جامعة نايف العربية للعلوم الأمنية

Naif Arab University For Security Sciences

تخزين الأسماء العربية في قواعد البيانات ومعالجتها حاسوبياً

أ.د. محمود إسماعيل صالح

٢٠٠٣م

تخزين الأسماء العربية
في قواعد البيانات ومعالجتها حاسوبياً

أ.د. محمود إسماعيل صالح

تخزين الأسماء العربية المكتوبة بالحرفين العربي واللاتيني في قواعد البيانات ومعالجتها حاسوبيا

أولا: التعامل مع الأسماء المكتوبة بالحرف اللاتيني

من المعروف أن رومنة الأسماء العربية وطباعتها عن طريق الحاسوب يواجه بعض المشكلات، من أهمها عدم توافر بعض الرموز الخاصة لتمثيل الأصوات العربية ضمن لوحات المفاتيح التقليدية. ومن هذه الرموز الحركات الطويلة (حروف لاتينية فوقها خطوط) ورموز حروف الحلق والمفخمات (مثل الحاء والصاد والطاء) التي تكتب غالبا على صورة حروف لاتينية تحتها نقاط.

هذا ونجد حلا لبعض هذه المشكلات في بعض برامج تنسيق الكلمات، مثل MS Word الذي يتيح للكاتب إدراج رموز من جداول بها حروف وأشكال مختلفة.

كذلك لابد لنا من التنبيه إلى مشكلة التمييز بين الهمزة والعين، وهو استخدام علامة ـ مع جعل الاتجاه نحو اليمين للعين ونحو اليسار للهمزة. وهو أمر لانستطيع التحكم فيه في الظروف الحالية.

أما المشكلة الكبرى فتبرز عند استخدام برامج الجداول وقواعد البيانات، مثل Exel وAccess، حيث لاتتوافر لمستخدميهما جداول الرموز ولا مرونة برامج تنسيق النصوص. فالمستخدم لايستطيع مثلا استخدام الخط السفلي تحت الحروف underscoring أوأية رموز ليست متوافرة في لوحة المفاتيح. وهذا ما حدث كما يبدو للعاملين في رومنة أسماء الأماكن في

المملكة العربية السعودية، حيث نجد تطبيقا جميلا للقواعد العلمية التي ذكرها الباحثون في مقدمة كتاب (دليل أسماء الأماكن في منطقة الرياض) عند تنفيذ الكتاب المطبوع (انظر الفصل المعنون: «الرومنة المستخدمة في هذا الدليل وخرائط المملكة العربية السعودية» (ص ص ٣٩ـ ٤٣) وباقي الكتاب. ولكن لما قامت هيئة المساحة الجيولوجية بتخزين البيانات في برنامج لقواعد البيانات اختفت جميع الرموز التي تختلف عن الحروف الرومانية التقليدية.

وقد وجدنا أن العاملين في بعض الجهات الرسمية لجؤوا لحل هذه المشكلة كما يبدو، إلى وضع الرمز الإضافي بعد الحرف، عوضا عن فوقه أوتحته مثلا. فلوفرضنا أن الطاء تكتب هكذا طباعة t أمكننا كتابته في قاعدة البيانات على شكل t_مثلا، وهكذا دواليك، نظرا لأن برنامج قاعدة البيانات لاتتوافر فيه إمكانية وضع خط تحت الحرف.

وجدير بالذكر أن استخدام أية رموز غير ألفبائية ستؤدي إلى مشكلات في ترتيب الكلمات sorting، حيث إن الكلمة التي تبدأ بمثل هذه الرموز سيتم ترتيبها وفقا للحركة التي تتبع الرمز. فنجد مثلا أن الأسماء «عمار، عمارة، عمر» سترد في مواقع مختلفة تماما، هكذا_Ammaar، _Imaarah، Umar، أي أن كل اسم سيرد تحت حرف مختلف. وتختلط هكذا الأسماء المبدوءة بهمزة بتلك التي تبدأ بالعين. فنجد أن كلمتي أمر وعمر وتحتلان مكانا واحدا، تحت الحرف a: Amr، Amr, .

ومن المشكلات الكبرى المتعلقة بترتيب الأسماء (ال) التعريف، حيث إن الحاسوب (أ) سيجعل جميع هذه الأسماء تحت الحرف a. وتزداد المشكلة تعقيدا عندما تكتب أداة التعريف بأشكال مختلفة، إذا ما التزمنا

الرومنة حسب النطق. فكلمة مثل الناس سترد في صورة an-naas بينما كلمة السماء ستكتب 'as-samaa، أي أن اللام الشمسية ستتحول إلى صورة الحرف الذي يليها.

حل مشكلة الرموز التي تتطلب علامات إضافية:

لحل هذه المشكلة، نجد أن مؤلف (معجم الألفاظ والتعابير الإسلامية) لجأ إلى استخدام الحروف الرومانية الكبيرة لتمثيل بعض الحروف العربية، مثل الحاء والصاد والضاد والطاء والظاء، حيث كتبت كما يلي:

Z T, D, S, H, . وذلك عند إعداد المعجم في برنامج Access لقواعد البيانات. ولكن ذلك قد يمثل مشكلة في كتابة أسماء الأعلام في النصوص العادية، حيث إن اسم العلم لابد أن يبدأ بحرف كبير capital letter فتختلط هذه الحروف مع مقابلاتها التالية: z t, d, s, h, فكلمتي هامد وحامد ستكتبان بالطريقة نفسها: Haamid. والحل هوعدم استخدام الحروف الكبيرة لغير الأغراض المشار إليها، بغض النظر عن موقع الحرف في الاسم. كما يمكننا استخدام الأرقام (كما فعل المؤلف مع العين حيث وضع لها الرقم ٩ ليميزها عن الهمزة).

عند إعداد القوئم لأغراض النشر مثلا يجري تحويل هذه الحروف والأرقام آليا إلى الرموز الأخرى المعروفة، مثل كتابة الحروف الكبيرة المذكورة في صورة الحروف المنقوطة أوالتي تحتها خط، كما في المثال التالي: محمد: muHammad التي تحول فيما بعد إلى Mu̲hammad.

أما الحركات الطويلة فقد عالجها مؤلف المعجم المذكور بتكرار الحركة القصيرة. فألف المد تكتب aa وواوالمد تكتب uu وياء المد كتبت ii. (انظر

المعجم المذكور في قائمة المراجع .) ونجد أن هذا الحل لاتنتج عنه مشكلات فنية . أما وجه اعتراض البعض عليها فهوغرابة تكرار رموز الحركات ، علما بأننا نفعل الشئ نفسه مع الحروف الساكنة دون اعتراض من أحد على ذلك .

توحيد كتابة الأسماء العربية:

أما مشكلة توحيد كتابة الأسماء (وهي ليست حاسوبية فقط) فيرى البعض أن حلها يكمن في بناء قاعدة بينانات لجميع الأسماء المعروفة مع تخزينها بكل من الحرف العربي واللاتيني . ومما لاشك فيه أن هذا حل مناسب في الحالات التي تكون فيها الأسماء معروفة أومحدودة (كما في شركة أرامكوالعربية أوجامعة الملك فهد للبترول) . فذاكرة الحاسوب تستوعب مئات الآلاف ، بل والملايين منها .

ولكن المشكلة تكمن في التعامل مع الأسماء الجديدة ، بل وغير المتوقعة أحيانا ، خاصة إذا نظرنا إلى القضية من زاوية تتعدى حدود البلد العربي الواحد . هنا نجد ضرورة ملحة إلى إيجاد خوارزميات لتكيكل الأسماء العربية قبل تخزينها بالحرف اللاتيني , حيث إن الكتابة بالحرف اللاتيني تتطلب إيراد جميع الحركات ، القصيرة منها والطويلة . وربما يمثل تخزين الأسماء في موسوعة السلطان قابوس للأسماء العربية منطلقا لقاعدة بيانات عربية للأسماء تتم الإضافة إليها وتحديثها بأسلوب مماثل لما طبقه الباحثون في جمع الأسماء لتلك الموسوعة ، إضافة إلى وسائل أخرى كجمع الأسماء من السجلات المدنية في البلدان العربية وتخزينها وفق قواعد موحدة يتفق عليها . أما الأسماء المستجدة فتخضع للبرنامج الحاسوبي الذي يقوم بشكيلها في ضوء نظرية الاحتمالات الإحصائية المبنية على قواعد البيانات المتوافرة بالأساليب التي ذكرنا سابقاً .

المشكلات الأخرى:

أما مشكلة ال التعريف فقد وضع لها مؤلف المعجم المذكور حلا مؤقتا يكمن في أمرين : (أ) كتابة أداة التعريف بصورة واحدة، أي حسب شكلها وليس نطقها . (ب) كتابة أداة التعريف بعد الاسم مسبوقا بشرطة . مثلا كلمة «المجيد» تكتب majiid-al .

أما الحل المتطور فيكمن في إعداد برنامج حاسوبي يجعل الحاسوب يتجاهل هذه ال المتبوعة بشرطة (al-) عند ترتيب الأسماء . وهنا تكمن أهمية وضع الشرطة بعد ال، حتى لايخلط الحاسوب بين ال التعريف وبين ال الأصلية في كلمات مثل : ألب وألماني Alb, Almaaniyy .

ولحل مشكلة ترتيب الأسماء التي تبدأ بالعين، فيكمن في استخدام حرف روماني، بدلا من رمز غير ألفبائي . فمثلا قد يرمز لها بالحرف g (التي ليس لها مقابل في العربية الفصحى) . بذلك سترد الأسماء التي تبدأ بالعين عند ترتيبها قبل الأسماء التي تبدأ بالغين، حيث إن الغين غالبا ما تكتب gh . فكملة عادل وغالب ستردان في موقعهما الصحيح ,gaadil ghaalib .

أما الهمزة في بداية الكلمة فالمتبع في أنظمة الرومنة المختلفة أن تسقط من الأسماء . مثلا أمية تكتب Umayyah وأمير تكتب Amiir وهكذا دواليك .

الحل الجذري:

يكمن الحل الجذري لمشكلة الرموز الخاصة برومنة الحروف العربية في

تصميم لوحة مفاتيح خاصة لرومنة الأسماء وغيرها من الكلمات العربية يمكن استخدامها مع برامج تنسيق الكلمات والجدولة وقواعد البيانات.

التعامل مع الأسماء المكتوبة بالحرف العربي:

في التعامل مع الأسماء المكتوبة بالحرف العربي، نلاحظ مايلي:

١ ـ الإدخال أوالتخزين:

ليست هناك كما يبدومشكلات معينة في إدخال أوتخزين الأسماء من حيث لوحة الحروف keyboard في الحواسيب. ولكن هناك مشكلات في عمليتي الترتيب والاسترجاع. وسنتحدث عن كل من هاتين المشكلتين بشيء من الإيجاز.

٢ ـ الترتيب:

من حيث الترتيب، نلاحظ مايلي:

ـ تمثل أل التعريف مشكلة من حيث إن الكلمات التي تبدأ بها يتم تصنيفها تحت حرف الألف, وليس الحرف الأول من الكلمة الأصلية. فكلمة الرحمن مثلا ترد قبل باسل، بدلا من بعدها (علما بأن الراء يجب أن ترد بعد الباء). وعلاج المشكلة من الناحية العملية هوالتعامل مع أل التعريف بالأسلوب الذي اقترحناه أعلاه في التعامل مع al ـ، من حيث كتابته ملحقا بالاسم بالشكل التالي: رحمن ـ ال، . بذلك سترد «الرحمن» في قسم الراء. وربما كان الحل هذا مناسبا أيضا لمشكلة أسماء العائلة التي تبدأ بكلمة آل، مثل آل سعود. غير أن المشكلة في هذا الحل تواجهنا عند كتابة اسم مثل «آل الشيخ» التي تجمع بين أل التعريف وبين كلمة آل في

وقت واحد. لذلك لابد لنا من نظرة أخرى لمعالجة هذا الأمر سوية مع كلمات مثل بن وأبووغيرها.

ب ـ هناك مشكلة ترتيب الأسماء التي تبدأ بالهمزة ، حيث إن الحاسوب يميز بين الكلمات المبدوءة بالهمزة المفتوحة (كذلك المضمومة) التي تكتب فوق الألف، مثل أحمد وأميمة، وبين الأسماء التي تبدأ بالهمزة المكسورة والتي تكتب تحت الألف، كما في إبراهيم. وتكمن المشكلة أنه عند ترتيب الأسماء يورد الحاسوب جميع الأسماء التي تبدأ بالهمزة فوق الألف (أبـ، أتـ, أثـ، أجـ، . . . ، أو، أيـ) قبل الأسماء التي تبدأ بالهمزة المكسورة. بذلك سنجد أن أحمد وأمين سيردان قبل إبراهيم وإجلال مثلا. أما الألف غير المهموزة فتأتي بعد الألفين المهموزتين. فلوفرضنا أن إبراهيم كتب بدون همزة فسنجد أنه يرد في موقع لاحق لجميع الأسماء التي تبدأ بالهمزة المكسورة (إبـ، . . . ، إيـ).

تبعا لهذه القاعدة التي يتبعها الحاسوب في الترتيب، سنجد أن الأسماء التي تكتب بالهمزة وبدونها سترد في مواقع مختلفة، فالاسم أحمد (بالهمزة) سيرد في مكان بينما يرد الاسم نفسه إذا كتب بدون الهمزة (احمد) في موقع مختلف تماما.

ويكمن الحل المؤقت لذلك في كتابة جميع الأسماء التي تبدأ بالهمزة بألف غير مهموزة، مما يخالف قواعد الإملاء العربية.

أما الحل الأمثل فأرى أنه يكمن في إعداد برنامج حاسوبي يحذف (أويتجاهل) الهمزة عند ترتيب الأسماء.

٣ ـ البحث والاسترجاع:

لاشك أن الهدف من تخزين البيانات في الحاسوب، إضافة إلى الحفاظ عليها، هوتوفيرها للباحث عنها. من هنا تأتي أهمية البحث والاسترجاع. فلا قيمة لبيانات أومعلومات مخزنة إذا لم تكن متاحة لطالبيها والباحثين عنها.

أما أهم مشكلات الاسترجاع للأسماء المكتوبة بالحرف العربي، فأرى أنها تكمن في حرفية الحاسوب ودقته، حيث إن أية علامة أوحركة، بل وفراغ أيضا، يؤثر في التعامل مع الحاسوب عند الاستعلام. فلوفرضنا أن الاسم تم تخزينه بشكل معين فإن الاستعلام عنه يتطلب عادة إدخال الاسم بنفس الشكل تماما بما في الفراغات بين مكونات الاسم المتعددة (مثلا اسم الشخص واسم أبيه). فلوكان الاسم «علي» مثلا تم تخزينه بهذا الشكل (أي بياء منقوطة) فلن نستطيع الاستعلام عنه بكتابة بالشكل التالي مثلا: على، كما يكتب في مصر مثلا. ولوكتبنا «محمد عبدالله» (بفراغ بين الدال والعين وبدون فراغ بين كلمتي عبد والله) فلا بد لنا من كتابتها بهذه الصورة تماما لنتمكن من استرجاع الاسم من قاعدة البيانات, في الظروف العادية. هنا تكمن مشكلة تعدد الصور التي تكتب بها بعض الأسماء، بالإضافة إلى طريقة إدخالها أوتخزينها. فالاسم «ريمة» يكتبها البعض بالتاء المربوطة ويكتبها آخرون بالألف. وعلى ذلك قس الأسماء مثل «رولا» التي تكتب «رولى» وغيرها.

إذا علمنا بمشكلة الفراغات التي نتركها بين الكلمات في اللغة العربية عند كتابتها بالحاسوب لأدركنا أنه يجب علينا ليس توحيد كتابة الاسم من حيث التهجئة فحسب, بل وحتى من حيث الفراغات كذلك. ثم اتباع

الشيء نفسه عند الاستعلام، حتى يتسنى لنا استرجاع البيانات المطلوبة. ولايفيدنا كثيرا أن نكتفي بإدخال ما يعرف باسم العائلة مثلا، ليعطينا الحاسوب ما لديه من أسماء تشتمل على ذلك الاسم، خاصة في ظل نظام التسمية المتبعة في المملكة العربية السعودية. فلوفرضنا أننا أردنا الاستعلام عن »«حمد صالح الحربي» وأدخلنا كلمة«»الحربي»، سنجد أن هناك عشرات الآلاف من الأسماء التي تحمل هذا اللقب أوتنتهي به. أما إذا لجأنا إلى الاسم الأول «محمد» فسنجد مئات الآلاف من الأسماء التي تبدأ به ولا شك.

لحل مثل هذه المشكلات لابد من إعداد برامج حاسوبية تتعامل معها، مثل قبول أكثر من صورة للاسم (أوتوحيد طريقة كتابتها إن أمكن) وتجاهل مسافات الفراغات وماشابه.

الخاتمـــــة:

مما سبق يتضح لنا أن هناك مشكلات في تخزين الأسماء في قواعد البيانات الحاسوبية سواء أكانت مكتوبة بالحرف اللاتيني أم الحرف العربي. وتختلف المشكلات تبعا للغة الحرف المستعمل.

فالتعامل مع الأسماء بالحرف اللاتيني له مشكلاته من حيث الرموز وإمكانات استخدام الرموز الإضافية diacritics من عدمها. ثم هناك المشكلات المتعلقة بترتيب الأسماء آليا sorting.

أما عند كتابة الأسماء بالحرف العربي فلا نتوقع مشكلات تذكر من حيث الإمكانات الحاسوبية. ولكن هناك مشكلات واضحة في ترتيب الأسماء وتصنيفها، كما وضحنا سابقا.

هناك أيضا مشكلات عند محاولتنا استرجاع المعلومات من قواعد البيانات والاستعلام عن الأسماء. وأرى أنه لابد لنا من اللجوء إلى البرمجة الحاسوبية للتعامل مع هذه المشكلات.

التوصيات

انبثقت عن الأوراق العلمية التي قدمت في الندوة وتقارير الوفود المشاركة، والمناقشات التي دارت خلالها التوصيات التالية:

١ـ الطلب من الأجهزة العربية المعنية التقيد باللغة العربية الفصحى عند كتابة أسماء الأعلام بالحرف العربي، مع الالتزام بتشكيل الحروف.

٢ـ دعوة الأجهزة الأمنية والعلمية العربية إلى دراسة الأنموذج المقترح المرفق لتوحيد معايير النقل الكتابي وإفادة شعبة اتصال مجلس وزراء الداخلية بالرياض بمقترحاتهم تمهيداً لإقراره وتعميمه.

٣ـ دعوة الأجهزة العربية المعنية إلى بناء قاعدة بيانات تحتوي على الأسماء العربية مكتوبة بالحرفين العربي والروماني وفقاً لنموذج موحد.

٤ـ دعوة الأمانة العامة لمجلس وزراء الداخلية العرب لدراسة بناء نظام حاسوبي لرومنة الأسماء العربية.

٥ـ دعوة أكاديمية نايف العربية للعلوم الأمنية إلى إنشاء موقع على شبكة الإنترنت تعمم من خلاله الأسس العلمية لكتابة الأسماء العربية ورومنتها.

٦ـ دعوة أكاديمية نايف العربية للعلوم الأمنية إلى تشكيل لجنة عربية متخصصة من الخبراء في مجالي اللسانيات والحاسوب تتولى إقرار ضبط الأسماء العربية بالشكل وتطوير الأنموذج المقترح عند الحاجة.

٧ـ دعوة أكاديمية نايف العربية للعلوم الأمنية إلى تنظيم دورات تدريبية للعاملين في الأجهزة الأمنية العربية المختصة على استخدام قواعد بيانات أسماء الأعلام العربية ورومنتها.

٨. رفع توصيات هذه الندوة إلى الأمانة العامة لمجلس وزراء الداخلية العرب عن طريق شعبة اتصال الرياض لعرضها على الاجتماع التنسيقي لأجهزة مجلس وزراء الداخلية العرب لاتخاذ الإجراءات المناسبة ورفع ما يتوصلون إليه إلى المجلس الموقر.

نموذج رومنة الأسماء

١ ـ يكتب الاسم متسلسلا (من الاسم الأول إلى الأب ثم الجد فاللقب) ويكتب رباعيا كلما أمكن ذلك ويوضع («ابن» أو «بنت» أو «ولد») بين الاسم الأول والثاني وذلك للتفريق بين الأسماء المركبة والمفردة وتكتب عند الرومنة bn, bt, wd .

٢ ـ عدم الأخذ بالإعراب عند كتابة الاسم . مثال : محمد ، Muhammad وليس Muhammadu ولا Muhammadun

٣ ـ تعامل الأسماء المركبة والكنى معاملة الاسم المفرد ، وذلك بضم مكونات الأسماء المركبة . مثال : أبو قيس ، شرفُ الدين ، Abuqays, Sharafuddeen

٤ ـ تعامل الـ التعريف حسب نطقها (تختلف الشمسية عن القمرية) : مثال : السالم والمسلم ، as-Saalim, al-Muslim

٥ ـ تعامل آل بطريقة مختلفة عن الـ التعريف : مثال : آل سالم و السالم ، aal Saalim, as-Saalim

٦ ـ تكتب التاء المربوطة حسب نطقها إما هاء أو تاء : مثال : هبة ، هبة الله ، Hibah, Hibatullah

٧ ـ تكرر الحروف المشددة : مثال : بَسَّام ، Bassaam

٨ ـ تعامل الألف المقصورة كالفتحة : مثال : ليلى ، Layla

٩ ـ تعامل الأسماء بحسب نطقها : مثال : طه ، Taahaa

١٠ ـ الحروف والحركات :

أ ـ الحروف العربية التي لها ما يقابلها صوتيا في الحرف .