Δικτυακά Προγράμματα

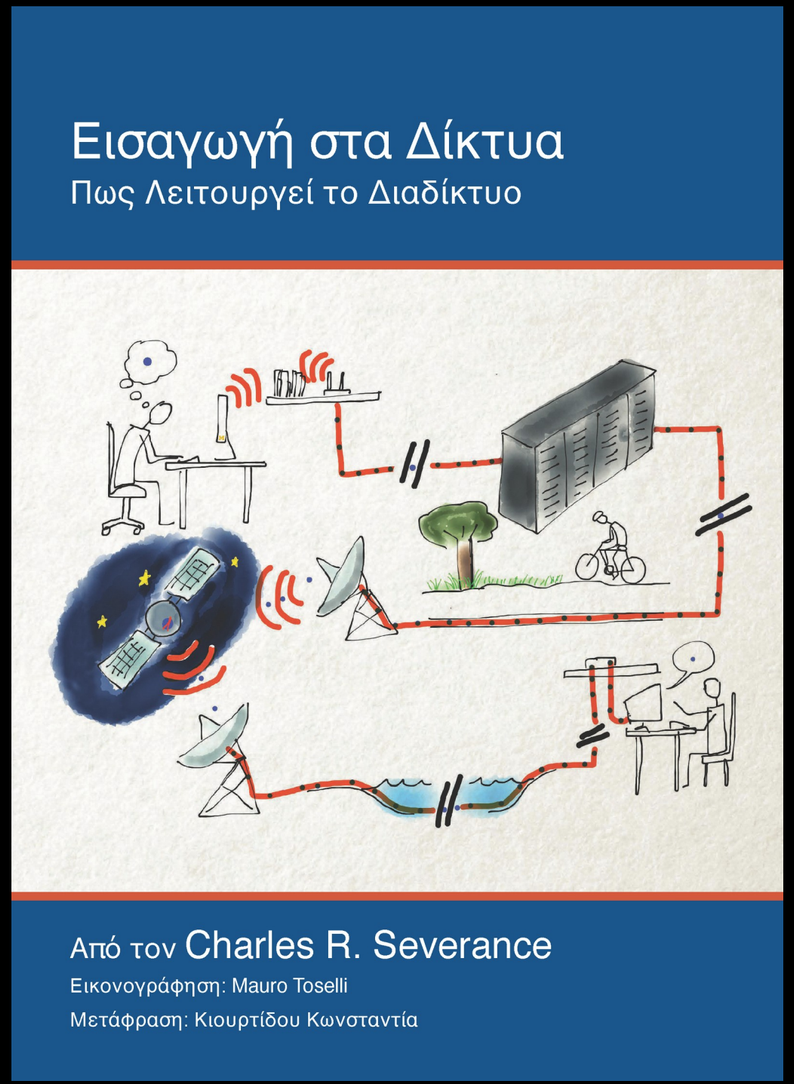
Κεφάλαιο 12

Python για Όλους

www.py4e.com

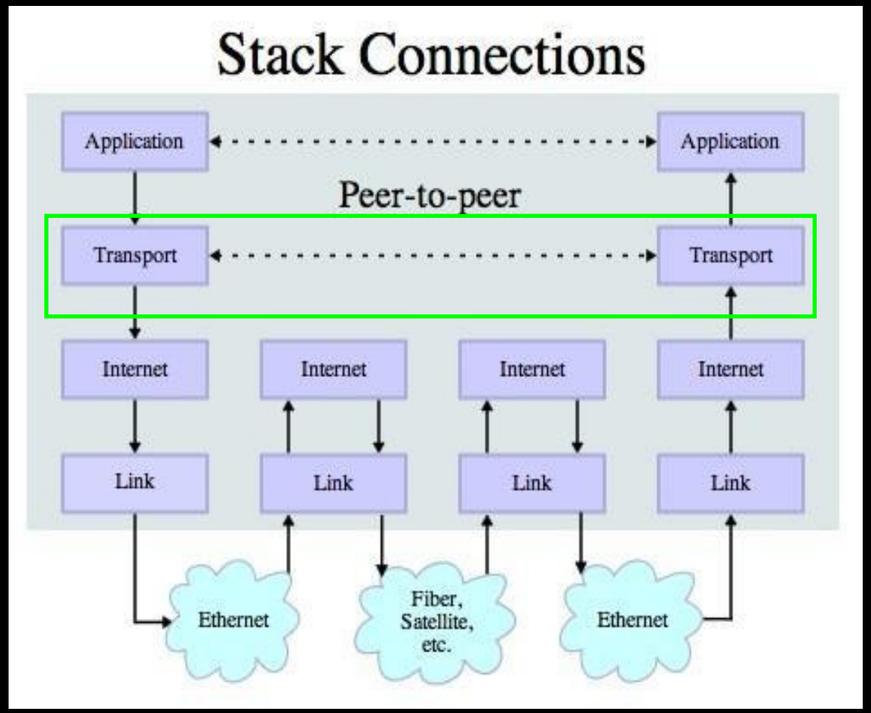
Ένα Δωρεάν Βιβλίο για την Αρχιτεκτονική Δικτύων

- Εάν βρίσκετε αυτό το θέμα ενδιαφέρον και/ή χρειάζεστε περισσότερες λεπτομέρειες
- www.net-intro.com



Επίπεδο Μεταφοράς – Transport Control Protocol (TCP)

- Χτισμένο πάνω από το IP (Επίπεδο Διαδικτύου)
- Υποθέτει ότι το IP μπορεί να χάσει ορισμένα δεδομένα - αποθηκεύει και αναμεταδίδει δεδομένα εάν φαίνεται να έχουν χαθεί
- Χειρίζεται τον «έλεγχο ροής» χρησιμοποιώντας ένα παράθυρο μετάδοσης
- Παρέχει έναν ωραίο αξιόπιστο αγωγό



Πηγή: http://en.wikipedia.org/wiki/Internet_Protocol_Suite

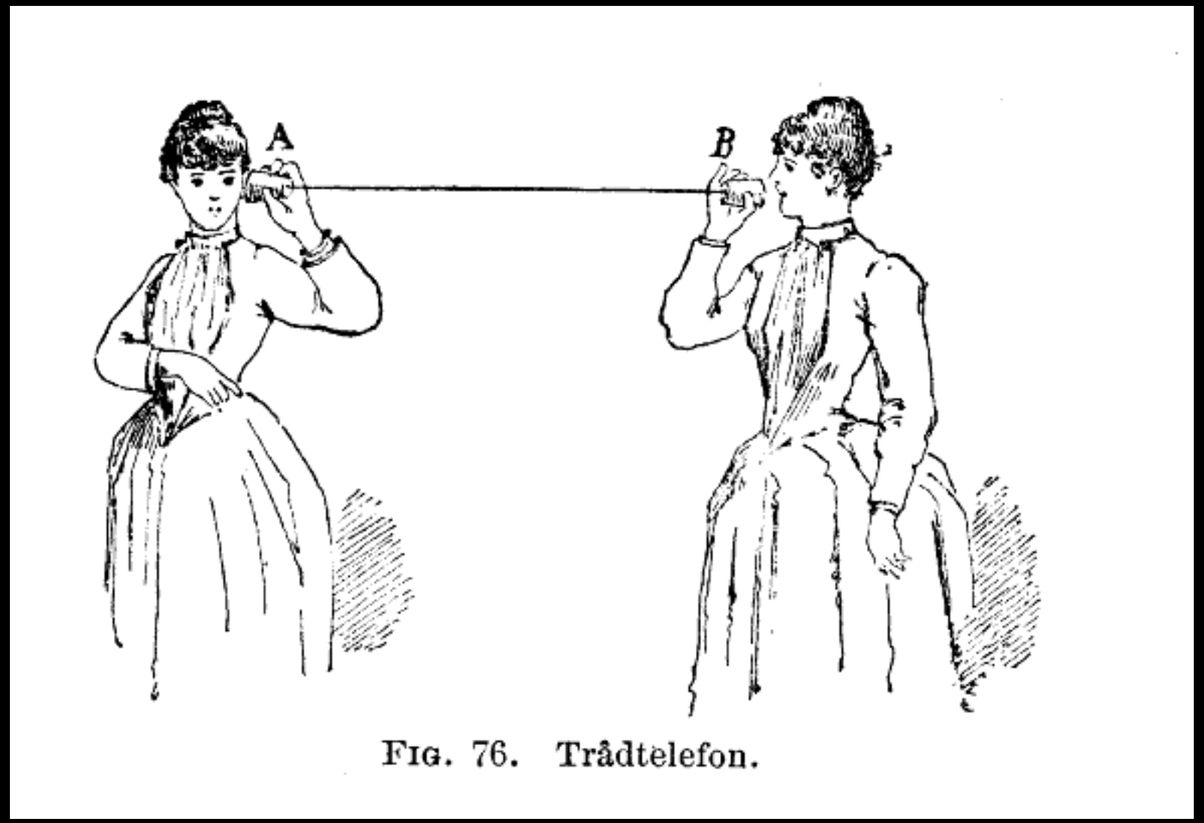


Fig. 76. Trådtelefon.

http://en.wikipedia.org/wiki/Tin_can_telephone

http://www.flickr.com/photos/kitcowan/2103850699/

Συνδέσεις TCP / Υποδοχές

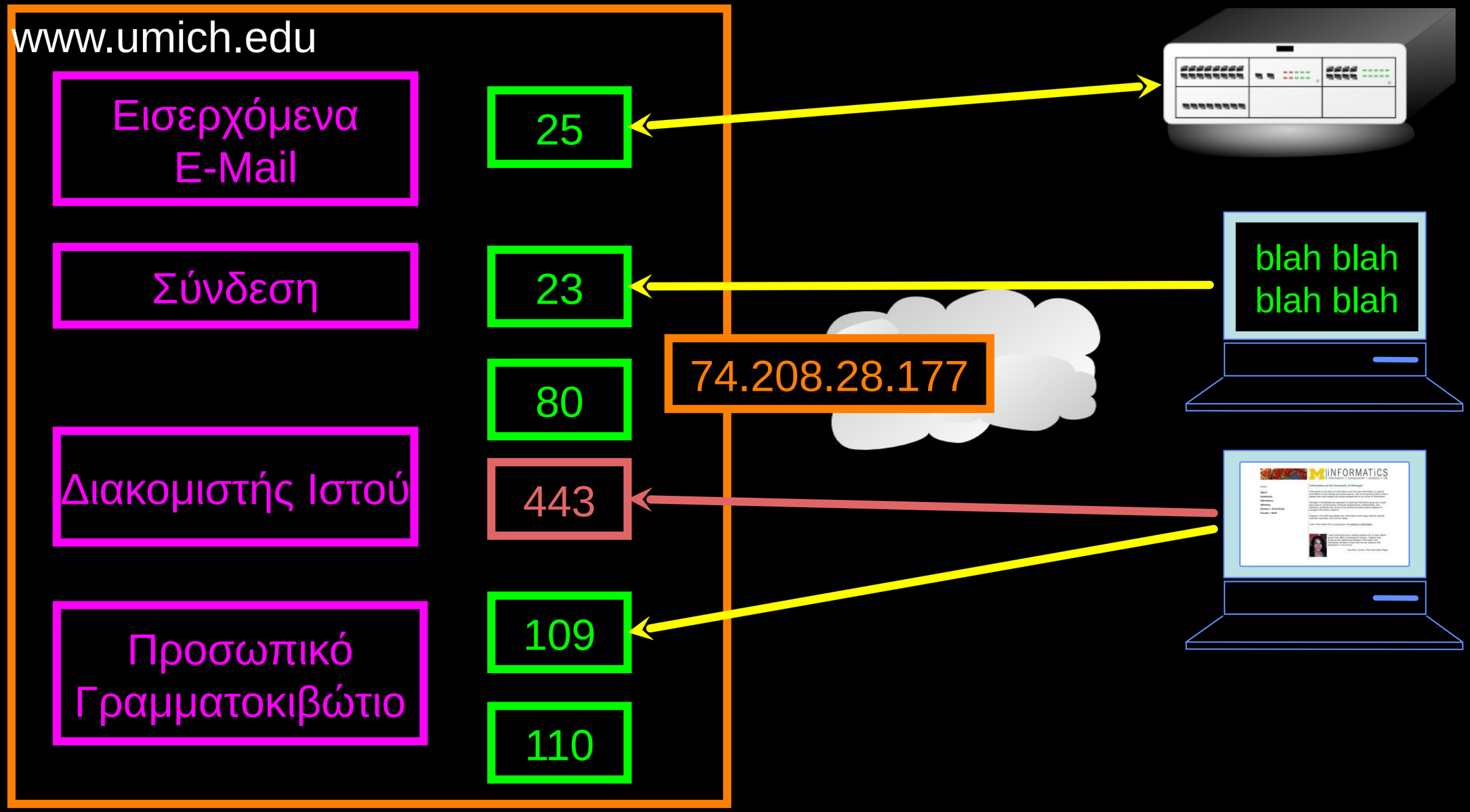
«Στη δικτύωση υπολογιστών, μια υποδοχή Internet ή μια υποδοχή δικτύου είναι ένα τελικό σημείο μιας αμφίδρομης διεργασίας ροής επικοινωνιών σε ένα δίκτυο υπολογιστών που βασίζεται σε Πρωτόκολλο Διαδικτύου, όπως το Διαδίκτυο»

http://en.wikipedia.org/wiki/Internet_socket

TCP Αριθμοί Θυρών

- Μια θύρα είναι ένα τελικό σημείο επικοινωνίας λογισμικού συγκεκριμένης εφαρμογής ή διαδικασίας
- Επιτρέπει τη συνύπαρξη πολλαπλών δικτυακών εφαρμογών στον ίδιο διακομιστή
- Υπάρχει μια λίστα με γνωστούς αριθμούς θυρών TCP

http://en.wikipedia.org/wiki/TCP_and_UDP_port

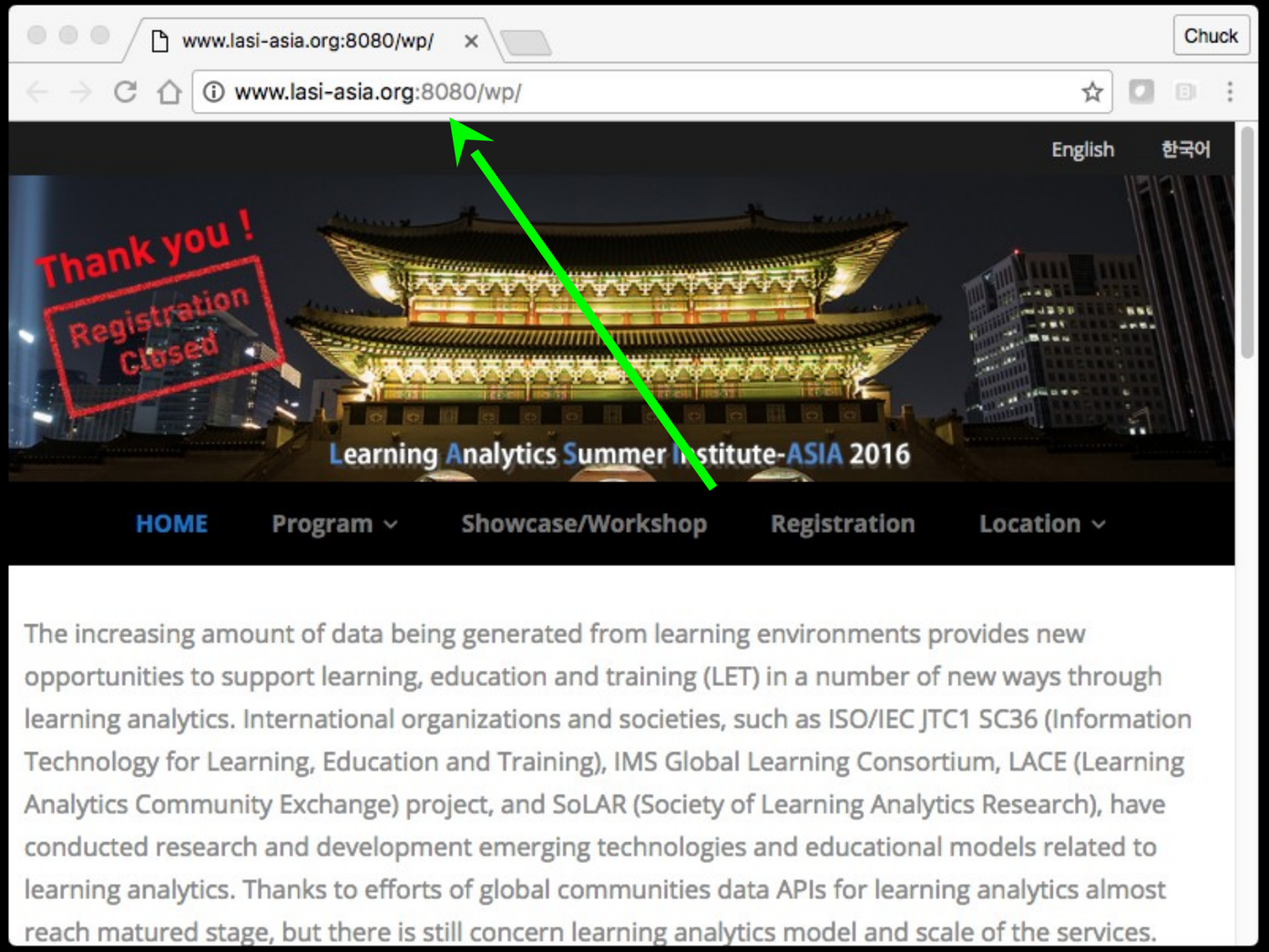


Εικονίδια: http://www.clker.com/search/networksym/1

Συνήθης Θύρες TCP

- Telnet (23) - Login
- SSH (22) - Secure Login
- HTTP (80)
- HTTPS (443) - Secure
- SMTP (25) (Mail)
- IMAP (143/220/993) - Mail Retrieval
- POP (109/110) - Mail Retrieval
- DNS (53) - Domain Name
- FTP (21) - File Transfer

http://en.wikipedia.org/wiki/List_of_TCP_and_UDP_port_numbers

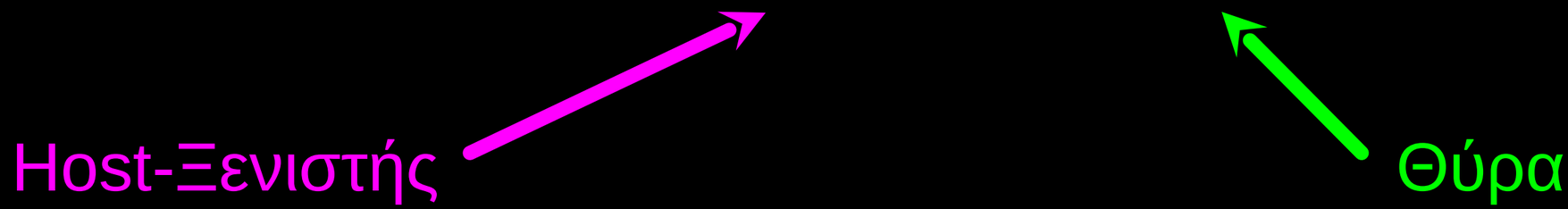


Μερικές φορές βλέπουμε τον αριθμό θύρας στη διεύθυνση URL εάν ο διακομιστής ιστού λειτουργεί σε μια «μη τυπική» θύρα.

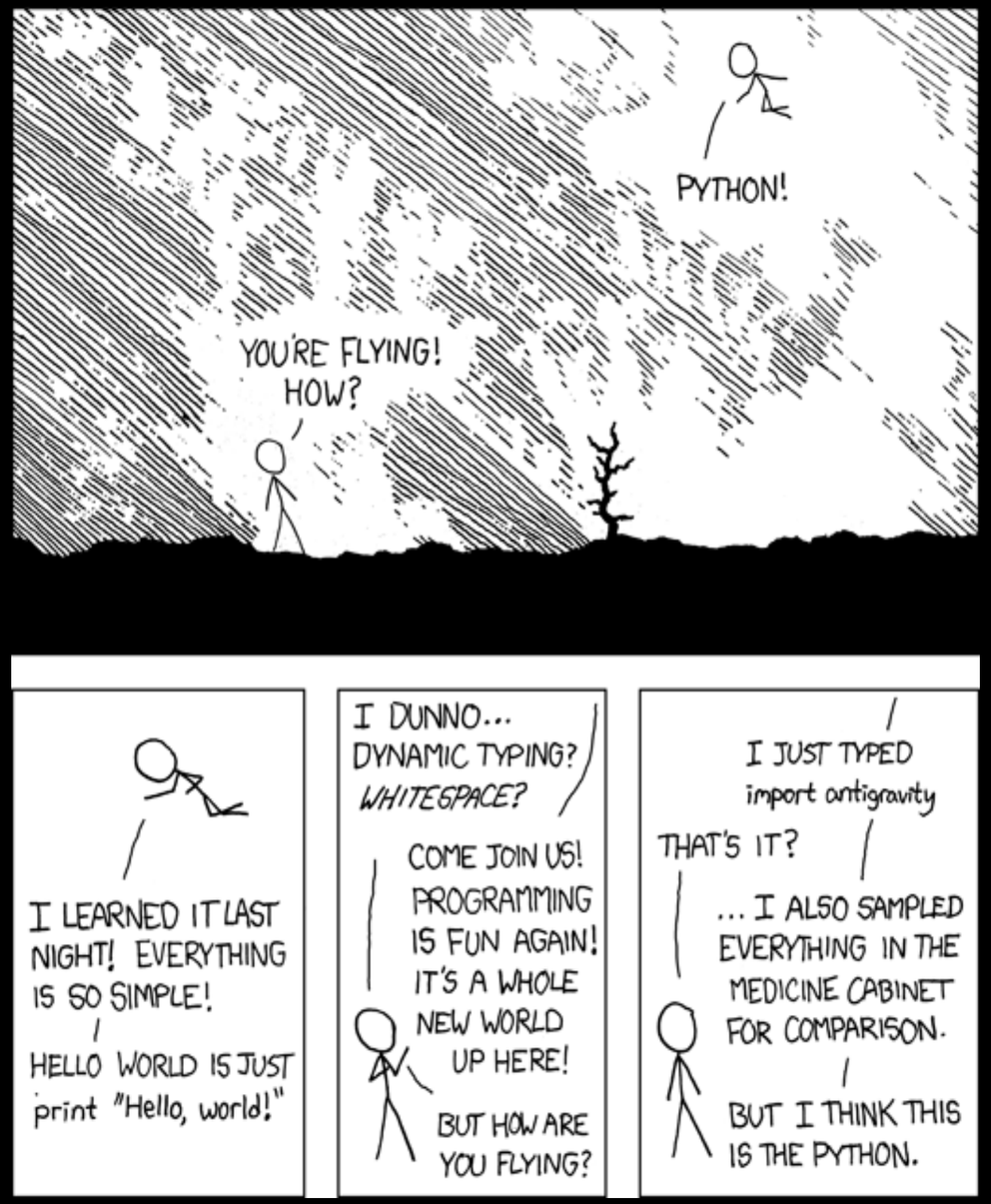
Υποδοχές στην Python

Η Python έχει ενσωματωμένη υποστήριξη για υποδοχές TCP

```
import socket
mysock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
mysock.connect( ('data.pr4e.org', 80) )
```



http://docs.python.org/library/socket.html



http://xkcd.com/353/

Επίπεδο Εφαρμογής

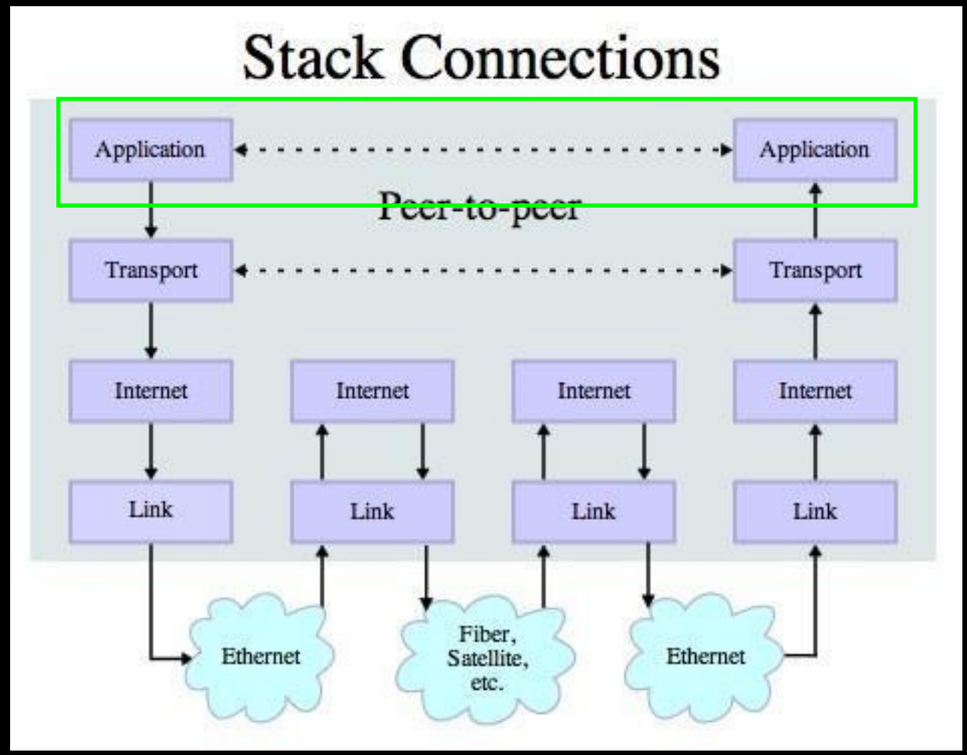
Επίπεδο Εφαρμογής

- Δεδομένου ότι το TCP (και η Python) μας δίνουν μια αξιόπιστη υποδοχή, τι θέλουμε να κάνουμε με την υποδοχή; Τι πρόβλημα θέλουμε να λύσουμε;

- Επίπεδο Εφαρμογής

  - Αλληλογραφία

  - Ιστός Παγκόσμιας Εμβέλειας



Πηγή: http://en.wikipedia.org/wiki/Internet_Protocol_Suite

HTTP – Πρωτόκολλο Μεταφοράς Υπερκειμένου

- Το κυρίαρχο Πρωτόκολλο Επιπέδου Εφαρμογής στο Διαδίκτυο
- Εφευρέθηκε για τον Ιστό - για ανάκτηση HTML, εικόνων, εγγράφων κ.λπ.
- Επεκτείνεται να ανακτά δεδομένα εκτός από έγγραφα - RSS, Υπηρεσίες Ιστού κ.λπ. Βασική Ιδέα - Πραγματοποίηση Σύνδεσης - Αίτημα Εγγράφου - Ανάκτηση Εγγράφου - Κλείσιμο Σύνδεσης

http://en.wikipedia.org/wiki/Http

HTTP

Το HyperText Transfer Protocol (Πρωτόκολλο Μεταφοράς Υπερκειμένου) είναι το σύνολο κανόνων που επιτρέπουν στα προγράμματα περιήγησης να ανακτήσουν έγγραφα ιστού από διακομιστές μέσω Διαδικτύου

Τί είναι ένα Πρωτόκολλο;

- Ένα σύνολο κανόνων που ακολουθούν όλα τα μέρη, ώστε να μπορούμε να προβλέψουμε τη συμπεριφορά του άλλου

- Και μην προσκρούουμε ο ένας στον άλλον

  - Σε δρόμους διπλής κατεύθυνσης στις ΗΠΑ, οδηγήστε στη δεξιά πλευρά του δρόμου

  - Σε δρόμους διπλής κατεύθυνσης στο Ηνωμένο Βασίλειο, οδηγήστε στην αριστερή πλευρά του δρόμου

http://www.dr-chuck.com/page1.htm

πρωτόκολλο host έγγραφο

http://www.youtube.com/watch?v=x2GylLq59rI

1:17 - 2:19

Λήψη Δεδομένων Από τον Διακομιστή

- Κάθε φορά που ο χρήστης κάνει κλικ σε μια ετικέτα αγκύρωσης με href = τιμή για να μεταβεί σε μια νέα σελίδα, το πρόγραμμα περιήγησης πραγματοποιεί σύνδεση με τον διακομιστή ιστού και εκδίδει ένα αίτημα «GET» - για να ΛΑΒΕΙ το περιεχόμενο της σελίδας στην καθορισμένη διεύθυνση URL

- Ο διακομιστής επιστρέφει το έγγραφο HTML στο πρόγραμμα περιήγησης, το οποίο μορφοποιεί και εμφανίζει το έγγραφο στο χρήστη

Διακομιστής Ιστού

80

Πρόγραμμα Περιήγησης

Διακομιστής Ιστού

80

Πρόγραμμα Περιήγησης

Κλικ

Αίτημα

Διακομιστής Ιστού

80

GET
http://www.dr-chuck.com/page2.htm

Πρόγραμμα Περιήγησης

Κλικ

Αίτημα

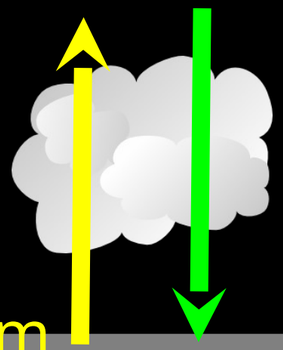
Διακομιστής Ιστού

80

Απάντηση

<h1>The Second Page</h1> If you like, you can switch back to the First Page.

GET
http://www.dr-chuck.com/page2.htm



Πρόγραμμα Περιήγησης

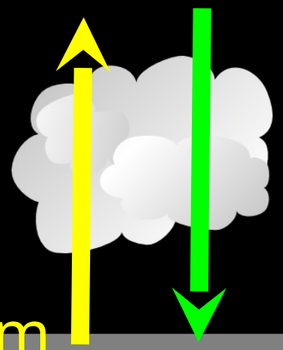
Κλικ

Αίτημα

Διακομιστής Ιστού

80

Απάντηση



GET
http://www.dr-chuck.com/page2.htm

<h1>The Second Page</h1> If you like, you can switch back to the First Page.

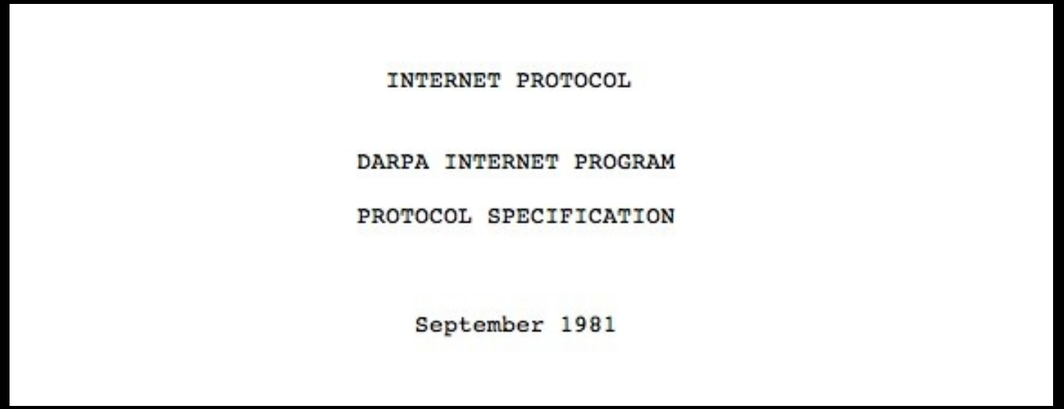
Πρόγραμμα Περιήγησης

Κλικ

Ανάλυση/
Απόδοση

Πρότυπα Διαδικτύου

- Τα πρότυπα για όλα τα πρωτόκολλα του Διαδικτύου (εσωτερικές λειτουργίες) αναπτύσσονται από έναν οργανισμό
- Internet Engineering Task Force (IETF) - Ομάδα Εργασίας Μηχανικής Διαδικτύου
- www.ietf.org
- Τα πρότυπα ονομάζονται «RFC» - «Αίτημα για σχόλια»



INTERNET PROTOCOL

DARPA INTERNET PROGRAM

PROTOCOL SPECIFICATION

September 1981

The internet protocol treats each internet datagram as an independent entity unrelated to any other internet datagram. There are no connections or logical circuits (virtual or otherwise).

The internet protocol uses four key mechanisms in providing its service: Type of Service, Time to Live, Options, and Header Checksum.

Πηγή: http://tools.ietf.org/html/rfc791

Network Working Group
Request for Comments: 2616
Obsoletes: 2068
Category: Standards Track

R. Fielding
UC Irvine
J. Gettys
Compaq/W3C
J. Mogul
Compaq
H. Frystyk
W3C/MIT
L. Masinter
Xerox
P. Leach
Microsoft
T. Berners-Lee
W3C/MIT

June 1999

Hypertext Transfer Protocol -- HTTP/1.1

Status of this Memo

This document specifies an Internet standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "Internet Official Protocol Standards" (STD 1) for the standardization state and status of this protocol. Distribution of this memo is unlimited.

Copyright Notice

Abstract

The Hypertext Transfer Protocol (HTTP) is an application-level protocol for distributed, collaborative, hypermedia information

http://www.w3.org/Protocols/rfc2616/rfc2616.txt

5 Request

A request message from a client to a server includes, within the first line of that message, the method to be applied to the resource, the identifier of the resource, and the protocol version in use.

```
Request       = Request-Line              ; Section 5.1
                *(( general-header        ; Section 4.5
                 | request-header         ; Section 5.3
                 | entity-header ) CRLF)  ; Section 7.1
                CRLF
                [ message-body ]          ; Section 4.3
```

5.1 Request-Line

The Request-Line begins with a method token, followed by the Request-URI and the protocol version, and ending with CRLF. The elements are separated by SP characters. No CR or LF is allowed except in the final CRLF sequence.

```
Request-Line   = Method SP Request-URI SP HTTP-Version CRLF
```

Υποβολή Αιτήματος HTTP

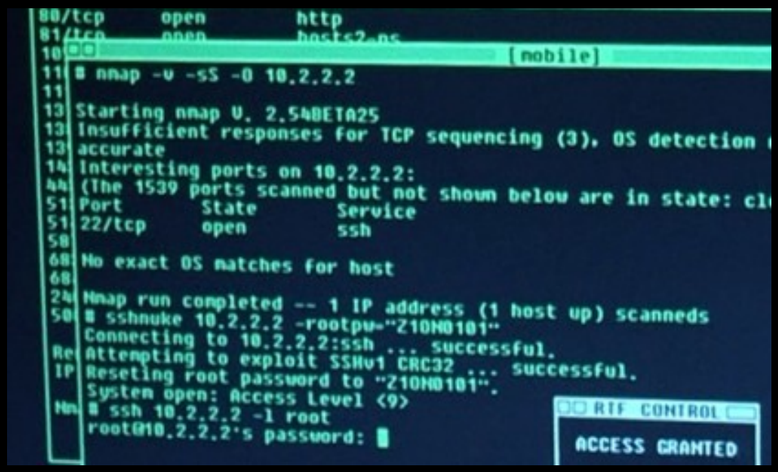
- Σύνδεση σε διακομιστή όπως ο «www.dr-chuck.com»
- Αίτημα ενός εγγράφου (ή του προεπιλεγμένου εγγράφου)
    - GET http://www.dr-chuck.com/page1.htm HTTP/1.0
    - GET http://www.mlive.com/ann-arbor/ HTTP/1.0
    - GET http://www.facebook.com HTTP/1.0

```
$ telnet www.dr-chuck.com 80
Trying 74.208.28.177...
Connected to www.dr-chuck.com.Escape character is '^]'.
GET http://www.dr-chuck.com/page1.htm HTTP/1.0

HTTP/1.1 200 OK
Date: Thu, 08 Jan 2015 01:57:52 GMT
Last-Modified: Sun, 19 Jan 2014 14:25:43 GMT
Connection: close
Content-Type: text/html

<h1>The First Page</h1>
<p>If you like, you can switch to
the <a href="http://www.dr-chuck.com/page2.htm">Second
Page</a>.</p>
Connection closed by foreign host.
```

Ακριβές Hacking στις Ταινίες

- Matrix Reloaded
- Bourne Ultimatum
- Die Hard 4
- ...

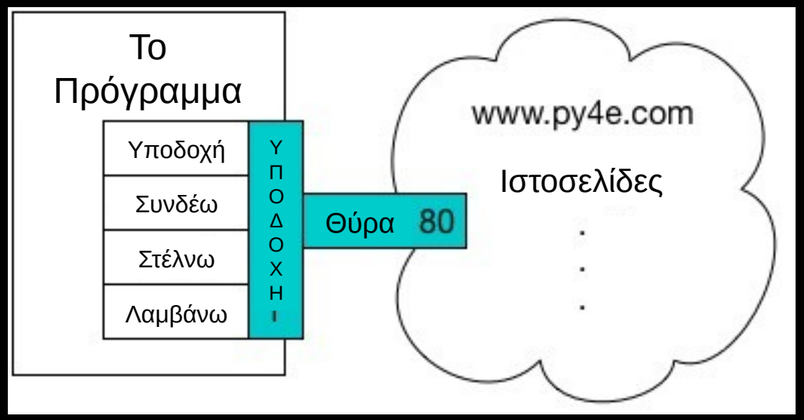
http://nmap.org/movies.html



Ας Γράψουμε ένα Πρόγραμμα Περιήγησης στον Ιστό!

An HTTP Request in Python

```
import socket

mysock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
mysock.connect(('data.pr4e.org', 80))
cmd = 'GET http://data.pr4e.org/romeo.txt HTTP/1.0\r\n\r\n'.encode()
mysock.send(cmd)

while True:
    data = mysock.recv(512)
    if (len(data) < 1):
        break
    print(data.decode(),end='')
mysock.close()
```



```
HTTP/1.1 200 OK
Date: Sun, 14 Mar 2010 23:52:41 GMT
Server: Apache
Last-Modified: Tue, 29 Dec 2009 01:31:22 GMT
ETag: "143c1b33-a7-4b395bea"
Accept-Ranges: bytes
Content-Length: 167
Connection: close
Content-Type: text/plain

But soft what light through yonder window breaks
It is the east and Juliet is the sun
Arise fair sun and kill the envious moon
Who is already sick and pale with grief
```

Επικεφαλίδα HTTP

```
while True:
    data = mysock.recv(512)
    if ( len(data) < 1 ) :
        break
    print(data.decode())
```

Σώμα HTTP

Σχετικά με Χαρακτήρες και Συμβολοσειρές …

ASCII

American
Standard Code
for Information
Interchange

Αμερικανικός
Τυποποιημένος
Κώδικας
Ανταλλαγής
Πληροφοριών

Dec	Hex	Oct	Bin	Char	Dec	Hex	Oct	Bin	Char	Dec	Hex	Oct	Bin	Char	Dec	Hex	Oct	Bin	Char
0	0x00	000	0000000	NUL	32	0x20	040	0100000	space	64	0x40	100	1000000	@	96	0x60	140	1100000	`
1	0x01	001	0000001	SOH	33	0x21	041	0100001	!	65	0x41	101	1000001	A	97	0x61	141	1100001	a
2	0x02	002	0000010	STX	34	0x22	042	0100010	"	66	0x42	102	1000010	B	98	0x62	142	1100010	b
3	0x03	003	0000011	ETX	35	0x23	043	0100011	#	67	0x43	103	1000011	C	99	0x63	143	1100011	c
4	0x04	004	0000100	EOT	36	0x24	044	0100100	$	68	0x44	104	1000100	D	100	0x64	144	1100100	d
5	0x05	005	0000101	ENQ	37	0x25	045	0100101	%	69	0x45	105	1000101	E	101	0x65	145	1100101	e
6	0x06	006	0000110	ACK	38	0x26	046	0100110	&	70	0x46	106	1000110	F	102	0x66	146	1100110	f
7	0x07	007	0000111	BEL	39	0x27	047	0100111	'	71	0x47	107	1000111	G	103	0x67	147	1100111	g
8	0x08	010	0001000	BS	40	0x28	050	0101000	(	72	0x48	110	1001000	H	104	0x68	150	1101000	h
9	0x09	011	0001001	TAB	41	0x29	051	0101001	)	73	0x49	111	1001001	I	105	0x69	151	1101001	i
10	0x0A	012	0001010	LF	42	0x2A	052	0101010	*	74	0x4A	112	1001010	J	106	0x6A	152	1101010	j
11	0x0B	013	0001011	VT	43	0x2B	053	0101011	+	75	0x4B	113	1001011	K	107	0x6B	153	1101011	k
12	0x0C	014	0001100	FF	44	0x2C	054	0101100	,	76	0x4C	114	1001100	L	108	0x6C	154	1101100	l
13	0x0D	015	0001101	CR	45	0x2D	055	0101101	-	77	0x4D	115	1001101	M	109	0x6D	155	1101101	m
14	0x0E	016	0001110	SO	46	0x2E	056	0101110	.	78	0x4E	116	1001110	N	110	0x6E	156	1101110	n
15	0x0F	017	0001111	SI	47	0x2F	057	0101111	/	79	0x4F	117	1001111	O	111	0x6F	157	1101111	o
16	0x10	020	0010000	DLE	48	0x30	060	0110000	0	80	0x50	120	1010000	P	112	0x70	160	1110000	p
17	0x11	021	0010001	DC1	49	0x31	061	0110001	1	81	0x51	121	1010001	Q	113	0x71	161	1110001	q
18	0x12	022	0010010	DC2	50	0x32	062	0110010	2	82	0x52	122	1010010	R	114	0x72	162	1110010	r
19	0x13	023	0010011	DC3	51	0x33	063	0110011	3	83	0x53	123	1010011	S	115	0x73	163	1110011	s
20	0x14	024	0010100	DC4	52	0x34	064	0110100	4	84	0x54	124	1010100	T	116	0x74	164	1110100	t
21	0x15	025	0010101	NAK	53	0x35	065	0110101	5	85	0x55	125	1010101	U	117	0x75	165	1110101	u
22	0x16	026	0010110	SYN	54	0x36	066	0110110	6	86	0x56	126	1010110	V	118	0x76	166	1110110	v
23	0x17	027	0010111	ETB	55	0x37	067	0110111	7	87	0x57	127	1010111	W	119	0x77	167	1110111	w
24	0x18	030	0011000	CAN	56	0x38	070	0111000	8	88	0x58	130	1011000	X	120	0x78	170	1111000	x
25	0x19	031	0011001	EM	57	0x39	071	0111001	9	89	0x59	131	1011001	Y	121	0x79	171	1111001	y
26	0x1A	032	0011010	SUB	58	0x3A	072	0111010	:	90	0x5A	132	1011010	Z	122	0x7A	172	1111010	z
27	0x1B	033	0011011	ESC	59	0x3B	073	0111011	;	91	0x5B	133	1011011	[	123	0x7B	173	1111011	{
28	0x1C	034	0011100	FS	60	0x3C	074	0111100	<	92	0x5C	134	1011100	\	124	0x7C	174	1111100	|
29	0x1D	035	0011101	GS	61	0x3D	075	0111101	=	93	0x5D	135	1011101	]	125	0x7D	175	1111101	}
30	0x1E	036	0011110	RS	62	0x3E	076	0111110	>	94	0x5E	136	1011110	^	126	0x7E	176	1111110	~
31	0x1F	037	0011111	US	63	0x3F	077	0111111	?	95	0x5F	137	1011111	_	127	0x7F	177	1111111	DEL

https://en.wikipedia.org/wiki/ASCII
http://www.catonmat.net/download/ascii-cheat-sheet.png

Αναπαράσταση Απλών Συμβολοσειρών

- Κάθε χαρακτήρας αντιπροσωπεύεται από έναν αριθμό μεταξύ 0 και 256 αποθηκευμένων σε 8 bit μνήμης
- Αναφερόμαστε σε «8 bits» μνήμης ως ένα «byte» μνήμης - (δηλαδή η μονάδα δίσκου μου περιέχει 3 Terabytes μνήμης)
- Η συνάρτηση ord() μας λέει την αριθμητική τιμή ενός απλού χαρακτήρα ASCII

```
>>> print(ord('H'))
72
>>> print(ord('e'))
101
>>> print(ord('\n'))
10
>>>
```

ASCII

```
>>> print(ord('H'))
72
>>> print(ord('e'))
101
>>> print(ord('\n'))
10
>>>
```

Στη δεκαετία του 1960 και του 1970, απλά υποθέταμε ότι ένα byte ήταν ένας χαρακτήρας

Dec	Hex	Oct	Bin	Char	Dec	Hex	Oct	Bin	Char	Dec	Hex	Oct	Bin	Char	Dec	Hex	Oct	Bin	Char
0	0x00	000	0000000	NUL	32	0x20	040	0100000	space	64	0x40	100	1000000	@	96	0x60	140	1100000	`
1	0x01	001	0000001	SOH	33	0x21	041	0100001	!	65	0x41	101	1000001	A	97	0x61	141	1100001	a
2	0x02	002	0000010	STX	34	0x22	042	0100010	"	66	0x42	102	1000010	B	98	0x62	142	1100010	b
3	0x03	003	0000011	ETX	35	0x23	043	0100011	#	67	0x43	103	1000011	C	99	0x63	143	1100011	c
4	0x04	004	0000100	EOT	36	0x24	044	0100100	$	68	0x44	104	1000100	D	100	0x64	144	1100100	d
5	0x05	005	0000101	ENQ	37	0x25	045	0100101	%	69	0x45	105	1000101	E	101	0x65	145	1100101	e
6	0x06	006	0000110	ACK	38	0x26	046	0100110	&	70	0x46	106	1000110	F	102	0x66	146	1100110	f
7	0x07	007	0000111	BEL	39	0x27	047	0100111	'	71	0x47	107	1000111	G	103	0x67	147	1100111	g
8	0x08	010	0001000	BS	40	0x28	050	0101000	(	72	0x48	110	1001000	H	104	0x68	150	1101000	h
9	0x09	011	0001001	TAB	41	0x29	051	0101001	)	73	0x49	111	1001001	I	105	0x69	151	1101001	i
10	0x0A	012	0001010	LF	42	0x2A	052	0101010	*	74	0x4A	112	1001010	J	106	0x6A	152	1101010	j
11	0x0B	013	0001011	VT	43	0x2B	053	0101011	+	75	0x4B	113	1001011	K	107	0x6B	153	1101011	k
12	0x0C	014	0001100	FF	44	0x2C	054	0101100	,	76	0x4C	114	1001100	L	108	0x6C	154	1101100	l
13	0x0D	015	0001101	CR	45	0x2D	055	0101101	-	77	0x4D	115	1001101	M	109	0x6D	155	1101101	m
14	0x0E	016	0001110	SO	46	0x2E	056	0101110	.	78	0x4E	116	1001110	N	110	0x6E	156	1101110	n
15	0x0F	017	0001111	SI	47	0x2F	057	0101111	/	79	0x4F	117	1001111	O	111	0x6F	157	1101111	o
16	0x10	020	0010000	DLE	48	0x30	060	0110000	0	80	0x50	120	1010000	P	112	0x70	160	1110000	p
17	0x11	021	0010001	DC1	49	0x31	061	0110001	1	81	0x51	121	1010001	Q	113	0x71	161	1110001	q
18	0x12	022	0010010	DC2	50	0x32	062	0110010	2	82	0x52	122	1010010	R	114	0x72	162	1110010	r
19	0x13	023	0010011	DC3	51	0x33	063	0110011	3	83	0x53	123	1010011	S	115	0x73	163	1110011	s
20	0x14	024	0010100	DC4	52	0x34	064	0110100	4	84	0x54	124	1010100	T	116	0x74	164	1110100	t
21	0x15	025	0010101	NAK	53	0x35	065	0110101	5	85	0x55	125	1010101	U	117	0x75	165	1110101	u
22	0x16	026	0010110	SYN	54	0x36	066	0110110	6	86	0x56	126	1010110	V	118	0x76	166	1110110	v
23	0x17	027	0010111	ETB	55	0x37	067	0110111	7	87	0x57	127	1010111	W	119	0x77	167	1110111	w
24	0x18	030	0011000	CAN	56	0x38	070	0111000	8	88	0x58	130	1011000	X	120	0x78	170	1111000	x
25	0x19	031	0011001	EM	57	0x39	071	0111001	9	89	0x59	131	1011001	Y	121	0x79	171	1111001	y
26	0x1A	032	0011010	SUB	58	0x3A	072	0111010	:	90	0x5A	132	1011010	Z	122	0x7A	172	1111010	z
27	0x1B	033	0011011	ESC	59	0x3B	073	0111011	;	91	0x5B	133	1011011	[	123	0x7B	173	1111011	{
28	0x1C	034	0011100	FS	60	0x3C	074	0111100	<	92	0x5C	134	1011100	\	124	0x7C	174	1111100	|
29	0x1D	035	0011101	GS	61	0x3D	075	0111101	=	93	0x5D	135	1011101	]	125	0x7D	175	1111101	}
30	0x1E	036	0011110	RS	62	0x3E	076	0111110	>	94	0x5E	136	1011110	^	126	0x7E	176	1111110	~
31	0x1F	037	0011111	US	63	0x3F	077	0111111	?	95	0x5F	137	1011111	_	127	0x7F	177	1111111	DEL

Code Charts

Home | Site Map | Search

Unicode 9.0 Character Code Charts

SCRIPTS | SYMBOLS | NOTES

http://unicode.org/charts/

Find chart by hex code: Go Related links: Name index Help & links

Scripts

European Scripts	African Scripts	South Asian Scripts	Indonesia & Oceania Scripts
Armenian	*Adlam*	*Ahom*	*Balinese*
Armenian Ligatures	*Bamum*	*Bengali and Assamese*	*Batak*
Caucasian Albanian	Bamum Supplement	*Bhaiksuki*	*Buginese*
Cypriot Syllabary	*Bassa Vah*	*Brahmi*	*Buhid*
Cyrillic	*Coptic*	*Chakma*	*Hanunoo*
Cyrillic Supplement	*Coptic in Greek block*	*Devanagari*	*Javanese*
Cyrillic Extended-A	Coptic Epact Numbers	Devanagari Extended	*Rejang*
Cyrillic Extended-B	*Egyptian Hieroglyphs (1MB)*	*Grantha*	*Sundanese*
Cyrillic Extended-C	*Ethiopic*	*Gujarati*	Sundanese Supplement
Elbasan	Ethiopic Supplement	*Gurmukhi*	*Tagalog*
Georgian	Ethiopic Extended	*Kaithi*	*Tagbanwa*
Georgian Supplement	Ethiopic Extended-A	*Kannada*	**East Asian Scripts**
Glagolitic	*Mende Kikakui*	*Kharoshthi*	*Bopomofo*
Glagolitic Supplement	*Meroitic*	*Khojki*	Bopomofo Extended
Gothic	Meroitic Cursive	*Khudawadi*	*CJK Unified Ideographs (Han) (35MB)*
Greek	Meroitic Hieroglyphs	*Lepcha*	CJK Extension-A (6MB)
Greek Extended	*N'Ko*	*Limbu*	CJK Extension B (40MB)
Ancient Greek Numbers	*Osmanya*	*Mahajani*	CJK Extension C (3MB)
Latin	*Tifinagh*	*Malayalam*	CJK Extension D
Basic Latin (ASCII)	*Vai*	*Meetei Mayek*	CJK Extension E (3.5MB)
Latin-1 Supplement	**Middle Eastern Scripts**	Meetei Mayek Extensions	(see also Unihan Database)
Latin Extended-A	*Anatolian Hieroglyphs*	*Modi*	*CJK Compatibility Ideographs*

Χαρακτήρες Πολλαπλών Byte

Για να αντιπροσωπεύσουμε το ευρύ φάσμα χαρακτήρων που πρέπει να χειρίζονται οι υπολογιστές, αντιπροσωπεύουμε χαρακτήρες με περισσότερα από ένα byte

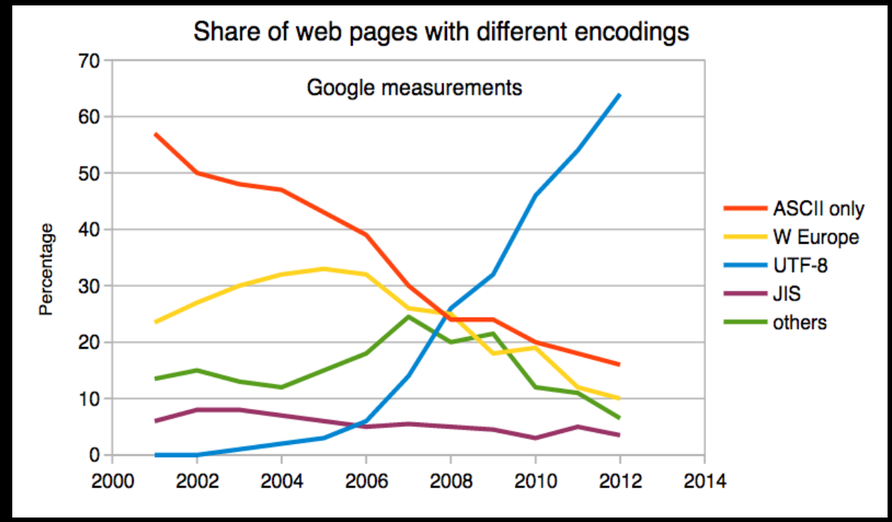
- UTF-16 – Σταθερό μήκος - Δύο bytes
- UTF-32 – Σταθερό μήκος - Τέσσερα Bytes
- UTF-8 – 1-4 bytes

  - Προς τα πάνω συμβατό με ASCII

  - Αυτόματη ανίχνευση μεταξύ ASCII and UTF-8

  - UTF-8 συνιστώμενη πρακτική για την κωδικοποίηση δεδομένων που ανταλλάσσονται μεταξύ συστημάτων

https://en.wikipedia.org/wiki/UTF-8



Δύο Είδη Συμβολοσειρών στην Python

```
Python 2.7.10
>>> x = '이광춘'
>>> type(x)
<type 'str'>
>>> x = u'이광춘'
>>> type(x)
<type 'unicode'>
>>>
```

```
Python 3.5.1
>>> x = '이광춘'
>>> type(x)
<class 'str'>
>>> x = u'이광춘'
>>> type(x)
<class 'str'>
>>>
```

Στην Python 3, όλες οι συμβολοσειρές είναι Unicode

Python 2 ένναντι Python 3

```
Python 2.7.10
>>> x = b'abc'
>>> type(x)
<type 'str'>
>>> x = '이광춘'
>>> type(x)
<type 'str'>
>>> x = u'이광춘'
>>> type(x)
<type 'unicode'>
```

```
Python 3.5.1
>>> x = b'abc'
>>> type(x)
<class 'bytes'>
>>> x = '이광춘'
>>> type(x)
<class 'str'>
>>> x = u'이광춘'
>>> type(x)
<class 'str'>
```

Python 3 και Unicode

- Στην Python 3, όλες οι συμβολοσειρές εσωτερικά είναι UNICODE
- Η εργασία με μεταβλητές συμβολοσειράς σε προγράμματα Python και η ανάγνωση δεδομένων από αρχεία συνήθως «απλά λειτουργεί»
- Όταν μιλάμε σε έναν πόρο δικτύου χρησιμοποιώντας υποδοχές ή μιλάμε σε μια βάση δεδομένων, πρέπει να κωδικοποιήσουμε και να αποκωδικοποιήσουμε δεδομένα (συνήθως σε UTF-8)

```
Python 3.5.1
>>> x = b'abc'
>>> type(x)
<class 'bytes'>
>>> x = '이광춘'
>>> type(x)
<class 'str'>
>>> x = u'이광춘'
>>> type(x)
<class 'str'>
```

Συμβολοσειρές Python σε Bytes

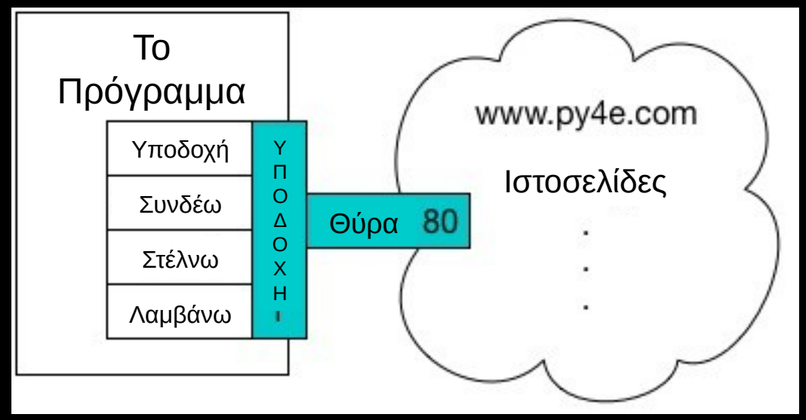
- Όταν μιλάμε με έναν εξωτερικό πόρο, όπως μια υποδοχή δικτύου, στέλνουμε byte, οπότε πρέπει να κωδικοποιήσουμε τις συμβολοσειρές της Python 3 σε δεδομένη κωδικοποίηση χαρακτήρων

- Όταν διαβάζουμε δεδομένα από έναν εξωτερικό πόρο, πρέπει να τα αποκωδικοποιήσουμε με βάση το σύνολο χαρακτήρων, ώστε να αναπαρίστανται σωστά στην Python 3 ως συμβολοσειρά

```
while True:
    data = mysock.recv(512)
    if ( len(data) < 1 ) :
        break
    mystring = data.decode()
    print(mystring)
```

Αίτημα HTTP στην Python

```
import socket

mysock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
mysock.connect(('data.pr4e.org', 80))
cmd = 'GET http://data.pr4e.org/romeo.txt HTTP/1.0\n\n'.encode()
mysock.send(cmd)

while True:
    data = mysock.recv(512)
    if (len(data) < 1):
        break
    print(data.decode())
mysock.close()
```



`bytes.`**`decode`**`(`*`encoding="utf-8", errors="strict"`*`)`

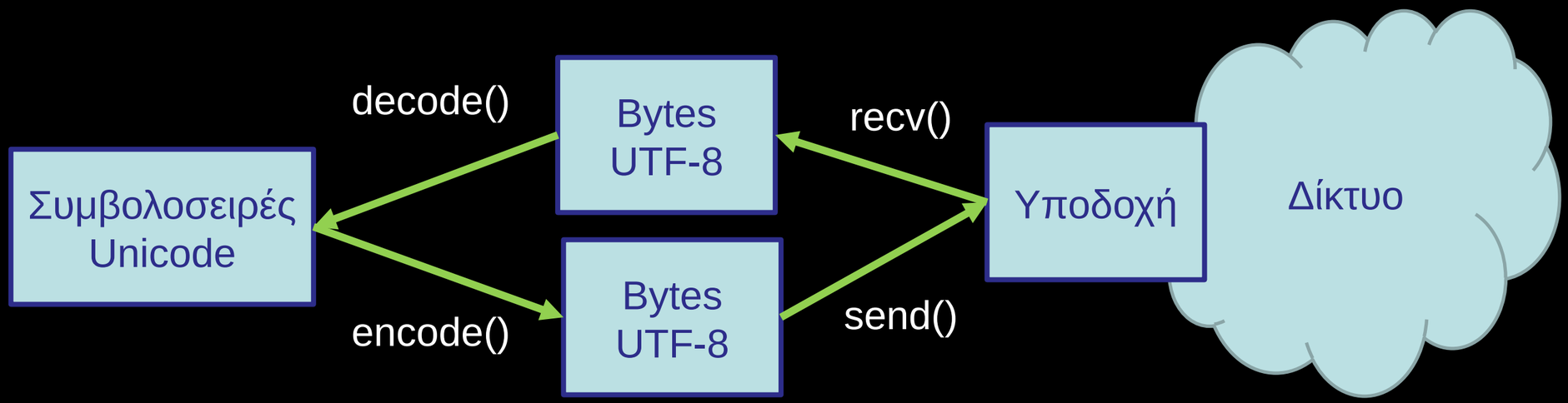
`bytearray.`**`decode`**`(`*`encoding="utf-8", errors="strict"`*`)`

Return a string decoded from the given bytes. Default encoding is `'utf-8'`. *errors* may be given to set a different error handling scheme. The default for *errors* is `'strict'`, meaning that encoding errors raise a `UnicodeError`. Other possible values are `'ignore'`, `'replace'` and any other name registered via `codecs.register_error()`, see section Error Handlers. For a list of possible encodings, see section Standard Encodings.

`str.`**`encode`**`(`*`encoding="utf-8", errors="strict"`*`)`

Return an encoded version of the string as a bytes object. Default encoding is `'utf-8'`. *errors* may be given to set a different error handling scheme. The default for *errors* is `'strict'`, meaning that encoding errors raise a `UnicodeError`. Other possible values are `'ignore'`, `'replace'`, `'xmlcharrefreplace'`, `'backslashreplace'` and any other name registered via `codecs.register_error()`, see section Error Handlers. For a list of possible encodings, see section Standard Encodings.

https://docs.python.org/3/library/stdtypes.html#bytes.decode
https://docs.python.org/3/library/stdtypes.html#str.encode




```
import socket

mysock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
mysock.connect(('data.pr4e.org', 80))
cmd = 'GET http://data.pr4e.org/romeo.txt HTTP/1.0\n\n'.encode()
mysock.send(cmd)

while True:
    data = mysock.recv(512)
    if (len(data) < 1):
        break
    print(data.decode())
mysock.close()
```

Κάνοντας το HTTP Ευκολότερο Με την urllib

Χρήση της urllib στην Python

Δεδομένου ότι το HTTP είναι τόσο συνηθισμένο, έχουμε μια βιβλιοθήκη που κάνει όλες τις υποδοχές για εμάς και κάνει τις ιστοσελίδες να μοιάζουν με αρχείο

```
import urllib.request, urllib.parse, urllib.error

fhand = urllib.request.urlopen('http://data.pr4e.org/romeo.txt')
for γραμμή in fhand:
    print(γραμμή.decode().strip())
```

urllib1.py

```
import urllib.request, urllib.parse, urllib.error

fhand = urllib.request.urlopen('http://data.pr4e.org/romeo.txt')
for γραμμή in fhand:
    print(γραμμή.decode().strip())
```

```
But soft what light through yonder window breaks
It is the east and Juliet is the sun
Arise fair sun and kill the envious moon
Who is already sick and pale with grief
```

urllib1.py

Όπως ένα Αρχείο...

```
import urllib.request, urllib.parse, urllib.error

fhand = urllib.request.urlopen('http://data.pr4e.org/romeo.txt’)

πλήθη = dict()
for γραμμή in fhand:
    λέξεις = γραμμή.decode().split()
    for λέξη in λέξεις:
        πλήθη[λέξη] = πλήθη.get(λέξη, 0) + 1
print(πλήθη)
```

urlwords.py

Διαβάζοντας Ιστοσελίδες

```
import urllib.request, urllib.parse, urllib.error

fhand = urllib.request.urlopen('http://www.dr-chuck.com/page1.htm')
for line in fhand:
    print(line.decode().strip())
```

```
<h1>The First Page</h1>
<p>If you like, you can switch to the <a
href="http://www.dr-chuck.com/page2.htm">Second
Page</a>.
</p>
```

urllib2.py

Ακολουθώντας Συνδέσμους

```
import urllib.request, urllib.parse, urllib.error

fhand = urllib.request.urlopen('http://www.dr-chuck.com/page1.htm')
for line in fhand:
    print(line.decode().strip())
```

```
<h1>The First Page</h1>
<p>If you like, you can switch to the <a
href="http://www.dr-chuck.com/page2.htm">Second
Page</a>.
</p>
```

urllib2.py

Οι πρώτες γραμμές κώδικα @Google;

```
import urllib.request, urllib.parse, urllib.error

fhand = urllib.request.urlopen('http://www.dr-chuck.com/page1.htm')
for line in fhand:
    print(line.decode().strip())
```

Ανάλυση HTML
(γνωστό και ως Ιστοσυγκομιδή)

Τί είναι η Ανίχνευση Ιστού;

- Όταν ένα πρόγραμμα ή σενάριο προσποιείται ότι είναι πρόγραμμα περιήγησης και ανακτά ιστοσελίδες, κοιτάζει αυτές τις ιστοσελίδες, εξάγει πληροφορίες και μετά κοιτάζει περισσότερες ιστοσελίδες

- Οι μηχανές αναζήτησης ανιχνεύουν ιστοσελίδες - το ονομάζουμε "spidering the web" ή "web crawling"

http://en.wikipedia.org/wiki/Web_scraping
http://en.wikipedia.org/wiki/Web_crawler

Γιατί Ιστοσυγκομιδή;

- Τράβηγμα δεδομένων - ιδιαίτερα κοινωνικά δεδομένα - ποιος συνδέεται με ποιον;
- Ανάκτηση των δεδομένων σας από κάποιο σύστημα που δεν δίνει «δυνατότητα εξαγωγής»
- Παρακολούθηση ενός ιστοτόπου για νέες πληροφορίες
- Ανίχνευση Ιστού για να δημιουργήσετε μια βάση δεδομένων για μια μηχανή αναζήτησης

Ιστοσυγκομιδή

- Υπάρχει κάποια διαμάχη σχετικά με την ιστοσυγκομιδή και ορισμένοι ιστότοποι είναι λίγο σκαιοί.
- Δεν επιτρέπεται η αναδημοσίευση πληροφοριών που προστατεύονται από πνευματικά δικαιώματα
- Δεν επιτρέπονται παραβιάσεις των όρων παροχής υπηρεσιών

Ο Εύκολος Τρόπος - Beautiful Soup

- Μπορείτε να κάνετε αναζητήσεις συμβολοσειράς με τον δύσκολο τρόπο
- Or χρησιμοποιήστε τη δωρεάν βιβλιοθήκη λογισμικού που ονομάζεται BeautifulSoup από www.crummy.com

https://www.crummy.com/software/BeautifulSoup/

Εγκατάσταση BeautifulSoup

```
# Για να το εκτελέσετε, μπορείτε να εγκαταστήσετε το BeautifulSoup
# https://pypi.python.org/pypi/beautifulsoup4

# Ή να κατεβάστε το αρχείο
# http://www.py4e.com/code3/bs4.zip
# και να το αποσυμπιέστε στον ίδιο φάκελο με αυτό το αρχείο

import urllib.request, urllib.parse, urllib.error
from bs4 import BeautifulSoup

...
```

urllinks.py

```
import urllib.request, urllib.parse,
urllib.error
from bs4 import BeautifulSoup

url = input('Enter - ')
html = urllib.request.urlopen(url).read()
soup = BeautifulSoup(html, 'html.parser')

# Ανακτά όλες τις ετικέτες αγκύρωσης
tags = soup('a')
for tag in tags:
    print(tag.get('href', None))
```

python urllinks.py
Enter - http://www.dr-chuck.com/page1.htm
http://www.dr-chuck.com/page2.htm

Σύνοψη

- The TCP/IP gives us αγωγούς / υποδοχές between applications
- Σχεδιάσαμε πρωτόκολλα εφαρμογών για τη χρήση αυτών των αγωγών
- Το HyperText Transfer Protocol (HTTP) είναι ένα απλό αλλά ισχυρό πρωτόκολλο
- Η Python έχει καλή υποστήριξη για υποδοχές, HTTP και ανάλυση HTML

Ευχαριστίες / Συνεισφορές

Αρχική ανάπτυξη : Charles Severance, University of Michigan School of Information

Απόδοση στα Ελληνικά: Κιουρτίδου Δ. Κωνσταντία

… Εισαγάγετε νέους Μεταφραστές και άτομα που έχουν συνεισφέρει εδώ

Συνέχεια…