

# Analisi del traffico su un sito Internet

---

- Perché è possibile
- Che cosa possiamo sapere
- Che cosa non possiamo sapere
- Gli strumenti di analisi
- Un'analisi dettagliata

# Perché è possibile analizzare il traffico di un sito

---

- Ogni sito ha un nome che corrisponde a un numero IP
- La navigazione di un utente su un sito corrisponde a una richiesta fatta al *web server* da parte del computer dell'utente che è identificato da un IP
- Il *web server* può memorizzare quale IP lo ha interrogato e conservare traccia di altri dati caratteristici: il *log*

# Che cosa possiamo sapere

---

- IP dell'utente
  - Posso dedurre qualcosa sulla località di provenienza della richiesta
- Data e ora
- In quale pagine c'era il collegamento che ha portato al nostro sito
- Quale ricerca sui motori ha portato al nostro sito
- Quanti dati sono stati trasmessi
- Il tempo impiegato
- Altri dati tecnici

# Che cosa non possiamo sapere

---

- L'identità del visitatore
- Il posto preciso da cui naviga
- Se ha letto il contenuto della pagina su cui ha navigato
- Il numero di visitatori effettivi distinti
  - Un visitatore può usare IP diversi in momenti diversi
- Il numero di visite effettivo
  - Ogni immagine scaricata può contare come una visita
  - Un utente può usare la cache del suo browser per rivedere una pagina
  - Ci sono cache intermedie gestite dagli ISP
- Quanto tempo ha dedicato al mio sito

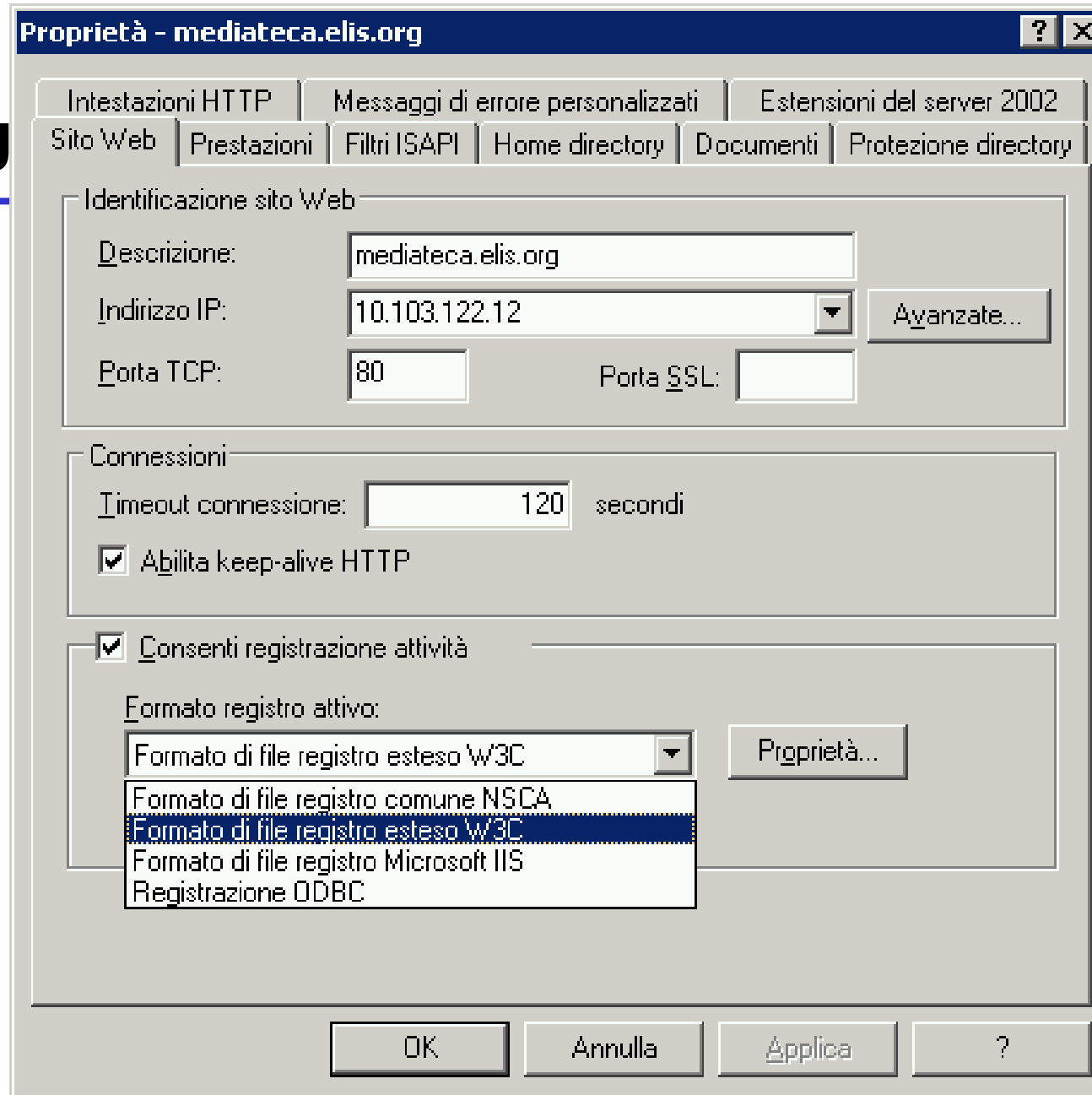
# Perché è importante l'analisi

---

- Studiare la tendenza
  - Stiamo acquisendo popolarità
- Quali sono le parole chiave che la gente cerca per trovare il nostro sito
  - Diverse per ogni motore di ricerca

# Esempio di log

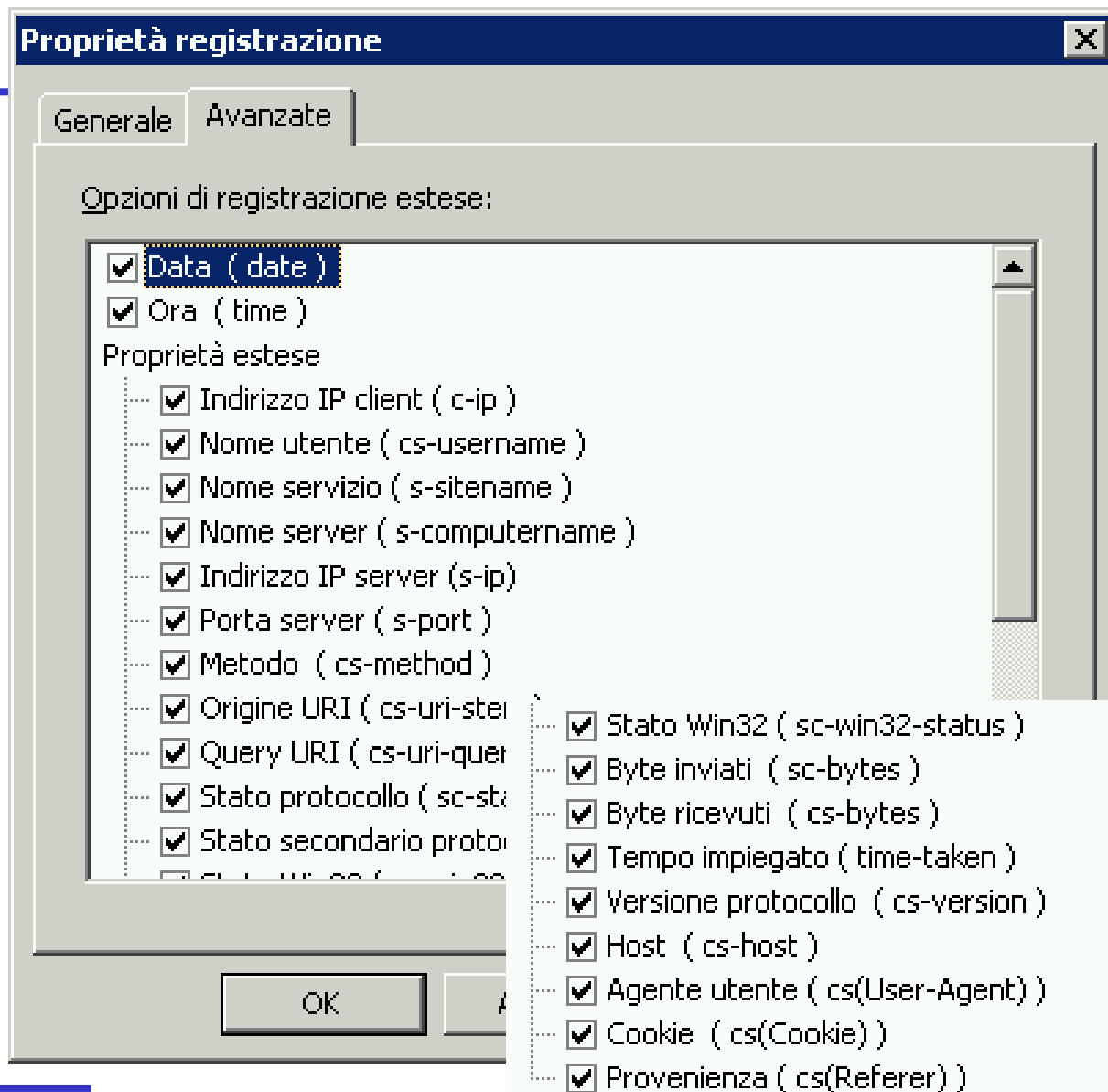
Con IIS 6 di Windows Server 2003 si può scegliere il formato di *log*



# IIS 6

Tutti i dati che si possono registrare con il *log* formato esteso W3C (con qualche modifica rispetto allo standard)

Azioni del  
c=client  
s=server  
cs=client verso il server  
sc=server verso il client



# Le definizioni dei campi registrati

---

date	La data
time	L'ora del server
c-ip	L'indirizzo IP del client che ha acceduto al server
cs-username	Il nome dell'utente (quando c'è di mezzo una autenticazione). Se è anonimo (caso più frequente) c'è un -
s-sitename	Il servizio Internet come identificato dal server
s-computername	Il nome del server
s-ip	L'indirizzo IP del server
s-port	Il numero di porta del servizio sul server
cs-method	Il metodo di azione del client (per esempio un <b>GET</b> ).
cs-uri-stem	La risorsa chiamata, per esempio la pagina HTML

# Le definizioni dei campi registrati

cs-uri-query	La query o interrogazione (quando è il caso) che il client fa al server
sc-status	La risposta del server HTTP (operazione riuscita, fallita, ecc.)
sc-win32-status	La risposta del server secondo Microsoft
sc-bytes	Il numero di byte inviati dal server
cs-bytes	Il numero di byte ricevuti dal server
time-taken	La durata dell'azione in millisecondi
cs-version	La versione del protocollo HTTP usato
cs-host	Il nome del sito web
cs(User-Agent)	Il browser usato dal client
cs(Cookie)	Il contenuto del cookie inviato o ricevuto dal server
cs(Referer)	Il sito sul quale c'è il link usato dal client per raggiungere il server

# Due record del log

---

Un utente scarica un file dalla pagina Download.asp della Mediateca-ELIS

date	31/03/2004	31/03/2004
time	22.48.30	22.49.01
s-sitename	W3SVC1159930175	W3SVC1159930175
s-computername	WEB-ELIS	WEB-ELIS
s-ip	10.103.122.12	10.103.122.12
cs-method	GET	GET
cs-uri-stem	/Download.asp	/software/Nm3.01.build3388.ita.exe
cs-uri-query	-	-
s-port	80	80
cs-username	-	-
c-ip	80.181.165.87	80.181.165.87
cs-version	HTTP/1.1	HTTP/1.1

# Due record del log

---

cs(User-Agent)	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+.NET+CLR+1.1.4322)	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+.NET+CLR+1.1.4322)
cs(Cookie)	-	ASPSESSIONIDAQDTBDAR=KGAABAGBEE NFOPDIAGKDAJAI
cs(Referer)	<a href="http://www.google.it/search?hl=it&amp;ie=UTF-8&amp;oe=UTF-8&amp;q=download+netmeeting+i+n+italiano&amp;lr=">http://www.google.it/search?hl=it&amp;ie=UTF-8&amp;oe=UTF-8&amp;q=download+netmeeting+i+n+italiano&amp;lr=</a>	<a href="http://mediateca.elis.org/Download.asp">http://mediateca.elis.org/Download.asp</a>
cs-host	mediateca.elis.org	mediateca.elis.org
sc-status	200	200
sc-substatus	0	0
sc-win32-status	0	0
sc-bytes	15250	1650764
cs-bytes	490	515
time-taken	421	25468

# Da che città navigava l'utente

- Non riusciamo a capire la città, ma abbiamo normalmente dati per sapere la nazione
- Telecom Italia può saperlo e lo rivela all'autorità giudiziaria se necessario

## Report per 80.181.165.87 [host87-165.pool80181.interbusiness.it]

Analisi: Pacchetti IP persi sulla rete "Interbusiness infrastruttura" al hop 11. Informazioni insufficienti nella cache per la determinazione della rete successiva al hop 12.

Hop	%Pers	Indirizzo IP	Nome nodo	Localazione	F.Ora	ms	Grafico	Rete
0		194.242.61.53	interhost.it	*			0 219	HostingSolutions.it
1		194.242.61.1	-	Florence, Italy	+01:00	0		HostingSolutions.it
2		62.94.32.145	-	(Italy)	+01:00	0		GENESYS INFORMAT
3		62.94.209.41	-	Arezzo, Italy	+01:00	7		Eutelia
4		62.94.31.38	a1-0-10.mi15.e	Arezzo, Italy	+01:00	15		Eutelia
5		217.29.66.35	interbusiness.	-		0		Milan Internet eXchang
6		151.99.98.225	-	(Netherlands)	+01:00	8		RIPE Network Coordin
7		151.99.75.157	r-mi258-vl3.op	Milan, Italy	+01:00	103		RIPE Network Coordin
8		151.99.98.97	r-rm213-mi258	Rome, Italy	+01:00	31		RIPE Network Coordin
9		151.99.101.19	r-fi63-rm213.oj	Florence, Italy	+01:00	22		RIPE Network Coordin
10		151.99.98.90	r-fi67-fi63.opb.	Florence, Italy	+01:00	28		RIPE Network Coordin
11		80.19.134.153	-	(Italy)	+01:00	24		Interbusiness infrastru
...								
?		80.181.165.87	host87-165.po	-				Telecom Italia S.p.A.

# Da che sito proveniva l'utente

---

- Lo deduciamo dal campo cs(Referer)
  - `www.google.it/search?hl=it&ie=UTF-8&oe=UTF-8&q=download+netmeeting+in+italiano&lr=`
- Possiamo sapere anche che cercava le parole
  - `download netmeeting in italiano`
- Referer è il dato registrato più interessante ai fini della promozione di un sito
  - Studiare gli effetti delle parole chiave che abbiamo impostato
  - Qual è il motore di ricerca dal quale provengono più richieste