RSS, PICS, CC/PP
Dichiarazione e negoziazione di contenuti sul Web

Matteo Cicognani

Indice della presentazione

- Problemi del Web attuale e classificazione autorevole dei contenuti: PICS e Medical-PICS
- Dichiarazione automatica della natura dei contenuti e del loro aggiornamento: RSS 0.9x, 1.0, 2.0
- Aspetti fondamentali di RSS, come creare un newsfeed, come utilizzare un newsfeed
- Separazione dei contenuti dalla presentazione, profilazione utente e caratteristiche dei dispositivi: CC/PP
- Il contesto di riferimento: *dichiarare e negoziare i contenuti presenti sul Web*

Problemi del Web attuale

- Il Web attuale: informazioni di qualità molto differente
  - Università, ricerca, industria, p.a.
  - Siti pornografici, pedofili, materiali inneggianti all'odio e alla violenza
  - Mezzo semplice ed immediato
- Chiunque può pubblicare qualsiasi cosa
  - Abusi e violazioni delle leggi e dell'etica
- Chiunque può leggere
  - Libero accesso a bambini e adolescenti

Classificazione autorevole dei contenuti

- La classificazione dei documenti si basa su elementi come: analisi dei metadati, analisi della presenza di link, lettura del testo contenuto tra gli elementi `<H1>`, rilevazione delle frequenze di accesso, etc. ma nessuno di questi strumenti è completamente affidabile.
- Infatti tutti questi strumenti possono essere facilmente "manipolati" dai creatori di siti per aggirare una eventuale censura o per rendere il sito più visibile, p.es. ad un motore di ricerca.
- Nel contesto della catalogazione delle informazioni sul Web, un maggior grado di affidabilità e di imparzialità è raggiungibile prevedendo la presenza di terze parti (persone o organizzazioni autorevoli) che operino una attenta, e soprattutto imparziale classificazione dei contenuti presenti sul Web in base a criteri precedentemente concordati, e del tutto indipendenti dalla volontà dei creatori dei siti.

PICS - Platform For Internet Content Selection

- Permette una classificazione dei documenti sul Web in base a criteri autorevoli stabiliti indipendentemente dagli interessi e dalla volontà dei creatori di siti.
- Nel 1996, a seguito di una inchiesta del NY Times, e a seguito della “sollevazione morale” di una parte dell’opinione pubblica americana contro la presenza sempre più ampia di materiale pornografico sul Web, il W3C reagì in modo rapido e compatto creando un modo per

Controllare senza censurare

i contenuti che gli utenti possono fruire sul Web

PICS - Platform For Internet Content Selection

- I principi fondamentali su cui PICS è nato e sui quali (fina dallo *Statement of Principles*) si è sempre basato sono:
  - **semplicità d'uso** affinché tutti (genitori e insegnanti) potessero utilizzarlo senza problemi**,**
  - **indipendenza dei creatori dei contenuti:** i creatori di siti non sono affatto vincolati dai criteri di classificazione, ma sono incoraggiati a fornire una classificazione di ciò che scrivono,
  - **indipendenza dei catalogatori dei contenuti:** i catalogatori agiscono in modo imparziale rispetto agli interessi privati dei creatori di siti
  - **indipendenza dei fruitori di contenuti:** i fruitori possono scegliere il metodo di classificazione delle informazioni più vicino al loro modo di pensare o ai loro costumi morali.

Concetti fondamentali di PICS (1)

- **Rating** (content label) cioè metainformazioni organizzate sotto forma di etichette, associate ai contenuti presenti su Internet (non solo sul Web) con lo scopo di offrire una catalogazione in base a criteri prefissati.
- **Rating Service** singole persone o organizzazioni (generalmente “trusted” da parte degli utenti finali) che si dedicano alla catalogazione delle informazioni (possono coincidere con gli autori) presenti sul Web associando le etichette.
- **Rating System** sono documenti (reperibili anche in formato HTML) che specificano e determinano le modalità e le dimensioni di catalogazione dei contenuti. Sono documenti emessi per esempio da organizzazioni come UNICEF o Simon Wiesenthal Center.

Concetti fondamentali di PICS (2)

- Fondamentalmente esistono tre meccanismi per la trasmissione delle etichette associate ad un contenuto:
  - HTTP (una semplice estensione del protocollo con metodi dalla semantica modificata per richiedere e ricevere etichette associate ai documenti)
  - Header equivalence (all'interno del tag `<meta>` della pagina HTML richiesta)
  - Label bureau (possono coincidere con i rating service) che distribuiscono su richiesta degli utenti finali le etichette associate a contenuti che risiedono su altri server

Concetti fondamentali di PICS (3)

- Poiché la presenza o meno delle etichette può decretare (almeno in via teorica) il fallimento o il successo di un sito Web, è stato necessario fornire dei meccanismi che garantissero **autenticità e integrità** delle label.
- La prima proprietà è garantita con meccanismi di firma elettronica delle etichette (per esempio tramite protocolli che utilizzano RSA).
- La seconda proprietà è garantita mediante l'utilizzo dell'algoritmo MD-5

Medical PICS

- Eysenbach e Diepgen: fornire metadati descrittivi (pubblico di riferimento) e di valutazione per classificare

  Siti Web con tenenti documenti di carattere medico–sanitario

- Usare label per filtrare le informazioni
- Sul Web è difficile distinguere i livelli di affidabilità (la variabilità della qualità delle info sul Web ne limita l'uso come fonte informativa affidabile) :
  - meno indicatori, autori anonimi, mancanza di contesto di riferimento
  - su carta era più semplice perché fenomeno più costoso e quindi più limitato alle fonti autorevoli

Dichiarazione automatica della natura dei contenuti e del loro aggiornamento (1)

- Fino a qualche anno fa la condivisione di risorse informative (contenuti) non aveva l'importanza che ha assunto oggi col fenomeno del "Content Providing"
- Questo non perché non fosse sentita la necessità di sviluppare dei sistemi che permettessero di ridurre i passaggi (e quindi i costi) dell'informazione in tutte le sue fasi di vita, dalla creazione alla pubblicazione su carta o qualsiasi altro mezzo; bensì perché
- Non era per nulla facile distribuire o scambiarsi informazioni in modo automatico, per esempio attraverso il Web e i suoi strumenti, poiché non esistevano vocabolari di metainformazioni comuni e neppure strumenti per definirli.

Dichiarazione automatica della natura dei contenuti e del loro aggiornamento (2)

- I documenti presenti sul Web sono milioni, e soprattutto variano rapidamente, risulta quindi difficile:
  - **Ricercarli** (Web Crawler e Semantic Search Engine) per ottenere le informazioni corrette e desiderate
  - **Condividerli** in modo automatico con vari scopi (uno dei quali è il syndication di informazioni)
    - ✓ **Syndication carta stampata**: distribuzione a pagamento di contenuti. In genere è effettuata da agenzie specializzate, e i principali fruitori sono le testate giornalistiche non specializzate in quel determinato campo
    - ✓ **Syndication su Web**: distribuzione automatica dei contenuti dal significato simile al syndication tradizionale, anche se la fase di pagamento deve essere realizzata con altri meccanismi.
- Spesso risulta difficile anche:
  - **Monitorare** più siti contemporaneamente se non si vuole operare un controllo diretto.

Soluzione ai problemi di Condivisione e Ricerca dei contenuti

- Per cercare di risolvere in parte questi problemi sono state sviluppate una serie di proposte, tutte molto valide, di cui le più importanti sono
  - XMLNews(META e STORY)
  - ICE – Information end Content Exchange(framework per il syndication di qualsiasi tipo di informazione sul Web, meccanismi per la sottoscrizione di contratti e formati per il delivery delle news del servizio)
  - RSS – Rich Site Summary, RDF Site Summary

  …e altre un po' meno:
  - CDF (da IE 4.0 in avanti)

RSS 0.9x, RSS 1.0, RSS 2.0 (1)

- Il termine RSS ha un significato molto vasto e comprende al suo interno almeno due formati parallelamente sviluppati.
- La versione RSS 0.90 era stata progettata da Netscape come formato per la costruzione di portali con i titoli delle notizie contenute all'interno di canali informativi.
- RSS 0.90 era una versione troppo complessa per l'epoca e così venne semplificata nella 0.91. La proposta inizialmente bene accolta perse subito di interesse a causa della perdita di interesse di Netscape nella realizzazione di portali commerciali.

RSS 0.9x, RSS 1.0, RSS 2.0 (2)

- Lo sviluppo continuò poi da una parte con Userland e dall'altra con un Working Group indipendente che voleva ristabilire i principi iniziali su cui RSS si era fondato.
- I primi continuarono uno sviluppo basato sul versioning, i secondi decisero di integrare RDF nelle specifiche di RSS, consentendo una facile estensione del protocollo per qualsiasi scopo che non fosse già contemplato nelle specifiche iniziali.
- Recentemente Userlend ha pubblicato le proprie specifiche della nuova versione di RSS, chiamata RSS 2.0

Significato di RSS

- L'acronimo RSS nella sua accezione originale stava per

  RICH SITE SUMMARY, richiamando nel nome lo scopo per cui era nato: la descrizione dei contenuti dei portali sviluppati da Netscape. Ad ogni aggiunta di tag e costrutti veniva rilasciata una nuova versione
- RSS sta anche per

  RDF SITE SUMMARY, e nel nome richiama quindi l'adozione dello standard RDF del W3C per la descrizione delle metainformazioni, contenute appunto nei file RSS. Le nuove aggiunte sono gestite con l'introduzione di moduli dedicati ciascuno ad argomenti specifici.

La famiglia RSS

Versione di RSS	Proprietario	Caratteristiche
0.90	Netscape	Reso Obsoleto dalla versione 0.91
0.91	Userland	Ufficialmente rimpiazzato da 2.0 ma ancora popolare, ha le funzioni di base per il syndication di informazioni
0.92, 0.93, 0.94	Userland	Descrizioni più ricche di metadati
1.0	RSS-DEV Working Group	Basato su RDF, estensibile con moduli e non controllato da un solo venditore
2.0	Userland	Estendibile con moduli, per syndication avanzato

Caratteristiche di RSS (1)

Le idee fondamentali su cui si basa RSS, sono:

- lo sviluppo di un vocabolario **comune estendibile** di metainformazioni per descrivere e condividere particolari tipi di contenuti
- l'associazione di un canale (Newsfeed o Channel) di informazioni a ogni sito Web, cioè un insieme di file RSS (uno o più) che in qualche modo descrivano il contenuto informativo del sito stesso, o di ciascuna pagina che lo compone (per esempio con un titolo, una breve descrizione, un link, o campi per il syndication)

Caratteristiche di RSS (2)

- Per garantire l'elaborazione automatica delle informazioni è necessario definire un vocabolario comune di metainformazioni da associare ai contenuti veri e propri.
- I vocabolari sono sviluppati in base alle esigenze e agli interessi dei gruppi che promuovono la standardizzazione, p.es. le agenzie stampa che offrono contenuti a pagamento alle testate informative sarebbero interessate allo sviluppo di vocabolari per il **Syndication** di informazioni

Struttura di un file RSS

- Ogni file RSS è un semplice documento XML,e al suo interno ha informazioni

  statiche

  (notizie sul sito di origine, come la data di creazione delle notizie, gli autori delle notizie, etc.)

  e

  dinamiche

  (le news vere e proprie con il loro contenuto, oppure i loro abstract, oppure un “summary” delle news)

RSS 0.9x: un esempio

```
<?xml version="1.0"?>
<rss version="0.91">
<channel>
  <title>Example Channel</title>
  <link>http://example.com/</link>
  <description>My example channel</description>
</channel>
<item>News for September the Second</item>
  <link>http://example.com/2002/09/02</link>
  <description>Things happened today</description>
</item>
</rss>
```

RSS 0.9x (1) vedi documenti originali

- **Description**: a plain text description of an item, channel, image or text input
- **Language**: it specifies the language of a channel
- **Link**: an url that a user is expected to click on, as opposed to a <url> that is for loading a resource, such as an image.
- **Title**: an identifying string for a resource. When used in an item, this is the name of the item's link. When used in an image, this is the Alt text for the image. When used in a channel, this is the channel's title. When used in a text input, this is the text input's title.

Optional metadata:

- **Image**: specifies an image associated with a channel
- **Item**: it describes an item that is associated with a channel. The item should represent a web page, or subsection within a web page. It should have a unique URL associated with it. Each item must contain a title and a link.
- **Last Build Date**: It indicates the last date the channel was modified
- **Pub Date**: it indicates the date of publication of the channel
- **Rating**: it can contain a recommended links to rating agencies, user actions (obtain a rating for your site from a well known rating agency; copy rating data into RSS file), and expected format

RSS 0.9x (2) vedi documenti originali

RSS 0.9x takes a versioned approach to extensibility; new features are added by declaring a new version of RSS in the 0.9 series. RDF references were removed.

In particular in the specifications for RDF 0.91 there is a note which explains that RDF references has been removed. In fact RSS was originally conceived as a metadata format providing a summary of a website. Two things have become clear: the first is that providers want more of a syndication format than a metadata format.

The structure of an RDF file is very precise and must conform to the RDF data model in order to be valid. This is not easily human understandable and can make it difficult to create useful RDF files. The second is that few tools are available for RDF generation, validation and processing. For these reasons, we have decided to go with a standard XML approach

RSS1.0: un esempio

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns="http://purl.org/rss/1.0/"
   xmlns:dc="http://purl.org/dc/elements/1.1/" >
<channel rdf:about="http://example.com/news.rss">
   <title>Example Channel</title>
   <link>http://example.com/</link>
   <description>My example channel</description>
   <items>
     <rdf:Seq>
       <rdf:li resource="http://example.com/2002/09/01/"/>
       <rdf:li resource="http://example.com/2002/09/02/"/>
     </rdf:Seq>
   </items>
</channel>
</rdf:RDF>
```

RSS 1.0 (1) vedi documenti originali

- **Channel**: the channel element contains metadata describing the channel itself, including a title, a brief description, and URL link to the described resource. The {resource}URL of the channel element's rdf:about attribute must be unique with respect to any other rdf:about attributes in the RSS document and is a URI which identifies the channel. Most commonly, this is either the URL of the homepage being described or a URL where the RSS file can be found.
- **title**: it indicates a descriptive title for the channel.
- **link**: it is the URL to which an HTML rendering of the channel title will link, commonly the parent site's home or news page.
- **description**: a brief description of the channel's content, function, source, etc.
- **Image**: it establishes an RDF association between the optional image element and this particular RSS channel. The RDF resource's {image_uri} must be the same as the image element" rdf:about{image_uri}
- **Items**: it is a table of contents, associating the document's item with this particular RSS channel. Each item's rdf:resource{item_uri} must be the same as the associated item element's rdf:about{item_uri}. An RDF Seq(sequence) is used to contain all the items rather than an RDF bag to denote item order for rendering and reconstruction

RSS 1.0 (2) vedi documenti originali

- **Item**: while commonly a news headline, item can just be anything: discussion posting, job listing, software patch, any object with an URI. There may be a minimum of one item per RSS document.
  - **Image**: it indicates an image to be associated with an HTML rendering of the channel. This image should be of the format supported by the majority of web browsers.
  - **Title**: the alternative text associated with a channel's image tag when rendered as HTML
  - **Url**: The url of the image to be used in the "src" attribute of the channel's image tag when rendered as HTML
  - **Link**: the url to which an HTML rendering of the channel image will link. This, as with the channel's title link, is commonly the parent site's home page.

Creare un NewsFeed(1)

Un canale informativo RSS può essere creato in diversi modi e con diversi risultati:

- **Self Scraping**: usare strumenti automatici per estrarre dalla pagina Web le parti rilevanti (si può ricorrere ad espressioni XPath o all'inserimento di markup minimale da aggiungere ai contenuti come hint per il software)
- **Gestione Integrata dei canali**: se la pagina è generata dinamicamente usando un linguaggio di script come Perl, o ASP, potrebbero esistere funzioni di libreria per integrare i canali all'interno del processo di pubblicazione.

Creare un NewsFeed(2)

- **Cominciare con il canale**: creare prima il canale e in un secondo momento generare le pagine Web (con strumenti tipo XSLT). Questo metodo garantisce canali sempre aggiornati sui contenuti effettivi del sito.
- **Scraping eseguito da terze parti**: non è mai affidabile, perché le terze parti (a pagamento) esaminano il contenuto senza conoscerlo a fondo.

Mettere a disposizione un canale (1)

- Alcuni programmi potrebbero cercare un link nella sezione head di una pagina scritta in linguaggio HTML, p.e.:

```
<head>
  <title>My Page</title>
  <link rel="alternate"
        type="application/rss+xml"
        href="feed.rss"
        title="RSS feed for my page">
</head>
```

Mettere a disposizione un canale (2)

- Sul Web esistono moltissimi canali RSS a disposizione degli utenti ma spesso è difficile trovarli e questo ne ostacola l'utilizzo.
- Le pagine Web che hanno associato un canale informativo RSS potrebbero indicarlo chiaramente con un link a quella risorsa di modo che chiunque desideri leggere direttamente i contenuti del canale RSS posso farlo attraverso un browser, p.es.

```
<a type="application/rss+xml"href="feed.rss">
      RSS Feed for this page
</a>
```

Fruire i canali RSS: gli Aggregators(1)

- Gli aggregators sono programmi software in grado di utilizzare i canali RSS messi a disposizione da un sito Web.
- Possono leggere le notizie in modo automatico senza obbligare l'utente a navigare il sito; sono in grado di trovare le novità o informazioni rispondenti agli interessi dell'utente; infine sono in grado di visualizzare le informazioni trovate secondo le necessità o le preferenze dell'utente.
- Gli aggregators esistono integrati nei browser, nei client di posta elettronica, nei desktop, o sono prodotti software standalone.

Fruire i canali RSS: gli Aggregators(2)

- Tra le loro capacità di visualizzazione vi è quella di ordinare temporalmente le informazioni da visualizzare, o di farle scorrere in fondo alla pagina sullo stile “breaking news”
- Un tempo gli aggregators erano grandi servizi centralizzati. Userland ha segnato una svolta proponendoli come software installabili su PC (Radio e la metafora delle stazioni radiofoniche)

Come dovete usare RSS

- Realizzare un canale per il sito e per tutte le pagine del sito
- Il WG deve stabilire
  - Criteri di accesso (es. URL standard)
  - Versione di RSS
  - Estensioni al vocabolario RSS scelto, e in particolare
  - Come esprimere il tesauro in RSS
  - Come specificare il match tra le metainformazioni di un documento e i summary in RSS dei vari canali
  - Accesso al source dei documenti XML
- I team dovrebbero
  - Realizzare il file RSS secondo i criteri del WG
  - Ogni volta che viene richiesto un documento ***X*** cercare negli RSS della concorrenza i summary che fanno match
  - In base al livello di match inserire un link oppure tutto l'abstract, oppure tutto il contenuto

Riferimenti

“RSS Tutorial for Content Publishers and Webmasters” 1999, http://ww.mnot.net/rss/tutorial

Libby D., “RSS 0.91 Specification Revision 3”, 1999, http://www.purl.org/

RSS 2.0 e altro materiale utile come aggregators, utilities, Validator, e specifiche. http://backend.userland.com/rss

RDF Site Summary (RSS) 1.0 http://purl.org/rss/1.0/spec

Riferimenti per PICS

Miller J., “PICS Statement of Principles”, 1997. http://www.w3.org/PICS/principles.html

Resnick P., Miller J., “PICS: Internet Access Control without Censorship”, *Communication Of the ACM,* 39(10), 1996, 87-93

Miller J., “Rating services and rating systems (and their machine readable description)”, W3C Reccomendation, 1996, http://www.w3.org/TR/REC-PICS_Services

CC/PP

Matteo Cicognani

Il Web aperto a tutti

"The power of the Web is in its universality. Access by everyone regardless of disability is an essential aspect."

Tim Berners-Lee

- Terminali braille, sintetizzatori vocali, etc.
- Dispositivi “non standard” con capacità di browsing, utili alle persone quando si spostano, per mantenersi informati

Avvento di nuovi dispositivi e fruizione dei contenuti\

- Nel 2002 i **dispositivi non standard** con capacità di browsing sono stati 1.3 miliardi rispetto ai 700 milioni di dispositivi classici

Hjelm,**Suryanarayana**

“CC/PP for content negotiation and contextualization

- PDA, telefoni cellulari, etc. hanno caratteristiche particolari:
  - I meccanismi forniti da HTTP 1.1 da soli non sono più sufficienti per descrivere accuratamente le caratteristiche di dispositivi sofisticati,e quindi per consentire Content Adaptation.
    - ✓ **User Agent** req-header field: consente di riconoscere in modo automatico gli user agent per favorire la creazione di risposte appropriate, e per venire incontro ai limiti di certi dispositivi.
    - ✓ E più in generale meccanismi di **Content Negotiation**

Il Content Adaptation: separare contenuto e presentazione

- **Due fasi distinte** nel processo di pubblicazione
  - Creazione dei contenuti e loro strutturazione (XML + DTD o Schema)
  - Creazione delle presentazioni (XSLT, CSS1, CSS2)
- **Riutilizzare** i contenuti per più scopi
  - Presentazione su media multipli (es. CSS2: video, carta, audio)
- HTML è stato molto importante per il successo del Web, ma col passare del tempo è diventato un'arma a doppio taglio: favorisce lo scambio e la pubblicazione ma crea appiattimento semantico

CC/PP Composite Capabilities/Preference Profiles (W3C)

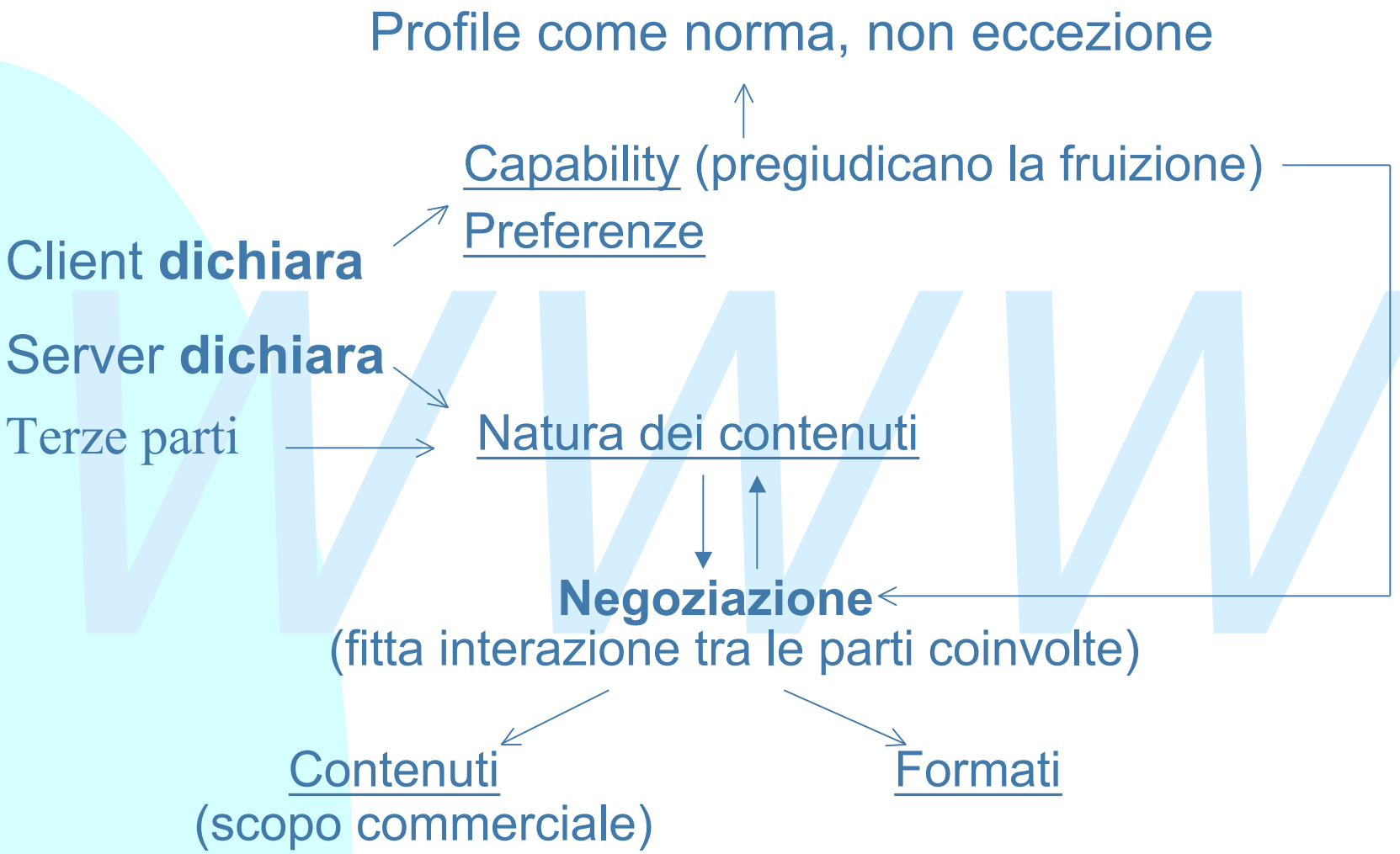
- Descrive le caratteristiche di ogni tipo di dispositivo (standard e non), attraverso dei **profile,** utilizzabili per creare presentazioni “ad hoc” e per consentire la fruizione dei contenuti:
  - Piattaforma Hardware
  - Sistema Operativo
  - Tipi di browser

  Coppie attributo-valore.
- Usa RDF.
- E’ verboso, ma si possono usare dei riferimenti a “repository” di profili (anziché inviare il profile per il proprio dispositivo all’interno di ogni singola richiesta, si inviano solo riferimenti a repository all’interno dei quali sono presenti le descrizioni complete delle caratteristiche dei dispositivi)

CC/PP: un esempio

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:ccpp="http://www.w3.org/2002/11/08-ccpp#"
        xmlns:ex="http://www.exa.com/schema#">
<rdf:Description
   rdf:about="http://www.example.com/profile#MyProfile">
   <ccpp:component>
   <rdf:Description
rdf:about="http://www.example.com/profile#TerminalHardware">
<rdf:type rdf:resource="http://www.exa.com/schema#HwPlatform">
        <ex:displayWidth>320</ex:displayWidth>
   <ex:displayHeight>200</ex:displayHeight>
</rdf:Description>
</ccpp:component>
```

Conclusioni



Riferimenti

CC/PP Working Group, “Composite Capabilities/

Preference Profiles(CC/PP): Structure and Vocabularies, 2001, W3C Working Draft

http//www.w3.org/TR/2001/WD-CCPP-struct-vocab-20010315/

Suryanarayana L., Hjelm J. “CC/PP for Content Negotiation And Contextualization”, LNCS 1987, 2001, 239-245