المستخدمة مع الكتابات اليدوية، بما يمثل تحليلا كاملا للنظام الفرعي المختص بالبحث والاسترجاع. أما الحديث هنا فسيكون مخصص عن جزء لا يتجزأ من عملية البحث وهو واجهة البحث **search interface**، التي تمثل انعكاسا لكل معايير وأدوات وأساليب البحث المستخدمة في نظام الكتابات اليدوية. فعلى واجهة النظام تتضح آليات البحث بصور ونماذج الكتابات اليدوية من خلال توفير مساحة لإدخال صورة كلمة البحث المطلوبة، أيضا تتضح آليات البحث بخصائص اللغة العربية واستخدام المعاجم الآلية في تطويع الكلمات المسترجعة من خلال البحث بجذر الكلمة وتصريفاتها اللغوية. وإذا كانت واجهة البحث تختص بهذا المسمى الوارد فيه كلمة البحث، فإن الشق الثاني من واجهة النظام هو قدرات عرض النتائج المستدعاة ومدى التوافق مع طبيعة

مواد المعلومات خاصة إذا ما كانت في الوسيط المصور، أو بتقنية الفلاش وغيرها من مستحدثات عرض الصور الثابتة والمتحركة.

## المبحث الرابع : واقع ومواقع الكتابات اليدوية العربية على الويب

ترصد الدراسة التطبيقية هنا واقع مواد الكتابات اليدوية في بيئة الويب، وتوضيح واقع المنظومة المتكاملة التي تعمل على تنظيم واسترجاع الكتابات اليدوية على الويب. وثمة ملاحظة أولية استقرأها الباحث في الإعداد للدراسة التطبيقية تتمثل في غياب أهم أشكال نظم استرجاع الويب من محركات وأدلة بحث ولم يوجد عند اختيار عينة الدراسة سوى قواعد بيانات تتبع مكتبات أو مؤسسات علمية وثقافية، بما يعكس واقع الويب في تأخرها عن مجاراة هذا الشكل تنظيما واسترجاعا. وستركز الدراسة التطبيقية هنا على مجموعة العناصر الفرعية المتكاملة المستخلصة من جملة ما سبق من تنظير حول الكتابات اليدوية، وهي العناصر التي تمثل مختلف جوانب بنية وإدارة ومعالجة واسترجاع الكتابات اليدوية في بيئة الويب، بما سيعكس اخضاعها للقياس التجريبي قراءة لواقعها واستشرافا لما ينبغي أن يكون عليه مستقبل الكتابات اليدوية وأدوات بحثها على الويب. وستنقسم الدراسة التجريبية لمحاور رئيسة تضم منظومة العناصر الفرعية كما يلي :-

### ٤ / ١ محور البنية والشكل

يستهدف هذا المحور قياس وجود وطبيعة العناصر الأولية الأساسية لمنظومة استرجاع الكتابات اليدوية، فحتى تكتمل هذه المنظومة ويحكم على واقعها تحتاج الى عناصر وجودية تتمثل في شكل محدد تتواجد عليه الكتابات اليدوية، التي تختلف طبيعتها وإدرتها وفقا لطبيعة نظام الاسترجاع. فضلا عن طبيعة أشكال الكتابات اليدوية التي اهتمت بها عينة الدراسة، ويمثل جدول رقم (٤) قياس محور البنية والشكل وفق للعناصر الفرعية المكون منها.

**جدول رقم (٤) قياس عناصر محور البنية والشكل في معالجة واسترجاع الكتابات اليدوية**

| محور البنية والشكل | | مواقع قواعد بيانات المخطوطات | | | | | |
|---|---|---|---|---|---|---|---|
| عناصر المحور | | العربية | | | الأجنبية | | |
| وتوصيفها | | bibalex | idsc | almajid | fihrist | westafrican | AUB |
| بنية | محركات بحث | --- | --- | --- | --- | --- | --- |
| أداة | قواعد | √ | √ | √ | √ | √ | √ |

| البحث | بيانات | | | | | | |
|---|---|---|---|---|---|---|---|
| بنية ملف المادة | OCR | --- | --- | --- | --- | --- | --- |
| | صورة | √ | √ | --- | --- | --- | √ |
| تعدد الأشكال | مخطوطات | √ | √ | √ | √ | √ | √ |
| | وثائق | √ | √ | --- | --- | --- | √ |
| | كتب تراثية | √ | √ | √ | | --- | --- |
| مستوى الاتاحة | بيانات | √ | √ | √ | √ | √ | مختصرة |
| | المادة الأصلية | √ | بعض | --- | --- | --- | √ |

يتضح من جدول رقم (٤) مجموعة من المشاهدات والملاحظات وفقا للعناصر التالية :-

**٤ / ١ / ١ بنية أداة البحث**

أولا : لم يجد الباحث بين كل مواقع تنظيم واسترجاع المخطوطات محركا واحدا لبحث المخطوطات كما يتوافق ومفاهيم محركات بحث مواد الويب. وعلى الرغم من اطلاق قاعدة بيانات افريقيا الغربية على نظام إدارتها محرك بحث مخطوطات، إلا أنه فعليا نظام إدارة قاعدة بيانات المخطوطات، ويعزي الباحث الاقتصار على قواعد البيانات كبنية لمواقع الكتابات اليدوية لخاصية التجميع والحفظ التي تمارس مع الكتابات اليدوية عامة والمخطوطات خاصة في مؤسسات ثقافية وتاريخية محددة، بما لا يدع مجالا لهذه المواد في انتشارها كمفردات سابحة في فضاء الويب مقل النصوص والصور وغيرها من مواد الويب.

**ثانيا** : نظرا لما للكتابات اليدوية من قيم تاريخية وثقافية، فإن مؤسسات رعاية الثقافة والتراث تعمل على حفظها واختزانها في قاعدة بيانات محكمة وفقا لمستوى الرقمنة المستخدم مع هذه المواد. ومن ثم لا يتوافق ذلك وأداء محركات البحث التي تعمل في بيئة النظم المفتوحة إضافة وتحكما وضبطا لمختلف مواد معلومات الويب.

**٤ / ١ / ٢ بنية ملف مادة الكتابات اليدوية**

**أولا** : اقتصرت عمليات الرقمنة في أعلى مستوياتها في عينة الدراسة على تحويل صفحات الكتابات إلى صور رقمية للنصوص دون الارتقاء لمستوى القراءة الآلية لتلك النصوص **OCR**. ويرجع الباحث ذلك لأن عمليات المعالجة والبحث المستهدف تقف عند حدود استخدام بيانات محدودة ووصف فني مختصر يمكن من التعريف بالمخطوط دون الطموح في التعامل مع نصوص الكتابات اليدوية بحثا واسترجاع وتحليلا.

**ثانيا** : يمثل استخدام الصور في بعض قواعد بيانات العينة وليس جميعها، سوء إدارة لمواد الكتابات اليدوية وخاصة المخطوطات في وقف استخدامها على قاعدة البيانات دون طرح النص للإتاحة والوصول من جانب الباحثين على مستوى العالم المهتمين بالنصوص المخطوطة. وقد يكون ذلك مبررا في حالة الوثائق القانونية، غير أنه لا يمكن القبول به في حالة المخطوطات والكتب التراثية، كما في حالة مواقع مركز جمعه الماجد والفهرست وافريقيا الغربية التي لم تضع من الأصل صورا لمجموعاتها.

**٤ / ١ / ٣ تعدد الأشكال**

لم تهتم مواقع العينة جميعها بكل أشكال الكتابات اليدوية سوى مكتبة الأسكندرية **bibalex** وشبكة المخطوطات العربية **idsc** بجمع المخطوطات والوثائق وكتب التراث، وهذا يرجع إلى تمثيل هذه المواقع لهيئات مثل مكتبة الأسكندرية وشبكة المخطوطات لمحاولات بناء فهارس رقمية لمجموعات المخطوطات التي توجد على المستوى الإقليمي بالنسبة لمكتبة الإسكندرية وعلى المستوى الوطني لشبكة المخطوطات العربية التي طورها مركز معلومات دعم واتخاذ القرار بمصر. فيما ركزت باقي قواعد بيانات عينة الدراسة على المخطوطات كأشهر أشكال الكتابات اليدوية المخطوطة. وما تستهدفه هذه الدراسة هو طرح تصور لأداة بحث لا تقتصر على شكل مخطوط بعينه، إنما تتعامل وكل الكتابات اليدوية الرقمية على الويب. فيما انفردت مكتبة الجامعة الأمريكية ببيروت بعرض مجموعة من الوثائق التاريخية التي تخص قضايا بعينها، وهو ما يعود إلى اهتمامها الإقليمي فيما يتعلق بتاريخ لبنان تحديدا.

**٤ / ١ / ٤ مستوى الاتاحة**

**أولا** : انعكس الواقع التقليدي الذي عجز عن اتاحة نص المخطوط الكامل ضمن عمليات المعالجة والضبط على واقع عينة الدراسة التي لم تصل بمستوى الإتاحة لعرض النصوص الكاملة للكتابات اليدوية على مستخدميها. فقد ركزت قواعد بيانات العينة جميعها على التعامل مع بيانات الضبط الببليوجرافي دون التعامل مع متون مجموعاتها

المخطوطة، واقتصر جهد مواقع مكتبة الإسكندرية والشبكة العربية للمخطوطات ومكتبة الجامعة الأمريكية ببيروت **AUB** على عرض صورا لنصوص المخطوطات للقراءة من خلال الصور وليس لتحليل النصوص المضمنة داخلها.

**ثانيا :** إن مستوى الإتاحة المستهدف لهذه الدراسة هو قدرة مستخدمي الويب في الوصول إلى نصوص المخطوطات العربية وتحقيق إتاحة النص الكامل من خلال توفير آليات البحث والمعالجة والإدراك اللازمة لذلك. وهو ما لم تقدمه أيا من مواقع عينة الدراسة، فأدوات بحث الكتابات اليدوية المأمولة هي نظم لإدارة المحتوى أكثر من كونها نظام لإدارة البيانات والمعلومات الببليوجرافية. وهي بذلك تمثل درجة أكثر تقدما في التعامل مع الكتابات اليدوية العربية، وتحقيق آفاق أوسع لرقمنة هذا التراث القومي.

**٤ / ٢ محور القراءة والإدراك**

يتناول هذا المحور واقع عمليات القراءة الآلية والإدراك للكلمات والحروف في نصوص مواد الكتابات اليدوية الرقمية ؛ حيث تمثل المرحلة التمهيدية لعملية التكشيف والمعالجة المعلوماتية داخل أدوات بحث الويب. فما لم تتم مرحلة الإدراك للنصوص لا يمكن إجراء مختلف العمليات المعلوماتية المستهدفة من هذه الدراسة. ويوضح جدول رقم (٥) واقع القراءة والإدراك للكتابات اليدوية على الويب.

**جدول رقم (٥) قياس عناصر محور القراءة والإدراك للكتابات اليدوية**

| محور القراءة والإدراك<br>عناصر المحور<br>وتوصيفها | | مواقع قواعد بيانات المخطوطات | | | | | |
|---|---|---|---|---|---|---|---|
| | | العربية | | | الأجنبية | | |
| | | bibalex | idsc | almajid | fihrist | westafrican | AUB |
| آليات قراءة النص | HMM | --- | --- | --- | --- | --- | --- |
| | Base-line | --- | --- | --- | --- | --- | --- |
| المعاجم الآلية | | --- | --- | --- | --- | --- | --- |

| عرض نصوص المخطوطات | نصي | --- | --- | --- | --- | --- | --- |
|---|---|---|---|---|---|---|---|
| | مصور | √ | √ | --- | --- | --- | √ |

يتضح من جدول رقم (٥) مجموعة من المشاهدات والملاحظات وفقا للعناصر التالية:-

**٤ / ٢ / 1 آليات قراءة النص**

**أولا :** لم تأت أي من قواعد بيانات عينة الدراسة بما يدلل على استخدام تقنيات التحليل والإدراك لنصوص الكتابات اليدوية حتى في قواعد البيانات التي تستخدم صورا رقمية لعرض محتوى المادة. وهذا يؤكد النهج البدائي لمواقع إدارة الكتابات اليدوية في استخدام الصور الجامدة للنصوص المخطوطة وعدم صبها في قوالب التحليل والإدارك الرقمي لمحتواها النصي، وهو ما تعكف الدراسة على تحقيقه في أدوات البحث المطروحة.

**ثانيا :** تستخدم نظم استرجاع الكتابات اليدوية آليات ماركوف غير المرئية والخط الرئيس **HMM, Base-line** في الوصول الى تحليل رسومات الكلمات والحروف المصورة بتقنية **OCR**، وكونها غير موجودة في هذه القواعد يعني انتفاء نية اتاحة النصوص الكاملة لمواد الكتابات اليدوية في هذه القواعد والاقتصار على التعامل مع البيانات الببليوجرافية وعناصر الميتاداتا، كاستمرار لما بدأت عليه عمليات الرقمنة الثابتة عند حدود رقمنة فهارس المخطوطات وليس المخطوطات أو الكتابات ذاتها.

**٤ / ٢ / ٢ المعاجم الآلية**

تمثل المعاجم امتدادا طبيعيا لآليات التحليل والإدراك المذكورة في العنصر السابق، لما لهما من تكامل في تحقيق القراءة والإدراك لنصوص وكلمات الكتابات اليدوية. وغياب أي من هاتين الأداتين يعني غياب الأخرى بالتبعية. ويرى الباحث أن ما تقدمه عينة الدراسة وقواعد بياناتها يمثل اختلافا نوعيا عما يحق لنظم استرجاع الكتابات اليدوية العربية أن تكون عليه في تقديم النصوص الكاملة قابلة للإدراك والمعالجة والبحث من جانب مستخدمي الويب.

**٤ / ٢ / ٣ عرض نصوص المخطوطات**

استمرار لعنصر التوافق بين قواعد بيانات العينة في محور القراءة والإداراك، فقد انتفى عرض النصوص الكاملة لمتن الكتابات اليدوية في قواعد بيانات عينة الدراسة. وذلك لما لاحتياج التعامل مع النص من الخروج من ضيق رقمنة البيانات إلى أفق رقمنة النصوص

ذاتها. وقد اقتصر التعامل مع نصوص المجموعات المختزنة في بعض قواعد البيانات على استخدام الصور الرقمية لنصوص الكتابات اليدوية وعرض النص كصورة أمام مستخدمي قاعدة البيانات. واستخدمت قواعد مكتبة الاسكندرية وشبكة المخطوطات العربية ومكتبة الجامعة الأمريكية ببيروت ملفات الصور الرقمية لعرض نصوص مجموعاتها.

### ٤ / ٣ محور المعالجة والتكشيف

يمثل هذا المحور قلب إدراة نظم استرجاع المعلومات لما له من انعكاس على بنية النظام ككل من حيث المعايير المطلوب توفيرها للعمل داخل النظام، فضلا عن أساليب وبنية التكشيف والتنظيم ومن ثم واقع أساليب البحث والاسترجاع التي تتوافق ومستخدمي الويب في استرجاع الكتابات اليدوي. ويوضح جدول رقم (٦) واقع التكشيف والتنظيم للكتابات اليدوية على الويب من حيث معايير الوصف وآليات التكشيف والتنظيم المتبعة حاليا.

**جدول رقم (٦) قياس عناصر محور المعالجة والتكشيف لكتابات اليدوية**

| محور المعالجة والتكشيف عناصر المحور وتوصيفها | | مواقع قواعد بيانات المخطوطات | | | | | |
|---|---|---|---|---|---|---|---|
| | | العربية | | | الأجنبية | | |
| | | bibalex | idsc | almajid | fihrist | westafrican | AUB |
| معيار الوصف | رقمي | √ | -- | --- | --- | --- | --- |
| | AACR | √ | √ | --- | --- | √ | --- |
| عناصر الوصف الفني المستخدمة لوصف الكتابات اليدوية (المخطوطات) من خلال التسجيلات النهائية لمواد | العنوان | √ | √ | √ | √ | √ | √ |
| | المؤلف | √ | √ | √ | √ | √ | √ |
| | الموضوع | √ | √ | √ | √ | √ | ملخص |
| | تاريخ الصورة | √ | -- | --- | --- | --- | --- |
| | تاريخ المسح | √ | -- | --- | --- | --- | --- |
| | تاريخ التسجيلة | √ | -- | --- | --- | --- | --- |
| | الناشر | √ | -- | --- | --- | --- | --- |
| | شكل | √ | -- | --- | √ | --- | --- |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **المخطوطات والمقترحة في الدراسة** | **الملف** | | | | | | |
| | **حجم الملف** | √ | -- | --- | --- | --- | --- |
| | **حقوق النشر** | √ | -- | --- | --- | --- | --- |
| | **الناسخ** | √ | √ | --- | --- | √ | --- |
| | **النوع** | √ | √ | --- | --- | √ | --- |
| | **النص الكامل** | **صور** | **صور** | --- | --- | --- | **صور** |
| | **الهيئة المسئولة** | √ | √ | √ | √ | √ | √ |
| | **URL** | **ID** | **ID** | **ID** | **ID** | **ID** | --- |
| **أسلوب التكشيف** | **محتوى** | --- | -- | --- | --- | --- | --- |
| | **نصي** | √ | √ | √ | √ | √ | --- |
| **آلية التكشيف** | **نص المادة** | --- | -- | --- | --- | --- | --- |
| | **الميتاداتا** | √ | √ | √ | √ | √ | --- |

يتضح من جدول رقم (٦) مجموعة من المشاهدات والملاحظات وفقا للعناصر التالية:-

**٤ / ٣ / ١ معيار الوصف الفني**

تدلل نوعية معيار الوصف الفني المستخدم في قاعدة البيانات على توجه النظام في التعامل مع الكتابات اليدوية من حيث المعالجة والتنظيم والبحث ؛ فالقواعد التي تحافظ على استخدام قواعد **AACR** ينظر إليها على أنها قواعد الرقمنة المحدودة التي تتوقف عند تحويل التسجيلات الببليوجرافية من شكلها التقليدي إلى الشكل الرقمي دون التطرق الى نص المادة المخطوطة. وقد تمثل ذلك في جل عينة الدراسة، ناهيك عن تلك التي اعتمدت على بيانات مختصرة لا ترتقي وقواعد الضبط الببليوجرافي. أما قاعدة بيانات مكتبة الإسكندرية **bibalex** فهي الوحيدة بين قواعد بيانات العينة التي طبقت مفهوم الميتاداتا ورقمنة المخطوطات من حيث الوصف الفني المطول والمقسم إلى أجزاء مصنفة تخص الشكل والبيانات الاساسية والمحتوى. وقد وصل الأمر إلى أدناه في مكتبة الجامعة الأمريكية ببيروت **AUB** التي لم تأت بشكل يذكر لمعايير الوصف، بل اعتمدت على ذكر العنوان والمؤلف مع شرح مختصر ثم صورة رقمية للمخطوط.

### ٤ / ٣ / ٢ عناصر الوصف المستخدمة

**أولا** : تتناسب نوعية وعدد العناصر المستخدمة لوصف الكتابات اليدوية الرقمية تناسبا طرديا مع مستوى الرقمنة المتبع داخل النظام ؛ حيث تتنوع العناصر المطلوبة لوصف ملفات الكتابات اليدوية كلما زادت عمليات المعالجة بذكر خصائص المادة التقليدية وخصائصها الرقمية وبيانات ملف حمل المادة المصورة مع البيانات الببليوجرافية التي توافق المخطوط كمادة معلومات. وقد حددت الدراسة مجموعة عناصر الوصف المقترحة لمعالجة الكتابات اليدوية كمقياس لعينة الدراسة، حيث شملت العناصر المقترحة مزيجا بين تلك التي تتوافق ومصادر الكتابات اليدوية العربية وبين العناصر المستخدمة لمعالجتها في البيئة الرقمية.

**ثانيا** : اتفقت مختلف قواعد بيانات عينة الدراسة في عناصر الوصف التي أخذت بها والعناصر التي لم تستخدمها، إلا مكتبة الاسكندرية التي نحت منحا رقميا لا يقارن مع باقي قواعد بيانات العينة ؛ حيث أخذت قواعد البيانات جميعها بعناصر المؤلف والعنوان والموضوع وزادت قاعدة بيانات افريقيا الغربية في عنصرين عن الناسخ ونوع المادة، بينما حققت مكتبة الإسكندرية **bibalex** كل عناصر الوصف المقترحة. ويعزي الباحث ذلك إلى دعم قاعدة بيانات مكتبة الإسكندرية للمفهوم الحقيقي للميتاداتا، فضلا عن استمرارية العمل والتحديث التقني والفني والمستمرين لقاعدة بياناتها، في حين تشير مواقع باقي عينة الدراسة الى توقف التحديث عند أعوام ٢٠٠٦ و٢٠١٢ في بعض قواعد البيانات الأخرى.

**ثالثا** : تمثل عناصر الوصف البنية الأساسية لملفات الكتابات اليدوية المقترح بناؤها رقميا، التي ستتعامل مع بيئة الويب المفتوحة وأدواتها تنظيما وبحثا بما سيجعل مادة المعلومات المخطوطة تنتشر سابحة بين قواعد بيانات محركات البحث حاملة مختلف عناصر الوصف الفني التي تحكي قصة المادة التاريخية بما في ذلك حقل **content** لحمل النص الكامل في شكل نصي. وليس في شكل ملف صورة كما تعاملت معه قواعد بيانات عينة الدراسة التي أتاحت نصوص الكتابات المصورة.

### ٤ / ٣ / ٣ أسلوب التكشيف

**أولا** : لم تحقق كل قواعد بيانات عينة الدراسة مفهوم التكشيف في بيئة الويب داخل مجموعاتها، حيث استخدمت عينة الدراسة جميعها مفهوم رؤوس الموضوعات للإشارة الى الموضوعات والأفكار الرئيسة الواردة داخل مجموعات المخطوطات. وقد زادت عن

ذلك مكتبة الإسكندرية **bibalex** فقط في عرض قائمة محتويات المخطوطات بما يمكن أن يمثل تكشيفا عاما وليس على مستوى النص الكامل. في حين لم تقترب مكتبة الجامعة الأمريكية ببيروت **AUB** من مفهوم التحليل الموضوعي إلا في عرض شرح مختصر لمحتوى المخطوطة.

**ثانيا** : ترتبط آلية تكشيف النص الكامل المستهدف في هذه الدراسة والمؤدي الى تحليل مختلف الكلمات والفقرات اعتمادا على قدرات المسح الرقمية وآليات الاشتقاق لجعل مختلف كلمات النص مداخل كشفية يمكن من خلالها بحث واسترجاع نصوص الكتابات اليدوية. ولم تستخدم آليات الإدراك **Recognotion** في الوصول الى نصية الكتابات اليدوية ومن ثم التمكن من تكشيفها بأسلوب التكشيف الاشتقاقي.

**ثالثا** : لم تحقق عينة الدراسة أسلوب التكشيف المبني على المحتوى المعتمد على التعامل مع ملفات الصور الرقمية وقراءة عناصرها الرسومية وأشكال الكلمات باستخدام آليات اللون والبنية والأبعاد في كتابات حروف اللغة العربية، وهو ما يمثل نوعا متقدما من التقنيات والبرمجيات التي مازالت قيد الدراسة والتجريب في العالم الغربي.

### ٤ / ٣ / ٤ آلية التكشيف

اعتمدت قواعد عينة الدراسة على التحليل والتكشيف من خلال التمثيل المكثف باستخدام الميتاداتا التي تصاحب المادة المخطوطة وليس النص الكامل للمادة، وهو ما يعكس مدى البعد عن الموضوعات والمحتوى الموضوعي الوارد في نصوص الكتابات اليدوية بما ينعكس نهاية بالسلب على تنظيم وبحث الكتابات اليدوية على الويب.

### ٤ / ٤ محور البحث والاسترجاع

تركز الدراسة في هذا المحور على النظر لواقع آليات وأساليب البحث والاسترجاع المستخدمة مع مواد الكتابات اليدوية حيث يمكن من خلالها الحكم على أسلوب بنية وعمل نظام استرجاع الكتابات ككل. فتوفير البحث الحر يعكس القراءة والادراك للنص الكامل لمادة الكتابات اليدوية من مخطوطات ووثائق، أيضا استخدام الحقول في البحث يعكس معيارية الوصف الفني للكتابات اليدوية. ويوضح جدول (٧) شكل وواقع البحث والاسترجاع للكتابات اليدوية.

**جدول رقم (٧) قياس عناصر محور بحث والاسترجاع الكتابات اليدوية**

| محور البحث والاسترجاع | مواقع قواعد بيانات المخطوطات | |
|---|---|---|
| | العربية | الأجنبية |

| عناصر المحور وتوصيفها | | bibalex | idsc | almajid | fihrist | westafrican | AUB |
|---|---|---|---|---|---|---|---|
| البحث عن المادة | النص | --- | -- | --- | --- | --- | --- |
| | الميتاداتا | √ | √ | √ | √ | √ | --- |
| أسلوب الاسترجاع | النصي | √ | √ | √ | √ | √ | --- |
| | المحتوى | --- | √ | --- | --- | --- | --- |
| أساليب البحث | الحـــــر | √ | √ | √ | √ | √ | --- |
| | التصفح | √ | √ | --- | √ | --- | √ |
| البحث بالحقول | العنوان | √ | √ | √ | √ | √ | --- |
| | المؤلف | √ | √ | √ | √ | √ | --- |
| | الموضوع | √ | √ | √ | √ | √ | --- |
| | الناسخ | √ | √ | --- | --- | --- | --- |
| | رقم الطلب | √ | -- | √ | --- | --- | --- |
| | التاريخ | √ | -- | --- | --- | √ | --- |
| | محددات البحث | --- | √ | √ | √ | --- | --- |
| آليات البحث | البولياني | √ | √ | √ | --- | √ | --- |
| | التطابق | √ | √ | √ | √ | √ | --- |
| | المقاطع | √ | √ | √ | --- | --- | --- |
| آليات بحث اللغة العربية | جذر الكلمة | --- | -- | --- | --- | --- | --- |
| | المقابل | --- | -- | --- | --- | --- | --- |
| | المعنى | --- | -- | --- | --- | --- | --- |
| | المترادفات | --- | -- | --- | --- | --- | --- |

يتضح من جدول رقم (٧) مجموعة من الملاحظات المستخلصة وفقا للعناصر التالية :-

**٤ / ٤ / ١ البحث في نص المادة**

هدفت الدراسة من خلال تحري واقع هذا العنصر معرفة التوجه العام لدى عينة الدراسة في بحثها لمواد الكتابات اليدوية ؛ حيث تختار نظم الكتابات اليدوية أحد خيارين ؛ فإما استخدام البحث في النص الكامل شأن محركات البحث مع صفحات الويب، وهو ما يدلل على تطويع النص الكامل للكتابات اليدوية لعملية البحث والاسترجاع، بما يعكس قدرات التحليل والمعالجة الكاملة للكتابات اليدوية. وإما أن تتجه النظم إلى البحث باستخدام الميتاداتا بمعزل عن نص ومحتوى الكتابات اليدوية، وهو ما يدلل على البعد التام عن محاولات المعالجة والبحث لنصوص الكتابات اليدوية. وقد أكد واقع عينة الدراسة ما تنتهجه قواعد الكتابات اليدوية في التركيز على المعالجة والبحث باستخدام الميتاداتا والبعد التام عن معالجة واسترجاع المحتوى والنصوص الكاملة للكتابات اليدوية، نظرا لما تحتاجه من قدرات إدراك وتحليل للنصوص المخطوطة. وقد ظهر ذلك جليا في مكتبة الجامعة الامريكية ببيروت التي اعتمدت على تصفح موادها فقط دون الاعتماد على خيارات البحث.

٤ / ٤ / ٢ أسلوب الاسترجاع

**أولا** : يعتمد بحث مواد المعلومات المصورة عامة على استخدام أسلوبي البحث بالنص والبحث بالمحتوى، ونظرا لطبيعة الكتابات اليدوية التي تتمثل في شكل نصوص مصورة بتقنية **OCR** فإن البحث بالمحتوى يعد أساسيا لنظم إدارة الكتابات اليدوية على الويب. واستمرار للمعالجة المحدودة غير المتعمقة للكتابات اليدوية ونصوصها بشكل مباشر، فقد اعتمدت عينة الدراسة على استخدام البحث بالنص من خلال حقول الميتاداتا لبحث واسترجاع الكتابات اليدوية. وانفردت شبكة المخطوطات العربية **idsc** وحدها بتبني أسلوب التصفح بصور المخطوطات كأحد أشكال البحث بالمحتوى من خلال عرض المجموعات للتصفح المرئي من جانب المستفيدين تحت مسمى ألبوم المخطوطات، على الرغم من كونه يشمل مجموعة محدودة فقط.

**ثانيا** : عكس واقع عينة الدراسة جميعها غياب آليات المعالجة والإدراك اللازمين لنصوص الكتابات اليدوية المصورة، فالكلمات والحروف في الكتابات اليدوية الرقمية تمثل رسوما وأشكالا غير مقروءة من أدوات بحث الويب. ومن ثم فإن بحث واسترجاع نصوص الكتابات المخطوطة على طبيعتها المصورة يتطلب استخدام آليات بحث الصور الرقمية وتحليل محتواها الشكلي. وباستخدام آليات تحليل صور محتوى الكتابات اليدوية

ستتمكن أدوات البحث من استخدام آليات البحث بشكل الكلمة والبحث بنماذج الحروف وصور الكلمات، فضلا عن البحث بتصفح صور النصوص المخطوطة.

**٤ / ٤ / ٣ أساليب البحث**

تتنوع أشكال بحث الكتابات اليدوية لتصب في أسلوبين أساسيين هما البحث الحر باستخدام الكلمات المفتاحية، والبحث بالتصفح باستخدام التوزيع الهرمي للمصطلحات أو للأشكال في حالة المواد المرئية. وقد أكدت عينة الدراسة على ثابت استخدام البحث بالكلمات المفتاحية في عينة الدراسة مع الاستعانة بأسلوب البحث بالتصفح في بعض قواعد بيانات الكتابات اليدوية. وقد اختلفت قواعد بيانات البيانات فيما بينها في شكل التصفح المستخدم مع الكتابات اليدوية ؛ حيث تبنت مكتبة الاسكندرية والفهرست استخدام التصفح بالمصطلحات التي تتنوع بين التصفح بالمؤلفين أو التصفح بتواريخ المخطوطات أو التصفح بالحروف الهجائية لعناوين الكتابات اليدوية، أما قاعدتا الشبكة العربية للمخطوطات ومكتبة الجامعة الأمريكية ببيروت فقد استخدمتا التصفح بصور المخطوطات في شكل لقطات مصورة. ولعل غياب الاهتمام بجانب تحليل وبنية ومعالجة ملفات الكتابات اليدوية ذاتها ونصوصها المصورة هو تفسير غياب التصفح والبحث بمحتوى مواد المعلومات المخطوطة.

**٤ / ٤ / ٤ البحث بحقول الوصف**

**أولا** : مثّل البحث بالحقول في عينة الدراسة انعكاسا لمدى الاعتماد على عناصر الميتادات في وصف الكتابات اليدوية وتنظيمها، وتزاداد قدرات البحث بالحقول في قواعد البيانات وفقا لعدد الحقول النشطة للبحث والاسترجاع من جملة حقول الوصف المكونة للتسجيلة النهائية لمادة المعلومات. وقد توافقت قواعد بيانات العينة على البحث استخدام حقول العنوان والمؤلف والموضوع في البحث عن الكتابات اليدوية، في حين زاد بعضها باستخدام حقول الناسخ والتاريخ ورقم الطلب **ID** مثل قاعدة بيانات مكتبة الاسكندرية الأكثر عددا لحقول البحث، ثم افريقيا الغربية وشبكة المخطوطات العربية

**ثانيا** : يرتبط باستخدام البحث بالحقول آلية محددات البحث، وذلك لضبط وتدقيق هذا استراتيجية البحث والتحكم التام في النتائج المسترجعة. وتمثل محددات البحث في طبيعتها حقولا إضافية للبحث لم تترك للمستخدم بوضع الكلمات المفتاحية بها، إنما بختار النظام اتاحتها من خلال قائمة محددة من خيارات البحث تساعد على ضبط النتائج المسترجعة، مثل محدد اللغة لاختيار لغة المجموعات المسترجعة دون غيرهان أو محدد التاريخ أو

محدد الموضوع والفئة العامة لنوعية الكتابات اليدوية. وقد عكس واقع الدراسة أن قواعد البيانات التي لم تستخدم حقولا للبحث كثير قد استعاضت عن ذلك بإتاحة محددات البحث باللغة وغيرها محقول بحث إضافية.

**٤ / ٤ / ٥ آليات البحث المستخدمة**

استخدمت قواعد البيانات العربية من عينة الدراسة آليات البحث مثل البحث بالتطابق والبحث بمقاطع الكلمات أكثر من نظيرتها الأجنبية. وذلك يرجع إلى وضع طبيعة البحث بالكلمات العربية التي تتسم بالميل إلى اللغة الطبيعية أكثر من البحث باللغة المقيدة، ومن ثم كانت آليات البحث للضبط والتحكم في الكلمات المفتاحية المدخلة في البحث. وترتبط آليات البحث بأساليب البحث المستخدمة لصياغة استراتيجيات وشكل البحث المعبر عن الحاجات المعلومات للمستفيدين، حيث يمكن من خلال هذه الآليات التحكم باستبعاد واستدعاء وتوسيع نتائج البحث وفقا للحاجة.

**٤ / ٤ / ٦ آليات بحث اللغة العربية**

خلت عينة الدراسة بقواعد بياناتها العربية والأجنبية من آليات البحث الدالة على معالجة النصوص العربية في مجموعات الكتابات اليدوية، وعلى الرغم من أن مجموعات المعلومات المخطوطة باللغة العربية، غير أنه لم تظهر في صفحات البحث لهذه النظم أي استخدام للبحث باللغة العربية مثل البحث بالمعنى أو بجذر الكلمة أو بالترادف. ويرجع الباحث ذلك إلى غياب تام لأدوات تحليل اللغة العربية في قواعد بيانات العينة من المعاجم الآلية والمحللات الصرفية. وهذا بدوره يعود إلى غياب مبدأ تحليل وإدراك نصوص الكتابات العربية التي يجب أن تعتمد في معالجتها وبحثها على أدوات اللغة للتوافق والبنية اللغوية التي أتت عليها هذه الكتابات. هذا فضلا عن استخدام قواميس لغوية تتوافق وطبيعة العصر والمكان اللذين شهدهما هذا الانتاج الفكري من المخطوطات والوثائق.

## المبحث الخامس : الرؤى والنتائج والتوصيات

**٥ / ٢ رؤى ونتائج الدراسة**

١) أن العالم العربي كوحدة ثقافية واحدة مشتركة الهوية لم تكن على قدر الاهتمام المطلوب في التعامل مع المواد التراثية العربية من مخطوطات وكتب تراثية ووثائق تاريخية كونها مواد معلومات تحمل قيمة معرفية وتاريخية وحضارية،

وتحتاج هذه المواد المخطوطة إلى التقانين والمعايير المتوافقة وبنيتها الرقمية وإختلاف تنظيمها ومعالجتها وبحثها في بيئة الويب عن البيئة التقليدية التي لبثت فيها عمرا سابقا. ويأت التقعيد والتقنين من خلال مجموعات علمية تاريخية ثقافية مشتركة عبر المؤسسات القومية العربية.

٢) أن الكتابات اليدوية مواد معلومات تقليدية النشأة ورقمية الإدارة والتنظيم بما يجعلها تحتاج الى آليات تكشيف وبحث واسترجاع تختلف عن طبيعة أدوات بحث الويب الآنية، بما يدفع القائمين على الويب **W3C** لتحديث أنماط وآليات أدوات بحث الويب في تعاملها مع مواد معلومات كالكتابات اليدوية– بالإضافة لاتخاذ التدابير التقنية والدراسات المكثفة في ضبط معالجة واسترجاع نصوص الكتابات اليدوية على الويب.

٣) أن الكتابات اليدوية العربية هي مواد تختلف بين العلمية والقانونية والاجتماعية، ومن ثم فإن طبيعة التكشيف المطلوب لتحليل هذه يتعامل مع مفردات وأساليب تعكس هوية وثقافة محددة. ولا يمكن الاعتماد هنا على استخدام تقنيات التحليل الرقمية وحسب، وإنما متابعة مخرجات النصوص الناتجة عن عملية التحليل لاقامة بنية المادة العلمية والسياق النصي كما أراد له كاتبوه.

٤) أن إدارة محتوى مواد الكتابات اليدوية العربية مازالت ترتكن إلى الباحثين والمتخصصين غير العرب لإجراء عمليات إدارك النصوص وتحليل المحتوى الفكري، مما أثر بطبيعته على معالجة وتكشيف النصوص العربية التراثية، فجعل حدود المعالجة لهذه المواد التراثية تقف عند حدود النواحي التقنية الرقمية والتعامل مع النصوص كصور وأشكال رسومية دون التعمق في فهم وإدراك المحتوى الفكري والتراث المعرفي بين سطور هذ الكتابات.

٥) أن مواد الكتابات اليدوية كمصادر معلومات لا تتوقف فقط عند المجموعات التاريخية ذات النشأة التقليدية، إنما تتعدى ذلك إلى المجموعات ذات النشأة الرقمية المنتجة بواسطة أدوات الكتابة اليدوية الرقمية الحالية. مما يعكس تنامي حجم وأشكال الكتابات اليدوية الرقمية ووقوف محركات وأدوات بحث الويب عند هذه المواد بالدراسة والمعالجة لضبط بحثها واسترجاعها على الويب. مما يجعل هذه الأدوات متأخرة في ملاحقة تداول وانتشار هذه المواد في بيئة الويب.

٦) أن مختلف محاولات الضبط والتنظيم لمواد الكتابات اليدوية التي ركزت على المخطوطات دون الوثائق قد عملت على استخدام أساليب العرض الرقمية المبسطة دون تكوين قواعد بيانات معيارية مهيكلة لمعالجة وإدراك محتوى الكتابات اليدوية. بما انعكس على مجموعات قواعد البيانات على الويب بعدم وجود آليات بحث للنصوص والمحتوى الفكري لها. ويتيح لأدوات بحث الويب فيما بعد التوافق وبحص هذه المواد.

٧) أن مواد الكتابات اليدوية مواد معلومات تعود في كتابتها وتأليف إلى عصور زمنية تختلف اجتماعيا وعلميا عن عصر الويب، مما جعلها تختلف في أسلوب الكتابة والصياغة والبناء اللغوي لمحتوى المخطوطات والوثائق. ومن ثم فإن عمليات تكشيف وتحليل محتوى نصوص هذه المواد يحتاج بناء معاجم لغوية ومحللات صرفية لا تتمع فقط بالقدرة على تحليل الكلمات لغويا، وإنما القدرة على فهم المفردات والمصطلحات التاريخية التي تنتمي لعصور زمنية راجعة. وذلك حتى يمكن الوقوف على محتواها الفكري والتراثي المستهدف ايصاله لقارئيه. وقد خلت عينة الدراسة من أساليب بحث اللغة العربية بما يدلل على غياب أدوات معالجتها.

٨) أن مواقع المخطوطات المعتمدة على قواعد بيانات تقدم معالجة وآليات لبحث مواد الكتابات اليدوية أكثر معيارية من تلك التي تتبعها محركات البحث التي تكاد على تقدم معالجة أو بحثل يدعم الكتابات اليدوية. وذلك لكون محركات البحث تتعامل مع فضاء الويب غير المحدود وما ينتشر فيه من مواد معلومات غير معيارية البناء الرقمي، بما انعكس على عدم وجود آليات لبحث الكتابات اليدوية، فضلا عن معياريتها وقدرتها على التعامل مع هذه النوعية من مواد المعلومات.

٩) أن عملية الإدراك **Recognotion** هي لب بناء نظم استرجاع الكتابات اليدوية ؛ فبدونها لا يمكن لعمليات التكشيف والتحليل أن تتم، فالتكشيف لا يتم الا على الكلمات النصية التي تعطي مفاهيم ومعان محددة يتم اختيارها واعتبارها مداخل كشفية أو مصطلخات دالة على المضمون الفكري للنص. ولا يمكن للنص أن يولد في شكل كلمات وحروف نصية إلا من خلال عملية الإدراك التي تحول الكلمات والحروف المصورة الى كلمات وحروف نصية تتوافق وعمليات التكشيف الآلي للنصوص.

١٠) أن بيئة الويب هي بيئة ديناميكية سريعة الدمج والتحول، وقد أفرز ذلك خروج مواد الكتابات اليدوية الرقمية على الويب التي جمعت في بنائها الرقمي بين خصائص المعلومات النصية والمعلومات المصورة في مادة معلومات واحدة. يمكن أن يطلق على توصيفها المعلومات نصوص مصورة أو صور نصية تحتاج بدورها الى آليات وأساليب معالجة وبحث ذات طبيعة مختلفة عن تلك التي اعتادت عليها أدوات بحث الويب.

١١) أثبتت الدراسة التجريبية أن مختلف قواعدة بيانات عينة الدراسة يرجع إلى بنية مؤسسية ذات صبغة علمية أو ثقافية، مما يعكس بدوره غياب الاتجاه العام للويب وادوات بحثها وتأخرها في معالجة وبخث واسترجاع مجموعات الكتابات اليدوية على الويب بشكل عام. وهو ما يؤكد على أن الاهتمام بمواد الكتابات اليدوية الرقمية على الويب مازال جزئيا أو مبتورا لا يتوافق وانتشار هذا الشكل الجديد من مواد حمل المعلومات الرقمية على الويب.

**٥ / ٣ توصيات الدراسة**

١. يجب على المؤسسات القومية العربية الاهتمام برقمنة ومعالجة التراث العربي ومواده المعلوماتية من الكتابات اليدوية في بيئة الويب، وذلك من خلال الهيئات الثقافية مثل المنظمة العربية للتربية والثقافة والعلوم أو المؤسسات المتخصصة في حفظ التراث العربي. ويتجلى هذا الاهتمام في طرح مشروعات قومية لرقمنة الكتابات اليدوية العربية والاشراف على بناء نظم استرجاع قومية تجمع التراث العربي الرقمي من الكتابات اليدوية.

٢. يجب أن تضطلع مؤسسات الويب مثل **W3C** بدورها في متابعة مستحدثات مواد المعلومات في بيئة الويب، والدأب على ابتكار أدوات وآليات بحث تتوافق ومواد المعلومات الرقمية ذات الخصائص المختلفة عما عليه بيئة الويب مثل مواد الكتابات اليدوية العربية.

٣. يجب على المؤسسات البحثية وأقسام علوم المعلومات الاهتمام بالتغيرات السريعة المتوالية في شكل ونوعية مواد المعلومات في بيئة الويب، ومن ثم ملاحقة هذه التطورات بتجهيزات الدراسة والبحث والتقعيد. وقد لمست دراستنا الحالية هذه الحاجة في حالة مواد الكتابات اليدوية العربية التي افتقدت الى معايير ومقاييس علمية مقننة.

٤. يجب على القائمين على تشريعات المعلومات وحماية التراث القومي العربي خاصة المضي في صياغة حزمة قوانين وتشريعات تتوافق واتاحة مواد الكتابات اليدوية العربية في بيئة الويب، وذلك لضبط تداولها واستخدامها وفي بيئة مفتوحة التعامل مع مواد المعلومات وأدوات بحثها.

**مصادر الدراسة :**


1 Khodadadzadeh , Iman. Recognition of Handwritten Arabic Characters, The University of Windsor, Canada, phd, 2010, UMI dissertation abstracts, 2011, Cited at 28/8/2014, cited at http://search.proquest.com/pqdtglobal/docview/753900059/fulltextPDF?source=fedsrch&accountid=37552

2 Boukharouba , Abdelhak and Bennia , Abdelhak. Recognition of Handwritten Arabic Literal Amounts Using a Hybrid Approach, Springer Science+Business Media, LLC 2010, cited at 28/8/2014, cited ar http://link.springer.com/article/10.1007/s12559-010-9088-6

3 Zaher al aghbari and Salama brook. HAH : A holistic paradigm for classifying and retrieving historical Arabic handwritten documents, expert systems and applications, ELSEVIER, 2009, cited at 21/10/2014, cited at http://www.sciencedirect.com/science?_ob=ArticleListURL&_method=list&_ArticleListID=630377253&_sort=r&_st=13&view=c&md5=ed939358c10c0d84af2f9e52664db0f3&searchtype=a

4 Ziaratban, Majid and Faez, Karim. Nona-uniform slant estimation and correction for Farsi/Arabic handwritten word, Springer-Verlag 2009, cited at 25/8/2014, cited at http://link.springer.com/article/10.1007/s10032-009-0092-x

5 Srihari, Sargur & Srinivasan, Harish & Pavithra Babu and Bhole, Chetan. Handwritten Arabic Word Spotting using the CEDARABIC Document Analysis System, State University of New York , 2010, cite at 25/8/2014, cited at http://www.sciencedirect.com/science/article/pii/S0890540185711054

6 Alkhateeb, Jawad H and Ren, jinchang. Offline handwritten Arabic cursive text recognition using hidden markov models and re-ranking, Elsevier B.V., 2011, cited at 1/9/2011, cited at http://www.sciencedirect.com/science/article/pii/S0167865511000432

7 Akinwonmi, A.E. Adewale, M.Tech., O.S. Design of a Neural Network Based Optical Character Recognition System for Musical Notes, The Pacific Journal of Science and Technology, 2008, cited at 20/10/2014, cited at http://link.springer.com/chapter/10.1007/978-3-540-77226-2_45

8 Stoliński, Sebastian. Application of OCR systems to processing and digitization of paper documents, Computer Engineering Department, Technical University of Łódź, Poland, 2006, cited at 20/10/2014, cited at http://www.google.com.eg/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0CCsQFjAB&url=http%3A%2F%2Fsstolin.kis.p.lodz.pl%2Fdane%2Fpub%2F10_isim_ocr.pdf&ei=f3oFVKLJHqX5yQPi3oCYAw&usg=AFQjCNGhTzYvivpt_sQG1uJALWfc9UsCgA

9 Bhatia, Neetu. Optical Character Recognition Techniques: A Review, International Journal of Advanced Research in Computer Science and Software Engineering, 2014, cited at 20/10/2014, cited at http://scholar.google.com.eg/scholar?q=arabic+handwritten+retrieval&hl=ar&as_sdt=0,5&as_vis=1

10 Hu, Zhijuan and Wu, Lu. Research on OCR Post-processing Applications for Handwritten Recognition Based on Analysis of Scientific Materials, springer, 2011, cited at 30/9/2014, cited at http://link.springer.com/chapter/10.1007/978-3-642-23777-5_22

11 سيد ربيع سيد ابراهيم. محركات بحث الصور الثابتة على الانترنت : دراسة تحليلية. الرياض : مكتبة الملك فهد الوطنية ، ٢٠٠٧.

12. نفس المصدر السابق

13 Li, Yi & Zheng, Yefeng and David Doermann. Script-Independent Text Line Segmentation in Freestyle Handwritten Document, transactions on paterin analysis and machine intelligence, 2008, cited at 10/10/2014, cited at http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4359385

14 Alma'adeed, Somaya and Higgins, Colin. Off-line recognition of handwritten Arabic words using multiple hidden Markov models, University of Nottingham, Nottingham NG8 1BB, UK, 2004, cited at 1/9/2014, cited at http://db5.eulc.edu.eg/eulc_v5/libraries/start.aspx?ScopeID=1.&fn=portal&DefaultLang=

15 Tagougui, Najiba , Kherallah, Monji and Alim, Adel M. Online Arabic handwriting recognition: a survey, Springer-Verlag 2012, cited at 12/10/2014, cited at http://link.springer.com/article/10.1007/s10032-012-0186-8

16 Alma'adeed, Somaya and Higgins, Colin.OP. CT.

17 Mozaffari, Saeed and Faez, Karim. Two-stage lexicon reduction for offline arabic handwritten word recognition, International Journal of Pattern Recognition, World Scientific Publishing Company, 2008, cited at 30/9/2014, cited at http://db5.eulc.edu.eg/eulc_v5/libraries/start.aspx?ScopeID=1.&fn=portal&DefaultLang=

18 Bernard, Bianne and Laure, Annev and Menasri, Fare `s and Mohamad, Rami Al-Hajj. Dynamic and Contextual Information in HMM Modeling for Handwritten Word Recognition, the IEEE Computer Societ, 2011, cited at 20/11/20134, cied at http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5708152

19 Abandah, Gheith A. And Khedher, Mohammed Z. Analysis of Handwritten Arabic Letters Using Selected Feature Extraction Techniques, International Journal of Computer Processing of Languages, World Scientific Publishing Company, 2009, cited at 14/10/2014, cited at http://db5.eulc.edu.eg/eulc_v5/libraries/start.aspx?ScopeID=1.&fn=portal&DefaultLang=

20 ˜a-Boquera, Salvador Espan and Castro-Bleda , Maria Jose. Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Model, IEEE transactions on paterin analysis and machine intelligence, 2011, cited at 30/9/2014, cited at http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5551147

21 Abandah, Gheith A. and Jamour, Fuad T. A Word Matching Algorithm in Handwritten Arabic Recognition Using Multiple-Sequence Weighted Edit Distances

IJCSI International Journal of Computer Science , 2014, cited at 13/10/2014, cited at http://link.springer.com/article/10.1007/s10032-013-0201-8

22 Ouwayed, Nazih and Abdel Bela¨ıd. Separation of Overlapping and Touching Lines within Handwritten Arabic Documents, Springer-Verlag Berlin Heidelberg, 2009, cited at 20/10/2014, cited at http://link.springer.com/chapter/10.1007/978-3-642-03767-2_29

23 Huang, Chen. Content-based handeritten document indexing and retrieval, the State University of New York, phd, UMI ProQuest Information and Learning Company, 2008, cited at 25/10/2014, cited at http://search.proquest.com/pqdtglobal/docview/304373118/fulltextPDF?source=fedsrch&accountid=37552

24 نظام إسترجاع المعلومات الكونى.- الرياض : مكتبة الملك فهد الوطنية ، ٢٠٠٩. : سيد ربيع سيد ابراهيم. الويب

25 . الرياض : مكتبة الملك فهد : سيد ربيع سيد ابراهيم. نظم إسترجاع قواعد بيانات الويب غير المرئية الوطنية، ٢٠١٠.

26 Mattison, david. Images of History on the Web. informataion today,inc, 2002, cited at: 15/10/2014, cited at http://www.infotoday.com/searcher/may02/mattison.htm

27 نظام إسترجاع المعلومات الكونى. مصدر سابق. : سيد ربيع سيد ابراهيم. الويب

28 AL-BADR , Badr H. Using the Internet in Arabic: Problems and Solutions , King Abdul aziz City for Science and Technology , [2009] , cited at 22/11/2014 , cited at http://www.isoc.org/inet98/proceedings/5f/5f_1.htm

29 Semmar, naserdine. A cross language information retrieval system, laboratory of multilanguage multimedia, 2005, cited at 11/11/2014, cited at www-list.cea.fr/.../gb/Alger2005_semmar_a_cross_language_information_retrieval_system.pdf

30 Srihari, Sargur & Huang, Chen and Srinivasan, Harish. Content-based Information Retrieval from Handwritten Documents, National Institute of Justice grant 2002, cited at 29/10/2014, cited at http://link.springer.com/article/10.1007/s10032-010-0124-6

31 Rodríguez-Serranoa, José A. And Perronnin , Florent. Handwritten word-spotting using hidden Markov model sand universal vocabularies, Elsevier Ltd. 2009, cited at 25/10/2014, cited at http://www.sciencedirect.com/science?_ob=ArticleListURL&_method=list&_ArticleListID=-

634286345&_sort=r&_st=13&view=c&md5=a9263c353414c46f7e0f68a7b773f681&searchtype=a

32 Jain , Anil K. and Namboodiri, Anoop M. Indexing and Retrieval of On-line Handwritten Documents , Michigan State University, IEEE, 2003, cited at 20/9/2014, cited at http://www.sciencedirect.com/science/article/pii/0167865595000941

33 Sarkar, Sayantan. Word Spotting in Cursive Handwritten Documents using Modified Character Shape Codes, Department of Electrical Engineering, NIT Rourkela, 2010, cited at 10/10/2014, cited at http://www.sciencedirect.com/science/article/pii/S0957417409001511

[34] سيد ربيع سيد ابراهيم. محركات بحث الصور الثابتة على الانترنت. مصدر سابق.

35 Cao, Huaigu. Indexing and retrieval of low quality handwritten documents, University of New York at Buffalo, PHD, author, UMI ProQuest LLC, 2008, cited at 24/10/2014, cited at http://search.proquest.com/pqdtglobal/docview/305521651/fulltextPDF?accountid=37552

36 Zaher al aghbari and Salama brook. HAH : A holistic paradigm for classifying and retrieving historical Arabic handwritten documents. OP. CT.

37 Srihari, Sargur & Srinivasan, Harish & Pavithra Babu and Bhole, Chetan. Content-based Information Retrieval from Handwritten Documents. OP.CT.