

# مشروع PHP واللغة العربية

خالد الشمعة [khaled.alshamaa@gmail.com](mailto:khaled.alshamaa@gmail.com)

<http://www.ar-php.com>

## مقدمة:

تعد لغة PHP أكثر لغات البرمجة في العالم شعبية في مجال تطوير مواقع الويب، لذا فقد عم انتشارها كوسيلة لبناء مواقع الويب المتوسطة والصغيرة ضمن ما بات يعرف ببيئة LAMP حيث يشير مصطلح أوائل الأحرف هذا إلى Linux كنظام تشغيل و Apache كمخدم ويب و MySQL كقاعدة بيانات فيما حرف P يشير إلى أي من Perl أو Python أو PHP، وجميع مكونات هذه البيئة مفتوحة المصدر بامتياز. إن النجاح المتميز لهذه البيئة وجدارتها كان مدخل الكثيرين إلى استخدام البرامج الحرة المفتوحة المصدر.

لقد استفاد الكثير من مطوري مواقع الويب في سوريا وأنا منهم (كما في العالم أجمع) من وفرة الأدوات والمنتجات المفتوحة المصدر المتاحة، لكن الطابع العام لهذا الاستخدام كان سلبيا استهلاكيا إلى حد بعيد، ولهذا السلوك حسب اعتقادي أسباب يمكن تعميمها على شريحة واسعة من المبرمجين في وطني منها عدم الإيمان بالعمل الجماعي وقدرته على الإنجاز، وضعف مهارات تنظيم وتنسيق الأعمال الجماعية، أضف إلى ذلك حقيقة الحاجة إلى التواصل مع الآخرين بلغة ليست هي لغتنا الأم، وهو ما وضع المزيد من العوائق. إن على المؤسسات التربوية والتعليمية أن تلعب دورا فاعلا في تذييل تلك العوائق والأخذ بأسباب النجاح وذلك من خلال توفير المناهج والممارسات الكفيلة بتأهيل الكوادر القادرة على الإندماج بسلاسة في أطر العمل الجماعي.

إلا أن الثقة تبقى هي العائق الأهم - بحسب وجهة نظري الشخصية - أمام التحول من مجرد مستخدمين سلبيين لنتاج البرمجيات الحرة المفتوحة المصدر إلى عناصر إيجابية فاعلة ومنتجة، فالشعور الدفين بأننا متأخرون زنيا ومهاريا عن اللحاق بركب من سبقنا إلى عالم البرمجيات الحرة المفتوحة المصدر هو هاجس مثبط لعزيمة الكثيرين وهمهم. لذا سأحاول من خلال ورقة البحث هذه إطلاعكم على خبرتي في العمل على أحد المشاريع الحرة المفتوحة المصدر من جهتي نظر مختلفتين، حيث سناقش بداية وجهة نظر المبرمج الشاب الذي يتهيب دخول هذا المعترك وما استلهمناه من دروس قد تفيد الآخرين في تجاربهم الخاصة، ثم تنتقل إلى وجهة نظر من يتطلع إلى نتائج هذا المشروع ومدى الفائدة التي يمكن أن يجنيها الآخرون منه أو من الإضافة إليه.

## القصة من البداية

إن مشروع PHP واللغة العربية هو مشروع شخصي صغير انطلق مع بداية العام ٢٠٠٦ يهدف إلى تقديم مجموعة متكاملة من الأدوات والمكتبات التي تعالج نصوص اللغة العربية وتتعامل معها بخصائصها وخصوصياتها، متيحا للمبرمجين العرب وسيلة تمكنهم من بناء تطبيقات ويب ذات طابع عربي أصيل تحت مظلة ترخيص البرمجيات الحرة المفتوحة المصدر، وذلك عوضا عما توفره لغة PHP من مكتبات وتقنيات صممت أصلا للتعامل مع نصوص اللغة الإنجليزية، وما يستتبع هذا من حشرها ولي عنقها للتعامل مع غيرها من لغات.

عنوان الصفحة الرئيسية للمشروع على شبكة الإنترنت هو <http://www.ar-php.com>

لقد كانت الفكرة الرئيسية للمشروع هي العامل الأساس في نجاحه، وهي التي حولت مخاوف الضعف إلى أسباب قوة وقلبت قواعد اللعبة لمصلحتنا، وقد كانت دعوة مفتوحة للمضي خطوة جديدة إلى الأمام متجاوزين حاجز تعريب برمجيات الآخرين وصولاً إلى المساهمة الفعلية في كتابة الشيفرات البرمجية، فمن يفوقنا معرفة بلغتنا الأم ومن أعلم منا بخفاياها؟ ومن سيحرمك التميز إن أنت قدمت حلولاً تناسب لغتك وثقافتك؟ إنها البداية الملائمة لفتح باب الإبداع واسعاً على مصراعيه دونما رهبة أو تردد حول أهمية المشاركة التي نضيفها إلى مجتمع البرمجيات الحرة المفتوحة المصدر.

تم اختيار العمل على تطوير أصناف Classes مفردة صغيرة في البدء نظراً لأن العمل على هذا المشروع كان يتم خلال أوقات الفراغ من جهة، ولكون الشريحة المستهدفة منه لم تكن المستخدم النهائي بقدر ما كانت المبرمج الذي يعمل على تطوير موقع ويب عربي ضمن بيئة LAMP من جهة أخرى. وكانت البداية مع صنف ArQuery الذي يختص بتحسين آلية البحث ضمن قاعدة MySQL للبيانات وذلك من خلال توظيف تقنيات التعابير النظامية Regular Expressions بغية الحصول على نتائج أفضل عند التعامل مع النصوص العربية. حيث أن معظم المواقع العربية المبنية باستخدام برمجيات طورت أساساً للتعامل مع اللغة الإنجليزية لا تقدم للأسف أكثر من وظيفة البحث الحرفي لزوارها ومستخدميها، لكن المشكلة في حالة اللغة العربية تكمن في أن ذات الكلمة يمكن لها أن تظهر بأشكال مختلفة بحسب سياق النص، وأضرب على ذلك مثلاً: إن كان المستخدم يبحث عن كلمة مغتربين فإنه لن يحصل في مجموعة نتائجه على تلك السجلات التي كانت تتضمن كلمة مغتربون! ليس هذا فحسب، فالأمور يمكن لها أن تسوء أكثر لتجعل من عملية البحث تلك منقوصة إلى حد بعيد.

يقوم صنف ArQuery بالبحث بدلالة ساق الكلمة مع ما يستلزمه ذلك من إجراءات لتقييس شكل الكلمة وتوحيده، وهو ما يتم من خلال التعامل مع التشكيل، والهزمة بأشكالها المختلفة، إضافة إلى موضوع السوابق كواو العطف التي تلصق أحياناً بالكلمة أو أل التعريف، واللواحق مثل إشارات التنثنية والجمع والتأنيث والضمائر المتصلة، وذلك دون أن ننسى ضرورة التعامل مع الأخطاء الشائعة من قبيل استبدال التاء المربوطة بالهاء أو الألف المقصورة بالممدودة أو بالعكس، وسواها من حالات أخرى. وقد طور هذا الصنف بطريقة تسهل دمجها مع أي تطبيق يستخدم لغة PHP ويحفظ معلوماته في قاعدة MySQL للبيانات، بحيث تعطى الكلمة أو مجموعة الكلمات التي أدخلها المستخدم كدخل لهذا الصنف، فيما نحصل منه بالنتيجة على تركيبة القسم الشرطي من عبارة SQL التي يجب تمريرها إلى قاعدة البيانات لتنتم عملية البحث بأسلوب ساق الكلمة آخذين بعين الاعتبار مجموعة النقاط التي سبق الإشارة إليها إضافة إلى المزيد غيرها من النقاط.

بعد إتمام بناء هذا الصنف وتجريبه تمت إضافته إلى موقع <http://www.phpclasses.org> وهو ذات الموقع الذي كنت أتوجه إليه فيما مضى بحثاً عن أصناف تؤدي وظائف أحتاج إليها في تطوير هذا التطبيق أو ذاك، لكنها كانت المرة الأولى بالنسبة لي التي ألعب فيها دور المؤلف الذي يضيف عملاً مفتوح المصدر ليصبح متاحاً للآخرين. إحدى النقاط الإيجابية في هذا الموقع هي أنه ليس مجرد مكنز للأصناف التي يطورها المبرمجون من جميع أصقاع العالم فقط، بل لهذا الموقع إدارة ومشرفين يعاينون البرامج التي ترسل إليه فإما أن يوافقوا عليها لتضاف، أو يعتذرون عن قبولها إن كانت دون المعايير التي يلتزمون بها (على الرغم من أنها معايير أقل صرامة من مكنز أخرى مثل PEAR المكنز الرسمي لأصناف PHP المعيارية)، ليس هذا فحسب، بل كانت

هنالك لجنة تنتقي أفضل عشرة أعمال تمت إضافتها إلى الموقع في كل شهر ليجري التصويت عليها طيلة أيام الشهر الذي يليه من قبل المبرمجين المنتسبين للموقع لاختيار أفضلها، وقد كان لاختيار ArQuery ومن ثم حلوله في المركز الثامن في شهر شباط / فبراير من عام ٢٠٠٦ بمثابة مفاجئة سعيدة عملت على تحويل فكرة غضة طفولية إلى مشروع حقيقي أخذ يستهلك شيئا فشيئا معظم أوقات فراغي. وهكذا توالى الأفكار والأعمال والجوائز، فأصبح هنالك حتى لحظة كتابة هذه الأسطر ١٧ صنفا نالت ٧ جوائز ٤ منها كانت للصدارة.

إن تلك الأعمال كانت سببا في التعرف واللقاء مع الكثير من المبرمجين العرب المتحمسين والذين أصبحوا فيما بعد أصدقاء عزيزين ساهموا بشكل أو بآخر في تحسين العمل الذي أقوم به سواء من خلال تنقيح الأخطاء والهفوات التي وقعت بها بين الحين والآخر، أو من خلال النصائح والأفكار التي ساعدتني أيما مساعدة على تطوير جودة العمل ومقدراته. ووصل المشروع بعد عام درجة من النضج سمحت بالحصول على فرصة تأليف كتاب يتحدث عن مفهوم البرمجيات الحرة المفتوحة المصدر بشكل عام، وعن أفكار وخوارزميات مجموعة منتقاة من هذه الأصناف بتفصيل تقني واف، ونشر هذا الكتاب من قبل دار شعاع للنشر مشكورة تحت عنوان "[PHP واللغة العربية](#)". كما تم إطلاق الموقع الرسمي الخاص بهذا المشروع على شبكة الإنترنت بعد أن كان لا يتعدى كونه طائفة من الأصناف المنشورة في هذا الموقع أو ذاك. وأخذ المبرمجون العرب رويدا رويدا يستخدمون هذه الأصناف في أعمالهم، ونذكر منهم على سبيل المثال لا الحصر [وزارة الأشغال والإسكان في مملكة البحرين](#)، و**[البوابة العربية](#)** وهي نظام إدارة محتوى عربي مفتوح المصدر، وكذلك موقع [إزالة الغابات، المصادر الإفريقية في هبوط](#) وهو موقع بيئي متعدد اللغات. إن كسب تلك الثقة لم يكن بالأمر السهل أو السريع على الإطلاق، إنما جاء نتاجا لمجهودات كبيرة خاصة فيما يتعلق بدعم المبرمجين في حل المشاكل التي يمكن أن تواجههم عند تطبيق هذه التقنيات للاستفادة منها على الوجه الأمثل.

### قصة جميلة، لكن ما الفائدة التي سيجنيها الآخرون؟

دعونا نعود للتكلم عن وجهة النظر الأخرى، والتي نتحدث من منظور الراغب في معرفة نتاج هذا المشروع حتى الآن، وكيفية الاستفادة منه. للوصول إلى هذه الغاية سنقوم باستعراض الأصناف التي تم تطويرها في هذا المشروع وعمل كل منها.

يقوم صنف ArAutoSummary بعملية التلخيص الآلي للنص العربي، فكل ما عليك القيام به هو تمرير النص الكامل إلى هذا الصنف إضافة إلى الإشارة إلى عدد الجمل الأقصى الذي تريده في الخرج، أو نسبة حجم التلخيص إلى حجم النص الكامل، لتحصل بالنتيجة على التلخيص المنشود. ليست هذه بالفكرة الجديدة، فهي موجودة منذ زمن في معالجات النصوص مثل MS-Word، كما أنها متاحة كمنتج تجاري لتطبيقات الويب من قبل شركة صخر ضمن حزمة [السراج](#) لتحليل النص العربي (تصنيف، تلخيص، كلمات مفتاحية، أعلام)، لكنها المرة الأولى التي يقدم فيها صنف مفتوح المصدر لمبرمجي PHP لتلخيص النص العربي.

تقوم الفكرة على أساس فرز كامل كلمات النص بحسب جذورها بعد حذف الكلمات الشائعة منها مثل أسماء الإشارة وأحرف الجر الخ... (في حقيقة الأمر لم أستخدم خوارزمية لاستخلاص الجذر اللغوي للكلمة، بل عوضا عن ذلك قمت بحذف الأحرف الأقل أهمية من الكلمات جميعا، وهي الأحرف الأكثر تكرارا في السوابق

واللواحق التي ترتبط عادة بالكلمات العربية، وقد كانت النتائج عند تطبيقها مع خوارزمية التلخيص الآلي التالي شرحها أكثر من مقبولة)، بعد ذلك يتم تقييم كل من تلك الجذور باستخدام عدد من النقاط تزيد بزيادة مرات تكرارها ضمن النص (وذلك لمحاولة التعرف على الكلمات المفتاحية في النص المدروس)، وبعدها تتم عملية إسناد النقاط لجمل النص الأصلي من خلال إيجاد متوسط نقاط الكلمات المؤلفة منها تلك الجمل بالإضافة إلى مجموعة أخرى من المعايير من قبيل مكان الجملة في الفقرة (فالجمل في بداية الفقرات ونهايتها عادة ما تكون ذات أهمية أعلى من تلك التي في قلب الفقرة)، وكذلك فإن طول الجملة يؤثر بدوره، فيتم استثناء الجمل فائقة القصر أو تلك الطويلة جداً، في نهاية المطاف يتم ترتيب الجمل التي حازت على أعلى نقاط ليستقر الخيار على استخدامها في الملخص، ومن ثم يعاد ترتيبها في الملخص بحسب ترتيبها في النص الأصلي بغض النظر عن مجموع النقاط الذي حازت عليه. إن تطبيق هذه الخوارزمية أتى بنتائج باهرة ومقبولة جداً بالنسبة للقارئ البشري.

أحد الأصناف الأخرى التي طورت في إطار مشروع PHP واللغة العربية هو الصنف ArSoundex المختص بكشف التشابه اللفظي بين الأسماء التي لا تشترك في ذات طريقة الكتابة، ففي البدء كانت اللغة، ومن ثم جاءت الكتابة لترمز الأصوات التي تقال بالأشكال التي ترسم (أو تكتب)، وقامت كل أمة بصياغة أحرف كتابة لغتها بما يتناسب والأصوات التي يطلقونها حين التحدث والكلام، لكن المشكلة تظهر حينما يراد كتابة اسم ما أو مصطلح بلغة تختلف عن لغته الأم، فحينها قد لا تسعف أحرف اللغة المنقول إليها في التعبير عن اللفظ الحقيقي لتلك الكلمة. إن الكلمات التي تكتب خطأ بأكثر من تهجئة عادة ما تشكل كابوساً حقيقياً لمصممي قواعد البيانات، ومنها الأسماء على سبيل المثال، خصوصاً تلك الأعجمية منها والتي تعود بأصولها إلى التركية أو الأرمنية أو الكردية أو سواها من مصادر ثقافات أخرى.

لحل هذه المشكلة برزت الحاجة إلى خوارزمية تستطيع العثور على الأسماء أو المصطلحات ذات الأصوات المتشابهة، إن مثل هكذا فئة من الخوارزميات تدعى SoundExes، حيث تقوم بتحويل كلمة ممررة إليها (كإسم شخص على سبيل المثال) إلى سلسلة نصية تحدد مجموعة الكلمات التي تقاربها لفظاً. تقوم الفكرة الأساسية لهذه الفئة من الخوارزميات على تصنيف الأحرف في مجموعات بحسب طريقة لفظها، وتمتلك لغة PHP تابعا خاصا بها لإجراء تلك المعالجة ويدعى Soundex، لكن كما هو واضح فإنه مخصص للكلمات المكتوبة باللغة الإنجليزية وديم الفائدة تماما في حالة اللغة العربية، ما قمنا به في هذا المشروع هو تطوير صنف مخصص للغة العربية، فإن أنت أدخلت أي من الأسماء التالية (كليتتون، كلينتن، كلينتون، كلنتن، كلنتون، كلايتون) لحصلت على الدوام على الرمز K453 كقيمة Soundex لهذه الكلمات، في حين أن اسم مثل "كليتزمان" له قيمة Soundex مختلفة وهي K452.

هنالك أيضا الصنفين ArTransliteration و EnTransliteration والذين يقومان بعملية توليد المكافئ اللفظي للكلمات فيما بين الأبجدية العربية والإنجليزية، والتي تعد حاجة شائعة في أنظمة استعادة المعلومات المتعددة اللغات، فأسماء الأشخاص والأماكن إضافة إلى المصطلحات العلمية والتقنية، كلها حالات تستدعي وجود آلية مؤتمتة لتنفيذ عملية التحويل تلك، حيث أن الحلول القائمة على أساس القواميس تبقى قاصرة ومكلفة لضخامة عدد الكلمات التي يراد التحويل فيما بينها.

ليس لهذا النوع من المسائل وجود فعلي حينما تتشارك اللغتان المراد التحويل فيما بينهما ذات الأبجدية (كما هي حال معظم اللغات الأوروبية)، لكن الحال ليست كذلك بين اللغة العربية والإنجليزية، لذا طورنا صنف EnTransliteration لتحويل الأسماء العربية إلى مكافئاتها اللفظية لكن باستخدام أحرف إنجليزية (فمثلا "نزار قَبَّانِي" تصبح Nizar Qab'bani، و "عُوطَة دِمَشَق" تصبح Ghutah Dimashq)، وكذلك الصنف ArTransliteration والذي يقوم بالدور المعاكس من خلال تحويل الأسماء والكلمات الإنجليزية إلى ما يكافئها لفظا بالعربية (فمثلا Internet تصبح "إنترنت"، و Arizona تصبح "أريزونه").

دعونا الآن نستعرض معا في عجالة ما تبقى من أصناف تم تطويرها حتى لحظة كتابة هذه الأسطر ضمن إطار ذات المشروع، فهناك الصنفين ArDate و ArMktime للتحويل فيما بين صيغة التاريخ الهجري والميلادي إضافة إلى وظائف إظهار التاريخ بالأسماء العربية للأشهر والأيام.

أما الصنف ArNumbers فيقوم بتحويل رقم صحيح معطى إلى ذات الرقم لكن كتابة باللغة العربية، أو ما اصطلح على تسميته عادة بالتفقيط، وهي وظيفة كثيرة الاستخدام في التطبيقات المصرفية أو التي تجري بها تعاملات مالية مثل الفواتير، حيث تراعي هذه المكتبة علامة إعراب المحدود سواء كانت بالرفع أم النصب أم الجر، إضافة إلى حالة المحدود سواء كان مذكرا أم مؤنثا.

بالحديث عن التذكير والتأنيث تجدر الإشارة إلى الصنف ArGender والذي يقوم بتمييز الكلمات العربية المؤنثة تأنيثا لفظيا، مثل الكلمات التي تنتهي بالتاء المربوطة أو المفتوحة، أو تلك التي تنتهي بالألف المقصورة أو الألف والهمزة. يمكن الاستفادة من هذا الصنف على سبيل المثال في مساعدة الصنف الخاص بالتفقيط للحصول على خرج أصح لغويا.

هنالك أيضا مجموعة من الأصناف المفيدة لأنظمة البحث عن المعلومات واسترجاعها، منها الصنف ArWordTag الذي يحدد التعابير الإسمية أو تلك التي تتضمن أرقاما أو تواريخ ضمن النص المعطى، وهي عادة المقاطع الأكثر أهمية من ذلك النص والتي تتضمن أكبر قدر من المعلومات فيه. يتم تحديد التعابير الإسمية باستخدام بعض القواعد اللغوية، فالأسماء على سبيل المثال هي التي تعرف بأل التعريف أو التي تتلو أحرف الجر. هنالك صنف آخر مفيد لأنظمة استرجاع البيانات هو الصنف ArIdentifier الذي يستخرج النصوص العربية الموجودة ضمن وثيقة متعددة اللغات تستخدم مجموعة المحارف العالمية الموحدة UTF-8 في ترميزها. كذلك فقد طورنا الصنف ArCharsetD ليقوم بمهمة التعرف على مجموعة المحارف المستخدمة في ترميز وثيقة عربية ما من خلال نموذج إحصائي يقارن ما بين أكثر الرموز تكرارا وما يجب أن يكافئها من حروف عربية بحيث يتم إنتقاء الترميز الذي يحقق أكبر مطابقة ما بين هذين العاملين. أما إن أردت التحويل فيما بين مجموعات المحارف العربية المختلفة التي تستخدم عادة في مواقع الويب العربية وهي Windows-1256 و UTF-8 مع وجود القليل من المواقع التي تعتمد مجموعة المحارف ISO 8859-6، في تلك الحالة يمكنك استخدام الصنف ArCharsetC لإنجاز عملية التحويل هذه.

من جهة أخرى تمتلك لغة PHP مكتبة مخصصة لتوليد ملفات PDF الديناميكية، لكنها للأسف لا تدعم اللغة العربية حتى الآن، لذا توجهنا ناحية استخدام مكتبة مطورة من قبل طرف ثالث تدعى UFPDF تستطيع التعامل

مع مجموعة الحروف العالمية الموحدة UTF-8، وهو ما سيمكننا ضمنا من التعامل مع النصوص العربية. لكن المشكلة ظهرت مع أول تجربة لنا حين ظهر النص العربي بأحرف مفصولة عن بعضها البعض.

ذات المشكلة كانت تظهر أيضا مع مكتبات PHP الأخرى التي تنشئ ملفات SWF Flash ديناميكية أو تتعامل مع النصوص ضمن بيئة VRML للواقع الافتراضي أو تكتب على ملفات الصور باستخدام مكتبة GD، فجميع تلك المكتبات تتعامل مع الأحرف العربية بشكل مصمت دون أي معالجة للسياق، حيث أن تلك المعالجة غير معروفة أو مستخدمة في حالة اللغة الإنجليزية، فأشكال الأحرف تبقى على حالها بغض النظر عما يسبقها أو يتلوها، وهذه ليست حال اللغة العربية، لذا قمنا بتطوير الصنف ArGlyphs والذي يجري معالجة أولية على النص العربي المدخل إليه بحيث يعطي كخرج له سلسلة من رموز مجموعة الحروف العالمية الموحدة الموافقة تماما للأشكال التي يجب أن تظهر عليها تلك الحروف بحسب مكانها من سياق الكلمة، وبذا أصبح ممكنا الحصول على نص عربي سليم المظهر مع أي من تلك المكتبات السابقة الذكر.

أخيرا، من منا لم يقع في خطأ نسيان تحويل لغة لوحة المفاتيح حين التعامل مع نص ثنائي اللغة مما ينتج عنه نص غير مفهوم باللغة الأخرى؟ لقد تم تطوير الصنف ArKeySwap لمعالجة مثل تلك الحالات بشكل آلي، حيث يتلقى هذا الصنف تلك السلسلة النصية ذات الأحرف العربية غير المفهومة ويعيد ما كان يجب أن تكون عليه تلك السلسلة النصية بأحرف إنجليزية، أو العكس بالعكس.

#### ما هي الخطوة التالية؟

كل صباح أستيقظ لأخاطب نفسي قائلا: حسنا، ماذا بعد؟ لكنني دوما ما أجد الأفكار الجديدة في هذه الأرض الخصبة العذراء، لكن وفي ذات الوقت فإن مهمة نقل جميع هذه الوظائف لبناء مكتبة PEAR موحدة لاحتياجات اللغة العربية هي إحدى المهام التي تقبع على رأس أولوياتي للعام القادم، وذلك رغبة في استكشاف آفاق جديدة من الاندماج في عالم تطوير البرمجيات الحرة المفتوحة المصدر ولعب دور إيجابي فاعل فيه، من يدري ربما نتحدث بعد عام عن تطوير مكتبة PECL أيضا!

لكن الخطوة الاستراتيجية التالية ستكون في التطبيقات التي تدمج ما بين اللغة من جهة والنماذج الرياضية الإحصائية من جهة أخرى، ومن أمثلة تطبيقات هذا التوجه أداة الترجمة الآلية بين اللغة العربية والإنجليزية التي طورتها شركة Google، أو المشكل الآلي المطور في معهد بحوث الحاسب والإلكترونيات في مدينة الملك عبد العزيز للعلوم والتقنية في المملكة العربية السعودية. إن العقبة الأساسية أمام مثل هكذا طموح هي حاجة جميع هذه الأنظمة المطورة على أساس النماذج الرياضية الإحصائية إلى مكنز مفتوح المصدر ومنقح وغزير للنصوص العربية الخام المترجمة أو المشكلة أو المصنفة ليتم تحليلها واستخدامها في بناء النموذج الإحصائي الرياضي المرغوب، وهو أمر يعز الحصول عليه حاليا ويحتاج إلى توفير الكثير من الموارد والوقت مما يجعل منه حكرا على المشاريع والشركات الكبيرة، لذا أطمح إلى التعاون في إطلاق مشروع مثل هكذا مكنز مفتوح المصدر ليكون متاحا للجميع ليسهل على الأفراد والشباب الاستفادة منه والتركيز على تطوير النموذج الإحصائي ذاته عوضا عن ضياع الكثير من الموارد في تأمين الاحتياجات اللوجستية التي بلغا ما بلغت تبقى منقوصة ومحدودة إن كان طابع العمل فرديا.