IRIs and IDNs: Problems of non-ASCII countries

Jakub Klímek

American
Standard
Code for
Information
Interchange

Since
1963

ASCII

IANA prefers the updated name **US-ASCII**, which clarifies that this system was developed in the **US** and based on the typographical symbols predominantly in use there

NUL SOH STX ETX EOT ENQ ■ BEL BS HT LF VT FF CR SS SI ■ DC1 DC2 DC3 DC4 NAK SYN EM ■ ■ ■ ■ FS GS RS US

!"#$%&'()*+,-./0123456789:;<=>?

@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]↑←

🟥 ACK 🟥 ESC DEL

Meanwhile, in other countries...

ありがとうございます

Где ты сейчас?

Είμαι καλά

איך בין פײַן

죄송합니다

කොහොමද සැප සනිප?

ငါကောင်းပါပြီ

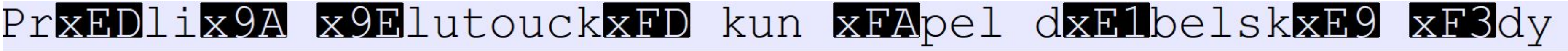
In Czechia, we are doing pretty good, but still...

Příliš žluťoučký kůň úpěl ďábelské ódy

What did we do when we wanted to use the computers?

Příliš žluťoučký kůň úpěl ďábelské ódy

PÅ™Ã-liÅ¡ Å¾luÅ¥ouÄ□kÃ½ kÅ¯Åˆ ÃºpÄ›l Ä□Ã¡belskÃ© Ã³dy



```
PrxEDlix9A x9ElutouckxFD kun xFApel dxE1belskxE9 xF3dy
```

Pý˘liç §lušouźkě k…Í ŁpŘl Ô belsk, ˘dy

ISO 8859-2

Windows CP1250

Kamenický encoding

...

...

...

What did we do when we wanted to use the computers?

Historically, all Czech geeks know that you **“just don’t use”** diacritics in computers.

File names, Folders, SMS messages, URIs, ...

Příliš žluťoučký kůň úpěl ďábelské ódy

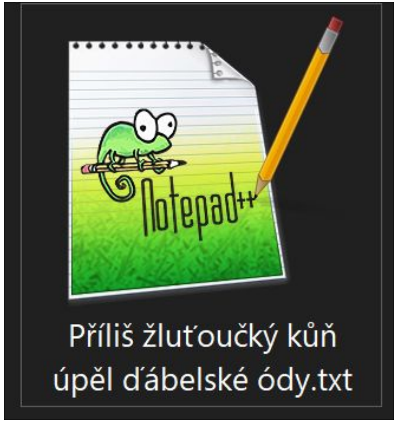
↓

Prilis zlutoucky kun upel dabelske ody

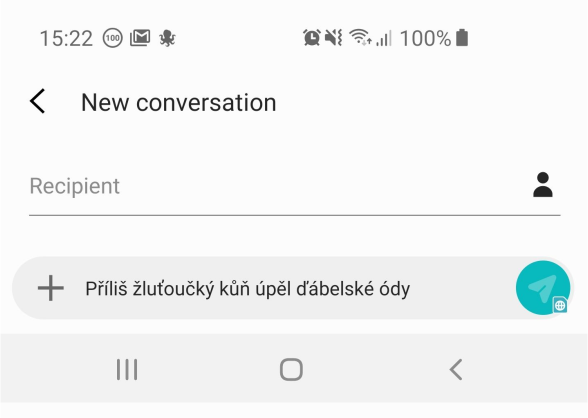
Unicode - UTF-8

Since 1993

Příliš žluťoučký kůň úpěl ďábelské ódy



Příliš žluťoučký kůň úpěl ďábelské ódy.txt



Příliš žluťoučký kůň úpěl ďábelské ódy

Unicode identifiers: We really need IDNs and IRIs

1. Linked Data: Identifying all things with unique identifiers - IRIs
    a. Pieces of legislation -> stripping of diacritics leads to conflicts, loss of meaning, hard to read https://slovník.gov.cz/legislativní/sbírka/111/2009/pojem/má-úplné-označení-ustanovení-včetně-označení-právního-předpisu
    b. Cities (https://cs.wikipedia.org/wiki/Šumperk)
    c. Companies (https://čokoláda.cz / https://cokolada.cz)
    d. ...
2. Publishing data about things identified by IRIs
    a. Web pages
    b. XML Schemas
    c. JSON Schemas
    d. CSV on the Web schemas
    e. API endpoints
    f. ...

IDN

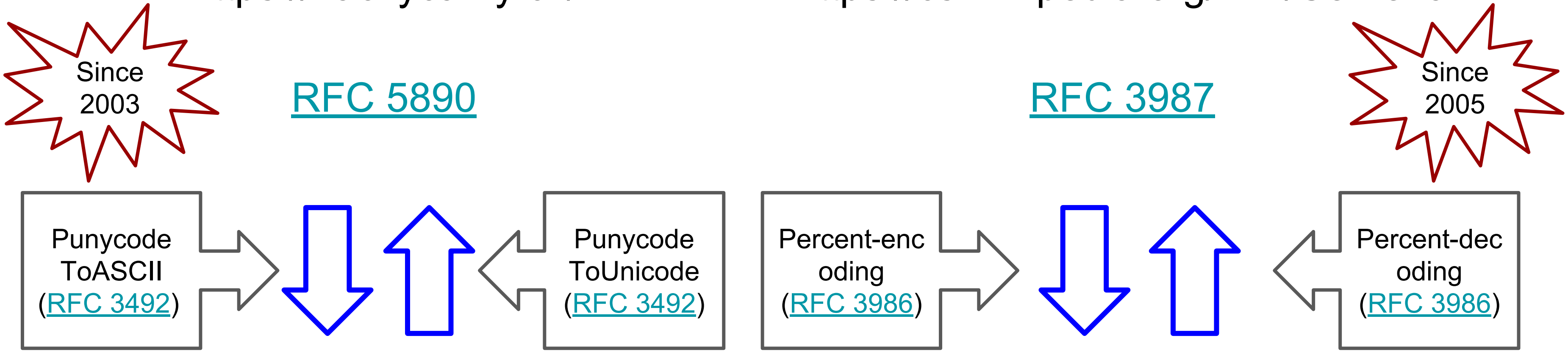
https://háčkyčárky.cz/

RFC 5890

IRI

https://cs.wikipedia.org/wiki/Červená

RFC 3987



Since 2003

Punycode ToASCII (RFC 3492)

Punycode ToUnicode (RFC 3492)

Percent-encoding (RFC 3986)

Percent-decoding (RFC 3986)

Since 2005

https://xn--hkyrky-ptac70bc.cz/

https://cs.wikipedia.org/wiki/%C4%8Cerven%C3%A1

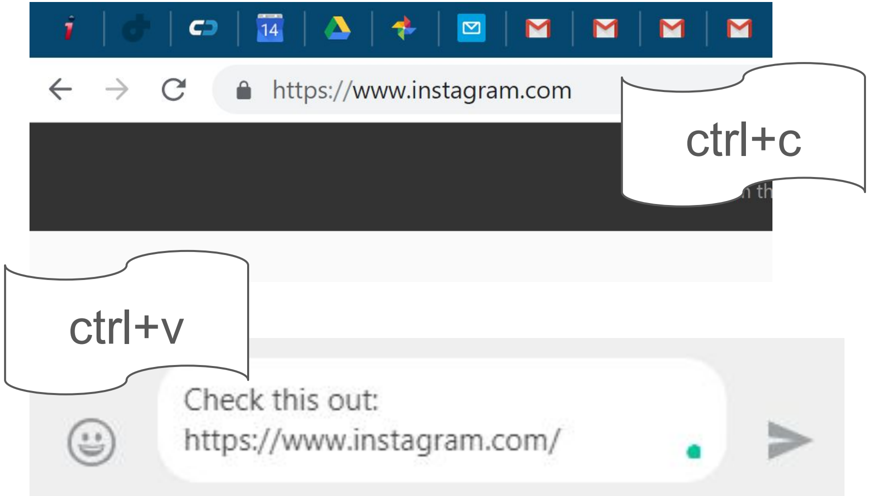
URI

URI

HTTP

IDNs and IRIs: So far so good, right?

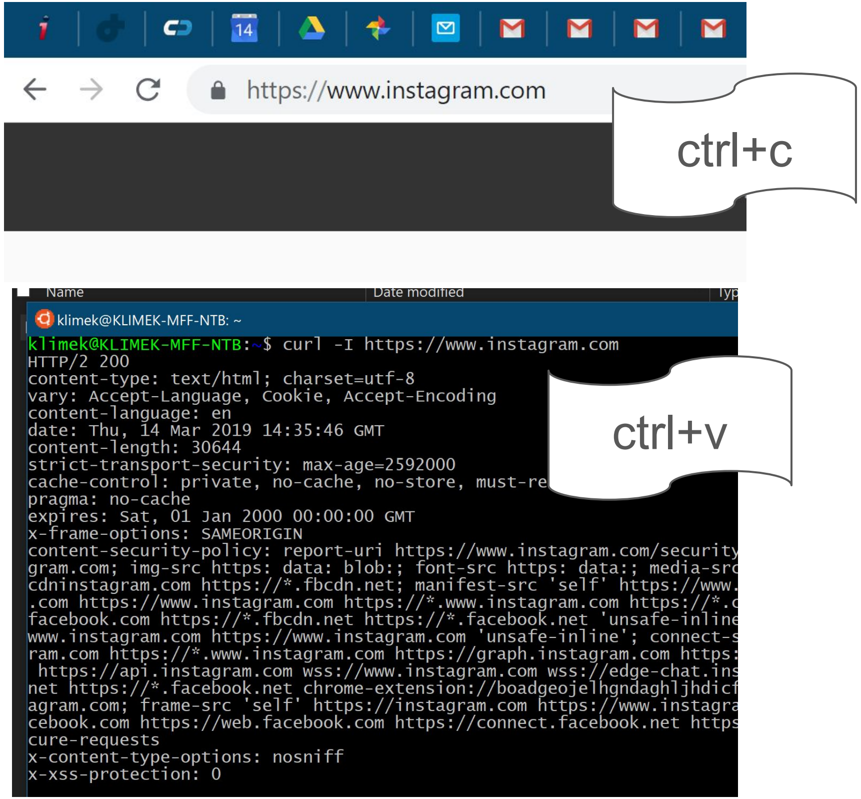
How do you use URIs today?



ctrl+c

ctrl+v

Check this out:
https://www.instagram.com/

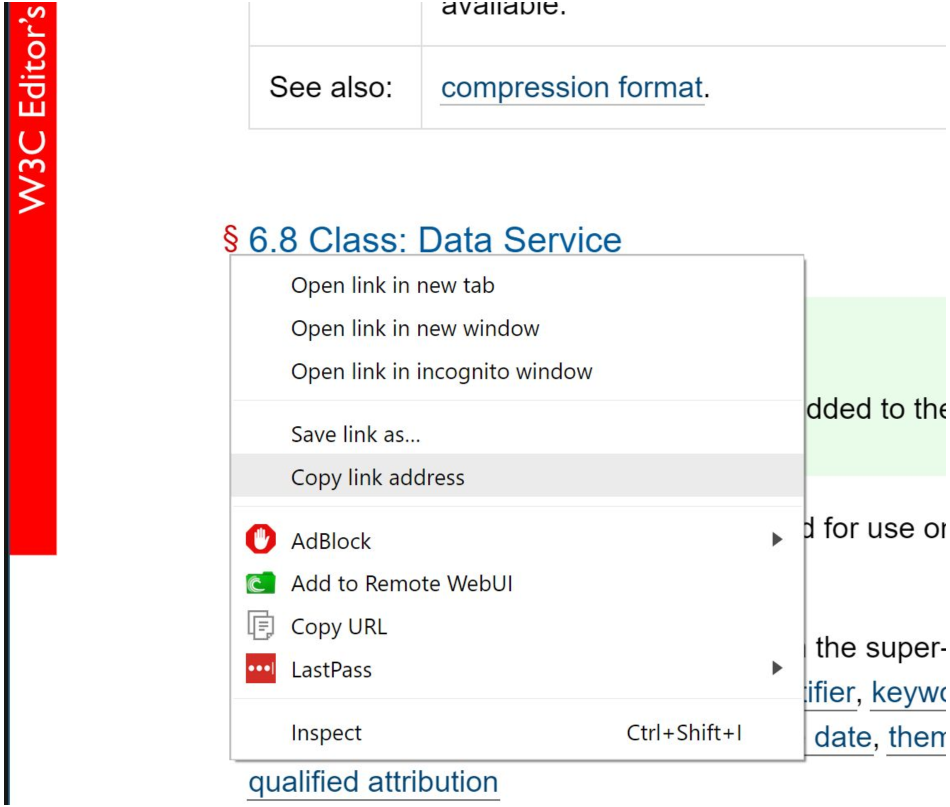


ctrl+c

ctrl+v

```
klimek@KLIMEK-MFF-NTB:~$ curl -I https://www.instagram.com
HTTP/2 200
content-type: text/html; charset=utf-8
vary: Accept-Language, Cookie, Accept-Encoding
content-language: en
date: Thu, 14 Mar 2019 14:35:46 GMT
content-length: 30644
strict-transport-security: max-age=2592000
cache-control: private, no-cache, no-store, must-re
pragma: no-cache
expires: Sat, 01 Jan 2000 00:00:00 GMT
x-frame-options: SAMEORIGIN
content-security-policy: report-uri https://www.instagram.com/security
gram.com; img-src https: data: blob:; font-src https: data:; media-src
cdninstagram.com https://*.fbcdn.net; manifest-src 'self' https://www.
.com https://www.instagram.com https://*.www.instagram.com https://*.c
facebook.com https://*.fbcdn.net https://*.facebook.net 'unsafe-inline
www.instagram.com https://www.instagram.com 'unsafe-inline'; connect-s
ram.com https://*.www.instagram.com https://graph.instagram.com https:
 https://api.instagram.com wss://www.instagram.com wss://edge-chat.ins
net https://*.facebook.net chrome-extension://boadgeojelhgndaghljhdicf
agram.com; frame-src 'self' https://instagram.com https://www.instagra
cebook.com https://web.facebook.com https://connect.facebook.net https
cure-requests
x-content-type-options: nosniff
x-xss-protection: 0
```

How do you use URIs today?



available.

See also: compression format.

§ 6.8 Class: Data Service

dded to the

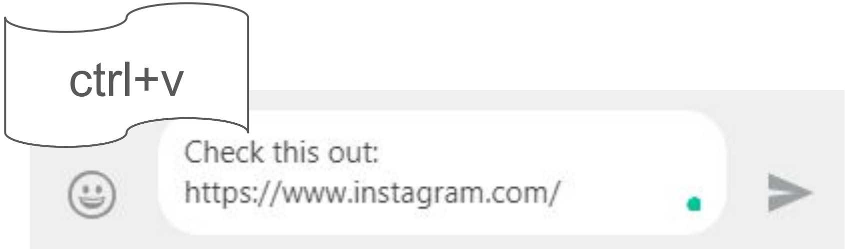
d for use or

the super-

ifier, keywo

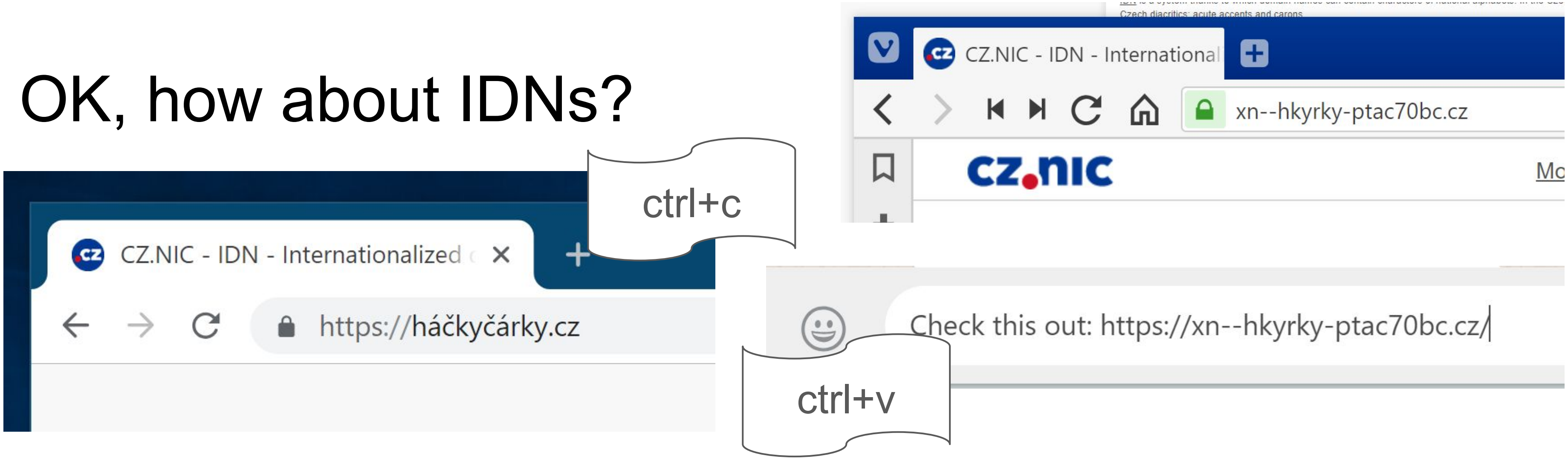
date, them

qualified attribution



Check this out:
https://www.instagram.com/

OK, how about IDNs?



https://háčkyčárky.cz

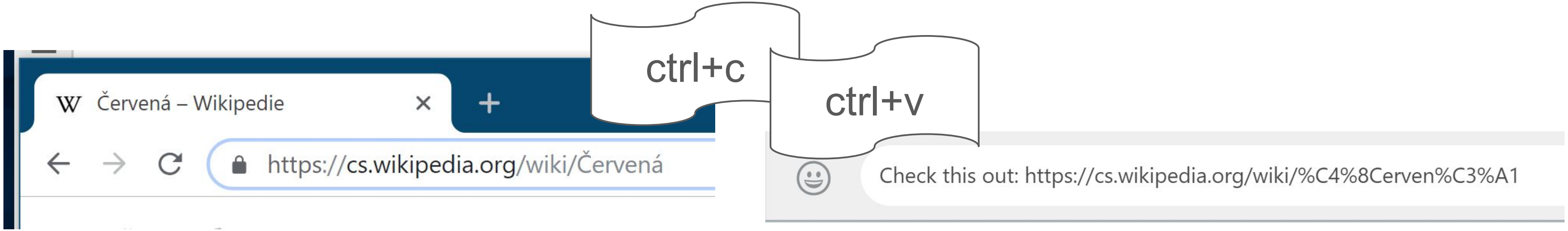
xn--hkyrky-ptac70bc.cz

Check this out: https://xn--hkyrky-ptac70bc.cz/

CZ.NIC - IDN - Internationalized
ctrl+c
Czech diacritics: acute accents and carons
CZ.NIC - IDN - International
cz.nic
ctrl+v

Google Chrome	https://xn--hkyrky-ptac70bc.cz/
MS IE 11	https://háčkyčárky.cz/
MS Edge	https://háčkyčárky.cz/
Firefox	https://háčkyčárky.cz/
Vivaldi	https://xn--hkyrky-ptac70bc.cz

OK, how about IRIs?



Červená – Wikipedie

https://cs.wikipedia.org/wiki/Červená

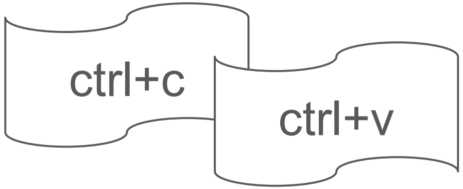
ctrl+c

ctrl+v

Check this out: https://cs.wikipedia.org/wiki/%C4%8Cerven%C3%A1

Google Chrome	https://cs.wikipedia.org/wiki/%C4%8Cerven%C3%A1
MS IE 11	https://cs.wikipedia.org/wiki/Červená
MS Edge	https://cs.wikipedia.org/wiki/Červená
Firefox	https://cs.wikipedia.org/wiki/%C4%8Cerven%C3%A1
Vivaldi	https://cs.wikipedia.org/wiki/Červená

IDNs and IRIs combined?



ctrl+c

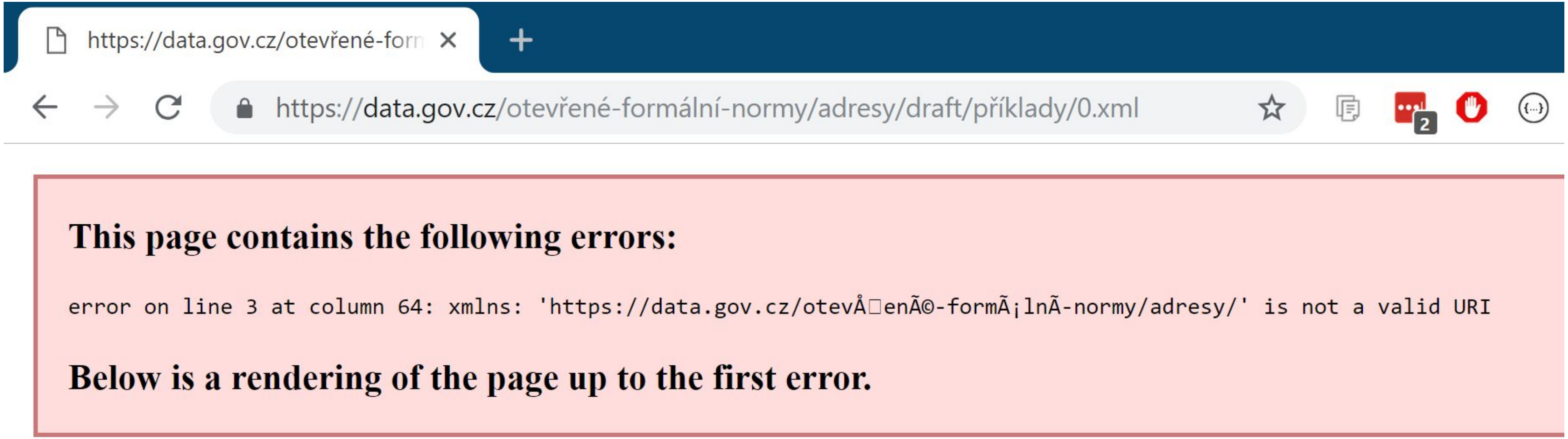
ctrl+v

https://háčkyčárky.cz/Červená

Browser	**IDN**	**IRI**
Google Chrome	xn--hkyrky-ptac70bc.cz	%C4%8Cerven%C3%A1
MS IE 11	háčkyčárky.cz	Červená
MS Edge	háčkyčárky.cz	Červená
Firefox	háčkyčárky.cz	%C4%8Cerven%C3%A1
Vivaldi	xn--hkyrky-ptac70bc	Červená

Support of IRIs elsewhere: XML 1.0

```
<?xml version="1.0" encoding="UTF-8" ?>
<adresa
     xmlns="https://data.gov.cz/otevřené-formální-normy/adresy/"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation="https://data.gov.cz/otevřené-formální-normy/adresy/
     https://data.gov.cz/otevřené-formální-normy/adresy/draft/schémata/adresa.xsd">
```

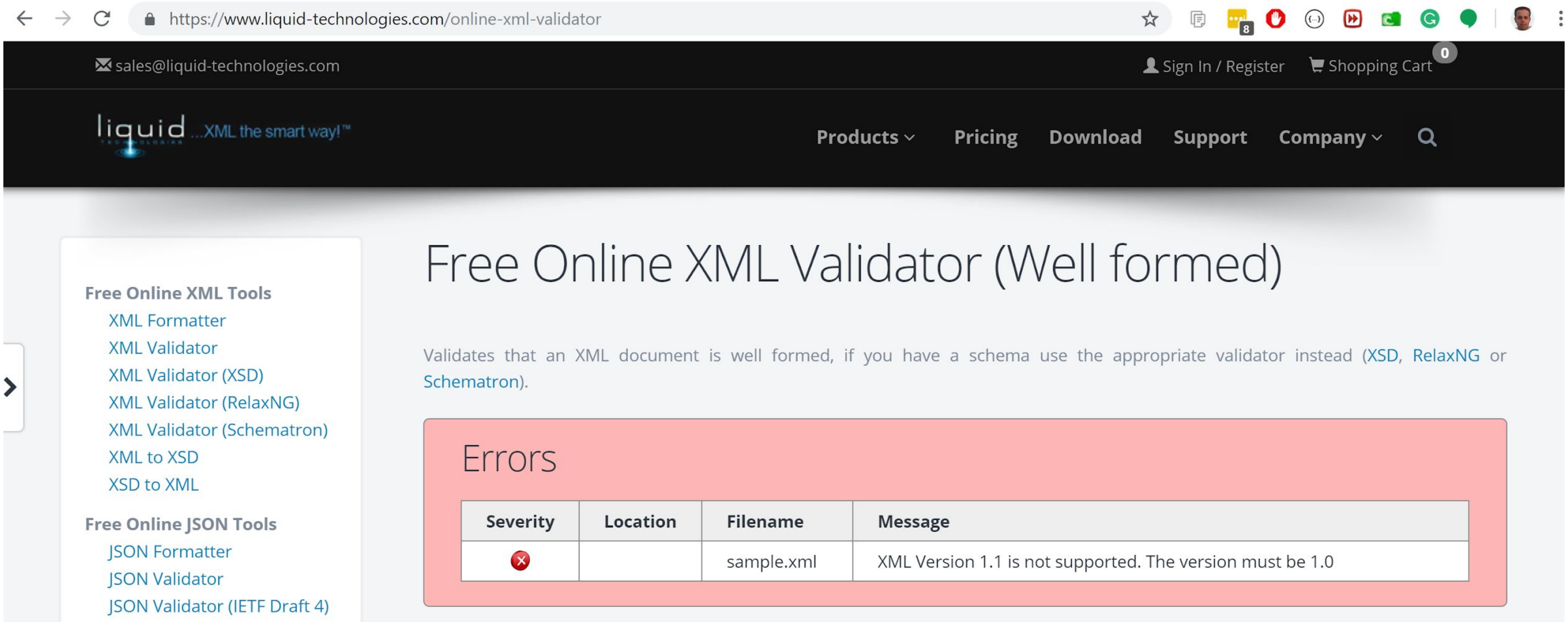

https://data.gov.cz/otevřené-forn
https://data.gov.cz/otevřené-formální-normy/adresy/draft/příklady/0.xml
2

This page contains the following errors:

```
error on line 3 at column 64: xmlns: 'https://data.gov.cz/otevÅenÃ©-formÃ¡lnÃ-normy/adresy/' is not a valid URI
```

Below is a rendering of the page up to the first error.

Support of IRIs elsewhere: XML 1.1



https://www.liquid-technologies.com/online-xml-validator

Free Online XML Tools
- XML Formatter
- XML Validator
- XML Validator (XSD)
- XML Validator (RelaxNG)
- XML Validator (Schematron)
- XML to XSD
- XSD to XML

Free Online JSON Tools
- JSON Formatter
- JSON Validator
- JSON Validator (IETF Draft 4)

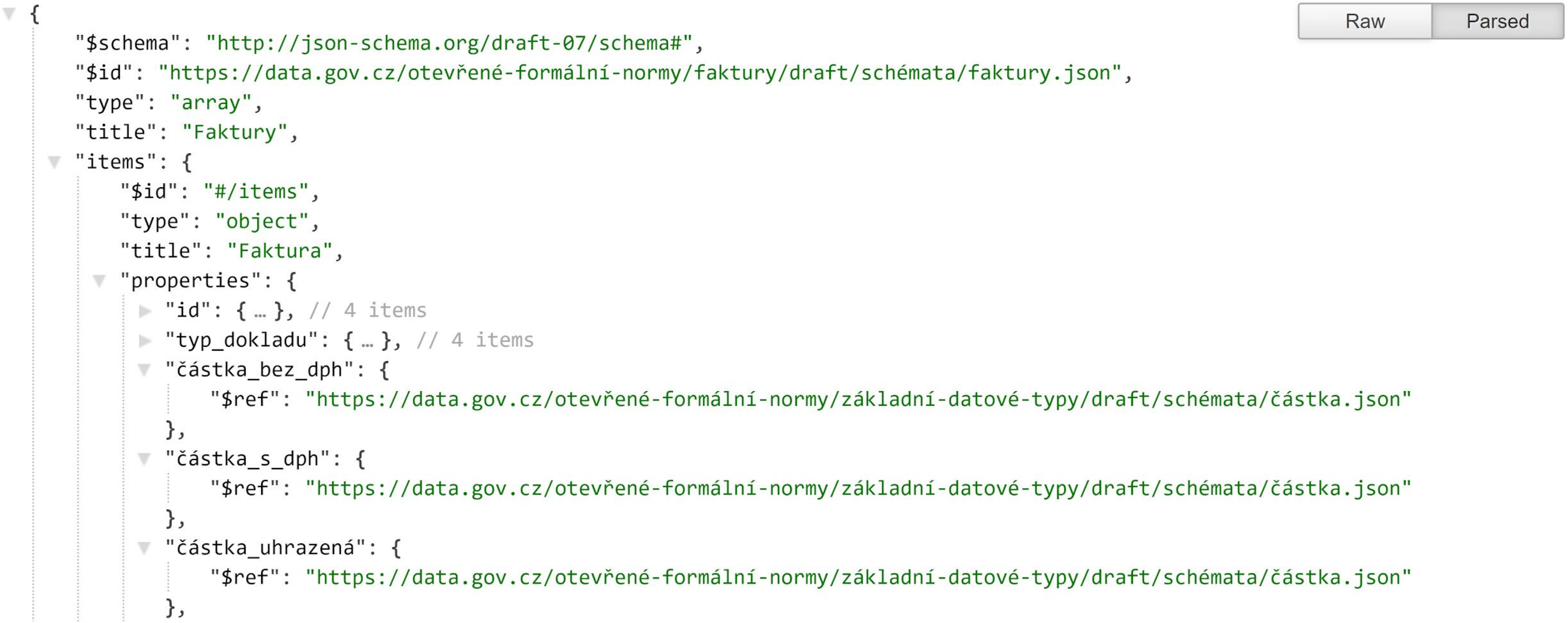
Free Online XML Validator (Well formed)

Validates that an XML document is well formed, if you have a schema use the appropriate validator instead (XSD, RelaxNG or Schematron).

Errors

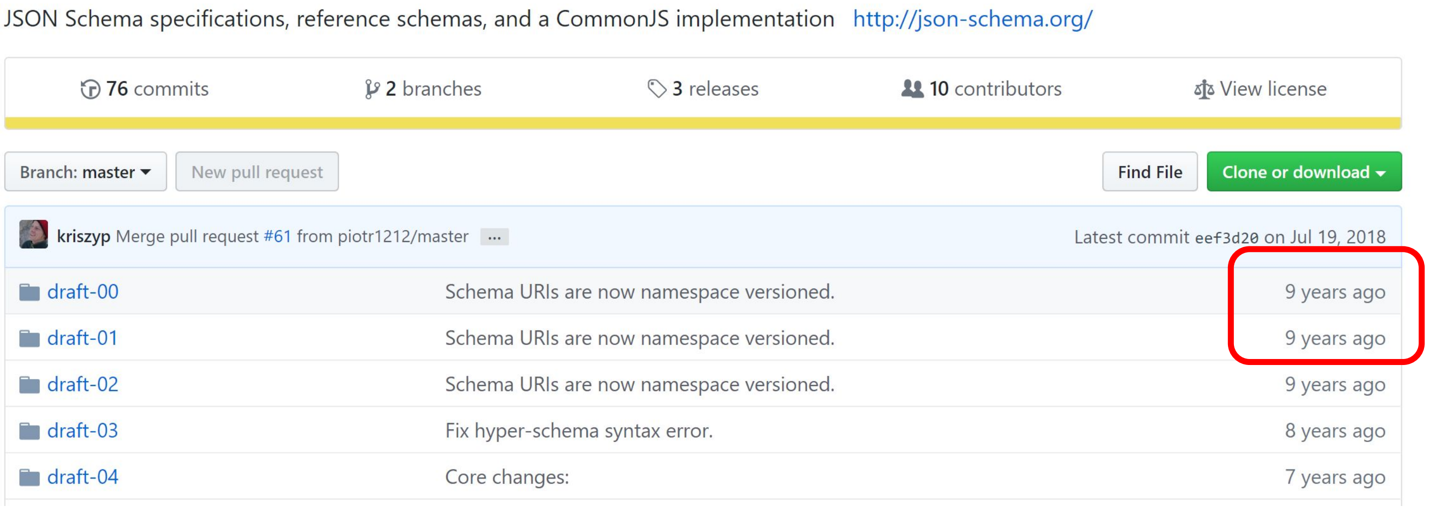
Severity	Location	Filename	Message
☒		sample.xml	XML Version 1.1 is not supported. The version must be 1.0

Support of IRIs elsewhere: JSON Schema



```
{
    "$schema": "http://json-schema.org/draft-07/schema#",
    "$id": "https://data.gov.cz/otevřené-formální-normy/faktury/draft/schémata/faktury.json",
    "type": "array",
    "title": "Faktury",
    "items": {
        "$id": "#/items",
        "type": "object",
        "title": "Faktura",
        "properties": {
            "id": { … }, // 4 items
            "typ_dokladu": { … }, // 4 items
            "částka_bez_dph": {
                "$ref": "https://data.gov.cz/otevřené-formální-normy/základní-datové-typy/draft/schémata/částka.json"
            },
            "částka_s_dph": {
                "$ref": "https://data.gov.cz/otevřené-formální-normy/základní-datové-typy/draft/schémata/částka.json"
            },
            "částka_uhrazená": {
                "$ref": "https://data.gov.cz/otevřené-formální-normy/základní-datové-typy/draft/schémata/částka.json"
            },
```

Support of IRIs elsewhere: JSON Schema

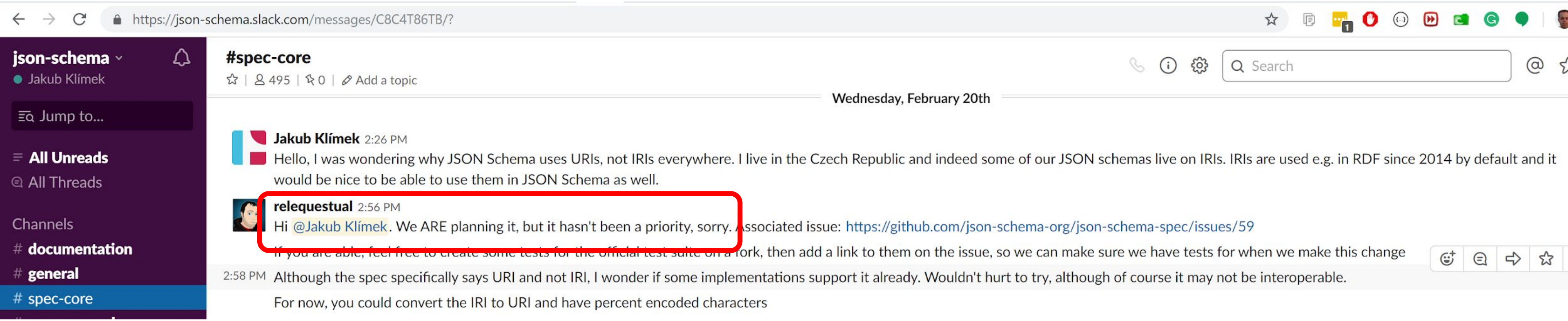


JSON Schema specifications, reference schemas, and a CommonJS implementation http://json-schema.org/

kriszyp Merge pull request #61 from piotr1212/master …

Latest commit eef3d20 on Jul 19, 2018

draft-00	Schema URIs are now namespace versioned.	9 years ago
draft-01	Schema URIs are now namespace versioned.	9 years ago
draft-02	Schema URIs are now namespace versioned.	9 years ago
draft-03	Fix hyper-schema syntax error.	8 years ago
draft-04	Core changes:	7 years ago



Wednesday, February 20th

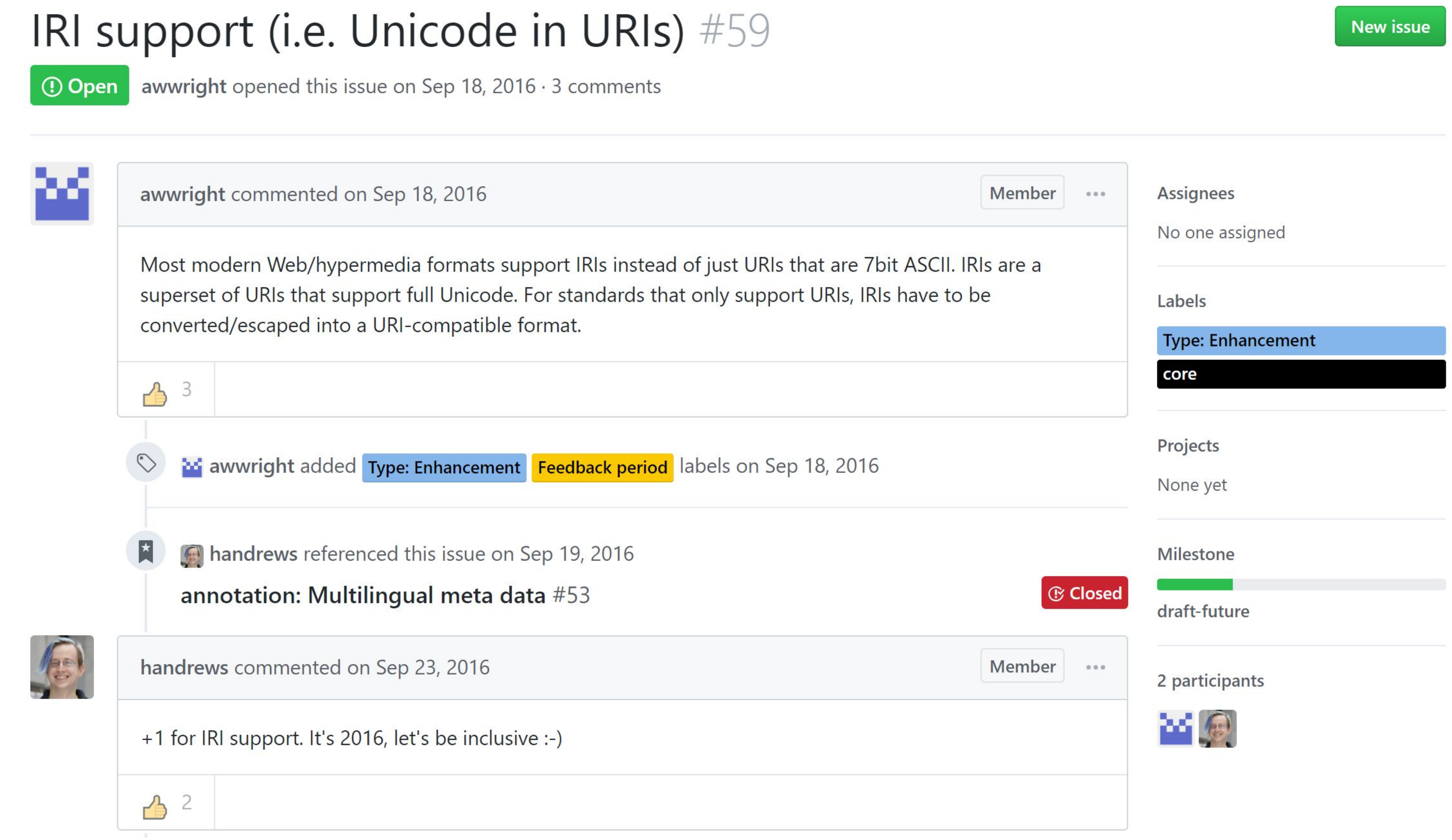
Jakub Klímek 2:26 PM
Hello, I was wondering why JSON Schema uses URIs, not IRIs everywhere. I live in the Czech Republic and indeed some of our JSON schemas live on IRIs. IRIs are used e.g. in RDF since 2014 by default and it would be nice to be able to use them in JSON Schema as well.

relequestual 2:56 PM
Hi @Jakub Klímek. We ARE planning it, but it hasn't been a priority, sorry. Associated issue: https://github.com/json-schema-org/json-schema-spec/issues/59

If you are able, feel free to create some tests for the official test suite on a fork, then add a link to them on the issue, so we can make sure we have tests for when we make this change

2:58 PM Although the spec specifically says URI and not IRI, I wonder if some implementations support it already. Wouldn't hurt to try, although of course it may not be interoperable.

For now, you could convert the IRI to URI and have percent encoded characters



IRI support (i.e. Unicode in URIs) #59

awwright opened this issue on Sep 18, 2016 · 3 comments

awwright commented on Sep 18, 2016

Most modern Web/hypermedia formats support IRIs instead of just URIs that are 7bit ASCII. IRIs are a superset of URIs that support full Unicode. For standards that only support URIs, IRIs have to be converted/escaped into a URI-compatible format.

👍 3

awwright added Type: Enhancement Feedback period labels on Sep 18, 2016

handrews referenced this issue on Sep 19, 2016

annotation: Multilingual meta data #53

handrews commented on Sep 23, 2016

+1 for IRI support. It's 2016, let's be inclusive :-)

👍 2

Assignees

No one assigned

Labels

Type: Enhancement

core

Projects

None yet

Milestone

draft-future

2 participants

And what about curl and IDN? (Ubuntu Linux)

```
klimek@KLIMEK-MFF-NTB:~$ curl -v -I https://háčkyčárky.cz/
*   Trying 217.31.205.51...
* Connected to háčkyčárky.cz (217.31.205.51) port 443 (#0)
*  subject: CN=xn--hkyrky-ptac70bc.cz
*  subjectAltName: host "háčkyčárky.cz" matched cert's "xn--hkyrky-ptac70bc.cz"

> HEAD / HTTP/2
> Host: xn--hkyrky-ptac70bc.cz
> User-Agent: curl/7.64.0
> Accept: */*
>

< HTTP/2 200
```

And what about curl and IDN? (Windows)

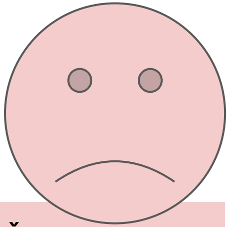
```
C:\Tools\curl-7.64.1-win64-mingw\bin>curl -v -I https://háčkyčárky.cz/

* Failed to convert háckycárky.cz to ACE;

* Closing connection -1

curl: (3) Failed to convert há
```

And what about curl and IRI? (Ubuntu Linux)

```
klimek@KLIMEK-MFF-NTB:~$ curl -v -I --http1.1 https://cs.wikipedia.org/wiki/Červená
> HEAD /wiki/Červená HTTP/1.1
> Host: cs.wikipedia.org
> User-Agent: curl/7.64.0
> Accept: */*
>
< HTTP/1.1 200 OK
```

```
klimek@KLIMEK-MFF-NTB:~$ curl -v -I --http1.1
https://opendata-mvcr.github.io/podmínky-užití/není-chráněna-zvláštním-právem-pořizovatele-databáze/
> HEAD /podmínky-užití/není-chráněna-zvláštním-právem-pořizovatele-databáze/ HTTP/1.1
> Host: opendata-mvcr.github.io
> User-Agent: curl/7.64.0
> Accept: */*
>
< HTTP/1.1 400 Bad Request
```

And what about curl and IRI? (Windows)



```
C:\Tools\curl-7.64.1-win64-mingw\bin>curl -v -I --http1.1 https://cs.wikipedia.org/wiki/Červená
> HEAD /wiki/Cervená HTTP/1.1
> Host: cs.wikipedia.org
> User-Agent: curl/7.64.1
> Accept: */*
>
< HTTP/1.1 404 Not Found
```

```
C:\Tools\curl-7.64.1-win64-mingw\bin>curl -v -I --http1.1
https://opendata-mvcr.github.io/podmínky-užití/není-chráněna-zvláštním-právem-pořizovatele-databáze/
> HEAD /podmínky-uzití/není-chránena-zvlástním-právem-porizovatele-databáze/ HTTP/1.1
> Host: opendata-mvcr.github.io
> User-Agent: curl/7.64.1
> Accept: */*
>
< HTTP/1.1 400 Bad Request
```

Every time something goes wrong with IRIs:

I told you so!

Everyone knows that you **“just don’t use”** diacritics in computers.

This should not be the message people get from using IRIs

These problems are not new

MY URL ISN'T YOUR URL

- DANIEL STENBERG

https://daniel.haxx.se/blog/2016/05/11/my-url-isnt-your-url/

IMHO: IRIs, when used outside of HTTP (in clipboard, when shared, …), should be HUMAN readable. Agents are responsible for implementing the necessary conversions (IDNs & IRIs, not http:////////////////////////////////////example.com)