Egységes erőforrás-azonosító (URI)

Jeszenszky Péter
Debreceni Egyetem, Informatikai Kar
jeszenszky.peter@inf.unideb.hu

Utolsó módosítás: 2024. szeptember 13.

URI (1)

- **Egységes erőforrás-azonosító** (***uniform resource identifier***) (**URI**):
  - Absztrakt vagy fizikai erőforrást azonosító tömör karaktersorozat.
    - Egy erőforrás nem feltétlenül érhető el a Weben.
    - URI-kat hozzá lehet rendelni akár a tárgyi világ objektumaihoz és fogalmakhoz is.
- A jelenleg aktuális szabvány:
  - Tim Berners-Lee, Roy Fielding, Larry Masinter. *RFC 3986: Uniform Resource Identifier (URI): Generic Syntax*. January 2005. https://www.rfc-editor.org/rfc/rfc3986

URI (2)

- Minden URI egy sémanévvel kezdődik, melyet egy `':'` karakter választ el a séma-specifikus résztől.
  - A séma-specifikus rész szintaxisát és jelentését sémaspecifikációk határozhatják meg bizonyos korlátok között.
- Az URI-sémákat (is) adminisztráló szerv:
  - *Internet Assigned Numbers Authority* (IANA) https://www.iana.org/
    - Lásd: *Uniform Resource Identifier (URI) Schemes* https://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml

Közismert URI sémák

- `file:`
  - Matthew Kerwin. *RFC 8089: The "file" URI Scheme*. February 2017. https://www.rfc-editor.org/rfc/rfc8089
- `http`/`https`:
  - Roy T. Fielding (ed.), Mark Nottingham (ed.), Julian F. Reschke (ed.). *RFC 9110: HTTP Semantics*. June 2022. https://www.rfc-editor.org/rfc/rfc9110
- `mailto`:
  - Martin Dürst, Larry Masinter, Jamie Zawinski. *RFC 6068: The 'mailto' URI Scheme*. October 2010. https://www.rfc-editor.org/rfc/rfc6068
- `about`:
  - S. Moonesamy (ed.). *RFC 6694: The „about" URI Scheme*. August 2012. https://www.rfc-editor.org/rfc/rfc6694

Hivatkozás-feloldás (*dereferencing*)

- Az URI által azonosított erőforráshoz való hozzáférést jelenti.
  - Ez leggyakrabban az erőforrás egy reprezentációjának letöltését jelenti.

URL vs URN (1)

- Történetileg kétfajta URI megkülönböztetése:
  - ***Uniform Resource Locator*** (**URL**): egységes erőforrás-helymeghatározó
    - Az erőforrások azonosítása az elérés módjával.
    - Tim Berners-Lee, Larry Masinter, Mark P. McCahill. *RFC 1738: Uniform Resource Locators (URL)*. December 1994. https://www.rfc-editor.org/rfc/rfc1738
  - ***Uniform Resource Name*** (**URN**): egységes erőforrás-név
    - Erőforrások helytől független tartós azonosítása.
    - Ryan Moats. *RFC 2141: URN Syntax*. May 1997. https://www.rfc-editor.org/rfc/rfc2141

URL vs URN (2)

- Ez a korábbi diszjunkt felosztás mára érvényét vesztette.
  - Michael Mealling (ed.), Ray Denenberg (ed.). *RFC 3305: Report from the Joint W3C/IETF URI Planning Interest Group: Uniform Resource Identifiers (URIs), URLs, and Uniform Resource Names (URNs): Clarifications and Recommendations*. August 2002. https://www.rfc-editor.org/rfc/rfc3305
  - *URIs, URLs, and URNs: Clarifications and Recommendations 1.0—Report from the joint W3C/IETF URI Planning Interest Group* (W3C feljegyzés, 2001. szeptember 21.) https://www.w3.org/TR/uri-clarification/

URL vs URN (3)

- A kortárs felfogás szerint egy URI lehet helymeghatározó, név, vagy egyszerre mindkettő.
  - Egy URI séma nem kell, hogy besorolható legyen a helymeghatározó vagy a név kategóriába.
- Az URL, mint nem hivatalos fogalom, olyan URI-k esetén használható, melyek az erőforrást az elérés módjával azonosítják.

URN

- Egy egységes erőforrás név (URN) egy olyan URI, melynek célja erőforrások helytől független tartós azonosítása.
- Az egységes erőforrás nevek az `urn` URI-séma alá tartozó URI-k.
- Lásd:
  - Peter Saint-Andre, John C. Klensin. *RFC 8141: Uniform Resource Names (URNs)*. April. 2017. https://www.rfc-editor.org/rfc/rfc8141

WHATWG szabvány (1)

- *URL Living Standard*
  https://url.spec.whatwg.org/
- Célok:
  - Az RFC 3986 és RFC 3987 specifikációk a kortárs implementációkhoz való igazítása és elavulttá tétele.
  - Az URL kifejezés szabványosítása.
  - A jelenlegi JavaScript URL API továbbfejlesztése.

WHATWG szabvány (2)

- URI-k és IRI-k egységes kezelése.
- Az URL-ek univerzális azonosítók.

URI vs URL (IETF vs WHATWG)

- Lásd:
  - Daniel Stenberg. *My URL isn't your URL*. May 11, 2016. https://daniel.haxx.se/blog/2016/05/11/my-url-isnt-your-url/
  - Daniel Stenberg. *One URL standard please*. January 30, 2017. https://daniel.haxx.se/blog/2017/01/30/one-url-standard-please/
  - Daniel Stenberg. *URL Interop*. https://github.com/bagder/docs/blob/master/URL-interop.md

URI példák

- https://www.rfc-editor.org/rfc/rfc3986.txt
- https://url.spec.whatwg.org/#references
- file:///usr/lib/R/library
- about:downloads
- mailto:jeszenszky.peter@inf.unideb.hu
- ldap://ldap.iplanet.com/dc=example,dc=com
- tel:+36-52-512-900
- news:comp.lang.c
- urn:isbn:0-395-36341-1
- urn:ietf:std:66
- urn:uuid:f81d4fae-7dec-11d0-a765-00a0c91e6bf6
- geo:47.5539464,21.6215658

URI karakterek (1)

- URI-kban megengedett karakterek:
  - Fenntartott karakterek az alábbiak:
    - `':', '/', '?', '#', '[', ']', '@', '!', '$', '&', ''', '(', ')', '*', '+', ',', ';', '='`
      - Határolójelként használt karakterek
  - Nem fenntartott karakterek:
    - `'A', …, 'Z', 'a', …, 'z'`
    - `'0', …, '9'`
    - `'-', '.', '_', '~'`
- A specifikáció nem határoz meg karakterkódolást.

URI karakterek (2)

- **Százalékos kódolás** (***percent-encoding***): nem megengedett karakterek használatához vagy fenntartott karakterek speciális jelentésének elnyomásához.
  - Tekintsük a karaktert az adott karakterkódolásban ábrázoló oktettsorozatot.
  - Az oktettsorozatot kódoljuk egy olyan karakterlánccal, melyben minden oktettet *%HH* módon ábrázolunk, ahol *HH* az oktett értékét reprezentáló két hexadecimális számjegy karakter.
    - Például a szóköz karaktert `%20` módon kell kódolni.
    - Használhatóak az `'A'`, …, `'F'` és az `'a'`, …, `'f'` hexadecimális számjegy karakterek is.
    - URI-k összehasonlításánál nem számít, hogy hexadecimális számjegyként kis- vagy nagybetűk szerepelnek-e.

URI karakterek (3)

- Példa a százalékos kódolás használatára:
  - `file:///media/Movies/What's Up, Tiger Lily? (1966)/` → `file:///media/Movies/What%27s%20Up%2C%20Tiger%20Lily%3F%20%281966%29/`
  - UTF-8 karakterkódolást feltételezve: `http://www.w3.org/People/Dürst/` → `http://www.w3.org/People/D%C3%BCrst/`

URI szintaxis (1)

- Hierarchikus felépítés.
  - A komponensek felsorolása balról jobbra haladva fontosság szerint csökkenő sorrendben történik.
- Általános szintaxis:

  séma `':'` *hierarchikus-rész* [`'?'` *lekérdezés*] [`'#'` *erőforrásrész*]

  - A hierarchikus rész egy *autoritás* (*authority*) és egy útvonal (*path*) komponenst tartalmazhat, szintaxisa:

    `'//'` *autoritás útvonal* vagy *útvonal*

    - Ha van *autoritás* komponens, akkor az útvonal üres kell, hogy legyen vagy a `'/'` karakterrel kell, hogy kezdődjön.
    - Ha nincs *autoritás* komponens, akkor az útvonal nem kezdődhet két `'/'` karakterrel.

URI szintaxis (2)

- Példa:

```
  https://wordery.com/search?term=scotland#results
  \___/   \_____________/\_____/ \_____________/ \_____/
    |            |         |           |           |
   séma         host     útvonal   lekérdezés  erőforrás-
(scheme)                (path)      (query)      rész
                                                (fragment)
```

- Példa:

```
  mailto:jeszenszky.peter@inf.unideb.hu?subject=URI
  \____/ \_____________________________________/ \_____/
    |                      |                       |
   séma                útvonal                 lekérdezés
(scheme)                (path)                   (query)
```

URI szintaxis (3)

- Érdekesség:
  - 2009 októberében egy *Times* cikkben Tim Berners-Lee elnézést kért a URI-k két perjeléért:
    - *„There you go, it seemed like a good idea at the time…”*
  - Idézve: *Berners-Lee 'sorry' for slashes*. 14 October 2009. http://news.bbc.co.uk/2/hi/technology/8306631.stm

Az autoritás komponens

- Nevét onnan kapta, hogy fennhatósága alá tartozik az URI további része által meghatározott névtér.
- Szintaxisa:

  [*userinfo* `'@'`] *host* [`':'` *port*]

  - Az URI sémák meghatározhatnak egy alapértelmezett portot.
    - Például a `http` sémánál 80 az alapértelmezés.

Az útvonal komponens

- Útvonal részek `'/'` karakterekkel elválasztott sorozata, amely lehet üres.
- Az első `'?'` vagy `'#'` karakterig, ezek hiányában pedig az URI végéig tart.
- Az állományrendszerekben megszokott módon használhatóak útvonal részként `'.'` és `'..'`.

A lekérdezés komponens

- A `'?'` karakter jelzi az elejét, a `'#'` karakterig, annak hiányában pedig az URI végéig tart.
- Nem hierarchikus adatokat tartalmaz.
- Gyakran *név* `'='` *érték* formájú, `'&'` karakterekkel elválasztott név-érték párokat tartalmaz.
  - A `http` és `https` URI sémák esetében ez űrlap adatok továbbítására szolgál (`application/x-www-form-urlencoded` kódolás).
  - Példa:
    - https://blackwells.co.uk/bookshop/search?keyword=sherlock+holmes&sortValue=DateDesc
  - Lásd: *HTML Standard – URL-encoded form data* https://html.spec.whatwg.org/multipage/forms.html#url-encoded-form-data

Erőforrásrész-azonosító (1)

- A '#' karakter jelzi az elejét, az URI végéig tart.
- Lehetővé teszi egy másodlagos erőforrás közvetett azonosítását egy elsődleges erőforrásra történő hivatkozáson keresztül.
  - A másodlagos erőforrás lehet például az elsődleges erőforrás egy része.
- Jelentését az elsődleges erőforrás elérése során kapott lehetséges reprezentációk határozzák meg, ezek média-típusa.
  - A média típusok megszorításokat szabhatnak az erőforrásrész-azonosító formájára, meghatározhatják az így azonosított másodlagos erőforrások jelentését.
- Hivatkozás-feloldás során mindig eltávolításra kerül.

Erőforrásrész-azonosító (2)

- A séma specifikációk olyan URI szintaxist kell, hogy meghatározzanak, melynek csak abszolút (erőforrásrész-azonosítót nem tartalmazó) URI-k felelnek meg.
  - Nem definiálnak erőforrásrész-azonosító szintaxist vagy használatot, tekintet nélkül arra, hogy az a sémán keresztül azonosítható erőforrásokra alkalmazható-e.

Az erőforrásrész-azonosító jelentése (1)

- `text/html` média típus:
  - Az erőforrásrész azonosító a dokumentum adott részét jelenti vagy állapot információt szolgáltat szkriptek számára. https://www.iana.org/assignments/media-types/text/html
  - Az erőforrásrész-azonosító feldolgozását részletesen a HTML5 specifikáció határozza meg.
    - Lásd: *Navigating to a fragment* https://html.spec.whatwg.org/multipage/browsing-the-web.html#scroll-to-fragid
  - Például a https://www.mozilla.org/hu/#colophon URI esetén az erőforrásrész-azonosító a `colophon` azonosítójú elemet jelenti.
  - Például a https://www.youtube.com/watch?v=w0ffwDYo00Q#t=77 URI esetén az erőforrásrész-azonosító azt jelzi, hogy mely pozíción kell elkezdeni a videó lejátszását (a 77. másodperctől).

Az erőforrásrész-azonosító jelentése (2)

- `application/xml`, `text/xml` és `*/*+xml` média típusok:
  - Az utóbbiba beleértendők például: `application/xhtml+xml`, `image/svg+xml`, `model/x3d+xml`, …
  - Az erőforrásrész-azonosító szintaxisa és jelentése az *XPointer Framework* specifikáción alapul. https://www.iana.org/assignments/media-types/text/xml
    - *XPointer Framework* (W3C ajánlás, 2003. március 25.) https://www.w3.org/TR/xptr-framework/
  - Például a https://www.w3.org/TR/xml/#sec-bibliography URI esetén az erőforrásrész-azonosító a `sec-bibliography` azonosítójú elemet jelenti a dokumentumban.

Abszolút URI, URI-hivatkozás, relatív hivatkozás

- **Abszolút URI**: olyan URI, amely nem tartalmaz erőforrásrész-azonosítót.
  - Bázis URI-ként csak abszolút URI használható.
- **URI-hivatkozás**: URI vagy relatív hivatkozás.
- **Relatív hivatkozás**: kb. egy URI séma-specifikus része, vagy annak egy megfelelő végszelete (lehet akár az üres karakterlánc is).
  - A „relatív URI” kifejezést a specifikáció egyáltalán nem használja!
  - Míg egy URI mindig a használat környezetétől függetlenül azonosít egy erőforrást, egy relatív hivatkozás egy adott környezetben értelmezett.
  - Egy úgynevezett bázis-URI alapján URI-vá lehet feloldani.
  - A relatív hivatkozások feloldásához egy algoritmust ír le a specifikáció.

Példák URI-hivatkozásokra

- `http://www.gnu.org/licenses/licenses.html`
- `http://www.w3.org/TR/xml/#abstract`
- `http://en.wikipedia.org/wiki/The_Beatles#History`
- `/pub/linux/kernel/v3.x/testing/`
- `../../images/bullet.png`
- `index.html#contents`
- `contacts.xml#element(/1/2)`
- `#nav`
- `gpl.html`
- ⟨*üres karakterlánc*⟩

Same-document reference

- Olyan URI hivatkozás, melynek feloldása az erőforrásrész-azonosítótól eltekintve a bázis-URI-val azonos URI-t eredményez.
  - Példa: ⟨*üres karakterlánc*⟩, #nav
  - Hivatkozás-feloldás nem eredményezheti az erőforrás újbóli letöltést.

Bázis-URI meghatározása

- Bizonyos média-típusoknál a relatív hivatkozások bázis-URI-ja beágyazható a tartalomba.
  - Így például dokumentumok definiálhatják a saját magukon belül érvényes bázis-URI-t.
    - XML: a bázis-URI-t az `xml:base` attribútum szolgáltatja (lásd később).
    - HTML: a bázis-URI-t a `base` elem szolgáltatja. https://html.spec.whatwg.org/multipage/semantics.html#the-base-element
- Ha nincs beágyazott bázis-URI és a dokumentumot egy másik entitás – például egy másik dokumentum – foglalja magában, akkor a bázis-URI ennek a bázis URI-ja.
- Ha nincs ilyen befoglaló entitás sem, akkor a bázis-URI az az URI lesz, amelyen a dokumentumot elérték (átirányítás esetén az utolsó használt URI).
- Egyébként a bázis-URI alkalmazásfüggő.

Példák relatív hivatkozások feloldására (1)

- Legyen a bázis-URI `http://example/a/b/c?q`

Relatív hivatkozás	Eredmény URI
`d`	`http://example/a/b/d`
`./d`	`http://example/a/b/d`
`/d`	`http://example/d`
`//localhost`	`http://localhost`
`?y`	`http://example/a/b/c?y`
`d?y`	`http://example/a/b/d?y`

Példák relatív hivatkozások feloldására (2)

- Legyen a bázis-URI `http://example/a/b/c?q`

Relatív hivatkozás	Eredmény URI
`#z`	`http://example/a/b/c?q#z`
`""` (üres karakteránc)	`http://example/a/b/c?q`
`.`	`http://example/a/b/`
`./`	`http://example/a/b/`
`..`	`http://example/a/`
`../d`	`http://example/a/d`
`../../d`	`http://example/d`

Példák relatív hivatkozások feloldására (3)

- Példa:
  - ```
    <!DOCTYPE html>
    <html lang="en">
      <head>
        <title>Example</title>
        <base href="http://example/docs/howto/">
        <link rel="stylesheet" type="text/css" href="theme.css">
      </head>
      <body>
        <a href="/about">
          <img src="../images/logo.png" alt="Logo">
        </a>
      </body>
    </html>
    ```
  - A relatív hivatkozások feloldása:
    - `theme.css` → `http://example/docs/howto/theme.css`
    - `/about` → `http://example/about`
    - `../images/logo.png` → `http://example/docs/images/logo.png`
    ```

URI-k összehasonlítása (1)

- A séma és a *host* komponensek kisbetű-nagybetű érzéketlenek.
- A többi komponensnél kisbetű-nagybetű érzékenységet kell feltételezni, hacsak nem ír elő kisbetű-nagybetű érzéketlenséget a séma.
- Tehát ekvivalensek például a http://www.w3.org/ és HTTP://www.W3.org/ URI-k.

URI-k összehasonlítása (2)

- Ekvivalencia egy lehetséges definíciója:
  - URI-k akkor ekvivalensek, ha ugyanazt az erőforrást azonosítják.
  - Ez a definíció gyakorlati szempontból használhatatlan, mivel általában nincs mód az erőforrások összehasonlítására.
- A gyakorlatban az ekvivalencia megállapítása az URI karakterláncok összehasonlításán alapul.
  - Az összehasonlítás során normalizálás (például nagybetű karakterek kisbetű karakterekké alakítása a kisbetű-nagybetű érzéketlen komponensekben).

JavaScript API

- A WHATWG *URL Living Standard* specifikációja határoz meg egy API-t URL-ek JavaScript-ben történő használatához és manipulálásához: https://url.spec.whatwg.org/#api
- Lásd még: https://developer.mozilla.org/en-US/docs/Web/API/URL
- Példa:

```
const url = new URL(
  "../images",
  "https://eg.com/docs/index.html"
);
console.log(url.href);
```

IRI (1)

- **Nemzetköziesített erőforrás-azonosító** (***internationalized resource identifier***) (**IRI**):
  - Az URI általánosítása, ASCII karakterek helyett Unicode/UCS karakterek használata.
  - A komponensek és a fenntartott karakterek ugyanazok, mint az URI specifikáció esetén.
  - A nem fenntartott karakterek halmazának kiterjesztése.
- A jelenleg aktuális szabvány:
  - Martin Dürst, Michel Suignard. *RFC 3987: Internationalized Resource Identifiers (IRIs)*. January 2005. https://www.rfc-editor.org/rfc/rfc3987

IRI (2)

- Példák:
  - https://en.wiktionary.org/wiki/γεια
  - http://www.öbb.at/

IRI (3)

- Minden IRI-hivatkozás átalakítható egy ekvivalens URI-hivatkozássá:
  - Az IRI-hivatkozás minden olyan karakterére hajtsuk végre az alábbi lépéseket, melyek URI-hivatkozásokban nem megengedettek:
    - Tekintsük a karaktert az UTF-8 karakterkódolásban ábrázoló oktettsorozatot.
    - Az oktettsorozatot kódoljuk egy olyan karakterlánccal, melyben minden oktettet %*HH* módon ábrázolunk, ahol *HH* az oktett értékét reprezentáló két hexadecimális számjegy karakter.
    - A nem megengedett karaktert helyettesítsük az oktettsorozatot kódoló karakterlánccal.
  - Példa: https://www.w3.org/People/Dürst/ → https://www.w3.org/People/D%C3%BCrst/

IRI (4)

- Előnyök:
  - Az URI-k használatának megkönnyítése olyan felhasználók számára, akik nem a latin ábécét használják.
- Kockázatok:
  - *Homograph attacks*: a felhasználó megtévesztése annak kihasználásával, hogy vannak olyan Unicode karakterek, melyek hasonlóan néznek ki.
    - Az URI-k esetén is fennáll a kockázat, lásd például `lame` és 1ame, broken és br0ken.

XML Base (1)

- *XML Base (Second Edition)* (W3C ajánlás, 2009. január 28.) https://www.w3.org/TR/xmlbase/
  - Mechanizmust biztosít bázis-URI-k definiálásához XML dokumentumokban.
  - A bázis-URI megadásához egy `xml:base` nevű attribútumot vezet be.
    - Az attribútumot öröklik a tartalmazó elem leszármazottai, értéke felülbírálható.

XML Base (2)

- Példa:

```
<book xml:base="http://example/books/untitled/"
  xmlns:xi="http://www.w3.org/2001/XInclude">
  <xi:include href="info.xml"/>
  <xi:include href="chapter1.xml"/>
  <xi:include href="../thanks.xml"/>
  <bibliography xml:base="/biblio/">
    <xi:include href="unsorted.xml"/>
  </bibliography>
</book>
```

  - A relatív hivatkozások feloldása:
    - `info.xml` → `http://example/books/untitled/info.xml`, `chapter1.xml` → `http://example/books/untitled/chapter1.xml`
    - `../thanks.xml` → `http://example/books/thanks.xml`
    - `unsorted.xml` → `http://example/biblio/unsorted.xml`

URL rövidítés (1)

- Hosszú `http(s)` URI-k rövidítése HTTP átirányítás révén.
- Célja ugyanarra az erőforrásra mutató, de esztétikusabb, helytakarékosabban megjeleníthető és kommunikációban könnyebben használható `http(s)` URI létrehozása.
  - Például a Twitter üzenetek maximális hossza eredetileg 140 karakter volt.
    - Lásd: *Giving you more characters to express yourself* https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html

URL rövidítés (2)

- URL rövidítő szolgáltatások: https://github.com/738/awesome-url-shortener
- Számos szolgáltatás biztosít funkciót:
  - *GitHub:* https://git.io/
  - *Twitter:* https://t.co/
  - *Wikipedia*: https://meta.wikimedia.org/wiki/Special:UrlShortener
  - *YouTube*: https://y2u.be/
  - …

URL rövidítés: TinyURL

Tulajdonos:	TinyURL, LLC.
Honlap:	https://tinyurl.com/
HTTP állapotkód:	`301` (`Moved Permanently`)
Regisztráció:	opcionális
Egyéni URI:	igen
URL információ:	igen
Követés:	igen
API:	https://tinyurl.com/app/dev
Példák:	https://tinyurl.com/y3rznxva https://preview.tinyurl.com/y3rznxva https://tinyurl.com/IntStream https://preview.tinyurl.com/IntStream

URL rövidítés: Bitly

Tulajdonos:	Bitly, Inc.
Honlap:	https://bitly.com/
HTTP állapotkód:	`308 (Permanent Redirect)`
Regisztráció:	igen
Egyéni URI:	igen
URL információ:	igen
Követés:	igen
API:	https://dev.bitly.com/
Példák:	https://bit.ly/2FlomT4 https://bit.ly/2FlomT4+ https://amzn.to/3h8qX00 https://amzn.to/3h8qX00+