ISSN - 2170- 0656

CERIST NEWS

Bulletin d'information trimestriel
Quatrième numéro - Décembre 2010

DOSSIER
INTERNET ET LA LANGUE ARABE

Les Langues sur la Toile

CENTRE DE RECHERCHE
SUR L'INFORMATION
SCIENTIFIQUE ET TECHNIQUE
cerist

JEBU'11

TROISIÈMES JOURNÉES D'ÉTUDE SUR LES BIBLIOTHÈQUES UNIVERSITAIRES

LE 30|31 MAI 2011 | CERIST - ALGER

La Langue Arabe à l'Ère d'Internet

Aliane Hassina
Chargée de Recherche - CERIST
Division
Théorie et
Ingénierie des
Systèmes
Informatiques

Parlez-vous Internet ? Non, ne froncez pas les sourcils, c'est une question sérieuse. En effet, aujourd'hui, à l'ère de l'information, tout le monde parle Internet ou doit le parler. La question est quelle est cette « nouvelle » langue et quels en sont les enjeux ? Si on part d'un point de vue technique, parler Internet c'est parler anglais et utiliser l'alphabet latin. C'est normal, question de filiation ; l'anglais est la langue maternelle de l'Internet. Néanmoins, Internet c'est la toile mondiale, c'est la société de l'information et de la connaissance, c'est la mondialisation avec tous ses nouveaux enjeux culturels et civilisationnels. L'enfant Internet a dû rapidement apprendre de nouvelles langues d'abord pour les besoins économiques et stratégiques de ses géniteurs mais aussi pour l'existence des autres peuples avec toutes les dimensions que comporte cette existence. La langue est au centre de la civilisation numérique car elle est vectrice de culture et de civilisation. L'UNESCO, parmi les résolutions qu'elle a adoptées prône la diversité culturelle qui passe nécessairement par le multilinguisme. D'ailleurs, on voit aujourd'hui émerger des langues comme le chinois, le hindi, le Coréen et l'Arabe sans oublier les langues slaves et d'autres. Que sera donc l'Internet Arabe ? Des statistiques récentes présentant le taux de pénétration de l'Internet par langue, montrent que la langue arabe vient en septième position sur les dix premiers pays en termes de nombre d'utilisateurs. En réalité la population arabe montre une réelle volonté d'appropriation de l'Internet. Cette volonté est accompagnée d'une volonté politique qui a consenti beaucoup d'efforts pour la généralisation des TIC. Mais le fossé numérique ne se dessine pas uniquement en termes d'équipements en TIC, mais surtout en matière de contenus.

En effet, sur la toile mondiale, l'enjeu ne s'exprime pas en termes de frontières, ces dernières étant les espaces où vous placez vos serveurs. L'enjeu est ailleurs, il est dans l'existence numérique qui se traduit en termes de production, autrement dit de contenus numériques ; reflet de l'évolution d'une société. Exister, c'est exister numériquement, donc produire et mettre à disposition une production de qualité. D'aucuns vont lier l'Internet arabe à une production en langue arabe. Une telle vision serait encore limitatrice, car comme nous l'avons dit, Internet est aujourd'hui incontestablement multilingue. Donc un Internet arabe ne serait pas uniquement un Internet en langue arabe mais un Internet où la production est arabe mais pouvant et devant parler plusieurs langues. La langue arabe prend un statut à part dans cette problématique dû aussi au retard accusé dans le domaine de son traitement technologique, mais ne constitue pas toute la problématique d'un Internet arabe.

A l'heure actuelle, nous avons les moyens technologiques de réaliser l'Internet arabe. La recherche et développement a un rôle important à jouer pour le traitement et la présentation du contenu aux utilisateurs. L'industrie des langues apporte des solutions et des outils pour toutes les tâches traitant des données linguistiques telles que l'indexation, la recherche, le résumé, la traduction, …. La technologie des services web prend aussi une importance primordiale pour la réalisation de l'Internet arabe car ce monde virtuel sera présenté à l'utilisateur sous forme de services avec des interfaces en langue arabe, de préférence multilingue. Le CERIST a toujours joué un rôle de précurseur dans la réalisation de cette société de l'information arabe qui prend forme de plus en plus. Et il a encore un rôle important à jouer à travers la R&D en particulier, en industrie de la langue arabe et en technologies des services web. Néanmoins, sans la production intellectuelle impliquant toute la société à différents niveaux, tous ces ingrédients resteraient un squelette attendant de se voir insuffler une âme (qui se veut arabe).

Sommaire

Journée d'information sur le programme « ERAWIDE & FP7 FINANCING SCHEMES »

Une journée d'information sur le programme « ERAWIDE & FP7 FINANCING SCHEMES » organisée par la Direction Générale de la Recherche Scientifique et du Développement Technologique et l'Union Européenne (DG-Recherche) a eu lieu au CERIST le 26 octobre 2010.

L'objectif de cette action était de renforcer les capacités de coopération des centres de recherche et d'améliorer les activités de recherche dans les thématiques du 7e PC.

Journées Universitaires Algéro-Françaises

Le CERIST a abrité les 17, 18 et 19 octobre 2011 une conférence Algéro-Française sur l'enseignement supérieur et la recherche. La délégation française participant à ces travaux était composée des représentants des ministères, des universités et des centres de recherche les plus impliqués dans cette coopération.

Cette conférence s'est déroulée en deux parties : Les deux premiers jours ont rassemblé les universités et établissements des deux pays qui développent des partenariats académiques renforcés. Le dernier jour a été consacré à l'examen, par les établissements et organismes de recherche français et algériens, d'un bilan conjoint des actions engagées afin d'envisager les partenariats à venir.

Par ailleurs, les différents centres et unités de recherche ayant pris part à ces journées ont pu exposer leurs activités dans des espaces réservés au niveau de la bibliothèque du CERIST.

Séminaire et table ronde sur la recherche d'information multilingue

Un séminaire suivi par une table ronde a eu lieu au CERIST, le 27 otobre 2010, sur la recherche d'information multilingue animée par M. Malek Boualem de France Telecom Orange Labs. Le séminaire a porté sur les thèmes suivants :

- Recherche d'information multimédia et multilingue
- Recherche d'information en langue arabe et traduction automatique
- Démonstration du service 2424 Actu d'orange
- Démonstration d'un moteur de recherche multimédia incluant la langue arabe

Le séminaire a abordé, également, d'autres aspects tels que :

- Sujets de recherche émergeants pouvant intéresser les équipes algériennes concernées par les technologies des langues naturelles écrites et orales
- Recommandations sur la gestion des projets de recherche
- Recommandations sur la recherche collaborative
- Exploration de pistes de collaboration

Visite de travail d'une délégation de Franche-Comté

Une délégation de Franche-Comté, conduite par le premier vice-président du conseil régional a visité le CERIST, le lundi 18 octobre 2010 pour discuter des axes de coopération dans les domaines des technologies de l'information.
La délégation était composée de :

- M. Denis Sommer, premier vice-président au conseil régional de Franche-comté
- M. Michel Stenta, Directeur Général de la SEM Numerica
- M. Frédéric Monnier, chargé d'affaires TIC et Multimédia
- M. Abbes Mokhnache, Ingénieur en opto-électronique

Conférence sur le partenariat entre les communautés scientifiques Algériennes aux USA et en Algérie

Le CERIST a abrité, les 3 et 4 décembre 2010, une conférence sur le partenariat entre les communautés scientifiques Algériennes aux USA et en Algérie, organisée par la Fondation Algéro-Américaine pour la technologie. Mise en place en janvier 2010, la Fondation Algéro-Américaine pour la Culture, l'Éducation, la Science et la Technologie (FAA-CEST) rassemble, notamment, des scientifiques et des universitaires algéro-américains. Cette fondation a pour mission essentielle le développement de la coopération entre l'Algérie et les Etats-Unis dans les domaines de la science, de la technologie et de la santé.

Ont pris part à cette conférence d'éminents spécialistes algériens dans la bio-engineering, les sciences médicales, la géophysique et la climatologie. Cette conférence avait pour objectif de mettre en place une plate-forme de partenariat dans le domaine combiné de la bio-engineering entre les académiciens algériens exerçant aux Etats-Unis et les chercheurs en Algérie et, également, de permettre une collaboration entre les experts dans la gestion et la prévention des catastrophes naturelles, essentiellement les séismes.

Pr. H. Aourag et son excellence l'ambassadeur des États-Unis David Pearce

Deuxième Workshop sur les Services Web dans les Systèmes d'Information - WWS'10

La deuxième édition du Workshop sur les Services Web dans les Systèmes d'Information organisée par la Division des Systèmes d'Information et Systèmes Multimédia a eu lieu, au CERIST, les 26 et 27 décembre 2010. Cette deuxième édition de l'atelier « services web » a été l'occasion de montrer les contributions récentes portant sur les nouveaux défis de recherche posés par la diversité croissante de paradigmes pour la spécification, le développement et la mise en œuvre de services Web. Les thématiques du workshop ont couvert un large spectre de problèmes liés à la représentation, la composition, la maintenance et l'intégration de services dans les applications traditionnelles ainsi que dans les domaines émergents comme les applications mobiles et le Web 2.0.

Pr Aourag, Directeur Général de la Recherche Scientifique, a présidé la cérémonie d'ouverture des travaux du workshop. Il a présenté le web comme une source d'information au centre des innovations permettant aux chercheurs algériens d'être à la page de ce qui se fait de par le monde. Il a aussi plaidé à cette occasion l'amélioration de la qualité de services des produits développés et la création de start-up en informatique.

Ce 2ème workshop sur les services web a regroupé d'éminents professeurs étrangers et algériens officiant au niveau de prestigieuses universités dans le monde pour débattre des innovations en matière des services web et des nouvelles applications. Le workshop a comporté également des conférences invités, des tutoriels et des tables rondes.

Pour plus d'informations a propos du WWS'10 veuillez consulter le lien ci-dessus :
http://www.cerist.dz/workshop/wws10.html

Internet Et la langue Arabe

« Sur le plan linguistique, le monde des TIC se trouve à un tournant, […] la langue arabe se trouve elle aussi à un tournant. Elle peut permettre aux pays arabes de combler leur retard dans la course au progrès de l'information ou contribuer à creuser le fossé linguistique séparant les arabes du reste du monde à différents niveaux ».

PNUD, RADH, 2002.

Document spécial de 20 pages : 10/29

« LES LANGUES SUR LA TOILE »

Un dossier élaboré par :

Hassina Aliane - Chargée de recherche

Équipe Web Sémantique et Langue Arabe
Division Théorie et Ingénierie
des Systèmes Informatiques - CERIST

Internet Et la langue Arabe

1. Introduction

D'après une étude effectuée par le British Council, qui établit des projections pour 2050 sur le statut respectif des langues du monde: à cette date, cinq langues seraient dominantes : le chinois, le hindi, l'anglais, l'espagnol et l'arabe. La promotion de l'arabe à cette place pose toutefois des questions. En effet, sous quel angle devrions-nous aborder cette question ? En réalité, la langue Arabe est la langue du Coran et malgré toutes les tares qui lui ont été associées à travers le temps -à tort ou à raison- elle est toujours là ; langue vivante, riche, à l'épreuve du temps. Mais au-delà du fait religieux, aujourd'hui, les défis posés dépassent le cadre d'existence si ce n'est qu'ils redéfinissent cette notion d'existence dans le nouveau contexte mondial fondé sur la société de l'information et de la connaissance. Dans cette nouvelle société, c'est la langue anglaise qui domine. En fait, ce qui caractérise la langue Anglaise, c'est qu'au-delà du bénéfice de positions économiques et politiques dominantes, cette langue s'est engagée dans un processus de désethnisation qui la rend moins marquée et qui fait qu'elle n'est perçue comme la propriété ni d'un groupe ethnique, ni d'une religion, ni d'une idéologie quelconque : ce faisant, elle se distingue, par exemple, de la politique de la francophonie, où la langue française reste très identifiée à la France. Ce caractère ouvert de la langue anglaise lui confère une flexibilité et une adaptabilité à des contextes culturels divers, ce qui fait d'elle une langue apprise actuellement par près de deux milliards de personnes.

La puissance économique de l'arabe place cette langue au 8 ème rang mondial. Néanmoins, L'économie reposant principalement sur la vente de pétrole ne suscite pas auprès des acheteurs un besoin d'apprendre cette langue. Au contraire, en dehors des sphères politiques et religieuses, la langue arabe connait de sérieux problèmes. Nous nous intéresserons dans ce dossier à l'existence numérique de la langue arabe que nous plaçons dans le contexte général d'un Internet Arabe.

2. Internet Arabe : quelle vision ?

L'Internet arabe est abordé aujourd'hui comme un slogan dans différents milieux, politiques, culturels ou techniques. Et ce qui est effrayant dans les slogans c'est leur nature éphémère. Les slogans ne durent pas longtemps, et on n'a pas eu le temps d'avoir œuvré pour un slogan qu'en survient déjà un nouveau. Par exemple et pour ne citer que les plus récents, nous avons déjà brandi le slogan de la société de l'information qui rapidement a évolué pour devenir la société de la connaissance et voilà qu'aujourd'hui, on revient au slogan de la langue que l'on a déjà brandi à maintes reprises dans d'autres contextes. Mais si tel est le lot des pays du tiers monde ou pour utiliser une expression plus sobre, des pays du sud, ce n'est peut-être pas une fatalité. La vraie question est : où sommes-nous dans cette évolution qui nous prend sans cesse de vitesse. Qu'y avons-nous produit ? Qu'y produisons-nous ? En réalité, jusqu'ici, nous n'avons fait que subir l'évolution et la différence aujourd'hui, ne peut venir qu'en termes de production intellectuelle et de contenus visibles et compétitifs. Oui, que veut- on dire par un internet arabe ? Un espace sur la toile proprement réservé aux arabes ? En réalité, sur la toile mondiale, l'enjeu ne s'exprime pas en termes de frontières, ces dernières étant les espaces ou vous placez vos serveurs. L'enjeu est ailleurs, il est dans l'existence numérique qui se traduit en termes de production, autrement dit de contenus numériques. Or les arabes n'ont pas de problème d'argent, et acheter les équipements les plus sophistiqués, ne devrait pas poser problème. Mais depuis la chute de l'Andalousie, le talon d'Achille des arabes est leur intellect. Et aujourd'hui internet nous met face à cette responsabilité civilisationnelle : exister, c'est exister numériquement donc produire et mettre à disposition une production de qualité.

D'aucuns, vont lier l'internet arabe à une production en langue arabe. Une telle vision serait encore limitatrice, car internet est aujourd'hui incontestablement multilingue. Donc un internet arabe ne serait pas uniquement un internet en langue arabe mais un internet où la production est arabe mais pouvant et devant parler plusieurs langues. La langue arabe prend un statut à part dans cette problématique dû aussi au retard accusé dans le domaine de son traitement technologique, mais ne constitue pas toute la problématique d'un internet arabe.

Internet Et la langue Arabe

En réalité à l'âge d'or des arabes c'est-à-dire à l'époque médiévale, la langue arabe a connu un statut qu'on peut comparer au statut actuel de la langue anglaise. En effet, l'arabe a été non seulement une langue d'échanges scientifiques et commerciaux mais elle était aussi la langue littéraire et religieuse des différentes communautés musulmane, chrétienne et juive ce qui indiquerait alors un processus de désethnisation de la langue arabe et pour certains auteurs, « même si le contexte est différent, on avait à peu près les mêmes conditions qui ont œuvré à l'émergence de l'anglais comme lingua-franca mondiale ». D'où la question, « la langue arabe peut-elle revenir comme langue mondiale majeure et renouer avec son passé de tolérance, d'ouverture et de grandeur ? » D'autre part, parler d'internet arabe, c'est se placer d'abord dans le cadre de la société de l'information et de la connaissance à travers des actions planifiées et entreprises, en premier lieu, par les gouvernements. Une fois cette volonté et cette vision établies et reconnues, techniquement la réalisation de l'internet arabe s'articule autour de la production de contenus numériques et la recherche & développement pour offrir les outils de traitement de ces contenus et leur mise à disposition pour l'utilisateur final.

2.1. La langue Arabe et la Société de l'Information

Le monde arabe est de plus en plus marqué par l'essor spectaculaire des (nouvelles) technologies de l'information et de la communication, même s'il reste encore loin des pays occidentaux, mais aussi des pays d'Asie ou d'Amérique latine. Néanmoins, le retard qu'il accuse peut être relativisé, et cela pour plusieurs raisons : le monde arabe constitue aujourd'hui une zone où les dépenses relatives aux TIC augmentent à un rythme deux fois plus rapide que partout ailleurs. En outre, les évolutions récentes témoignent d'une réelle démocratisation de l'usage d'internet, qui touche des fractions de plus en plus larges des populations. Enfin, les états eux-mêmes ont pris conscience de l'importance des enjeux et promeuvent des politiques publiques de développement des TIC. A l'évidence, les enjeux industriels et économiques, mais aussi sociaux, culturels et même politiques, sont pourtant considérables. Mais bien qu'on ne puisse pas encore évaluer les effets de transformations qui jouent, pour nombre d'entre elles, sur le moyen terme, on peut déjà observer l'engouement de la société arabe et sa forte volonté d'appropriation des TIC. Des conférences et des ateliers sont organisés régulièrement pour définir une stratégie arabe commune pour la société de l'information : le sommet mondial de la société de l'information SMSI, des ateliers sur la fracture numérique et le monde arabe, le workshop sur la stratégie arabe TIC à l'horizon 2012…

Internet Et la langue Arabe

Le Sommet mondial sur la société de l'information (SMSI) a été mis en place par l'ONU qui a chargé l'UIT (Union internationale des télécommunications) de coordonner le développement des nouvelles technologies de l'information et des communications dans le monde. Le SMSI est un sommet tripartite, ouvert aux gouvernants de tous les pays, aux firmes multinationales, et à la Société civile (organisations non-gouvernementales, collectifs citoyens, syndicats). Le SMSI s'est déroulé en 2 phases. La première phase, accueillie par le Gouvernement suisse, a eu lieu à Genève du 10 au 12 décembre 2003 et avait pour but d'adopter une déclaration de principe et un plan d'action. La deuxième phase, accueillie par le Gouvernement tunisien, a eu lieu à Tunis du 16 au 18 novembre 2005. L'objectif de cette deuxième phase était de mettre en œuvre le plan d'action de Genève et aboutir à des solutions, à des accords sur la gouvernance de l'Internet, les mécanismes de financement, le suivi et la mise en œuvre des actions adoptées.

Les grandes orientations du SMSI

- C1. Le rôle des instances publiques chargées de la gouvernance et de toutes les parties prenantes dans la promotion des TIC pour le développement.
- C2. L'infrastructure de l'information et de la communication
- C3. L'accès à l'information et au savoir
- C4. Le renforcement des capacités.
- C5. Établir la confiance et la sécurité dans l'utilisation des TIC.
- C6. Créer un environnement propice.
- C7. Les applications TIC:

Administration Électronique
Commerce Électronique
Télé-Enseignement
Télésanté
Cybertravail
Cyberécologie
Cyberagriculture
Cyberscience

- C8. Diversité et identité culturelles, diversité linguistique et contenus locaux.
- C9. Média
- C10. Dimensions éthiques de la société de l'information.
- C11. Coopération internationale et régionale.

Si on ne peut pas prédire, les conséquences de l'usage généralisé des TIC dans des domaines précis pour chaque pays arabe. En revanche, il est parfaitement possible d'observer que les mutations qu'elles entraînent font partie désormais des facteurs qui contribuent à modeler le paysage social et politico-économique de cette région du monde. Il suffit pour s'en convaincre de constater que le développement des TIC, leur contrôle et leur exploitation figurent aujourd'hui au premier plan des préoccupations des États arabes eux-mêmes.

Internet Et la langue Arabe

2.2 La fracture numérique

D'une manière générale, le fossé numérique peut être défini comme une inégalité face aux possibilités d'accéder et de contribuer à l'information, à la connaissance et aux réseaux, ainsi que de bénéficier des capacités majeures de développement offertes par les TIC. Ces éléments sont quelques-uns des plus visibles du fossé numérique, qui se traduit en réalité par une combinaison de facteurs socio-économiques plus vastes, en particulier l'insuffisance des infrastructures, le coût élevé de l'accès, l'absence de formation adéquate, le manque de création locale de contenus et la capacité inégale de tirer parti, aux niveaux économique et social, d'activités à forte intensité d'information.

Elie Michel in,
« Le fossé numérique. L'Internet, facteur de nouvelles inégalités ? »,
Problèmes politiques et sociaux - La Documentation Française, N° 861, août 2001.

Pour les gouvernements des pays Arabes, fournir aux citoyens un contenu de TIC utile est aussi important que la fourniture de l'accès lui-même aux TIC. Autrement dit, si les portails de l'information n'offrent pas un contenu utile aux citoyens, ils seront incapables de tirer profit des TIC. Il faut également s'assurer que le contenu est présenté en Arabe. Une minorité parle des langues étrangères, et c'est certainement cette minorité qui a accès aux outils des TIC. Donc, en présentant le contenu dans la langue locale, la majorité de la population qui n'est pas familière avec les TIC aura ainsi accès à internet. A côté des portails éducatifs et informatio-nnels, la priorité doit être accordée au lancement des portails qui offrent des services publics en ligne ayant pour objectif de répondre aux demandes économiques et sociales de la population.

Internet Et la langue Arabe

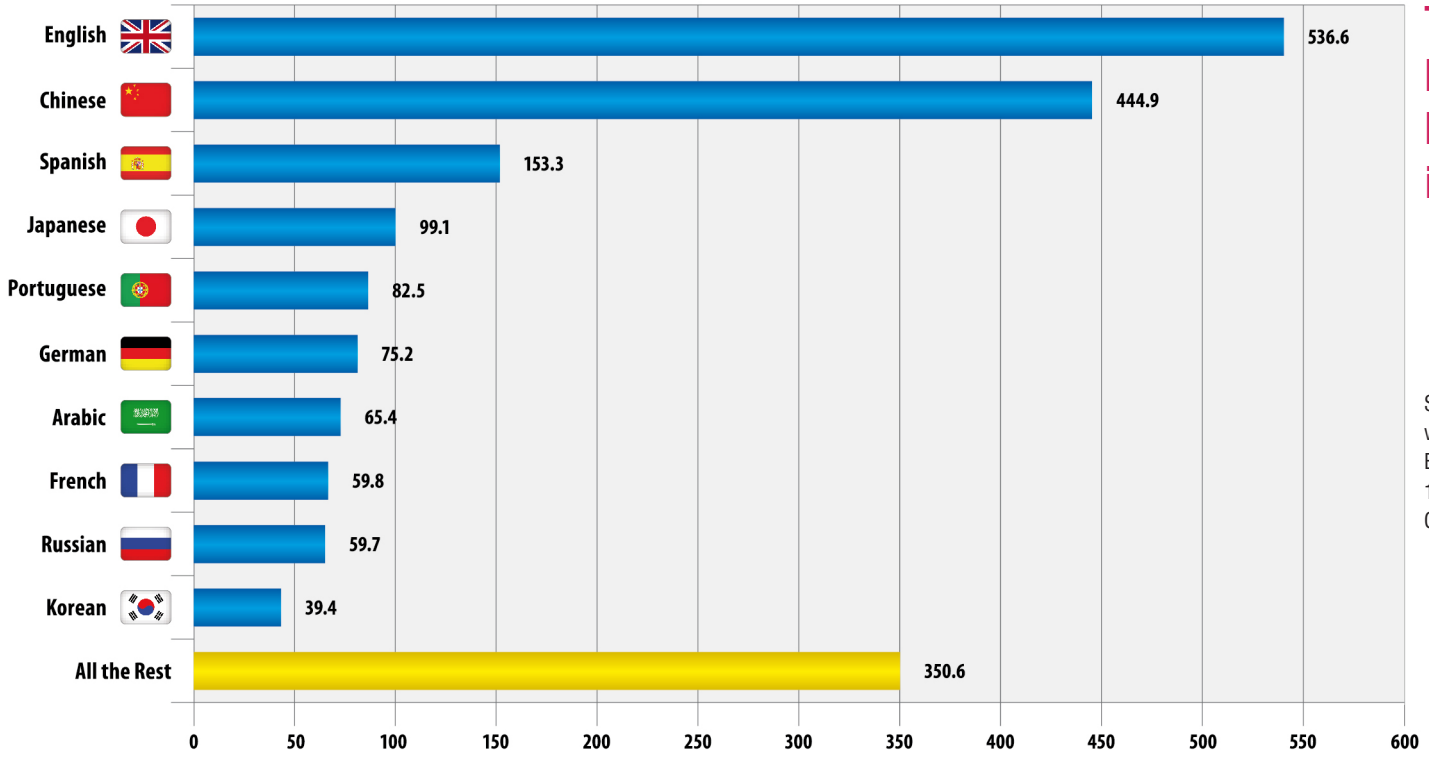
La société de l'information, se définit par l'utilisation de l'information et des connaissances, qui devient le moteur de la croissance. Le contenu devient alors le pivot de cette société. Dans un internet arabe, le citoyen arabe devrait avoir accès à des services, informations et connaissances en arabe et dans d'autres langues. Encourager la diversité culturelle et le multilinguisme est aussi la devise de la SI. Des études ont même montré que parler de la langue arabe ne s'arrête pas à parler de l'arabe classique dit aujourd'hui Arabe Standard Moderne (ASM) mais également de toutes les formes dialectales de cette langue. En effet, ces études montrent que l'appropriation du numérique (Internet et même le téléphone mobile) par une majorité de la population qui utilisent ces médias de communication se fait par l'intermédiaire des dialectes pouvant utiliser pour la transcription aussi bien l'alphabet arabe que latin. D'où la nécessité de prendre en considération cette dimension linguistique dans la communication numérique. En effet, la population utilisant l'ASM où cherchant l'information en ASM constitue une minorité des utilisateurs arabes du net. L'ASM est utilisé principalement pour le codage et la transmission orale ou écrite de la connaissance (écrits sous toutes ses formes, discours, …) académique, politique, religieuse. Mais la grande population de la toile veut discuter, communiquer, accéder à des services au quotidien, et ce serait une erreur que de négliger la prise en charge technologique d'une langue (en réalité plusieurs) qui est en train de s'imposer qu'on le veuille où non. L'enjeu de cette prise en charge aura en particulier pour impact, de garder l'alphabet arabe dans la communication et s'affranchir de l'alphabet latin, tout en redonnant une place et une visibilité à ces dialectes qui constituent une dimension importante de la société et de la culture arabe.

Enfin, et pour reprendre des termes proposés récemment, on passe d'une économie au capital tangible à une économie au capital intangible. Or, si les technologies de l'information et de la communication sont des outils indispensables à cet égard, l'essentiel n'est pas là : il est dans les personnes humaines qui créent et utilisent le savoir, qui représentent le capital intangible. Les informations les plus pertinentes, les meilleures bases de données, les logiciels les plus performants, les programmes d'enseignement les plus perfectionnés ne servent à rien s'il n'y a pas les ressources humaines suffisamment formées pour les utiliser de manière productive. Mais surtout l'économie de la connaissance, dans sa globalité, exige des populations de plus en plus formées pour améliorer la productivité et maintenir la compétitivité. Donc, comme dans le reste du monde, la formation de la ressource humaine est aussi un élément clé dans la société arabe de l'information.

Maintenant que nous avons un peu dessiné les contours de la société arabe de l'information centrée notamment sur l'utilisation de l'Internet, nous allons voir dans le paragraphe suivant la situation actuelle de l'internet dans les pays arabes.

3. Pénétration de l'Internet en langue Arabe

Le graphe et le tableau suivants présentent une estimation de l'Internet World Stat de l'utilisation de l'Internet par langue. Ce tableau a été publié en 2010.



Top Ten Languages In The Internet 2010 in millions of users

Source: Internet World Stats
www.internetworldstats.com/stats.htm
Estimated Internet users are
1,966,514,816 on June 30,2010

Internet Et la langue Arabe

TOP TEN LANGUAGES USED IN THE WEB (Number of Internet Users by Language)					
TOP TEN LANGUAGES IN THE INTERNET	Internet Users by Language	Internet Penetration by Language	Growth in Internet (2000 - 2010)	Internet Users % of Total	World Population for this Language (2010 Estimate)
English	536,564,837	42.0 %	281.2 %	27.3 %	1,277,528,133
Chinese	444,948,013	32.6 %	1,277.4 %	22.6 %	1,365,524,982
Spanish	153,309,074	36.5 %	743.2 %	7.8 %	420,469,703
Japanese	99,143,700	78.2 %	110.6 %	5.0 %	126,804,433
Portuguese	82,548,200	33.0 %	989.6 %	4.2 %	250,372,925
German	75,158,584	78.6 %	173.1 %	3.8 %	95,637,049
Arabic	65,365,400	18.8 %	2,501.2 %	3.3 %	347,002,991
French	59,779,525	17.2 %	398.2 %	3.0 %	347,932,305
Russian	59,700,000	42.8 %	1,825.8 %	3.0 %	139,390,205
Korean	39,440,000	55.2 %	107.1 %	2.0 %	71,393,343
TOP 10 LANGUAGES	1,615,957,333	36.4 %	421.2 %	82.2 %	4,442,056,069
Rest of the Languages	350,557,483	14.6 %	588.5 %	17.8 %	2,403,553,891
WORLD TOTAL	1,966,514,816	28.7 %	444.8 %	100.0 %	

4. Problèmes techniques liés à la langue Arabe

En dehors des problèmes socio-économiques qui sont considérés comme des facteurs de retard pour la pénétration d'Internet dans les pays arabes ou encore les résistances que l'on peut rencontrer au niveau des administrations et organisations où se doter en matière d'équipements ne pose nullement problème, il existe des problèmes techniques que rencontre un utilisateur qui ne connaitrait que la langue arabe et pour lequel la manipulation des caractères arabes relève d'un vrai parcours du combattant. En effet, bien avant de parler d'internet, quiconque a eu à utiliser des caractères arabes se rappelle de la difficulté d'utiliser les caractères arabes ne serait ce que dans un traitement de texte. Le traitement par des programmes et l'affichage est encore une autre histoire. La codification ascii se voulait être la codification universelle pour tout le monde. Néanmoins, des efforts ont été consentis à différents niveaux pour la codification des caractères autres que latins. La diffusion d'internet a rendu le problème encore plus crucial et aujourd'hui nous pouvons situer le problème de la codification des caractères arabes (et en général non latin) à deux niveaux, d'abord le codage des caractères pour les contenus et les applications et plus récemment les adresses internet.

4.1 Le codage des caractères arabes

La langue arabe possède plusieurs caractéristiques qui demandent des traitements particuliers pour qu'elle soit implémentée dans un programme, intégrée dans un PC ou dans un téléphone mobile. D'abord, la direction d'écriture est de droite à gauche. En plus, les caractères arabes changent leurs formes selon la position qu'ils occupent dans un mot. Il y a aussi la possibilité de transformer deux caractères en une seule forme.Le standard de codage ASCII a montré son insuffisance, puisqu'il fonctionne sur 8 bits, c'est-à-dire qu'il ne permet que 128 positions de codage. L'évolution du codage des caractères a suivi celle des possibilités des microprocesseurs qui sont conçus aujourd'hui pour fonctionner sur 64 bits. Les premiers systèmes se sont donc attachés au codage de l'alphabet latin, puis la tendance s'est engagée vers la communication multilingue. Les codes utilisés pour le traitement électronique des caractères se sont succédés, sans que l'on parvienne à un véritable consensus international pendant plusieurs années. Au milieu des années 80, l'ISO (International Standard Organisation) et un certain nombre de constructeurs informatiques, inaugurent l'idée d'un système universel où chaque caractère est codé sur 1, 2, 3, ou 4 octets, ce qui aboutit à l'élaboration de deux systèmes différents.

Internet Et la langue Arabe

L'un de l'ISO, et l'autre d'un consortium de vendeurs : Unicode. Finalement, un consensus s'est établi entre les deux parties. La norme ISO/IEC 10646 parue en mai 1993, propose un jeu universel de caractères codés sur 4 octets, sous la forme d'une table multilingue à 4 dimensions. Normalisation du caractère arabe : L'expérience du contexte arabe a généré dans ce sens une surabondance de normes qui font objet de références dans le développement d'applications bilingues. Seulement, à part les différences enregistrées entre ASMO 449 & ASMO 709, la BMP de l'UNICODE ou de l'ISO 8859-6, des douzaines de normes sont encore utilisées par les concepteurs d'applications bilingues en arabe/latin.

Unicode

Unicode est un code universel développé par le Consortium Unicode. Il est fait pour résoudre ces problèmes de pluralité des caractères, en créant un ensemble unique de codes, Universal Character Set, UCS, sur plus d'un octet qui peut aller jusqu'à 1 million de codes et qui inclut tous les systèmes d'écriture du monde. L'équivalent ISO de Unicode est ISO-10646. Unicode définit non seulement le code d'un caractère mais aussi ses propriétés sémantique comme la direction d'écriture dans le cas de l'arabe. Unicode est devenu un choix stratégique pour Internet et les développeurs logiciels. Unicode a introduit une nouvelle façon de considérer un caractère. Unicode considère une lettre A ou B ou K, comme une entité abstraite à laquelle il fait correspondre un nombre appelé code point. Ce nombre, il peut y'en avoir jusqu'à 65.536, est unique et est noté U+valeur. Ainsi, à la lettre A correspond U+0041, et au Alef correspond U+0627. Les valeurs inférieures à 127 (U+0000 à +007F) sont restées affectées aux caractères anglais comme dans l'ASCII. Ainsi, le multilinguisme est théoriquement possible en Unicode, comme le montre la figure suivante, puisque chaque lettre de chaque langue, dispose d'un code point distinct.

Internet Et la langue Arabe

UTF-8

Si on veut stocker sur un support informatique un texte constitué de caractères Unicode, il faut choisir un procédé transformant une suite de caractères Unicode en une suite d'octets et réciproquement avoir un procédé pour retrouver la suite de caractères à partir d'une suite d'octets bien formée, ces données constituent ce que l'on appelle un encodage. Le plus connu des ces encodage et aussi le plus simple est UTF-8, qui transforme un code point Unicode en une suite d'octets (bytes sequence). Les valeurs sur 16 bits se prêtent mal au transport par octets utilisés classiquement dans les protocoles et aux traitements associés. UTF-8 répond à ce besoin. Il a aussi l'avantage d'être compatible avec l'ASCII. En UTF-8, chaque code point dans [0-127], est représenté sur un seul octet. Ainsi toutes les chaînes ou fichiers ASCII peuvent être reversés dans UTF-8. Les autres code points , 128 et plus, utilisent deux octets ou plus. L'arabe utilise deux octets en UTF-8. « d8a8 » est la lettre Ba et « d98a » est la lettre Ya.

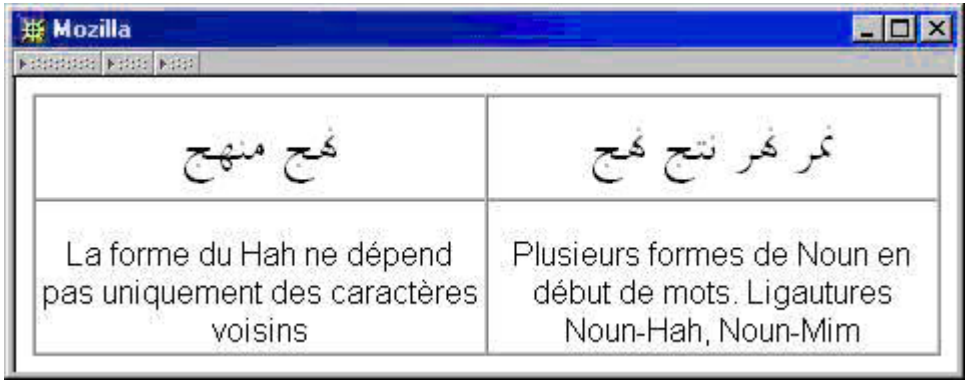
On peut maintenant essayer de penser à un texte arabo-anglais par exemple. L'anglais sera représenté comme en ASCII et l'Arabe avec son code propre; pourvu, que le système reconnaisse Unicode (autrement on voit affiché un point interrogation ?) et dispose des polices qu'il faut. Il y a d'autres représentations comme UTF-16 (dite aussi UCS-16) utilise deux octets pour tous les caractères, UTF-32 qui utilise 32 bits et UTF-7 qui est comme UTF-8, mais avec le 8ieme bit à 0.

Affichage de Texte

A l'affichage d'un texte la notion de caractère disparaît au profit de sa représentation visuelle, appelée glyphe. Un mot constitué d'une séquence de lettres, devient à l'affichage une suite de glyphes. Ceci est particulièrement vrai pour l'Arabe, langue cursive et où la forme d'un caractère (glyphe donc) dépend de l'endroit où il se trouve dans la séquence: au début, au centre et en fin de mot. Il y aussi la forme caractère isolé. Ci-dessous, la lettre Hah, et ses quatres formes.

Dans cet exemple, chaque forme est la même lettre Hah de valeur U+ 0647. Il incombe à l'outil client, browser, éditeur ou autre de prendre en charge l'affichage du bon glyphe. Mais le contexte d'un caractère n'est pas nécessairement la lettre qui suit ou qui précède. L'arabe est une langue très décorative, et il est souvent fait appel à des ligatures, glyphe composé de deux caractères ou plus. Cela est possible grâce à des polices très stylisées comme la police «Naskh» ou «TraditionalArabic» (pour les PCs) pas très courantes malheureusement.

Remarque: Unicode affecte aussi un code à ces ligatures, mais on ne doit l'utiliser qu'exceptionnellement. Les ligatures sont une affaire d'affichage et non pas de caractères dans le texte.



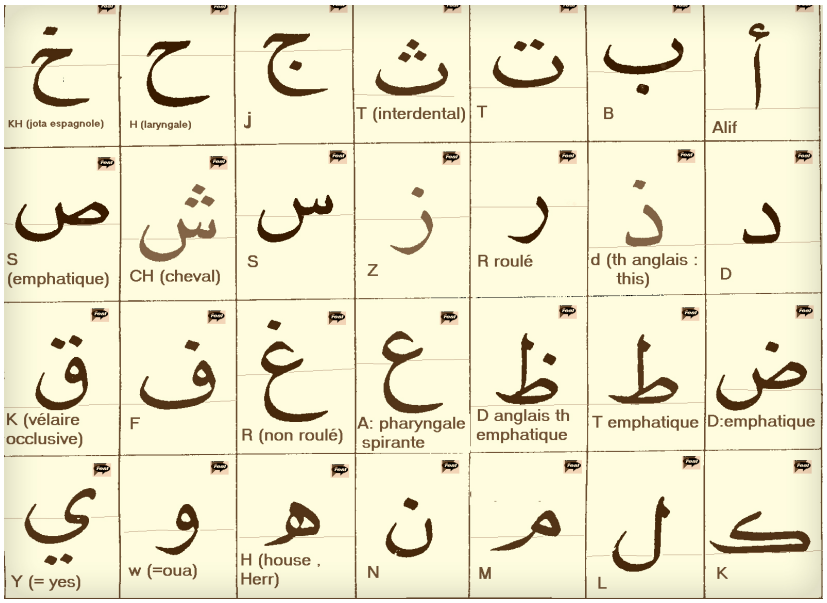
4.2 Le codage des adresses internet

Aujourd'hui, on va de plus en plus vers un internet non latin, c'est un fait que démontre l'évolution récente et la volonté inflexible des langues telles que l'Arabe, le Chinois, le Coréen, le Hindi - pour n'en citer que ceux-là de conquérir la toile. Or, si cette évolution profonde s'est effectuée plutôt d'une façon silencieuse et presque inaperçue, pour un grand nombre d'internautes, elle tient presque de la révolution pour de nombreux autres. Il devient possible pour ces derniers d'accéder à leurs sites préférés, d'envoyer des emails sans devoir utiliser le jeu de caractères latin, sans accent et baptisé Ascii par les anglo-saxons. Depuis déjà de nombreuses années, des expériences ont été menées pour introduire les accents dans les noms de domaine, porte d'entrée obligatoire du réseau des réseaux. Mais l'exploitation complète d'une adresse internet en caractères non latin restait jusqu'à présent impossible puisque l'extension, la partie à droite du point est depuis la création du web obligatoirement en ascii. Sous l'égide de l'ICANN, le régulateur mondial du nommage sur Internet, un programme de développement technologique a été initié il ya quelques années dans le but de permettre l'écriture d'une adresse internet complète en caractères non-latins y compris l'extension. Ce programme a livré ses fruits il y a peu avec l'insertion par l'ICANN dans le système internet de quatre extensions non latines. Il s'agit obligatoirement d'extensions demandées par des pays, en l'occurrence, l'Arabie Saoudite, les Émirats, la Russie et l'Égypte qui sont les pionniers de l'internet non-latin.

Internet Et la langue Arabe

Sans surprise donc, la demande est forte en provenance de pays dont l'alphabet est différent de l'anglo-saxon, les pays arabes mais aussi des pays comme la Chine ou la Corée sans oublier ceux qui écrivent en cyrillique comme la Russie ou la Bulgarie. Les extensions non-latines devraient aider à diminuer le nombre de personnes ne pouvant pas accéder au web dans ces contrées. Rod Beckstrom, le PDG de l'ICANN a assisté personnellement au lancement des extensions de l'Égypte et des Émirats. « Cinq des dix langues les plus utilisées sur Internet ont des alphabets non-latins, le Chinois, le Japonais, l'Arabe, le Russe et le Coréen » a-t-il rappelé lors de la cérémonie Égyptienne. « Cela représente 647 millions d'Internautes. Nous estimons en plus à 310 millions le nombre d'Internautes dont la langue maternelle, bien que hors du top 10 de l'Internet, n'utilise pas l'alphabet latin. Sans l'obligation de maîtriser les 29 lettres de l'alphabet arabe pour aller sur Internet, ces millions de personnes vont enfin pouvoir surfer en version originale. Gageons que très rapidement, ils en viennent à préférer les adresses écrites dans leur alphabet, plutôt que de se forcer à en apprendre un autre ». L'enjeu est la démocratisation du web dans les contrées où la barrière de l'alphabet anglo-saxon est trop forte.

« À voir ce chiffre augmenter économiquement, il serait sot de sous-estimer l'ampleur de ce nouveau phénomène. Le potentiel de captation de nouveaux Internautes offert par les extensions non latines est gigantesque. Depuis 10 ans, le nombre de personnes parlant l'arabe utilisant l'Internet a augmenté de 2300 % » a souligné Rod Beckstrom, toujours en Égypte.



L'alphabet Arabe

5. Le rôle des industries de la langue

La langue est aujourd'hui au cœur du numérique et ce qu'on appelle les technologies de la langue deviennent incontournables pour la réalisation de la Société de l'Information. En effet, l'Information en tant que telle à l'état brut ne constitue pas une finalité. La finalité d'existence d'une information est son utilisation. L'information doit donc être traitée, transformée. Ce sont justement les industries ou les technologies de la langue qui nous fournissent les outils de traitement de l'information comme les moteurs de recherche, les outils de traduction… La rencontre de l'informatique et de la linguistique a donné lieu depuis plus de 50 ans déjà à de nombreuses recherches, développements technologiques et intégration dans des services applicatifs. Des applications traditionnelles telles que la traduction, l'interrogation des bases de données en langage naturel, le dialogue homme/machine ainsi que la reconnaissance et la synthèse de la parole existent déjà bien avant le net.

La croissance de l'Internet et sa convergence avec les autres médias devenus numériques (téléphone, radio, télévision, radio, presse) ont suscité l'apparition de nouveaux besoins. Des techniques de traitement des données textuelles sont apparues pour sélectionner, classer, structurer, … l'ensemble des informations disponibles sous forme numérique qu'elles soient orales ou multimédia, monolingues ou multilingues. Intégrées à une multitude d'applications qui rendent des services tels que la veille informationnelle, la gestion documentaire, la traduction, … les technologies de la langue sont de plus en plus sollicitées pour assurer les tâches complexes que sont la production, l'accès, l'analyse, la transformation et la diffusion des contenus numériques multimédia et multilingues présents sur les réseaux. Les outils développés grâce aux technologies de la langue font naître le besoin et rendent toujours plus nécessaire une approche multilingue du monde. Il faut pouvoir accéder et comprendre (être accédé et se faire comprendre) à l'immensité numérique du monde.

Il est donc essentiel, d'outiller et d'équiper la langue arabe afin qu'elle puisse avoir le statut de langue de communication internationale, de travail, de services et non pas seulement une langue savante ou idéologique. Concrètement, une fois les politiques et les cadres définis, le défi pour la langue se situe donc aujourd'hui dans le développement de cette industrie de la langue pour la mise à disposition d'outils, de ressources, de services et d'applications en langue arabe. Or, si l'industrie de la langue arabe comme d'autres langues non latines n'est pas encore visible à l'échelle internationale, Internet est sans aucun doute l'occasion de booster cette industrie avec les nouveaux besoins chaque jour suscités.

Internet Et la langue Arabe

A l'heure actuelle, il existe presque dans chaque université où centre de recherche arabe, des laboratoires ou des équipes qui travaillent sur des projets ayant trait à l'industrie de la langue arabe. En attendant, Google (pourquoi ne sommes nous pas surpris !) occupe le devant de la scène.

Google à la conquête du monde Arabe

Dans un article datant du 24 septembre 2010, la Compagnie Google prévoit une forte hausse des utilisateurs d'Internet en langue arabe d'ici trois ans, au Maghreb et au Moyen-Orient. Une augmentation notamment due à la progression de la pénétration d'Internet dans la région. L'entreprise compte tirer profit de cette croissance en multipliant outils et services spécifiques aux pays arabes. Début 2010, la société de recherches dans le domaine des TIC, Madar Research estimait à 56 millions les utilisateurs Internet en arabe dans la région MENA, contre 45,6 millions en 2009. D'ici à 2013, ils seront près de 82 millions à naviguer sur le Web dans la langue d'Al Moutanabbi.

Internet Et la langue Arabe

Seulement 1% du contenu Web disponible en arabe

Le succès de Google dans la région est toutefois à relativiser. Selon la Banque Mondiale, plus de 320 millions de personnes parlent arabe dans le monde, alors que moins de 1% du contenu disponible sur Internet est en arabe. Pour y remédier, l'entreprise s'efforce de rendre ses services (tels que Google News, Google Chrome, Traduction, Gmail, etc.) disponibles dans la langue arabe, dans les 90 jours suivant leur lancement. L'entreprise s'efforce également de développer des produits locaux, spécifiques à la région arabe. Parmi ces outils figure Google Ta3rib, un système de translittération qui permet aux internautes ne disposant pas de clavier arabe de rédiger dans cette langue ; le site Web Ahlan, lancé en avril 2010, qui permet aux nouveaux utilisateurs d'apprendre à utiliser certains aspects du Web (comme le Chat, l'e-mailing et le partage d'informations) en regardant des clips vidéo sur YouTube ; ou encore le site Ejabat, une sorte foire aux questions destinée au Moyen-Orient et qui a attiré plus de 100 000 utilisateurs.

Dans ses efforts pour améliorer le volume du contenu arabe en ligne, Google a également lancé des domaines de recherche qui permettent de fournir des informations ciblées. Treize pays sont à ce jour concernés :Algérie, Maroc, Égypte, Territoires palestiniens, Libye, Émirats Arabes Unis, Bahreïn, Qatar, Arabie Saoudite, Jordanie, Koweït, Liban et Oman.

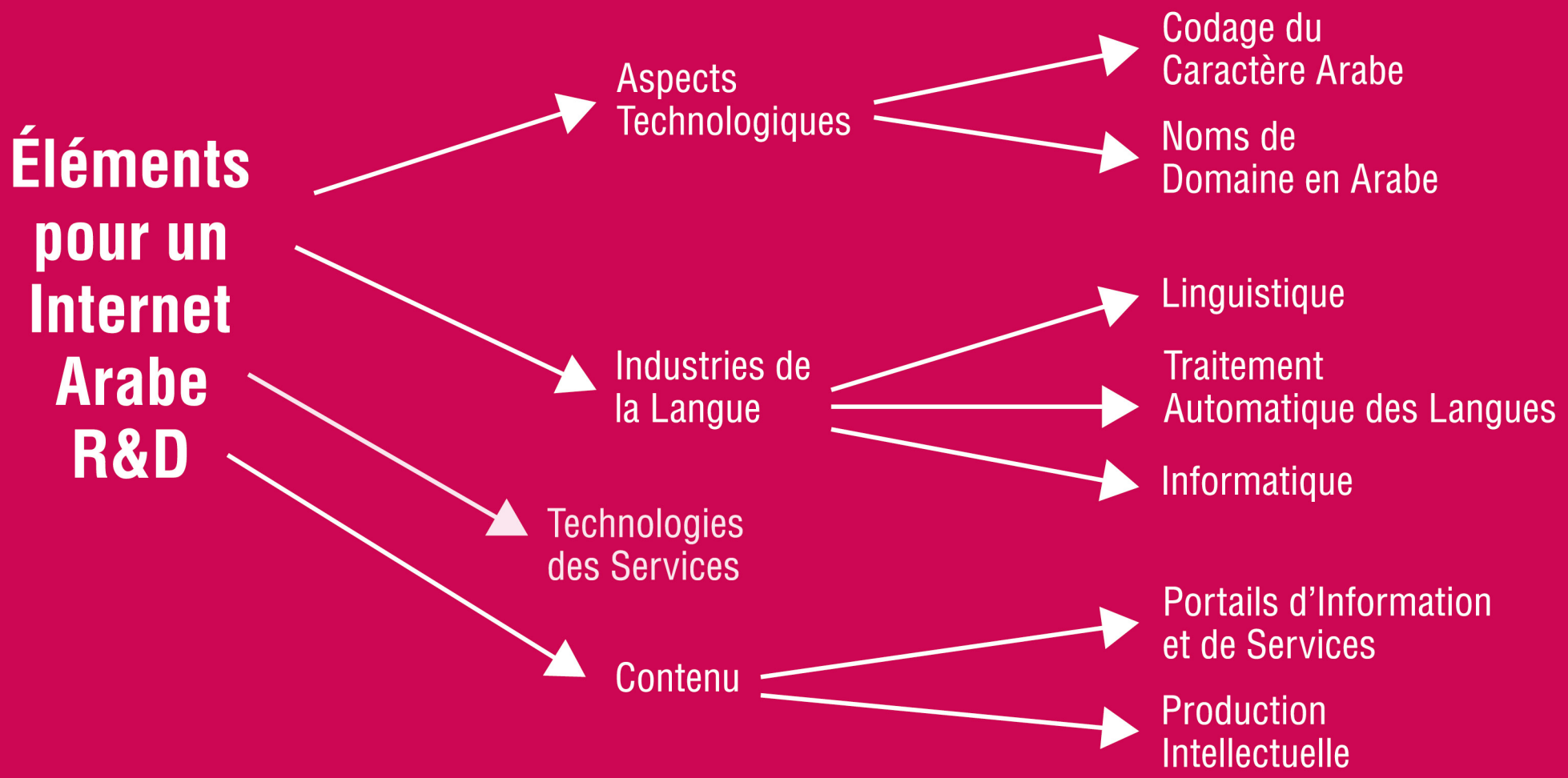
Le géant américain n'est cependant pas seul maître à bord dans le monde arabe. En août 2009, son rival Yahoo avait annoncé le rachat de Maktoob.com, un portail Internet très populaire dans les pays arabes. La firme compte, à travers cette acquisition estimée à 82 millions de dollars, bénéficier des 16,5 millions de visiteurs uniques de Maktoob, et ainsi étendre sa présence dans ces marchés émergents. Comme d'autres pays arabes, en 2008, le CERIST, a lancé un accord avec la firme californienne pour Google.dz qui a été opérationnel le 12 Août de la même année pour la version française, la version arabe a suivi quelques semaines après. L'Algérie a été le 163 ème pays à bénéficier d'un domaine local dans ce moteur de recherche. L'Algérie devient ainsi le 27 ème pays africain à intégrer cette liste et le 11 ème pays arabe après, par ordre alphabétique, l'Arabie Saoudite, le Bahreïn, Djibouti, l'Égypte, les Émirats arabes unis, la Jordanie, la Libye, le Maroc, le Sultanat d'Oman et le Qatar.

Par ailleurs, le célèbre mesureur d'audience Alexa, lui aussi américain, a déjà attribué un classement à ce site en dotant Google version Algérie, du numéro 3.355.539. Ce rang est appelé à régresser au regard du nombre toujours croissant des internautes algériens. Et bien que google occupe le devant de la scène, ceci n'empêche pas des nouveaux moteurs de recherche prenant en charge l'arabe d'apparaitre chaque jour...

Voir le tableau des moteurs de recherche Arabe en page 28

Internet Et la langue Arabe

Internet Et la langue Arabe



Quelques moteurs de recherche prenant en charge l'arabe

Algérie - MarWeb http://arabic.marweb.com Bilingue anglais/arabe	Algérie - Arabji http://www.arabji.com/Arabic/algeria/ Bilingue anglais/arabe	Égypte - Ajeeb http://www.ajeeb.com
Égypte - Masrawy http://www.masrawy.com	Égypte - EgyptWWW : http://www.egyptwww.com/ar/ Bilingue anglais/arabe	Égypte - Egypt.com : http://www.egypt.com Bilingue anglais/arabe
Iran - Parseek : http://www.parseek.com	Iran - Janane : http://www.janane.com	Iran - Google Iran : http://www.google.com/intl/fa/
Annuaire Iraq - Iraq http://www.iraqdirectory.com/ar/ Bilingue anglais/arabe	Anya : http://www.ayna.com/	Koweït - Q8Y2B http://www.q8y2b.com
Arabie Saoudite - Naseej : http://www.naseej.com	Iraq - Fahmi : http://www.fahmi.com	Iran - IranMehr : http://www.iranmehr.com
Arabie Saoudite - Google Arabe http://www.google.com/intl/ar/	Arabie Saoudite - Alnokhba http://www.alnokhba.com	Syrie - Syria Gate : http://www.syriagate.com
Émirats Arabes Unis - AME Info http://www.ameinfo.com/arabic/	Émirats Arabes Unis - Alkhaleej http://www.alkhaleej.co.ae	Arab Wide Web http://www.arabwideweb.com
Émirats Arabes Unis - Albahhar http://www.albahhar.com	Arabo : http://www.arabo.com	MSN Arabe http://arabic.arabia.msn.com
Access GCC : http://www.accessgcc.com	Albawaba : http://www.albawaba.com anglais/arabe	Ame Info : http://www.ameinfo.com/arabic/ Bilingue anglais/arabe
Ajeeb : http://www.ajeeb.com	Arab Info Seek : http://www.arabinfoseek.com	Khayma : http://www.khayma.com
Wahaweb : http://www.wahaweb.com	Raddadi : http://www.raddadi.com	Yasalaam : http://www.yasalaam.net
Arabji http://www.arabji.com/Arabic/ Bilingue anglais/arabe	Arab Data Net : http://www.arabdatanet.com	123 Arab http://www.123arab.com/arabic/ Bilingue anglais/arabe

Conclusion

Le paysage de l'internet arabe se dessine peu à peu, avec une connectivité croissante, de nouveaux besoins apparaissent chaque jour. En dehors de l'élite intellectuelle maitrisant les langues latines français et/ou anglais, le reste de la population arabe qui constitue une majorité, parle l'arabe et a besoin de parler arabe (sous ses différentes formes) sur la toile. Il s'agit donc pour les fournisseurs de service d'offrir des versions arabe, de préférence multilingue de leurs services sous peine de voir une bonne tranche de la société leur tourner le dos. La recherche & développement en technologies de la langue arabe a aussi un rôle central à jouer dans la mise à disposition des contenus en langue arabe. Mais là encore sans les contenus produits, ces recherches ne pourraient être rentabilisées et cela reviendrait à bâtir un moulin mais n'avoir pas du blé à moudre.

Méfiez-vous du Phishing

Le phishing ou hameçonnage est une technique d'ingénierie sociale, qui consiste à exploiter non pas une faille informatique mais la « faille humaine ». Cette technique d'escroquerie consiste à vous subtiliser des données personnelles (mots de passe de connexion à un service, numéro de compte en banque ou de carte bancaire...) en vous piégeant avec un faux courrier électronique qui reprend le logo, la mise en page, l'adresse de votre banque, de votre fournisseur d'accès à Internet, d'un service de messagerie …etc . Le message prétexte un problème lié à votre compte et vous invite à cliquer sur un lien pour donner vos coordonnées. Vous basculez en réalité sur un faux site et ainsi l'attaquant récupère les informations saisies. Les conséquences sont diverses selon le type de renseignements que vous avez fournis. Cela va du pillage de votre compte en banque, en passant par des achats effectués avec votre numéro de carte bancaire. Autre arnaque très en vogue, l'usurpation de votre identité sur des sites de réseaux sociaux comme Facebook ou MySpace.

Quelques règles pour éviter le Phishing

- Ne cliquez pas directement sur le lien contenu dans le mail, ouvrez plutôt votre navigateur et saisissez vous-même l'URL d'accès au service.
- Méfiez-vous des formulaires demandant des informations bancaires. Il est en effet rare (voire impossible) qu'une banque vous demande des renseignements aussi importants par un simple courrier électronique. Dans le doute contactez directement votre agence par téléphone !
- N'envoyez jamais vos mots de passe, identifiants de connexion ou toutes autres informations personnelles par courrier électronique. Méfiez-vous toujours des messages vous invitant à saisir des informations personnelles, même si la demande semble légitime. Si vous pensez avoir été victime de phishing en donnant vos identifiants de connexion et/ou vos mots de passe, changez vos données d'authentification au plus vite.
- Assurez-vous, lorsque vous saisissez des informations sensibles, que le navigateur est en mode sécurisé, c'est-à-dire que l'adresse dans la barre du navigateur commence par https et qu'un petit cadenas est affiché dans la barre d'état au bas de votre navigateur.
- Lorsque vous cliquez sur le lien d'un courriel, vérifiez une fois le navigateur ouvert que l'adresse du site est bien orthographiée. Les attaquants utilisent parfois la même charte graphique d'un site légitime et modifient un ou plusieurs caractères dans l'url afin de faire croire à la victime qu'elle est bien sur le site sollicité.

Pour plus d'informations, veuillez consulter notre site Internet : **www.wikayanet.dz**

Zoom Sur Un Projet

Hassina Aliane
Chargée de Recherche
Équipe Web Sémantique et Langue Arabe
Division Théorie et Ingénierie des Systèmes Informatiques

Le Projet AL-KHALIL : Contribution à l'Internet Arabe

Le Projet AL-KHALIL : Contribution à l'Internet Arabe

Introduction

Bien que la langue Arabe soit la langue de centaines de millions de personnes à travers le monde, dans le monde de l'informatique, elle dispose de bien peu de ressources, outils et applications. A l'âge de l'Internet et des enjeux de l'existence numérique, les langues sont en train de se frayer un chemin, et on voit de plus en plus émerger des langues comme le chinois, l'hindou, le coréen et l'arabe car à l'image de l'humanité, l'internet parle et parlera toutes les langues du monde. La langue arabe a accusé un retard technologique que tout le monde connait et beaucoup reste affaire. Le projet que nous présentons ici, se veut une contribution pour apporter une pierre à l'édifice qui prendra de plus en plus forme, nous en sommes convaincus, en particulier grâce à la volonté des utilisateurs de plus en plus nombreux à s'approprier la toile et les exigences qui s'en suivent pour la Recherche & Développement.

Objectifs du Projet

Notre projet consiste à construire une infrastructure centrée ontologie pour des ressources, des applications et des services en langue arabe et se situe à la jointure du domaine du traitement automatique des langues naturelles (TAL) et des technologies du web sémantique tout en se fondant sur les résultats de nos recherches sur la langue Arabe. Notre objectif est de contribuer à combler le vide numérique dans ce domaine de deux manières :

- Donner plus de visibilité à la linguistique Arabe et faire connaitre les travaux dans le domaine,
- Notre ontologie servira d'infrastructure (ouverte) pour construire différentes applications pour les linguistes et la communauté TAL et en même temps constituera le noyau qui pourra supporter des outils et des applications utilisateurs tels, la recherche d'information, l'extraction de connaissances, …

Méthodologie

Pour atteindre notre objectif, la méthodologie que nous avons adoptée s'articule autour des étapes suivantes :

1. Délimiter le contenu et les utilisateurs de l'ontologie,
2. Choisir une approche motivée pour le développement de cette ontologie ainsi que les outils de développement,
3. Validation de l'ontologie,
4. Mise en ligne de l'Ontologie,
5. Définir les besoins en ressources et applications nécessaires et utiles qui utiliseront cette ontologie.
6. Définir un planning de priorité et lancer le développement d'applications et de services
7. Mise à disposition et évaluation.

AL-KHALIL : Une infrastructure centrée ontologie pour la langue Arabe

A l'heure actuelle, avec la profusion de données textuelles sur internet et le besoin de différentes applications pour l'utilisation de ces données, le chercheur dans différents domaines faisant intervenir les données textuelles (RI, indexation, …) se trouve confronté au besoin de certains outils de traitement de données linguistiques dont la disponibilité lui ferait gagner beaucoup de temps et lui permettrait de se concentrer sur sa problématique de base. Avant le développement d'internet, chacun (y compris nous-mêmes) développait ses outils au besoin et ces outils étaient donc intégrés dans le système (par exemple de recherche ou d'indexation lui-même). Aujourd'hui, la tendance est à la mise à disposition séparément d'outils tels que les lemmatiseurs, les taggers, les analyseurs morphologiques, .., D'ailleurs, il en existe même un assez grand nombre de tels outils pour les langues française et anglaise. Par contre, il est de notoriété que la langue arabe manque de tels outils disponibles. Nous développons nous-même ces outils dans le cadre du projet « AlKhalil ; une ontologie de la langue arabe et aussi dans le projet indexation, extraction et RI multilingues ».

Les ontologies linguistiques connaissent aujourd'hui un intérêt grandissant pour la communauté TAL, comme le Wordnet par exemple et ce pour différentes applications basées sur le contenu telles, l'indexation conceptuelle, la désambigüisation sémantique et la recherche d'information. Récemment, l'intérêt grandissant de la communauté TAL pour les ontologies a conduit au développement d'ontologies pour le TAL à différentes fins. Les ontologies sont aussi au cœur de notre projet.

Nous avons baptisé notre projet AL-KHALIL du nom de l'ancien grammairien arabe, Al-khalil ibn Ahmad Al-farahidi dont le fameux « KITAB AL'AYN» constitue en quelque sorte la première ontologie de la langue arabe. KITAB AL'AYN signifie le livre de la lettre "AYN" car le dictionnaire suivait un ordre phonétique commençant par le son pharyngal "AYN".

Zoom Sur Un Projet

Bien que nous allions avoir d'autres ontologies dans notre infrastructure, nous avons choisi une ontologie linguistique comme noyau car les autres ontologies ainsi que les applications que nous escomptons feront appel à diverses étapes à de l'information linguistique. Ce projet a fait l'objet d'une communication à la conférence internationale sur les ressources langagières en 2010. L'ontologie a déjà été développée et en phase de tests et validation, elle a été développée selon l'approche suivante :

- Initialisation manuelle en choisissant les concepts de la linguistique arabe et en les relatant à une ontologie générale de linguistique appelée Gold.
- Utilisation d'un algorithme d'extraction de connaissances pour enrichir l'ontologie. L'architecture générale de ce système est décrite par la figure 1.

Les outils et les ressources linguistiques autour de l'ontologie

Nous avons adopté dans notre projet une approche modulaire et ouverte de sorte à offrir entre autres une boite à outils pour le TAL ainsi que des lexiques et dictionnaires qui pourraient être utilisés de façon autonome par d'autres chercheurs et d'autres projets.

Une boite à outils pour le TAL Arabe comprenant :

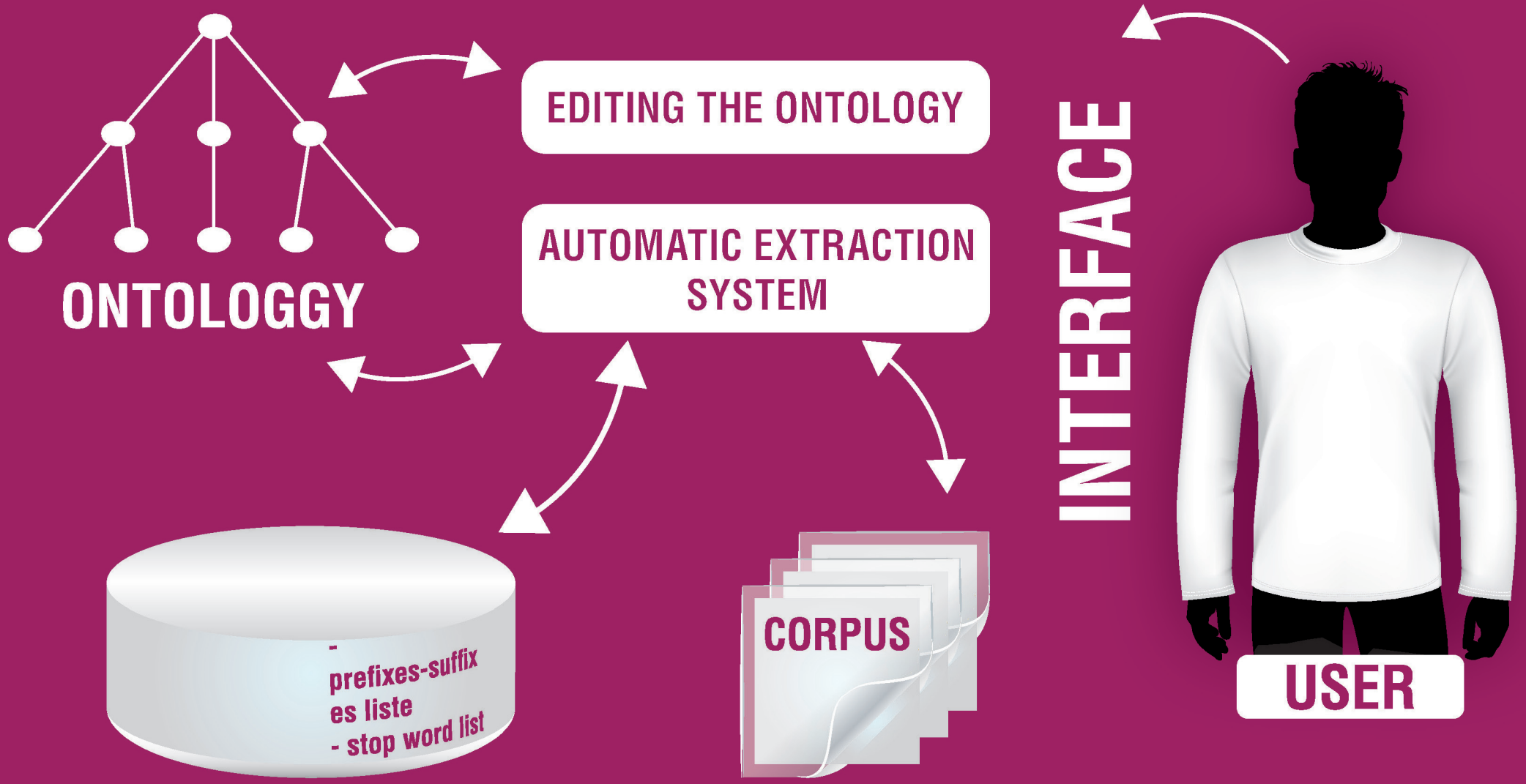
- un outil de segmentation et de lemmatisation
- un étiqueteur utile pour l'analyse syntaxique et l'annotation de corpus ou la création de métadonnées.
- Un analyseur morpho -lexical
- Un analyseur syntaxique

Un ensemble de ressources comprenant :

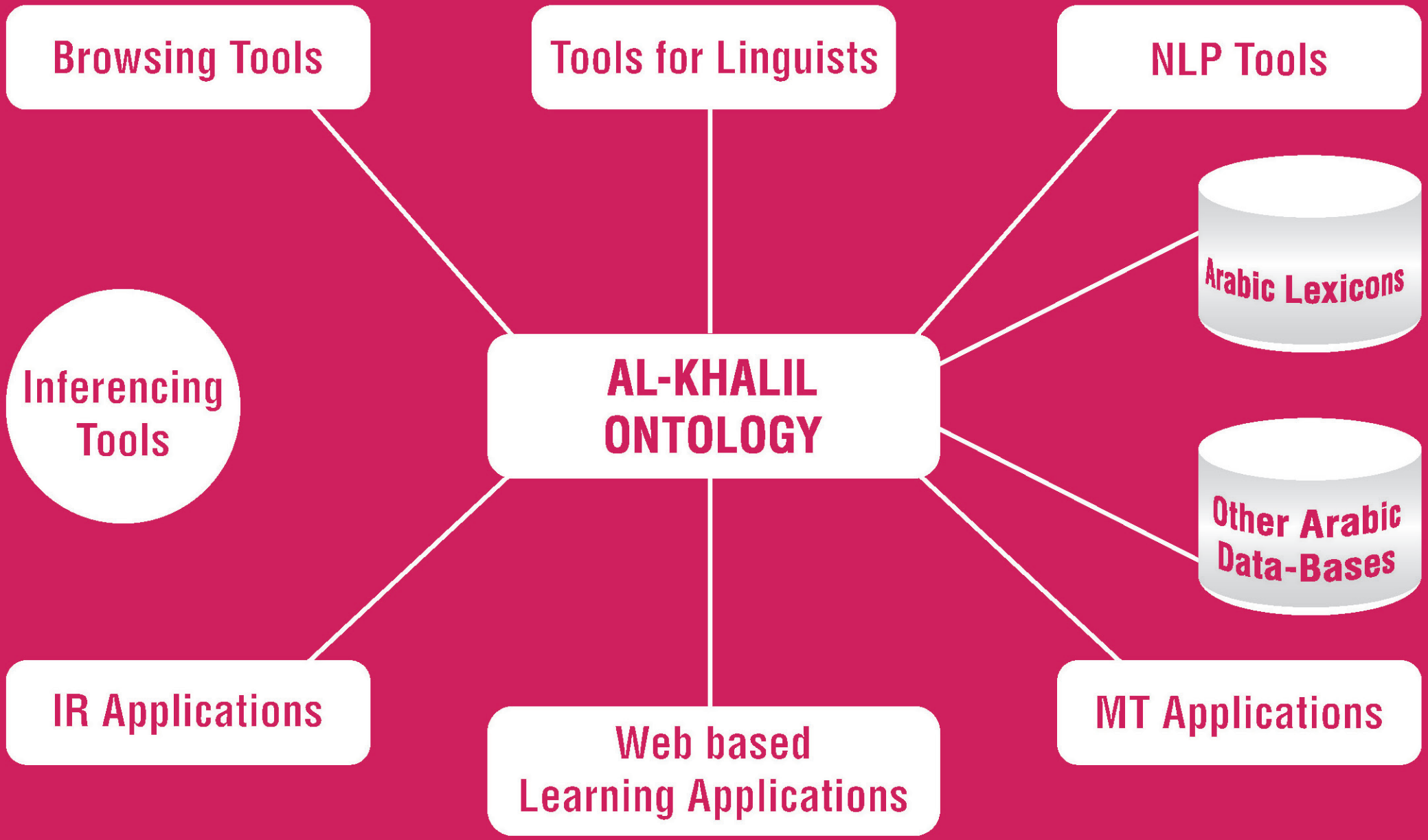
- Une ontologie pour la linguistique arabe (c'est le noyau de notre plateforme comme nous l'avons mentionné) utile aussi bien au linguiste qu'au chercheur en TAL Arabe.
- Un dictionnaire électronique de la langue Arabe qui sera aussi supporté par une ontologie
- Des ontologies de domaine créées au besoin : une ontologie pour le journal officiel est actuellement en cours de développement

Applications envisagées

- E-Learning pour la langue arabe
- Utilisation des outils dans le cadre de notre projet indexation et RI multilingue *(voir bulletin CERISTNEWS numéro 2)*.
- Développement de services autour des ontologies de domaines.


ONTOLOGGY
EDITING THE ONTOLOGY
AUTOMATIC EXTRACTION SYSTEM
INTERFACE
USER
-
prefixes-suffix
es liste
- stop word list
CORPUS

AL-KHALIL ONTOLOGY PROJECT



FORMATION

Une formation portant sur l'initiation à l'utilisation du logiciel de gestion des bibliothèques « SYNGEB » a été assurée au CERIST, du 22 au 25 novembre 2010, par des ingénieurs du département Information Scientifique et Technique au profit des cadres de l'Office National des Statistiques, de l'École des banques, de l'Institut National de Formation du Personnel de l'Éducation, de l'École Nationale des Sciences et Technologies de Rouïba ainsi qu'aux gestionnaires de la maison de culture MILA et du Musée National de l'Antiquité.

RAPPORTS DE RECHERCHE INTERNES

Benmeziane Souad, Badache Nadjib, Tor Network Limits. Alger: CERIST, 2010. ISRN CERIST-DTISI/RS--10-000000021—dz. http://www.cerist.dz/publication/index.php?option=com_content&task=view&id=575&Itemid=52

Bouchama Samira, Hamami Latifa, Aliane Hassina, Watermarking of compressed video based on DCT coefficients and watermark preprocessing. Alger: CERIST, 2010. ISRN CERIST-DTISI/RR--10-000000022--dz. http://www.cerist.dz/publication/index.php?option=com_content&task=view&id=576&Itemid=52

CERIST

Bases de données documentaires

Accessibles sur : www.cerist.dz

Le CERIST permet l'accès à la documentation scientifique et technique à travers des bases de données et sources d'information internationales.

L'accès est établi par reconnaissance de l'adresse IP du proxy du CERIST ou via le réseau ARN.

ACM Digital Library

ACM Digital Library propose un accès à 50 ans d'archives et 1,4 million de pages de texte issues des : Journals, Magazines, Transactions, Proceedings, Newsletters, Publications by Affiliated Organizations, Special Interest Groups (SIGs).

INIS

Le système d'information INIS collecte depuis 1970 la littérature scientifique et technique du monde entier sur les applications pacifiques des sciences et technologies nucléaires. Il offre plus de 3 millions de notices indexées.

CHICAGO JOURNAL

La base de données en ligne de l'Université de Chicago publie plus de 50 revues en sciences sociales et humaines, en éducation, biologie et sciences médicales, ainsi qu'en physique.

JSTOR

Un site d'archives électroniques donnant accès en texte intégral à plus de 500 périodiques dés leur première édition jusqu'aux numéros récents.

SPIE Digital Library

La Bibliothèque Numérique SPIE fournit un accès sans précédent à plus de 275,000 articles des revues SPIE (SPIE journals) et des Actes de conférences datant de 1990 à ce jour. Plus de 17000 nouveaux articles de recherche sont ajoutés annuellement.

Directeur de publication

Pr. BADACHE Nadjib

Dossier : INTERNET ET LA LANGUE ARABE. Réalisé par :

Hassina Aliane

Rubrique : Les Conseils de DZ - CERT

L'ÉQUIPE DZ-CERT

Rubrique : Zoom sur un Projet

Hassina Aliane - Chargée de Recherche
Équipe Web Sémantique et Langue Arabe
Division Théorie et Ingénierie
des Systèmes Informatiques - CERIST

Comité de communication et de rédaction

BEBBOUCHI Dalila

BENNADJI Khedidja

Photographies

ALIMIHOUB Dahmane

Réalisation graphique

BOUDIA Nacer

Publié par le CERIST

5, rue des 3 Frères Aissou. Ben Aknoun. BP 143, 16030 - Alger
Tél : +213 (21) 91 62 05 – 08 / Fax : +213 (21) 91 21 26

E - mail : vrr@mail.cerist.dz

www.cerist.dz

Impression

ANEP

ISSN : 2170- 0656 / **DÉPÔT LÉGAL :** 2690-2010

Le Bulletin CERISTNEWS

CENTRE DE RECHERCHE SUR L'INFORMATION SCIENTIFIQUE ET TECHNIQUE - CERIST
5, Rue des Trois Frères Aissou, Ben - Aknoun - BP 143. 16030 - Alger
Tél : +213 (21) 91 62 05 - 08 / Fax : +213 (21) 91 21 26

www.cerist.dz