SaudiNIC’s Notes and Thoughts on

Presentations of Arabic Web/Email Addresses

SaudiNIC
10/2/2018

CONTENTS

SaudiNIC's Notes and Thoughts on Presentations of Arabic Web/Email Addresses

EXECUTIVE SUMMARY

There should be no mixing, apart from numbers, of right to left and left to right scripts in domain names, mailbox names, and email addresses.

I. INTRODUCTION

The Arabic script is the writing system used for writing Arabic and several other languages of Asia and Africa, such as Azerbaijani, Pashto, Persian, Kurdish, Lurish, Urdu, Mandinka, and others. It is the second-most widely used writing system in the world by the number of countries using it and the third by the number of users, after Latin and Chinese characters. The Arabic script is written from right to left in a cursive style. *(Source Wikipedia)*

Using Arabic script in domain names (i.e., Internationalized Domain Names - IDN) or in email addresses raise a serious concern, that is mixing RTL (Right-To-Left) and LTR (Left-To-Right) scripts within domain names and email addresses.

In fact, mixing RTL with LTR code-points within a label or a domain name add yet another complexity (security and user confusability). In this short report we will concentrate on problems due to mixing RTL with LTR code-points in domain names or email addresses. They are presented from Arabic-speaking user expectation viewpoint.

In nutshell, we strongly believe that mixing RTL with LTR code-points (except digits) in labels, domain names, and email addresses (mailbox) will be confusing, illogical, unacceptable, and un-useful to the Arabic-speaking communities. We also believe it will create a playground for domain/email phishing.

In this report we illustrate acceptable (from our standpoint) and widely used domains or email addresses that are mainly based on RTL code-points. Additionally, we will demonstrate how mixing RTL with LTR code-points in domain names or email addresses will cause confusion and awkward usage. Finally, we will provide our conclusions and recommendations.

Before we discuss mixing RTL and LTR in domain names and email addresses, it is worth to spend some time to see experiences and expectations of Arabic speakers with respect to mixing RTL and LTR in their writing. Next Section will demonstrate some of these experiences and expectations.

II. Arabic Users Expectations Regarding Mixing RTL/LTR

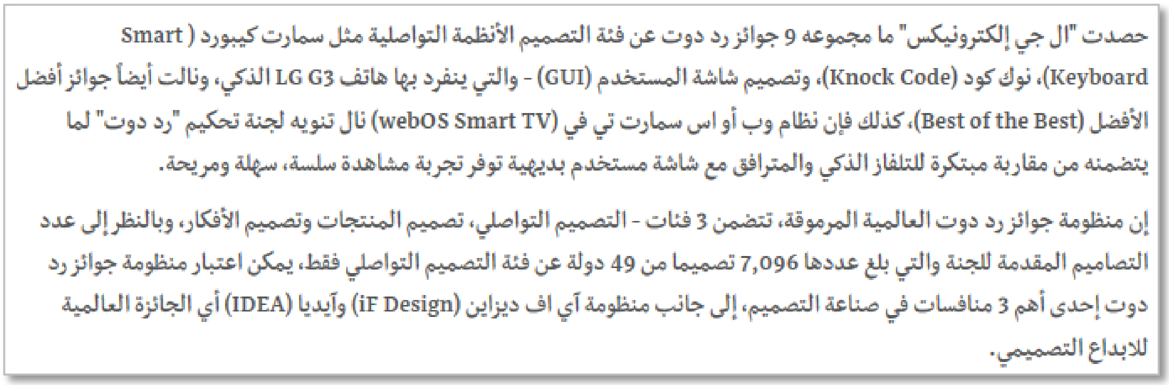
In this section we will illustrate the widely used practice in dealing with mixing LTR and RTL texts. Arabic speakers have been dealing with LTR and RTL in their writing for a long time (before Computers where invented). Consider the following examples:

Person Type	Fist Name	Last Name	Name in English (LTR)	Name in Arabic (RTL)
English Person	John	Smith	John Smith	جون سميث
Arabic Person	صالح	الفلاني	Saleh AlFulani	صالح الفلاني

Text Type	Example
LTR Paragraph	Mr. *Saleh* AlFulani *organized a workshop and met Mr. Jon Smith and then signed the agreement.*
LTR Paragraph with RTL text	Mr. *Saleh* AlFulani *(صالح الفلاني) organized a workshop and met Mr. Jon Smith and then signed the agreement.*
RTL Paragraph	*نظم السيد صالح الفلاني ورشة عمل واجتمع مع السيد جون سميث ومن ثم وقعا الاتفاقية.*
RTL Paragraph with LTR text	*نظم السيد صالح الفلاني ورشة عمل واجتمع مع السيد جون سميث (John Smith) ومن ثم وقعا الاتفاقية.*

As you can see, English names (in an RTL paragraph) or Arabic names (in an LTR paragraph) are written without changing their direction (original structure).

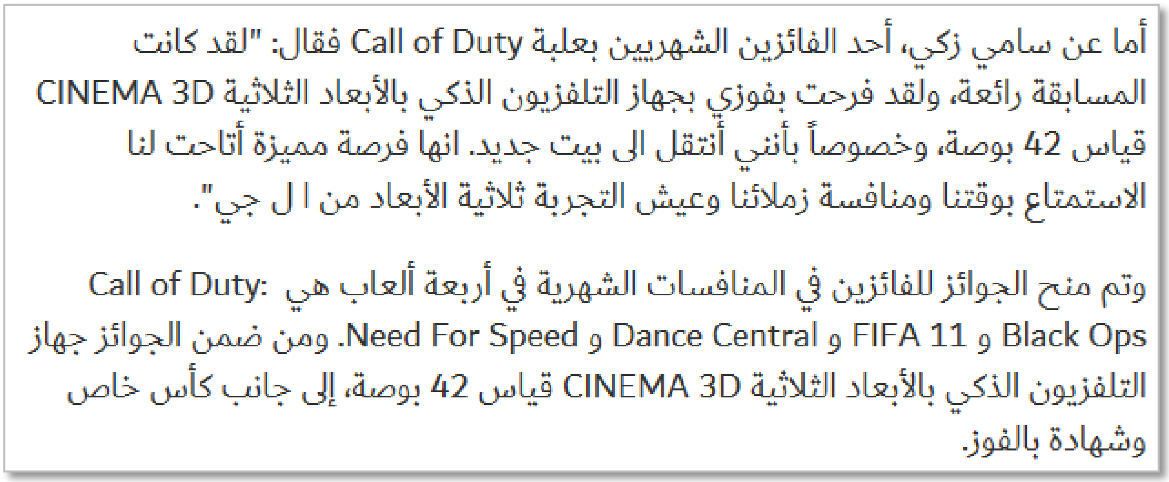
Here are some samples from newspapers inside the Arabic world:



حصدت "ال جي إلكترونيكس" ما مجموعه 9 جوائز رد دوت عن فئة التصميم الأنظمة التواصلية مثل سمارت كيبورد (Smart
Keyboard)، نوك كود (Knock Code)، وتصميم شاشة المستخدم (GUI) - والتي ينفرد بها هاتف LG G3 الذكي، ونالت أيضاً جوائز أفضل
الأفضل (Best of the Best)، كذلك فإن نظام وب أو اس سمارت تي في (webOS Smart TV) نال تنويه لجنة تحكيم "رد دوت" لما
يتضمنه من مقاربة مبتكرة للتلفاز الذكي والمترافق مع شاشة مستخدم بديهية توفر تجربة مشاهدة سلسة، سهلة ومريحة.

إن منظومة جوائز رد دوت العالمية المرموقة، تتضمن 3 فئات - التصميم التواصلي، تصميم المنتجات وتصميم الأفكار، وبالنظر إلى عدد
التصاميم المقدمة للجنة والتي بلغ عددها 7,096 تصميما من 49 دولة عن فئة التصميم التواصلي فقط، يمكن اعتبار منظومة جوائز رد
دوت إحدى أهم 3 منافسات في صناعة التصميم، إلى جانب منظومة آي اف ديزاين (iF Design) وآيديا (IDEA) أي الجائزة العالمية
للابداع التصميمي.

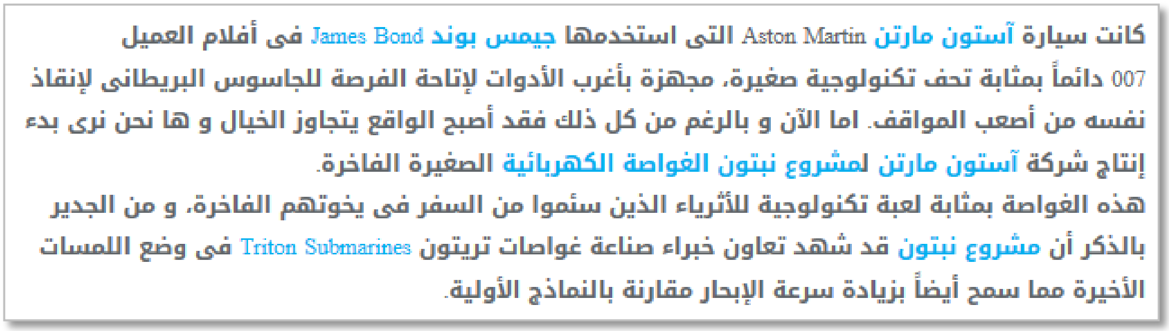
Source: http://www.alriyadh.com/975687



أما عن سامي زكي، أحد الفائزين الشهريين بعلبة Call of Duty فقال: "لقد كانت
المسابقة رائعة، ولقد فرحت بفوزي بجهاز التلفزيون الذكي بالأبعاد الثلاثية CINEMA 3D
قياس 42 بوصة، وخصوصاً بأنني أنتقل الى بيت جديد. انها فرصة مميزة أتاحت لنا
الاستمتاع بوقتنا ومنافسة زملائنا وعيش التجربة ثلاثية الأبعاد من ا ل جي".

وتم منح الجوائز للفائزين في المنافسات الشهرية في أربعة ألعاب هي: Call of Duty
Black Ops و FIFA 11 و Dance Central و Need For Speed. ومن ضمن الجوائز جهاز
التلفزيون الذكي بالأبعاد الثلاثية CINEMA 3D قياس 42 بوصة، إلى جانب كأس خاص
وشهادة بالفوز.

Source: http://www.albayan.ae/economy/last-deal/2011-12-09-1.1551679



كانت سيارة آستون مارتن Aston Martin التى استخدمها جيمس بوند James Bond فى أفلام العميل
007 دائماً بمثابة تحف تكنولوجية صغيرة، مجهزة بأغرب الأدوات لإتاحة الفرصة للجاسوس البريطانى لإنقاذ
نفسه من أصعب المواقف. اما الآن و بالرغم من كل ذلك فقد أصبح الواقع يتجاوز الخيال و ها نحن نرى بدء
إنتاج شركة آستون مارتن لمشروع نبتون الغواصة الكهربائية الصغيرة الفاخرة.
هذه الغواصة بمثابة لعبة تكنولوجية للأثرياء الذين سئموا من السفر فى يخوتهم الفاخرة، و من الجدير
بالذكر أن مشروع نبتون قد شهد تعاون خبراء صناعة غواصات تريتون Triton Submarines فى وضع اللمسات
الأخيرة مما سمح أيضاً بزيادة سرعة الإبحار مقارنة بالنماذج الأولية.

Source: http://aitmag.ahram.org.eg/News/99214.aspx

III. Normal (Expected) Web/Email Address Presentations

As mentioned above, the Arabic script is written from right to left. Therefore, names (first name, last name) and, in particular, domain names and email addresses are typically written in RTL, e.g.,

Type	Example	Components	Comments
Personal name	صالح الفلاني	First Name: صالح Last Name: الفلاني	
Domain Name	سجل.السعودية	2nd level domain: سجل TLD: السعودية	The Arabic IDN TLD is always expected to be the last label from right-to left.
Email Address	الدعم@رسيل.السعودية	User: الدعم Domain: رسيل.السعودية	The user part is always to the right-side of the (@) sign

In the Latin-world, the following web and email address formats are used:

web address: domain.tld
email address: user.mailbox@domain.tld

This can be easily deciphered by normal users to the following:

1. The user part is always to the left-side of the sign (@): user.mailbox
2. The domain name is always to the right-side of the sign (@): domain.tld
3. A domain name is arranged in a well-defined label hierarchy where a TLD is always the rightmost label of the domain name: .tld

Hence, the email address (user.mailbox@domain.tld) will be used without altering its direction or swapping between its parts regardless of the direction of the text. For example, see the following English and Arabic tests:

Text direction	Text
LTR	*Please download the form from our site: domain.tld, then fill it and send it to us in using the following email address: user.mailbox@domain.tld*
RTL	*آمل تنزيل النموذج من موقعنا على الإنترنت: domain.tld ، وتعبئته ومن ثم إرساله إلينا على البريد الإلكتروني التالي: user.mailbox@domain.tld*

As you can see, regardless of the text direction (LTR or RTL) the email address maintains its original form and integrity (i.e., user.mailbox@domain.tld). This will allow users to easily and confidently construct and deconstruct email addresses correctly without confusion or mistakes. Now, let us demonstrate this using more complicated email addresses. Consider the following examples:

care.sa@car.com
car.com@care.sa

They are straightforwardly interpreted as follows (no confusion whatsoever):

Email Address	User Part	Domain	TLD
care.sa@car.com	care.sa	car.com	.com
car.com@care.sa	car.com	care.sa	.sa

Now let us move to the Arabic-world and construct/deconstruct Arabic web and email addresses. The Arabic email address format is expected to be as follows (RTL):

Web Address	2nd Level Domain	TLD
اسم-الجهة.نطاق-علوي	اسم-الجهة	.نطاق-علوي

Email Address	User Part	Domain	TLD
صندوق.البريد@اسم-الجهة.نطاق-علوي	صندوق.البريد	اسم-الجهة.نطاق-علوي	.نطاق-علوي

A native Arabic-speaking user would expect the following regardless of the text direction:

1. The user part is always to the right-side of the sign (@): صندوق.البريد
2. The domain name is always to the left-side of the sign (@): اسم-الجهة.نطاق-علوي
3. A domain name is arranged in a well-defined label hierarchy where an Arabic TLD is always the leftmost label of a domain name: نطاق-علوي.

Therefore, a given Arabic email address, such as:

صندوق.البريد@اسم-الجهة.نطاق-علوي

should be used without altering its direction or swapping between its parts regardless of the text direction (LTR or RTL) so that the email address maintains its original form and integrity and hence remove any confusion or misinterpretation. For example, see the following English and Arabic texts:

Text direction	Text
LTR	*Please download the form from our site: اسم-الجهة.نطاق-علوي, then fill it and send it to us using the following email address: صندوق.البريد@اسم-الجهة.نطاق-علوي*
RTL	*آمل تنزيل النموذج من موقعنا على الإنترنت: اسم-الجهة.نطاق-علوي ، وتعبئته ومن ثم إرساله إلينا على البريد الإلكتروني التالي: صندوق.البريد@اسم-الجهة.نطاق-علوي*

As you can see, regardless of the text direction (LTR or RTL) the Arabic email address maintains its original form. This will allow users to easily and confidently construct and deconstruct email addresses correctly without confusion or mistakes. Now, let us demonstrate this using a more complicated email addresses. Consider the following examples:

منتج.السعودية@خدمة.عرب

خدمة.عرب@منتج.السعودية

They are straightforwardly interpreted as follows (no confusion whatsoever):

Email Address	User Part	Domain	TLD
منتج.السعودية@خدمة.عرب	منتج.السعودية	خدمة.عرب	.عرب
خدمة.عرب@منتج.السعودية	خدمة.عرب	منتج.السعودية	.السعودية

IV. AWKWARD DOMAIN/EMAIL ADDRESS PRESENTATIONS

With many variations in how mixing RTL and LTR in domains and email addresses. However, some of them undergo major concerns from users' expectation viewpoint. Here is a summary and conclusion related to mixing RTL and LTR in domains and email addresses (mailboxes):

1) Mixing RTL and LTR within a label of a domain name or across all the labels:

   The entire label(s) (as part of a domain name or across the whole domain) should be formulated from a single script and a single direction (RTL or LTR) with the exception of digits (LTR) that can be in the middle or at the end of that label, i.e., no mixture of Arabic (RTL) and ASCII (LTR) code points within a domain name label or across all the domain labels. Thus, the following examples are not accepted:

Format	Example
<FirstName><Arabic-LastName>	*Saleh*الفلاني
<Arabic-LastName><FirstName>	الفلاني*Saleh*
<FirstName>.<Arabic-LastName>	*Saleh*.الفلاني
<TLD>.<Arabic-Domain>	sa.رسيل
<Domain>.<Arabic-TLD>	raseel.السعودية

2) Mixing RTL and LTR within the user part of an email address (EAI)

   It is the same as the previous point (mixing in domain labels), no mixing should be allowed. Thus, the following examples are not accepted:

Format	Example
<FirstName>.<Arabic-LastName>	*Saleh*.الفلاني
<Arabic-FirstName>.<LastName>	صالح.alfulani

3) Mixing RTL and LTR between domain and mailbox

   The entire domain name part (i.e. all labels, e.g., domain.tld) and the entire user part (the mailbox name, e.g. FirstName.LastName@) should be formulated from a single script (with the exception of digits with a condition (that are LTR)), i.e., no mixture of Arabic (RTL) and ASCII (LTR) code points at all. Thus, some of the following examples are clear and understandable by Arabic users while others are not:

User Direction	Domain Direction	Real example	Clear to Arabic users?
LTR	LTR	raed.alfayez@raseel.sa	Yes
RTL	RTL	رائد.الفايز@رسيل.السعودية	Yes
RTL	LTR	رائد.الفايز@raseel.sa	No
LTR	RTL	رسيل.السعودية@raed.alfayez	No

Please note, the last two rows are not easy to deal with, to implement, or to differentiate between mailbox and domain parts from a reader's point of view.

4) Display issues when having an RTL domain or email in LTR context (e.g. inserting an Arabic domain/email in an English article or vice versa):
   - RTL text should remain intact and in correct order all the time regardless of the context. The RTL mailbox part should be always at the right of (@) sign and should maintain the right order of its components (e.g., صالح.الفلاني). The RTL domain name part should be always at left of (@) sign and should maintain the right order of subdomains (if exists), domain, and TLD (e.g., رسيل.السعودية).

V. DIGITS

Users who use an Arabic script to write Arabic-based languages (e.g., Arabic, Urdu, Persian …) use one or more set of digits in their normal writing without mixing them together in writing numbers. These set are (according to Unicode terminologies):

1.	European digits	U+0030 .. U+0039	(0123456789)
2.	Arabic-Indic digits	U+0660 .. U0669	(٠١٢٣٤٥٦٧٨٩)
3.	Eastern Arabic-Indic digits	U+06F0 .. U+06F9	(۰۱۲۳۴۵۶۷۸۹)

The three sets of digits mean the same (zero to nine) despite their differences in shape, and they all written LTR.

Even in one language community such as the Arabic speaking community, users are using different digits. For example, eastern Arab region (e.g., Egypt, Syria, Sudan, Iraq, all GCC countries, Lebanon, Palestine, Jordan, …) are mainly using Arabic-Indic digits while the western Arab region (e.g., Libya, Tunis, Algeria, Morocco, Mauritania, …) mainly use European digits. But never mixing them together while writing numbers.

#	Display	Input/Action									
1	مؤتمر2056	**Input string order**	1	2	3	4	5	6	7	8	9
		Code point	م	ؤ	ت	م	ر	2	0	5	6
		Acceptable: Pure European digits									
2	مؤتمر٢٠٥٦	**Input string order**	1	2	3	4	5	6	7	8	9
		Code point	م	ؤ	ت	م	ر	٢	٠	٥	٦
		Acceptable: Pure Arabic-Indic digits									
3	مؤتمر۲۰۵۶	**Input string order**	1	2	3	4	5	6	7	8	9
		Code point	م	ؤ	ت	م	ر	۲	۰	۵	۶
		Acceptable: Pure Eastern Arabic-Indic digits									
4	مؤتمر٢٠56	**Input string order**	1	2	3	4	5	6	7	8	9
		Code point	م	ؤ	ت	م	ر	٢	٠	5	6
		Not-Acceptable: Mix between European digits & Arabic-Indic digits									
5	مؤتمر٢٠٥۶	**Input string order**	1	2	3	4	5	6	7	8	9
		Code point	م	ؤ	ت	م	ر	٢	٠	٥	۶
		Not-Acceptable: Mix between Arabic-Indic digits & Eastern Arabic-Indic digits									
6	مؤتمر20۵۶	**Input string order**	1	2	3	4	5	6	7	8	9
		Code point	م	ؤ	ت	م	ر	2	0	۵	۶
		Not-Acceptable: Mix between European digits & Eastern Arabic-Indic digits									

It seems that this issue has been resolved in the current IDNA2008 implementation and it is a good practice since:

- It ensures that non-mixing of digits is obligatory and not left to individual registry decisions.
- Numbers are different from letters in this specific case. The 3 number sets are distinct and separation could be easily addressed on the protocol level, whereas in visual confusion of letters, there's no clear grouping for confusingly similar characters, the existing overlap between language tables makes it hard to separate, in addition, communities using the Arabic script have different needs and different user experiences.

- Protocol-level solution would limit the maximum number of labels to be registered to 3, independent of the number of digits contained in a given label.

In the context of Arabic IDN, the label separator (the Dot) will be considered as a decimal point when it is proceeded and followed by digits. In this case, the display sequence of the digits will get mixed up. Additionally, some browsers (e.g., MS IE) replace the decimal point (".") with the decimal coma (",") in the URL. Therefore, we recommend using digits only in the middle or end of a label and not in the beginning of a label.

VI. CONCLUSION

Mixing RTL with LTR code-points within a label or a domain name raises some issues with respect to security and user confusability. We strongly believe that mixing RTL with LTR code-points (except digits) in labels, domain names, and email addresses (mailbox) will be confusing, illogical, unacceptable, and un-useful to the Arabic-speaking user communities. We also believe it will be a playground for domain/email phishing. Therefore, a given Arabic email address, such as:

صندوق.البريد@اسم-الجهة.نطاق-علوي

should be used without altering its direction or swapping between its parts regardless of the text direction (LTR or RTL) so that the email address maintains its original form and integrity and hence remove any confusion or misinterpretation.

Our recommendations:

- Registries enforce a non-mixing of scripts at a level that they can (e.g., the 2nd level)
- Mailbox administrators enforce a non-mixing of scripts in the mailbox name

Additionally, we recommend using symbols for protocol names (such as http, https, mailto, ..etc.) instead of clear text so that it will not add more complexity due to mixing RTL with LTR.

VII. FOR MORE INFORMATION

https://www.w3.org/TR/charmod/

https://unicode.org/reports/tr9/

https://www.w3.org/TR/html-bidi/

https://tools.ietf.org/html/rfc5893