

Новые технологические возможности для кириллической почты

Юрий Каргаполов
Юрий Гончарук

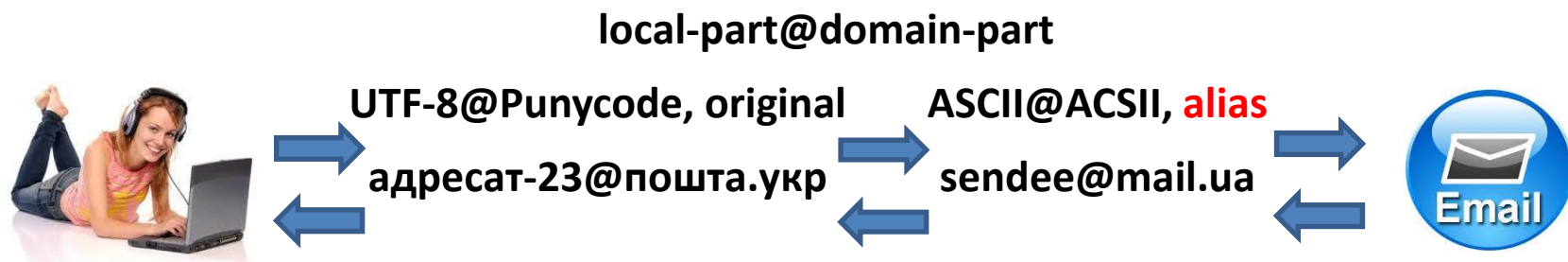
2015 UADom - 2015
4 декабря 2015, Киев



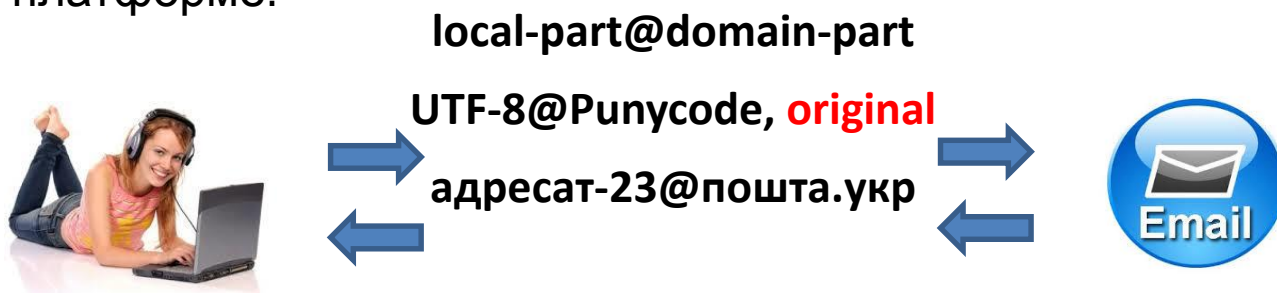
2 подхода к решению IDN email

.email .biz
.org .info
.net .com
.укр

Подход #1. Для EAI-адресов создаются псевдонимы в ASCII. IDN-адрес электронной почты в формате Unicode связан через WEB-платформу электронной почты с адресом электронной почты в формате ASCII.

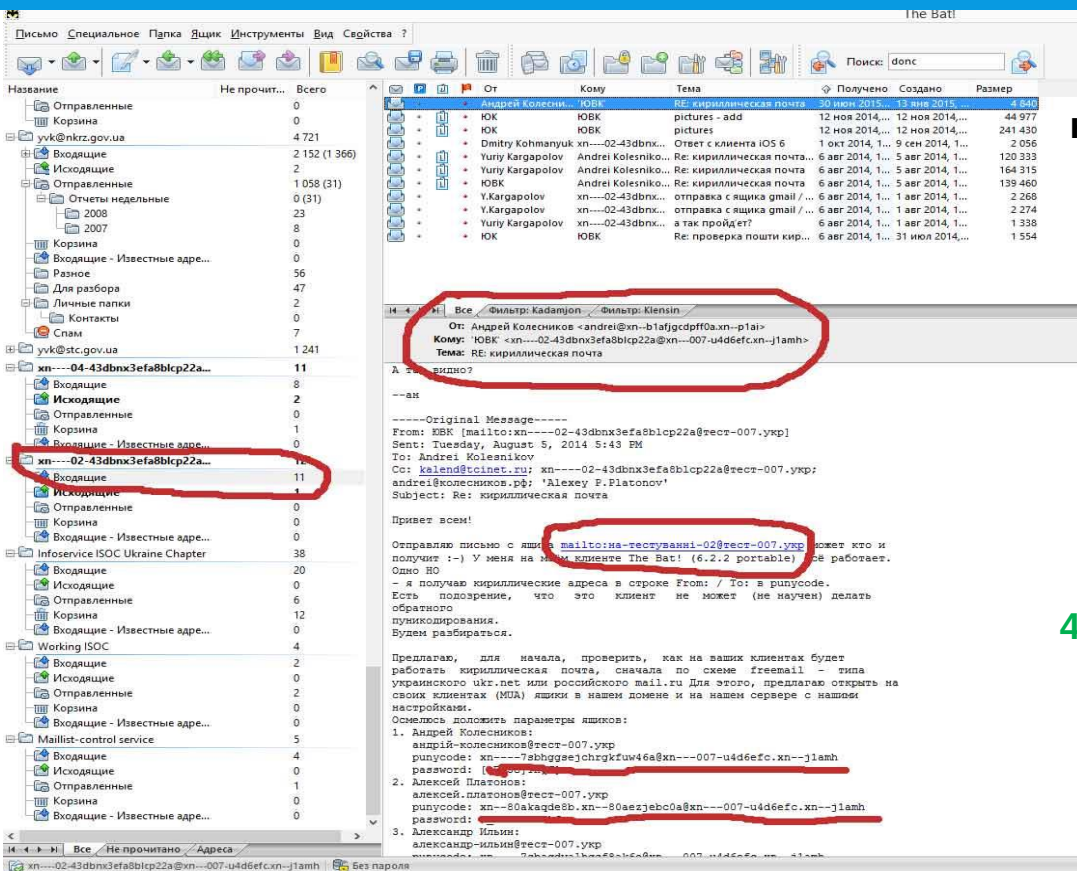


Подход #2. «Честное» обслуживание EAI, когда абонент использует Unicode для почтовых адресов вместе с ASCII без специальной специфической привязки к какой либо платформе.



Punycode

.email .biz
.org .info
.net .com
.укр



на-тестуванні-02@тест- andrei@колесников.
007.укр рф

PUNYCODE



xn---02-
43dbnx3efa8blcp22a@xn---
007-u4d6efc.xn--j1amh

andrei@xn--
b1afjgcdpff0a.xn--
p1ai

адресат-23@пошта.укр

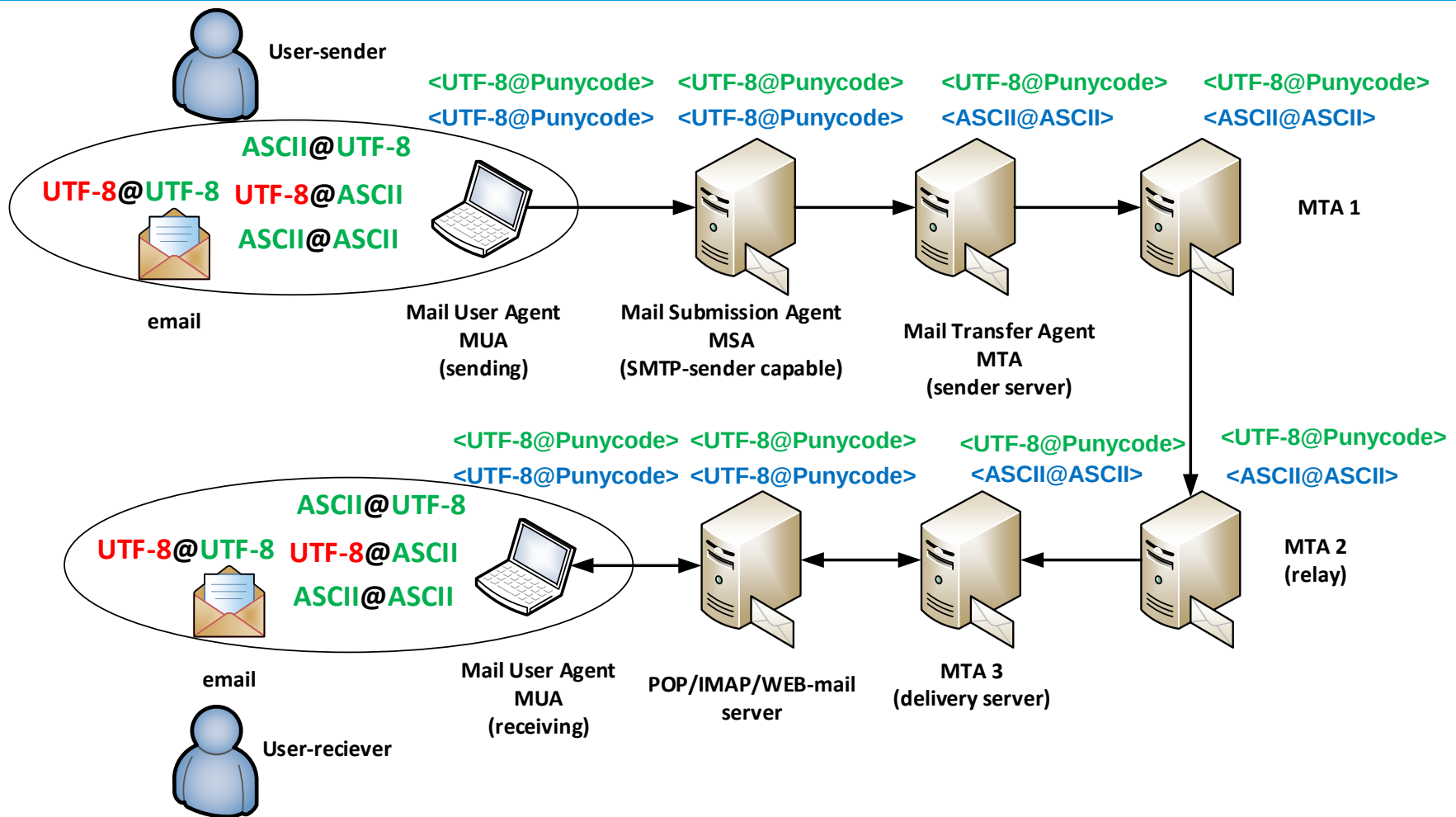
PUNYCODE

xn---23-5cdaxh1gkl@xn--80a1acn3a.xn--j1amh
correct

адресат~23@пошта.укр

Unassigned 5F@xn--80a1acn3a.xn--j1amh
incorrect

Подходы есть, проблемы остаются...



RFC 6530 "Overview and Framework for Internationalized Email" The base SMTP specification ([Section 2.3.11 of RFC 5321](#) [RFC5321]) states that "due to a long history of problems when intermediate hosts have attempted to optimize transport by modifying them, the local-part MUST be interpreted and assigned semantics only by the host specified in the domain part of the address". This is not a new requirement; equivalent statements appeared in specifications in 2001 [RFC2821] and even in 1989 [RFC1123].

Из чего состоит обмен сообщениями электронной почты в Интернет?

1. **Почтовый клиент** (почтовая программа на локальном компьютере пользователя)
2. **Почтовый сервер 1** (почтовый сервер с учетной записью Отправителя)
3. **Почтовый сервер 2** (почтовый сервер с учетной записью Получателя)
4. **Почтовый сервер транзитный**

Из чего состоит сообщение электронной почты?

1. **Заголовки (конверт) сообщения**
2. **Тело сообщения**

Из чего состоит адрес электронной почты?

1. **LOCAL-PART** – это имя учётной записи на почтовом сервере
2. **DOMAIN-PART**, из которой мы понимаем адрес почтового сервера

Сообщение электронной почты

.email .biz
.org .info
.net .com .укр

Заголовок (конверт) сообщения согласно RFC 822

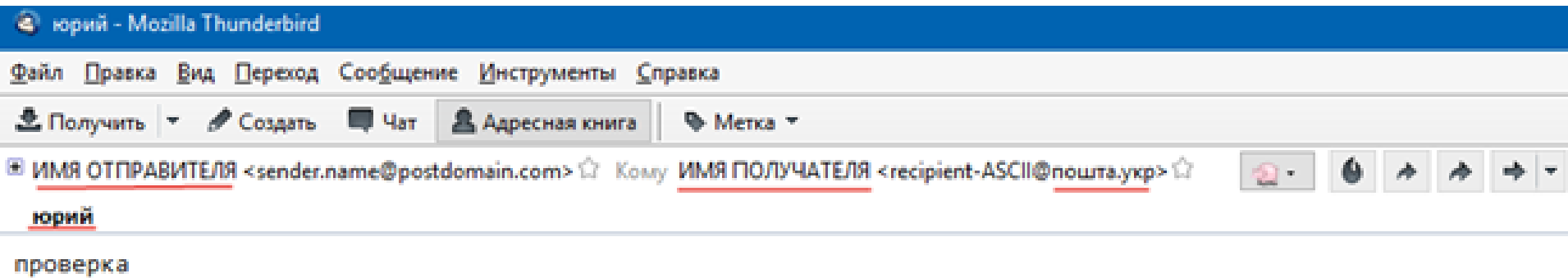
```
Return-Path: <sender.name@postdomain.com>
Received: from mx1.server.ua ([194.93.168.163] verified)
  by mx2.server.ua (Mailer Pro SMTP 5.1.10)
  with ESMTP id 5850452 for sender.name@postdomain.com@mx2.server.ua; Wed, 11 Nov 2015 14:39:40 +0200
Received: from [62.149.8.98] (HELO postoffice.net.ua)
  by mx1.server.ua (DigitalGate Pro SMTP 5.1.10)
  with ESMTP id 8963317 for sender.name@postdomain.com; Wed, 11 Nov 2015 14:50:00 +0200
Received: from [46.164.143.170] (account sender.name@postdomain.com HELO [192.168.15.129])
  by postoffice.net.ua (DigitalGate Pro SMTP 5.2c3)
  with ESMTPSA id 420813792 for sender.name@postdomain.com Wed, 11 Nov 2015 14:32:59 +0200
To: =?UTF-8?B?0JjQnNCvINCf0J7Qm9Cj0KfQkNCi0JXQm9Cv?= <recipient-ASCII@xn--80alacn3a.xn--1lamh>
From: =?UTF-8?B?0JjQnNCvINCe0KLQn9Cg0JDQktCY0KLQldCb0K8?= <sender.name@postdomain.com>
Subject: =?UTF-8?B?0Y7RgNC40Lk=?=
Organization: UANIC
Message-ID: <564336D0.2000905@postdomain.com>
Date: Wed, 11 Nov 2015 14:38:40 +0200
User-Agent: Mozilla/5.0 (Windows NT 10.0; WOW64; rv:38.0) Gecko/20100101
  Thunderbird/38.3.0
MIME-Version: 1.0
Content-Type: text/plain; charset=utf-8; format=flowed
Content-Transfer-Encoding: 8bit
```

проверка

Попробуем подготовить к отправке? ☺

Что скажет почтовый клиент?

.email .biz
.org .info
.net .com .укр



А если заменить LOCAL-PART с латиницы на кириллицу, т.е. recipient-ASCII → тест?

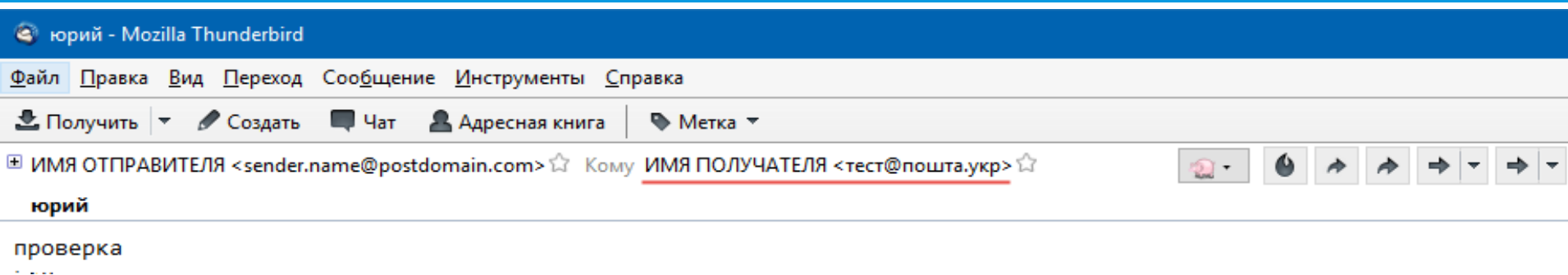
=?UTF-8?B?0JjQnNCvINCf0J7Qm9Cj0KfQkNCi0JXQm9Cv?= =?UTF-8?B?0YLQtdGB0YI=?@xn--80a1acn3a.xn--j1amh

```
Return-Path: <sender.name@postdomain.com>
Received: from mx1.server.ua ([194.93.168.163] verified)
  by mx2.server.ua (Mailer Pro SMTP 5.1.10)
  with ESMTP id 5850452 for sender.name@postdomain.com@mx2.server.ua; Wed, 11 Nov 2015 14:39:40 +0200
Received: from [62.149.8.98] (HELO postoffice.net.ua)
  by mx1.server.ua (DigitalGate Pro SMTP 5.1.10)
  with ESMTP id 8963317 for sender.name@postdomain.com; Wed, 11 Nov 2015 14:50:00 +0200
Received: from [46.164.143.170] (account sender.name@postdomain.com HELO [192.168.15.129])
  by postoffice.net.ua (DigitalGate Pro SMTP 5.2c3)
  with ESMTPSA id 420813792 for sender.name@postdomain.com Wed, 11 Nov 2015 14:32:59 +0200
To: =?UTF-8?B?0JjQnNCvINCf0J7Qm9Cj0KfQkNCi0JXQm9Cv?= <=?UTF-8?B?0YLQtdGB0YI=?@xn--80a1acn3a.xn--j1amh>
From: =?UTF-8?B?0JjQnNCvINCf0KlQn9Cg0JDQktCY0KLQldCb0K8=? <sender.name@postdomain.com>
Subject: =?UTF-8?B?0Y7RgNC40Lk=?
Organization: UANIC
```

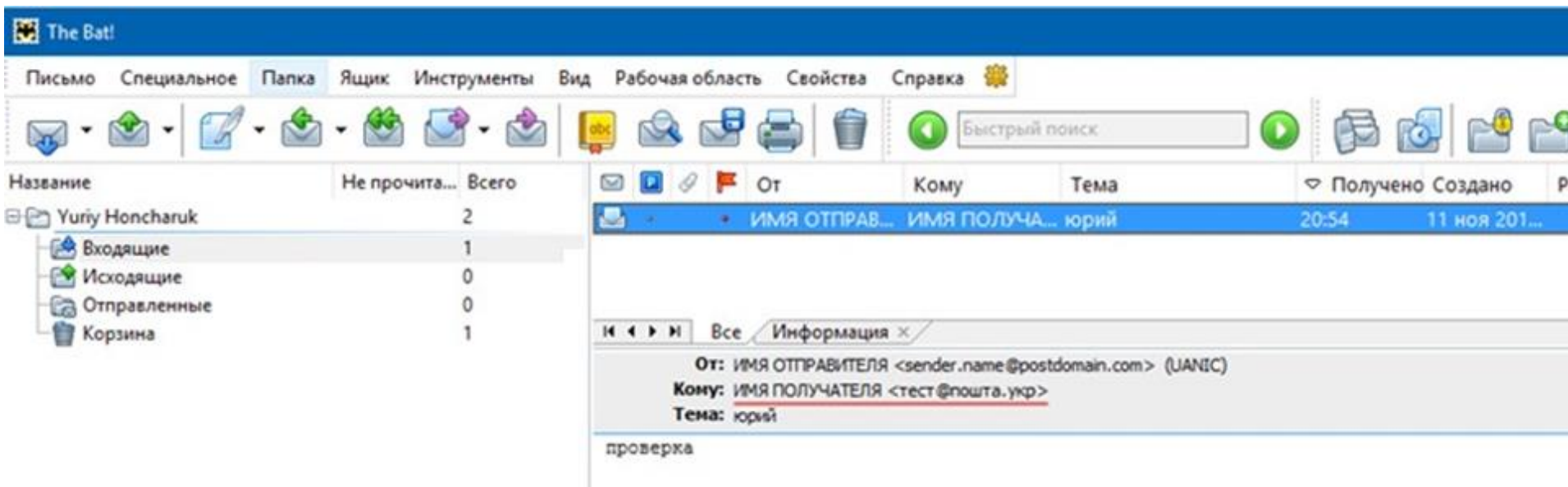
Что отобразится в почтовом клиенте?

Что скажет почтовый клиент?

.email .biz
.org .info
.net .com .укр



А в другой программе?



local-part отображается корректно

Что скажет почтовый клиент?

.email .biz
.org .info
.net .com .укр

Братец Outlook ты с нами?

Язык



- ☐ Использовать английский язык в подписях отметок сообщений
- ☐ Использовать английский язык в заголовках сообщений при ответах и пересылке, а также в уведомлениях о пересылке.

☒ Автовыбор кодировки исходящих сообщений

Кодировка исходящих сообщений: Юникод (UTF-8)

☒ Автовыбор кодировки исходящих vCard

Кодировка исходящих vCard: Юникод (UTF-8)

- ☒ Разрешить международные доменные имена в адресах электронной почты
- ☒ Включить поддержку кодировки UTF-8 для протокола mailto:

Предупреждение



В имени почтового ящика в адресе получателя тест@пошта.укр имеются не-ASCII символы. Это пока не поддерживается. Пожалуйста, замените этот адрес и попробуйте снова.

OK

Так что же сказал почтовый клиент?

Почтовые программы (MUA):

1. уже сегодня умеет обрабатывать (парсить) кириллицу в заголовках сообщения электронной почты,
2. без проблем **отображают корректно в надлежащих заголовках кириллицу как в LOCAL-PART так и в DOMAIN-PART**

Следовательно....

Почтовые серверы (MTA) могут нормально обрабатывать заголовки сообщения, содержащие кириллицу закодированную как в **UTF-8** в виде **=?UTF-8?B?0YlQtdGB0YI=?=**, так и закодированную в **Punycode** в виде **xn--80a1acn3a.xn--j1amh**

Есть ли ограничения?

Да. Существует ограничение – некоторые почтовые клиенты, руководствуясь существующим RFC для протокола SMTP не отображают то, что другие почтовые клиенты всё-таки отображают.

А если это ограничение убрать?!



Не всё так просто с Unicode

.email .biz
.org .info
.net .com .укр

IAB Statement on Identifiers and Unicode 7.0.0

<https://www.iab.org/documents/correspondence-reports-documents/2015-2/iab-statement-on-identifiers-and-unicode-7-0-0/>

What is peculiar about these cases, as distinct from other confusable cases, is that the decomposed and precomposed forms are in the same script and cannot be distinguished visually by users, even in large fonts designed for clarity. It is only by knowing the language that it is possible to detect whether a use of the character is the correct one. The reasoning for this may be peculiar to Arabic script; in the absence of an exhaustive analysis of all scripts — those completely encoded, partially encoded, and so far unencoded — we do not have the necessary background to make a determination. In any case, after considerable analysis, we believe that, whatever the reasoning, these sorts of cases are problematic for IETF-related uses of Unicode for identifiers.

The character that caused this issue to become plain is ARABIC LETTER BEH WITH HAMZA ABOVE (U+08A1). It looks like this:

بْ

U+08A1

In versions of Unicode since at least 3.2 and including 7.0.0, it is possible to write the letter *beh* and a combining character *hamza* to get a sequence, ARABIC LETTER BEH (U+0628) and ARABIC HAMZA ABOVE (U+0654):

ب + ء = بْ

U+0628

U+0654

It is not unusual to have multiple ways to generate "the same" character in the same script. For instance, in the Latin script, there is the character *a* with a diaeresis above it. It can be generated as a stand-alone character LATIN SMALL LETTER A WITH DIAERESIS (U+00E4). It looks like this:

ä

U+00E4

In versions of Unicode since at least 3.2 and including 7.0.0, it is possible to write *a*-diaeresis by writing the letter *a* and a combining diaeresis to get a sequence: LATIN SMALL LETTER A (U+0061), and COMBINING DIAERESIS (U+0308):

a + ¨ = ä

U+0061

U+0308

What is different between the "a" and "BEH" cases is the way these characters are treated. When one applies Normalization Form C (see UAX #15 at <http://www.unicode.org/reports/tr15/>) to the series U+0061 U+0308, it produces U+00E4; similarly, when one applies Normalization Form D to U+00E4, it produces U+0061 U+0308. These normalization rules reflect the Unicode observation that these are two different ways of writing the same character. The same is not true of U+08A1 and U+0628 U+0654, however. These pairs are not treated as being the same character. The reasons for this are discussed in section 9.2 of the Unicode Standard, version 7.0.0, pp362 ff (available at <http://www.unicode.org/versions/Unicode7.0.0/ch09.pdf>). The Standard says roughly that, because one approach is used to write a glottal stop whereas the other is written as a diacritic, the two characters are not the same thing linguistically. (The details are considerably more subtle than this. We refer the interested reader to the original in order to understand the details.)

Что дальше?

.email .biz
.org .info
.net .com .укр

Network Working Group
Internet-Draft
Updates: 5892, 5894 (if approved)
Intended status: Standards Track
Expires: September 11, 2015

J. Klensin
P. Faltstrom
Netnod
March 10, 2015

IDNA Update for Unicode 7.0.0
draft-klensin-idna-5892upd-unicode70-04.txt

“... This document, in its present form, is not a proposal for a solution.

Instead, it is intended to be (or evolve into) a comprehensive description of the issues and problems and to outline some possible approaches to a solution....

“...that one needs to rely on per-code point information in the Unicode Character Database rather than on rules or principles...”

Unicode-32?

“Каждый символ должен иметь свой код”

Спасибо! Вопросы?

Юрий Каргаполов
yvk@uanic.net

Юрий Гончарук
seo@uanic.net

