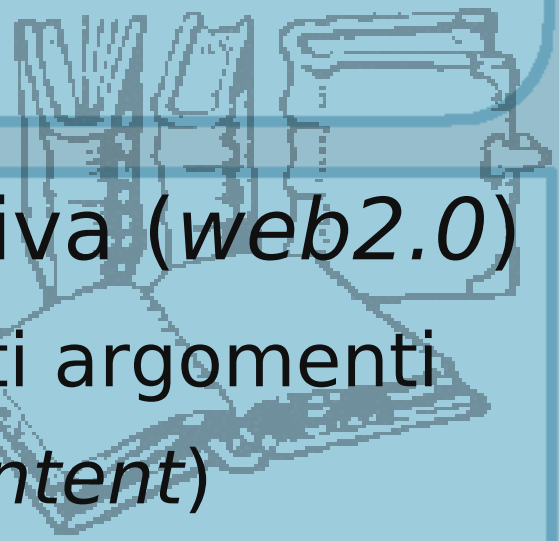
Sviluppo di un ambiente software per la consultazione offline di Wikipedia

Relatori: Prof. Marco Mezzalama
Prof. Juan Carlos De Martin
Ing. Alessandro Ugo
Candidato: Emanuele Richiardone

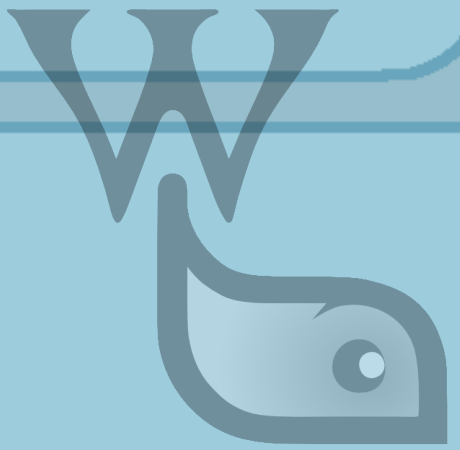
Novembre 2007
Laurea Magistrale in Ingegneria Informatica
III Facoltà di Ingegneria - Politecnico di Torino

Wikipedia



- Vasta enciclopedia collaborativa (*web2.0*)
  - Multilingue, buona qualità, molti argomenti
  - Fonte libera di sapere (*open content*)
  - Crescita più che lineare
- *wiki* – *wikitext* – MediaWiki
  - PHP e MySQL, estensioni
  - Voci: normali/redirect/categorie/template
  - Non solo testo
  - Licenze d'uso libere

Progetto WaNDA



- Accesso offline ai contenuti:
  - Digital divide
  - Educational
  - Enciclopedia tascabile
  - Open content
- Wikipedia è strutturata per l'accesso online
- Progetto WaNDA (linux@studenti)
- Studio di meccanismi e implementazioni

Caratteristiche

- Definizione di un processo di costruzione:
  - Mantenibile nel tempo
  - Utilizzo di software opensource
- Definizione di un formato:
  - Contenente testi enciclopedici e immagini
  - Per supporto eterogeneo “hot plug” (DVD, schede di memoria, pendrive, microdrive, ...) e non
  - Accessibile da generica postazione (multipiattaforma)
  - Consultazione intuitiva senza richiedere installazione di software apposito

Formato del supporto

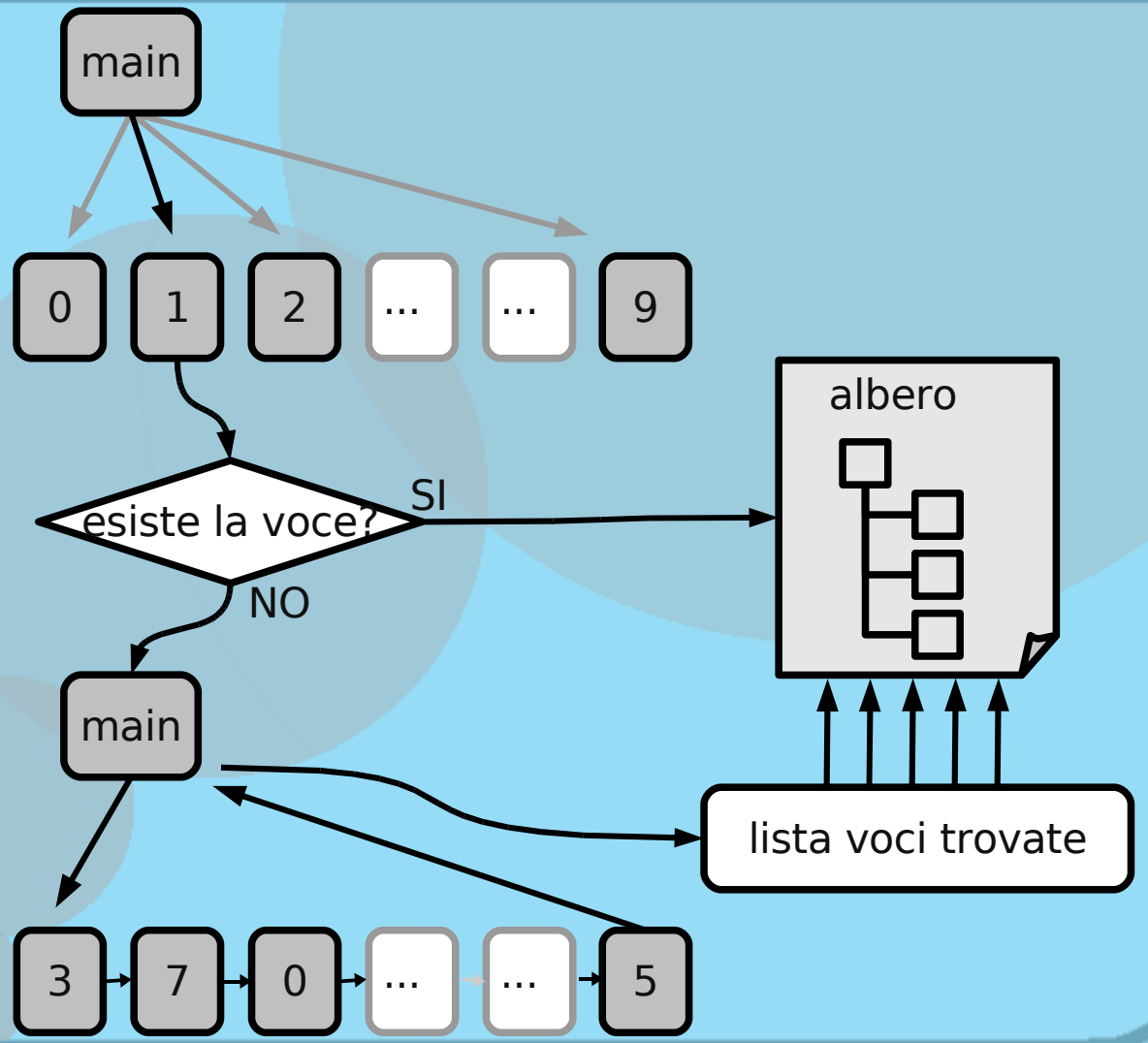
- Gerarchia di directory, detto *albero*:
  - *Ossatura* comune alle diverse versioni:
    - componenti statiche (HTML, CSS, ...)
    - componenti dinamiche (JavaScript, GET HTTP)
  - Voci normali, categorie e redirect costruite dal processo sotto forma di pagine HTML
  - Immagini immesse dagli utenti e generate
- Impacchettamento per la distribuzione:
  - Immagine ISO-9660:1999
  - Archivio per FAT32

Componenti dinamiche

- Impiego di tecnologie standard e diffuse:
  - JavaScript (*ECMAscript* v3)
  - *Percent-encoding*, *GET HTTP*
- Redirezione delle pagine
- Motore di ricerca offline
  - Basato su vettore associativo frammentato
  - Ricerca *esatta*, ricerca *estesa*
  - Ricerca completa
  - Estrazione casuale

Ricerca completa

Ricerca esatta

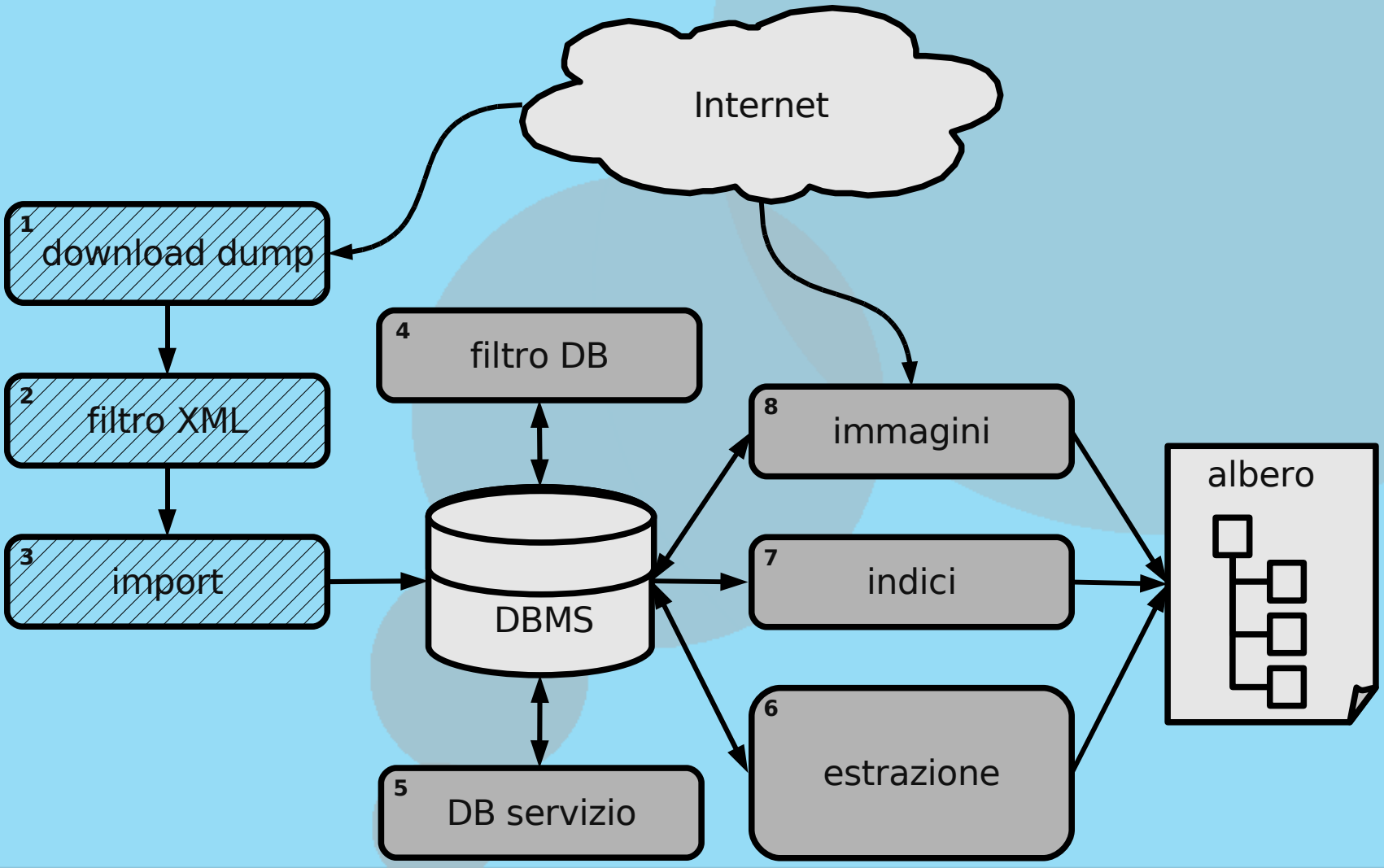


Ricerca estesa

Processo di conversione

- Software opensource e multipiattaforma
- Piattaforma x86 con UNIX derivato (massima compatibilità con Wikipedia)
- DBMS MySQL, PHP-CLI, JVM, mawk
- Due gruppi di operazioni:
  - Download dei contenuti e importazione nel DBMS
  - Esportazione del *wikitext* e costruzione dell'*albero*

Processo di conversione



Prima fase

- Contenuti enciclopedici testuali disponibili sotto forma di dump XML compressi, aggiornati ogni 1-2 mesi
  - Download dei dump
  - Decompressione e filtro AWK, per ridurre la mole di dati da gestire di un fattore 1:20 (~20 ore)
  - Traduzione dell'XML in comandi SQL ed importazione nel DBMS locale

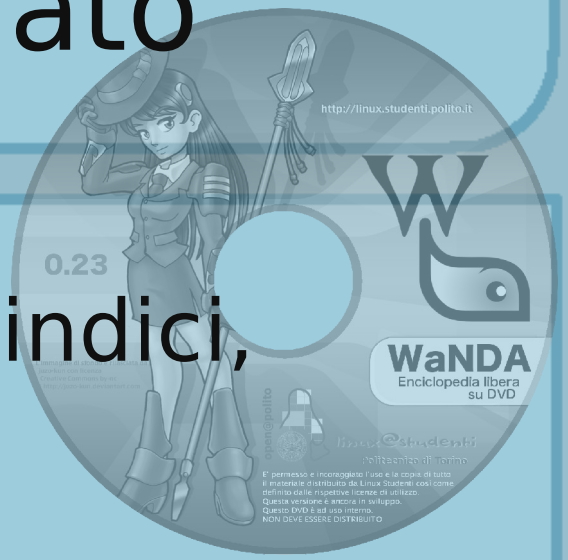
Seconda fase

- Utilizza un'installazione di MediaWiki con estensioni e il DBMS
- Totalmente gestita da un codice PHP:
  - Filtri SQL sui contenuti (voci segnate, pagine non enciclopediche, case-insensitive)
  - Creazione di una tabella di servizio
  - Estrazione delle voci (normali, categorie, redirect) e traduzione del *wikitext* utilizzando MediaWiki, con conseguente scrittura dei file HTML nell'albero (~40 ore)

Seconda fase

  - Lista delle pagine e vettori per la ricerca
  - Integrazione delle immagini utente secondo licenza e di quelle generate dal wikitext
  - Impacchettamento nei due formati
- Le due fasi in totale durano alcuni giorni
- Le due operazioni più lente sono state ottimizzate
- Banda di accesso alla memoria secondaria, cache del processore

Controllo del formato



- Verifica automatica in PHP di indici, collegamenti e immagini
- Comunità di testing
  - Più proficua data la natura degli eventuali problemi
- Modifiche ai contenuti vanno eseguite sul DBMS prima della seconda fase
- Inesattezze implicite del sistema

Stato del progetto

- Obiettivi tecnologici raggiunti:
  - Buona qualità del formato
  - Soluzione mantenibile
  - Supporto testato su diverse piattaforme
- Applicabile ad altri contenuti:
  - Wikimedia (altre lingue, Wiktionary, ...)
  - Qualsiasi installazione di MediaWiki per ottenere versioni offline di un *wiki*
- Software rilasciato sotto GNU GPLv2

Ostacoli legali

- Diffusione online dei pacchetti
  - Coperta da limitazione di responsabilità se il soggetto è Wikimedia Foundation, altri?
  - D.Lgs 70/2003 (commercio elettronico)
- Distribuzione fisica di supporti
  - Editoria? (legge 47/1948)
- Ambito del progetto non proprio contemplato: licenze d'uso libere con nuove tecnologie

Conclusioni

- Ottimo progetto di diffusione dell'open content e dell'opensource
- Innovative soluzioni multipiattaforma
- Pesanti ostacoli giuridici sulle licenze e alla diffusione open content
- Futuro del progetto:
  - Miglioramenti?
  - Scalabilità del supporto