

Egységes erőforrás-azonosító (URI)

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

Utolsó módosítás: 2025. szeptember 20.

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. *Uniform Resource Identifier (URI): Generic Syntax*. RFC 3986, January 2005. <https://www.rfc-editor.org/rfc/rfc3986>

URI (2)

- Minden URI egy sémanévvvel kezdődik, melyet egy ' : ' karakter választ el a sémaspecifikus résztől.
 - A sémaspecifikus 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. *The "file" URI Scheme*. RFC 8089, February 2017. <https://www.rfc-editor.org/rfc/rfc8089>
- `http/https`:
 - Roy T. Fielding (ed.), Mark Nottingham (ed.), Julian F. Reschke (ed.). *HTTP Semantics*. RFC 9110, June 2022. <https://www.rfc-editor.org/rfc/rfc9110>
- `mailto`:
 - Martin Dürst, Larry Masinter, Jamie Zawinski. *The 'mailto' URI Scheme*. RFC 6068, October 2010. <https://www.rfc-editor.org/rfc/rfc6068>
- `about`:
 - S. Moonesamy (ed.). *The „about” URI Scheme*. RFC 6694, August 2012. <https://www.rfc-editor.org/rfc/rfc6694>

Hivatkozásfeloldás (*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. *Uniform Resource Locators (URL)*. RFC 1738, December 1994. <https://www.rfc-editor.org/rfc/rfc1738>
 - ***Uniform Resource Name (URN)***: egységes erőforrásnév
 - Erőforrások helytől független tartós azonosítása.
 - Ryan Moats. *URN Syntax*. RFC 2141, May 1997. <https://www.rfc-editor.org/rfc/rfc2141>

URL vs URN (2)

- Ez a korábbi diszjunkt felosztás mára érvényét veszítette.
 - Michael Mealling (ed.), Ray Denenberg (ed.). *Report from the Joint W3C/IETF URI Planning Interest Group: Uniform Resource Identifiers (URIs), URLs, and Uniform Resource Names (URNs): Clarifications and Recommendations*. RFC 3305, 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 erőforrásnév (URN) egy olyan URI, melynek célja erőforrások helytől független tartós azonosítása.
- Az egységes erőforrásnevek az urn URI-séma alá tartozó URI-k.
- Lásd:
 - Peter Saint-Andre, John C. Klensin. *Uniform Resource Names (URNs)*. RFC 8141, 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) (1)

- 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 vs URL

(IETF vs WHATWG) (2)

- Az IETF és a WHATWG eltérő terminológiája:

IETF-terminológia	WHATWG-terminológia
URI-hivatkozás	URL
URI	abszolút URL
relatív hivatkozás	relatív URL
abszolút URI	–

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%20%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]`
- A hosztkomponens formái:
 - hosztnév (például <http://www.iana.org>)
 - IPv4-cím (például <http://192.0.46.8>)
 - IPv6-cím (például [http://\[2001:500:88:200::8\]](http://[2001:500:88:200::8]))
- Az URI-sémák meghatározhatnak egy alapértelmezett portot.
 - Például a `http` sémánál `80` az alapértelmezés.

Az útvonal komponens

- Útvonalrészecskék '/' 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 médiatípusa.
 - A médiatípusok megszorításokat szabhatnak az erőforrásrész-azonosító formájára, meghatározhatják az így azonosított másodlagos erőforrások jelentését.
- Hivatkozásfeloldás során mindig eltávolításra kerül.

Erőforrásrész-azonosító (2)

- A sémaspecifikációk 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 HTML-specifikáció határozza meg.
 - Lásd: *Navigating to a fragment*
<https://html.spec.whatwg.org/multipage/browsing-the-web.html#scroll-to-fragment>
 - 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 melyik 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 sémaspecifikus 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
 - Hivatkozásfeloldás 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
<code>d</code>	<code>http://example/a/b/d</code>
<code>./d</code>	<code>http://example/a/b/d</code>
<code>/d</code>	<code>http://example/d</code>
<code>//localhost</code>	<code>http://localhost</code>
<code>?y</code>	<code>http://example/a/b/c?y</code>
<code>d?y</code>	<code>http://example/a/b/d?y</code>

Példák relatív hivatkozások feloldására (2)

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

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

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">
        
      </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)

- Ekvivalencia egy lehetséges definíciója:
 - URI-k akkor ekvivalensek, ha ugyanazt az erőforrást azonosítják.
- A gyakorlatban az ekvivalencia megállapítása az URI karakterláncok összehasonlításán alapul.
 - Az összehasonlítás előtt normalizálás történik.
 - Például a nagybetű karakterek kisbetű karakterekké alakítása a kisbetű-nagybetű érzéketlen komponensekben.
 - A normalizálás sémaspecifikus átalakításokat tartalmazhat.
 - Például a `http` és `https` URK-knál a port alkomponens elhagyásra kerül, ha megegyzik a séma alapértelmezett portjával.

URI-k összehasonlítása (2)

- 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.

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. *Internationalized Resource Identifiers (IRIs)*. RFC 3987, 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 oktetsorozatot.
 - Az oktetsorozatot 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 oktetsorozatot 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 1ame 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 felülbíráható.

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 `ht tp(s)` URI-k rövidítése HTTP átirányítás révén.
- Célja ugyanarra az erőforrásra mutató, de esztétikusabb, helytakarékosabban megjeleníthető és kommunikációban könnyebben használható `ht tp(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 webhely biztosít funkciót:
 - OECD: <https://oe.cd/>
 - *Twitter*: <https://t.co/>
 - *Wikipedia*:
<https://meta.wikimedia.org/wiki/Special:UrlShortener>
 - ...

URL rövidítés: Bitly

Tulajdonos:	Bitly, Inc.
Honlap:	https://bitly.com/
HTTP állapotkód:	301 (Moved Permanently)
Regisztráció:	igen
Egyéni URI:	igen
URL információ:	igen
Követés:	igen
API:	https://dev.bitly.com/
Példák:	https://bit.ly/3Kjkd3D https://bit.ly/3Kjkd3D+ https://amzn.to/3h8qX00 https://amzn.to/3h8qX00+