POLITECNICO DI TORINO

III Facoltà di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Informatica

Tesi di Laurea Magistrale

Sviluppo di un ambiente software per la consultazione offline di Wikipedia

Relatori:
Prof. Marco Mezzalama
Prof. Juan Carlos De Martin
Ing. Alessandro Ugo (co-relatore esterno)

Candidato:
Emanuele Richiardone

Novembre 2007

Sommario

Uno fra i progetti Internet di informazione collaborativa nata all'inizio del "web 2.0" sicuramente più utili è *Wikipedia*, una vasta enciclopedia *libera* con contenuti multilingue inseribili e accessibili da chiunque. Wikipedia è oggi una fonte di sapere *open content* nota e apprezzata in ogni ambiente. Le voci enciclopediche, caratterizzate da una buona qualità e una sorprendente vastità di argomenti, sono suddivise in diversi settori a seconda della lingua; la parte italiana si situa fra prime dieci lingue per numero di voci.

La consultazione e la modifica delle voci enciclopediche è resa possibile grazie ad un software *wiki* di nome MediaWiki, che offre a chiunque un facile e intuitivo mezzo per la creazione e la modifica dei contenuti; il sito è organizzato come una collezione illimitata di pagine HTML, ognuna contenente un voce enciclopedica oppure un contenuto multimediale. L'utente controlla la formattazione del testo grazie ad un intuitivo linguaggio indicato con il termine *wikitext*. Il software MediaWiki è scritto in linguaggio PHP e memorizza le pagine in un database MySQL, e spesso amplia il linguaggio wikitext con delle estensioni. Le pagine possono anche contenere un rimando ad un'altra pagina (pagina di redirect), un raccoglitore di collegamenti (categoria) oppure una parte di testo da includere in altre voci (template); per ognuna è inoltre possibile estrarre facilmente ogni singola versione precedente con l'indicazione dell'autore.

Wikipedia è gestita dalla Wikimedia Foundation, che si occupa di molti altri progetti collaborativi; essa delega la gestione delle versioni linguistiche di Wikipedia ad associazioni locali, che raccolgono amministratori volontari; questi si dedicano al controllo e pulizia di eventuali testi spuri, oltre a bloccare l'accesso anonimo a voci che vengono incessantemente devastate. Tutti progetti collaborativi di Wikimedia sono memorizzati in server collocati in Florida, USA.

Qualsiasi testo inserito viene rilasciato sotto un licenza d'uso permissiva di nome GNU FDL; essa garantisce agli altri utenti la libertà di diffondere e di apportare modificare, a patto di riportare gli autori precedenti e di redistribuire i contenuti derivati con la medesima licenza d'uso. Per i contenuti non testuali l'autore indica la licenza d'uso che può essere FDL o più lasca, di dominio pubblico, oppure soggetta a diritti d'autore esclusivi secondo alcune condizioni.

Per esigenze di backup e per seguire le indicazioni della licenza FDL, i contenuti del database di Wikipedia sono a disposizione di tutti, presenti sotto forma di dump XML.

Il progetto WaNDA

La consultazione dell'enciclopedia è subordinata alla disponibilità di un accesso ad Internet e i tempi di risposta hanno latenze spesso elevate a causa della bassa diffusione della banda larga e dei lenti sistemi di accesso ai dati. Poter consultare Wikipedia da qualsiasi postazione anche non connessa, riduce il "digital divide" e propone uno strumento prezioso per la didattica. Disporre localmente dei contenuti di Wikipedia permette inoltre di possedere uno supporto multimediale molto comodo. La struttura dell'enciclopedia è ottimizzata per l'accesso online e necessita quindi di una conversione per la consultazione offline dei contenuti. Lo scopo del progetto *WaNDA* è proprio quello di definire ed implementare dei meccanismi per la conversione del formato dei contenuti della parte italiana di Wikipedia, e una struttura offline per la memorizzazione su supporti eterogenei e l'accesso da generiche postazioni.

Il progetto rispetta i seguenti requisiti: il supporto offline è "hot plug", utilizzando tecnologie quali i dischi ottici, le schede di memoria e chiavi USB, microdrive e dischi portabili; il formato del supporto è multi-piattaforma, permettendo di essere consultato

da calcolatori di vario tipo, inclusi palmari e smartphone; il supporto contiene testo e le immagini con licenze libere che ne permettono la redistribuzione; il processo di conversione ed il formato del supporto sono mantenibili nel tempo; i software implicati nella conversione e nella consultazione sono opensource e di libero utilizzo; la consultazione è facile ed immediata, senza richiedere l'installazione di software in modo da facilitare il requisito di genericità; il formato finale è indipendente dal supporto di memorizzazione. Il software implementato, sia esso utilizzato per la conversione che per la consultazione, è rilasciato sotto licenze opensource.

Supporto per la consultazione

Il supporto contiene i dati organizzati in modo tale da non richiedere particolari applicativi per la consultazione; il suo formato è infatti composto da una gerarchia di directory. Una struttura base, contenente le componenti fisse per ogni versione del formato quali alcune pagine di servizio ed un motore di ricerca offline, viene staticamente distribuita assieme al software di conversione dei contenuti. La soluzione è valida per molti sistemi dato che la presenza di un browser web è in genere garantita.

I contenuti sono organizzati in due formati differenti. Il primo è un'immagine ISO-9660:1999 per dischi ottici, uno standard affermato che fornisce la convenzione per i nomi dei file ideale. Per gli altri supporti il filesystem scelto è FAT32, molto diffuso a causa delle sua semplicità implementativa e meno restrittivo rispetto all'ISO-9660. Tutti i file dell'albero sono composti da lettere minuscole della codifica ASCII a 7 bit, permettendo di essere compatibile con il maggior numero di piattaforme possibili, che siano case-sensitive oppure case-insensitive. Tuttavia spesso sono presenti caratteri "estesi", rappresentati con una codifica interpretata nativamente dai browser di nome *percent-encoding*; tale codifica viene utilizzata non solo per i nomi dei file ma anche per i collegamenti e per il motore di ricerca.

La presentazione dei contenuti è effettuata grazie a componenti statiche, composte dalle pagine enciclopediche e dalle pagine di servizio, formattate con HTML e CSS, e dinamiche, ovvero un motore di ricerca offline in JavaScript. È stata inoltre prevista l'incorporazione di un browser opensource per la consultazione dell'enciclopedia sulle postazioni proprietarie.

Il motore di ricerca offline, implementato in JavaScript utilizzando funzioni e oggetti presenti su tutte le implementazioni, permette la ricerca di una stringa di testo sull'insieme dei titoli delle voci e dei redirect; si basa sull'estrazione di chiavi da un vettore associativo. Effettuare la ricerca implica vari passaggi di dati da una pagina HTML ad un'altra, compiuti utilizzando delle GET HTTP. Possono essere compiute quattro operazioni: una ricerca "esatta" che determina velocemente se la stringa cercata è presente come titolo e restituisce l'eventuale pagina trovata; una ricerca "estesa", che cerca un riscontro all'interno di tutte le stringhe presentando una barra di avanzamento e concludendo con una lista di risultati; una ricerca "completa" (quella di solito utilizzata) che prima effettua una ricerca esatta e se fallisce passa a quella estesa; una semplice e veloce estrazione casuale di una voce.

Processo di conversione

Il processo di conversione parte dai contenuti di Wikipedia tratti dai dump XML pubblicamente rilasciati; è un'operazione computazionalmente pesante che dura molte ore a causa della grande quantità di dati da elaborare. La piattaforma su cui eseguire il processo è stata scelta considerando l'ambito opensource del progetto scegliendo quindi un sistema operativo di derivazione UNIX, che inoltre offre la massima compatibilità software con i server di Wikipedia.

La conversione si compone di due fasi. La prima fase prevede la replicazione in un database locale dei contenuti prettamente enciclopedici della parte italiana di Wikipedia. Questa fase si occupa di decomprimere e filtrare sommariamente i contenuti inutili dei dump XML, utilizzando un programma AWK apposito, allo scopo di rendere più gestibili le successive elaborazioni; tale operazione, che dura circa 20 ore, riduce di un fattore 1:20 la dimensione dei dump. Segue una traduzione dell'XML in comandi SQL utilizzando un programma Java, per poi poter importare i contenuti nel database MySQL locale.

La seconda fase estrae dal database e scrive sul formato finale le voci enciclopediche, oltre ad integrare i contenuti grafici; questa fase viene interamente gestita da un programma PHP sviluppato, che utilizza un'installazione di MediaWiki per garantire la piena compatibilità nel tradurre il wikitext. La prima operazione effettuata è l'esecuzione di alcuni filtri sotto forma di query SQL, per eliminare con maggiore precisione i contenuti indesiderati. Eventuali modifiche ai contenuti vanno infatti effettuate nel database, poiché diventano definitive nel formato finale essendo il wikitext elaborato secondo le informazioni qui disponibili. Sono quindi estratte tutte le voci enciclopediche traducendo il wikitext in pagine HTML, le pagine di redirect e le pagine delle categorie; le pagine sono scritte nella struttura finale integrandosi alla struttura base contenente le componenti fisse. Alla fine di questa conversione, che dura circa 40 ore, viene stilata una lista delle voci esportate per costruire un elenco di tutte le voci e riempire il vettore associativo per la ricerca. È ancora necessario elaborare le immagini, che secondo la licenza vengono scaricate oppure copiate localmente se sono generate dalle estensioni di MediaWiki. Infine si conclude con l'impacchettamento del formato finale, che si aggira sui 3.5 GB, nei due formati presentati.

Conclusione

Il controllo della validità del formato finale può essere effettuato con un programma PHP, che verifica i collegamenti tra pagine e alle immagini, oppure con l'aiuto di una comunità di persone, più proficua data la tipologia di problemi relativi principalmente alla presentazione o ai contenuti.

Tra gli ampliamenti futuri del progetto potrebbe aver senso lo sviluppo di un motore di ricerca avanzato meno universale da affiancare a quello JavaScript, implementato in un linguaggio di programmazione completo, come per esempio in Java. Inoltre il progetto WaNDA definisce un percorso applicabile non solo a Wikipedia Italia, ma a qualsiasi altro progetto di Wikimedia; con poche modifiche il percorso può essere applicato a qualsiasi installazione di MediaWiki, come ad esempio un wiki accademico da cui potrebbero essere estratte delle versioni offline.

Lo sviluppo di un procedimento per la conversione, dopo aver implementato diverse soluzioni, è infine giunto a buoni risultati: il meccanismo attuale ha raggiunto gli obiettivi prefissati, i risultati del processo sono buoni e la soluzione è mantenibile nel tempo.

Lo studio di meccanismi per la diffusione del prodotto, che ha definito i due formati finali, non ha ancora avuto la sua realizzazione. Ciò è dovuto a molti fattori esterni al progetto, tra i quali l'attuale legislazione italiana, che non contempla le licenze d'uso libere. La distribuzione del supporto fisico ricade nell'editoria che non inquadra le moderne tecnologie di diffusione; la distribuzione online potrebbe ricadere nella normativa sul commercio elettronico se chi pubblica fosse Wikimedia Foundation; un soggetto differente potrebbe invece incorrere in pesanti responsabilità civili e forse penali. Il progetto ha però buone speranze, essendo il software finito, pubblico ed applicabile ad altri contesti.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

Nella società dell'informazione odierna sono progressivamente sorti un numero crescente di servizi web che offrono notevoli capacità espressive, facilitando la comunicazione e la collaborazione degli utenti di Internet e la diffusione di contenuti. Questo fenomeno, reso in gran parte possibile dallo sviluppo di tecnologie innovative, rappresenta una seconda generazione di Internet, comunemente indicata con l'espressione "web 2.0".

Uno dei progetti di informazione collaborativa più riusciti è sicuramente *Wikipedia*, una vasta enciclopedia *libera* con contenuti multilingue inseribili e accessibili da qualsiasi utente della rete. Wikipedia è oggi una fonte di sapere *open content* nota e apprezzata in ogni ambiente, tanto che i suoi accessi superano quelli dell'autorevole "Encyclopaedia Britannica". La consultazione e la modifica delle voci enciclopediche è resa possibile grazie ad un software *wiki*, che offre un facile ed intuitivo mezzo per la creazione e la modifica dei contenuti. Le voci, caratterizzate da una buona qualità e una sorprendente vastità di argomenti, sono suddivise in diversi settori a seconda della lingua; la parte italiana si situa fra prime dieci lingue per numero di argomenti.

La consultazione dell'enciclopedia è subordinata alla disponibilità di un accesso ad Internet e i tempi di risposta hanno latenze spesso elevate a causa della relativamente bassa diffusione della banda larga e dei lenti sistemi di accesso ai dati. Poter consultare Wikipedia da qualsiasi postazione anche non connessa, riduce il "digital divide" e propone uno strumento prezioso per la didattica.

Per superare i vincoli citati è quindi necessario disporre localmente dei contenuti di Wikipedia, permettendo inoltre di possedere un supporto multimediale molto comodo. Allo scopo di trasferire i contenuti enciclopedici online su un supporto fisico è nato il progetto *WaNDA*; esso definisce ed implementa dei meccanismi per la conversione del formato dei contenuti in modo che siano memorizzabili ed accessibili tramite DVD, memorie solide, dischi portatili o dispositivi mobili. Il progetto si è occupato della parte italiana, anche se il suo utilizzo può essere esteso alle altre lingue e ad altri progetti di tipo wiki.

Il progetto ha prodotto un'implementazione software per compiere la conversione, composta da varie tecnologie (PHP, AWK, Java, scripting, SQL, XML) per effettuare diverse fasi necessarie all'operazione. Inoltre è stato anche curato il risultato della conversione,

occupandosi della gestione, presentazione ed accessibilità dei contenuti per l'utente finale; fra le diverse funzionalità è stato implementato un motore di ricerca offline, toccando un'altra serie di tecnologie (JavaScript, FORM HTTP, HTML e CSS).

Il prodotto offline è multipiattaforma, poichè non ha specifici requisiti software o hardware, e non richiede l'installazione di programmi da parte dell'utente per l'accesso ai contenuti (è per così dire "hot plug"). Esso include sia i testi enciclopedici che parte delle immagini; infatti il testo è rilasciato sotto licenze libere, mentre i contenuti multimediali, fra le quali le immagini, possono essere esemplari di opere con diritti d'autore esclusivi e quindi non redistribuibili.

Tutto il software implementato, sia esso utilizzato per la conversione che per la consultazione, è rilasciato sotto licenze opensource; anche le tecnologie utilizzate dal software del progetto WaNDA sono libere. L'ambito del progetto è infatti quello dell'*open content* poichè utilizza i contenuti liberi di Wikipedia e nasce nell'ambito del gruppo *open@polito*, il centro di competenza per l'opensource e il software libero dell'ateneo.

Durante lo sviluppo del progetto sono anche state affrontate tutte le tematiche giuridiche riguardo il diritto d'autore, le opere collaborative, le licenze d'uso libere e la loro applicazione pratica, e le responsabilità legali nel redistribuire i contenuti.

Il capitolo 2 presenta il progetto Wikipedia, soffermandosi sugli aspetti tecnologici, sul contenuto e sulle licenze utili alla comprensione del resto del testo. Viene anche introdotto l'ambiente in cui nasce e si sviluppa, descrivendo gli altri progetti simili per tecnologia o contenuto. Una cospicua parte è dedicata all'analisi del software wiki utililizzato per la gestione di Wikipedia, descrivendo sia il funzionamento del codice che in particolare le strutture dati; si riporta inoltre lo schema della rete interna di Wikipedia, complementare alla comprensione del funzionamento. L'analisi è proficua all'illustrazione del progetto WaNDA poichè il software di conversione in parte si basa su questo, e da qui provengono i contenuti enciclopedici.

Il capitolo 3 introduce il progetto WaNDA, descrivendone gli obiettivi da raggiungere e le linee guida prefissate. Poichè il progetto ha prodotto diverse soluzioni tecnologicamente differenti, vengono analizzate le diverse soluzioni possibili e, in dettaglio, l'attuale soluzione definitiva, che meglio soddisfa gli obiettivi. È presente una breve descrizione delle componenti del progetto, assieme all'introduzione della terminologia utilizzata per definire le fasi e le parti del formato finale. Oltre ad una valutazione sulle scelte tecnologiche del progetto, vengono riportati alcuni progetti in parte affini, commerciali e non.

Il capitolo 4 completa il precendente definendo l'architettura del progetto WaNDA. Il capitolo inizia analizzando le componenti sviluppate per il processo di conversione e quelle sia statiche che dinamiche per la presentazione dei contenuti. Prosegue con la definizione di varie tecniche impiegate per la gestione dei contenuti, approfondendo la questione dei collegamenti fra le voci. È poi presente un'approfondita descrizione delle tecnologie multipiattaforma scelte e sviluppate per la presentazione dei contenuti, illustrando fra l'altro il motore di ricerca offline ed i formati utilizzabili per i supporti fisici.

La parte più pratica è raccolta nel capitolo 5, che descrive la piattaforma software e hardware utilizzata per effettuare la conversione; sono elencate in dettaglio le fasi necessarie all'estrazione, soffermandosi sul carico del sistema. Sono riportati i programmi

utilizzati ed utilizzabili per costruire l'ambiente dove effettuare la conversione, tutti rilasciati con licenze opensource. Poichè la conversione è un'operazione pesante, si riporta uno studio minuzioso sulle risorse occupate e sugli accorgimenti possibili per ridurre il tempo di elaborazione totale.

Infine il capitolo 6 presenta una valutazione sul prodotto finale. Sono esaminati gli errori che possono sorgere ed i meccanismi sviluppati per la verifica. Sono inoltre presentate varie idee per un possibile miglioramento futuro del prodotto finale, in particolare tenendo in conto la continua crescita di volume dei testi, e un'analisi sull'estensibilità ad altri contenuti online di tipo wiki. Sono in seguito affrontate le questioni giuridiche riguardanti il progetto WaNDA e Wikipedia; vengono trattate sia le licenze d'uso dei contenuti collaborativi, sia la diffusione del prodotto finale che esso avvenga online o su supporto fisico, considerando principalmente l'ambito italiano ma anche quello internazionale.

Capitolo 2

Il progetto Wikipedia

Che cos'è esattamente Wikipedia? *Wikipedia* è un vasto progetto enciclopedico multilingue a cui si accede tramite Internet; le voci enciclopediche sono scritte in modo collaborativo da volontari sparsi per il mondo e possono essere modificate da chiunque sia dotato di un accesso ad Internet. I contenuti sono generalmente rilasciati sotto licenze libere, anche se possono essere presenti eccezioni.

Il progetto è gestito dalla *Wikimedia Foundation*, un'organizzazione senza scopo di lucro, che vive grazie ad annuali donazioni. Essa provvede a mantenere sia i sistemi che contengono e permettono l'accesso alle informazioni, che i contenuti del progetto Wikipedia in inglese. La manutenzione dei contenuti nelle diverse lingue di Wikipedia è affidata in modo abbastanza autonomo ad enti locali, come ad esempio per l'Italia è l'associazione *Wikimedia Italia*.

Con il termine *wiki* si indica un sito o altro meccanismo su Internet che permette un approccio collaborativo alla stesura di contenuti [1]; la parola "wikipedia" deriva quindi dalla fusione di questo termine con la parola inglese "encyclopedia".

In settembre 2007 il progetto Wikipedia include oltre 8 milioni di voci in 250 lingue, per un totale di quasi un miliardo e mezzo di parole [3]. La parte italiana di Wikipedia è giunta a 320 mila voci; è fra le dieci maggiori. La popolarità di Wikipedia ed il suo numero di voci è cresente di giorno in giorno; il dominio `wikipedia.org` si situa fra i dieci siti più visitati al mondo [2].

A causa della natura libera dei contenuti, che deriva dall'approccio *opensource* allo sviluppo di software (infatti si parla di *open content*), è stata spessa criticata l'attendibilità delle informazioni ivi contenute. Il progetto è infatti soggetto ad attacchi di vandalismo o alla presenza di informazioni non veritiere oppure opinabili. Tuttavia alcuni studi [4] hanno dimostrato che tali contenuti negativi non permangono a lungo e che in genere l'enciclopedia converge verso un'accuratezza maggiore.

2.1 Analisi di Wikipedia

Lo scopo ultimo di Wikipedia, nell'intento dei fondatori, è quella di creare un'enciclopedia libera il più vasta possibile con una qualità migliore a quella delle equivalenti enciclopedie

commerciali. Si riporta un riassunto della storia di Wikipedia, utile alla comprensione dell'ambiente in cui sorge il progetto.

Il primo progetto di tipo wiki nasce a metà degli anni '90 con lo scopo di raccogliere algoritmi di programmazione; il nome del progetto è "Portland Pattern Repository" [5]. Il software che permette agli utenti di aggiungere e modificare gli algoritmi presenti è chiamato *WikiWikiWeb* [6] ed è stato scritto da H. G. Cunningham; il termine "wiki" deriva da una parola hawaiana che significa "veloce", è utilizzato per indicare la velocità e facilità con la quale è possibile interagire con i contenuti.

Con il passare del tempo questo approccio va diffondendosi e nascono altri software di questo tipo, che vengono indicati come "wiki software".

Nel frattempo attorno al 2000 presso la società che gestisce il portale Bomis [7], l'allora CEO Jimmy Whales e Larry Sanger valutano l'idea di creare un'enciclopedia gratuita basata su un software wiki. L'enciclopedia prende il nome di *Nupedia* [8] e prevede la stesura delle voci in sette passi, secondo una successione di revisioni da parte di esperti. Il software alla base è NupeCode; gli articoli vennero inizialmente rilasciati con una licenza apposita (Nupedia Open Content License) e poi, sotto le pressioni di R. M. Stallman che l'aveva da poco redatta, con licenza GNU Free Documentation License.

Il progetto però non andò in porto, producendo in due anni 24 voci e circa una settantina ancora da revisionare. L'idea di un'enciclopedia di questo tipo modificabile da chiunque, non solo da esperti, piacque a Stallman tanto che iniziò il progetto GNUPedia nell'ambito della Free Software Foundation [9]. Quest'idea venne applicata anche da Whales e Sanger, che diedero vita nel 2001 al progetto Wikipedia in parallelo a Nupedia per facilitare l'approvazione delle voci da revisionare. Poiché la struttura di Nupedia e Wikipedia era già in piedi mentre quella di GNUPedia no, e dato che i mezzi e le finalità erano le stesse, il progetto GNUPedia terminò senza aver prodotto nulla di fatto.

Il progetto Wikipedia iniziò a crescere rapidamente, diventando ben presto il progetto principale. Esso era basato su *UseModWiki* [10], un software wiki personalizzabile scritto in PERL, sviluppato nel 2000 e rilasciato sotto licenza GPL.

Nel 2002 il progetto Nupedia quindi termina e poco dopo Wikipedia divenda un progetto autonomo da Bormis e gestito da una fondazione appositamente creata, la Wikimedia Foundation [11]. Inizia inoltre lo sviluppo di un nuovo software wiki, di nome *MediaWiki*, pensato appositamente per gestire Wikipedia. È anche in questo periodo che Sanger lascia il progetto, per poi tempo dopo dedicarsi ad un altro progetto enciclopedico derivato da Wikipedia, però maggiormente verificato, di nome Citizendium [12].

Wikipedia è probabilmente stato il primo servizio *web 2.0*, seguiti da *YouTube* e *MySpace* [13]; con tale termine si intende indicare una piattaforma Internet che permette una completa interazione fra creatori e usufuitori di un servizio, nell'ottica di ampliare le capacità comunicative del web.

2.1.1 Altri progetti

La diffusione di Wikipedia ha portato Wikimedia Foundation ad investire in altri progetti basati sul suo funzionamento; tutti i contenuti di questi progetti sono quindi inseriti da chiunque. Attualmente sono mantenuti i progetti:

- Wiktionary (Wikizionario) [15]: un dizionario multilingue.
- Wikibooks [16]: una raccolta di libri di testo, manuali e altri testi educativi a contenuto aperto.
- Wikiquote [17]: una raccolta di citazioni di qualsiasi genere.
- Wikisource [18]: una raccolta di testi e documenti di dominio pubblico.
- Wikispecies [19]: un catalogo aperto e libero di tutte le specie viventi.
- Wikinews (Wikinotizie) [20]: una fonte di notizie a contenuto aperto.
- Wikiversity (Wikiversità) [21]: una raccolta di risorse e attività didattiche.

Ognuno di questi progetti esiste in diverse lingue, a seconda dell'interesse dell'ente locale che rappresenta Wikimedia Foundation; per l'italiano sono tutti presenti. Esistono altri due progetti, che sono utili alla gestione di tutti i progetti di Wikimedia e sono quindi presenti in sola lingua inglese:

- Commons [22]: un contenitore per le risorse multimediali comune ai progetti.
- Meta-Wiki [23]: contiene le pagine per il coordinamento dei progetti.

Per convenzione e praticità, visto il grande numero di progetti, nell'ambito di Wikimedia si indica con *wiki** l'insieme dei progetti della Wikimedia Foundation. È anche possibile utilizzare il termine **wiki* per indicare l'insieme delle lingue per un dato progetto; per esempio con "itwiki" si intende di solito la parte italiana di Wikipedia [1].

2.1.2 Licenza dei contenuti

Il progetto Wikipedia segue le ideologie del sapere libero, applicando dove possibile questi concetti. Questo si applica sia ai contenuti dell'enciclopedia che sul software utilizzato per erogare il servizio (descritto in sezione 2.2) [2].

I contenuti in questione sono di libero accesso e riproducibili; più specificatamente si parla quindi di "contenuto libero" o "*open content*". L'idea del contenuto libero deriva dall'idea dell'opensource per i software: perché non rendere liberamente utilizzabile e modificabile non solo software ma anche contenuti testuali e multimediali? Come nel caso del software, con contenuto libero non si allude ad una licenza in particolare ma che soddisfa una certa serie di requisiti.

Si introduce un altro concetto, il *copyleft* [125], una forma di concessione di diritti che si contrappone al diritto d'autore classico (notare il gioco di parole con il termine inglese "copyright") poiché non prevede la necessità di chiedere all'autore il permesso per

[1] Tale convenzione sarà utilizzata nel resto del testo.

[2] Alla fine del testo, in sezione 6.4, è presente un'ampia trattazione sull'aspetto giuridico delle licenze libere, applicato sia ai contenuti di Wikipedia (riproposti dal progetto WaNDA), sia sulla distribuzione dei supporti del progetto WaNDA.

riutilizzare o modificare i suoi contenuti. Tuttavia diversamente dall'open content di più ampio respiro, il copyleft richiede che i contenuti modificati debbano essere rilasciati con la stessa licenza con la quale l'autore originale li ha rilasciati.

Si noti che il copyleft non solo denota un classe di licenze libere, ma anche un ambiente culturale di pensiero.

Le licenze di utilizzo con le quali sono in genere rilasciati i contenuti di Wikipedia definiscono alcuni elementi:

- Attribuzione: la licenza indica se è necessario includere o meno le informazioni sull'autore; nel caso non vi sia attribuzione, in genere non si tratta di una vera e propria licenza ed il contenuto è presente nel pubblico dominio.
- Utilizzo commerciale: la licenza indica se i contenuti possono essere utilizzati per scopi prettamente commerciali, ovvero se vengono generati introiti dalla sola vendita del contenuto.
- Alterazione: la licenza indica se è permesso modificare il contenuto di un altro autore. Se questo fosse possibile la licenza può indicare se il contenuto modificato deve essere rilasciato con la stessa licenza originale o meno; ciò si traduce in una licenza di tipo copyleft o no.

La licenza con la quale i contenuti di Wikipedia vengono generalmente rilasciati è la *GNU Free Documentation License* (*GNU FDL*) [126]. Questa licenza si può applicare sia a contenuti testuali che multimediali; essa è di tipo copyleft poiché prevede la libera circolazione e rielaborazione dei contenuti, indicandone la fonte originale, ma tutte le opere derivate (anche solo in parte) devono essere rilasciate con licenza sempre FDL.

Su Wikipedia un autore che decide di scrivere una voce, rilascia forzatamente il testo sotto licenza GNU FDL. Un autore che modifica il testo di una voce già presente segue così i principi della licenza GNU FDL. D'altra parte un utente che vuole redistribuire o modificare un contenuto di Wikipedia è libero di farlo alla condizione di rilasciarlo sotto medesima licenza ed indicarne la fonte.

2.1.2.1 Contenuti non liberi

Tutti i testi di Wikipedia, essendo presenti sotto forma di voci, sono quindi presenti con licenza GNU FDL. Questo non vale però per i contenuti multimediali, prevalentemente immagini ma anche musiche. L'autore di una voce può decidere di includervi, allo scopo illustrativo, contenuti multimediali; tuttavia non è detto che egli sia il possessore dei diritti su quel contenuto.

Il contenuto multimediale potrebbe essere stato rilasciato con un'altra licenza copyleft oppure open content, nel qual caso l'autore deve verificare la compatibilità fra le licenze.

Inoltre la maggior parte delle immagini e delle musiche presenti non hanno neppure una licenza compatibile, ma sono incluse secondo il concetto di *fair use*; con questo si intende l'utilizzo non autorizzato o l'incorporazione di materiale protetto da copyright nel lavoro di un altro autore sotto certe specifiche condizioni. Sebbene esso sia in vigore nella

legislazione statunitense e non in quella italiana, è utilizzata come misura precauzionale per limitare i contenuti sfacciatamente illegali. D'altronde non tutte le lingue di Wikipedia ammettono l'utilizzo di materiale altrui con licenza non libera; un esempio è la versione francese, che accetta soltanto materiale multimediale rilasciato sotto licenza compatibile con la FDL oppure appartenenti all'autore che immette il contenuto.

Wikipedia ha redatto una policy per descrivere i casi in cui il concetto di fair use si applica, detta "Politica di Dottrina dell'Esonero", dall'inglese "Exemption Doctrine Policy" (*EDP*) [14]. Essa permette l'immissione di materiale protetto di altri a condizione che:

- Non sia ragionevolmente possibile ottenere un contenuto equivalente dal punto di vista illustrativo e dotato di una licenza libera.
- Il contenuto rientri in una serie di tipologie, tra i quali l'utilizzo autorizzato a Wikipedia per soli fini non commerciali, immagini di stemmi, enti o altro, e screenshot a dimensioni ridotte.
- Il contenuto sia utilizzato da una voce enciclopedica che deve riguardare l'argomento del contenuto multimediale. Nel caso di immagini essa non può fare parte di una galleria di immagini.
- Il contenuto sia corredato di un riepilogo in cui si indica la fonte, la licenza, il detentore dei diritti e infine una spiegazione sul motivo per cui si utilizza il contenuto.

  Si noti che anche per i contenuti rilasciati con una licenza libera è necessario indicare la licenza specifica e la fonte.

Come esempio si possono considerare le copertine degli album musicali, che non sono sostituibili con altro materiale e devono essere presenti sulla pagina con la loro descrizione. Un altro esempio multimediale sono gli estratti musicali, che per convenzione non devono superare i trenta secondi. Anche i testi possono eventualmente essere riportati in Wikipedia pur non essendo stati rilasciati con licenza libera: è questo il caso delle citazioni, che possono essere state estratte da materiale con diritti. Questo è permesso dalle policy di Wikipedia.

Questo tipo di inclusione a scopo illustrativo, derivato dal fair use statunitense, potrebbe essere ricondotto all'articolo 70 della legge sul diritto d'autore (legge n. 633 del 22 aprile 1941, aggiornato con il D.Lgs n. 72 del 22 marzo 2004), indicato con il nome di *diritto di corta citazione* [131], che permette la riproduzione dell'opera entro limiti e regole precise, e prevede come requisiti l'indicazione della fonte. Non è chiaro se tale diritto sia applicabile all'ambito di WaNDA, poiché potrebbe essere inteso con fini commerciali o comunque che generi una qualche sorta di profitto; tale articolo è inteso per essere applicato con fini escluvisamente didattici. La giurisdizione in questo campo riserva tuttavia in via esclusiva la gestione dei diritti, indicato nell'articolo 180 della suddetta legge, ad un organo specifico (la Società Italiana degli Autori ed Editori, SIAE) che valuta autonomamente l'uso e le sue finalità di opere con diritti. Pertanto le regole dell'EDP sono utili per regolamentare gli utenti di Wikipedia, ma non hanno un chiaro valore legale in Italia poiché non proteggono chi riporta le opere.

2.1.3 Punto di vista neutrale

Wikipedia è un wiki scritto da chiunque; tuttavia essendo un'enciclopedia questo comporta che i contenuti devono essere scritti con oggettività. Il principio alla base di Wikipedia è quindi il *punto di vista neutrale* (NPOV da "neutral point of view"), poiché chi scrive deve evitare l'opinabilità delle informazioni; eventuali opinioni di personaggi illustri possono essere riportati con rigore scientifico. Il progetto si basa sull'idea che il punto di vista neutrale non è completamente raggiungibile dal singolo autore ma dalla collettività. È possibile riscontrare imprecisioni o dichiarazioni non oggettive; la natura aperta di Wikipedia tuttavia non solo provoca la presenza di queste imprecisioni, ma anche la loro correzione, dato che con un numero sufficientemente grande di autori è probabile che il punto di vista neutrale venga raggiunto.

In estremi casi in cui non si giunge ad una versione stabile è possibile ricorrere ad una sorta di limitazione nella modifica dei contenuti da parte degli amministratori, che permettono soltanto agli utenti registrati di modificare il contenuto. Si noti che la registrazione a Wikipedia è gratuita e senza impegni: il suo scopo è quello di indentificare in modo migliore gli autori.

Gli amministratori hanno il compito di vigilare sullo svolgimento corretto dell'enciclopedia; essi si occupano per esempio di gestire i casi di pagine con dichiarazioni evidentemente non oggettive, oppure impedire che atti di vandalismo provochino danni, o decidere di limitare i permessi di modifica senza registrazione, o anche ammonire gli utenti che ripetutamente hanno influito negativamente sull'evoluzione del progetto. Il compito di amministratore è assegnato ad una cerchia di persone il più variegato possibile, per evitare una guida in qualche modo "di parte" di Wikipedia; le decisioni vengono prese in gruppo tramite votazione.

2.1.4 Conoscenza collettiva

Il progetto Wikipedia è portato avanti con l'idea che l'approccio collaborativo alla redazione dei testi porta inevitabilmente alla miglioria degli stessi, come succede con il software opensource.

Tuttavia le due situazioni non sono proprio identiche per diversi motivi. Per esempio chi modifica le voci di Wikipedia può davvero essere chiunque, mentre chi modifica un software rilasciato con licenza libera di solito appartiene già ad una cerchia più ristretta di individui dotata di competenze specifiche; questo perché il linguaggio di programmazione del software è sicuramente meno conosciuto della lingua italiana. Inoltre di solito chi effettua modifica sul codice non modifica automaticamente il software stesso, ma di solito propone le proprie modifiche all'autore o agli autori iniziali che decidono il da farsi, effettuando implicitamente una sorta di selezione dei contenuti. Un altro elemento da tenere in mente è il rigore maggiore di un codice software che del testo di una voce; nel secondo caso vi è molta più libertà di scelta e l'opinabilità di un concetto è sicuramente maggiore.

Questo fenomeno si può riscontrare anche tra diversi argomenti di Wikipedia: per esempio gli articoli di un ambito specifico tendono ad essere meno soggetti a discussioni e

convergere più velocemente ad una versione stabile. Un articolo di più ampio respiro che lascia maggiore spazio alla soggettività può richiede più tempo a raggiungere una versione stabile.

Il meccanismo di modifica ed immissione di testo su Wikipedia è estremamente più semplice che per un qualsiasi codice software. Questo costringe anziché ad una verifica preventiva dei contenuti ad una verifica a posteriori non sempre del tutto efficace. Ciò lascia molta libertà e spazio agli autori, che per un argomento non sempre trovano tutti d'accordo; accade quanto descritto prima, in cui il testo di una voce è oggetto di lunghe discussioni e può dover essere negata la possibilità di modifica agli utenti non registrati o a tutti, in attesa di "calmare" la situazione: questo fenomeno viene indicato con l'espressione "guerre di edizione".

Nel caso in cui vi siano alterazioni successive di diversi punti di vista è possibile effettuare a posteriori una fusione dei contenuti, poiché è presente per ogni voce lo storico di tutte le modifiche. Questo permette anche di visitare rapidamente la storia di una pagina per verificare i singoli contributi.

Sebbene una voce abbia raggiunto un ottimo grado di neutralità e sembri abbastanza completa, ovvero una versione stabile, essa non cessa di poter essere modificata. Ciò permette un possibile continuo miglioramento della qualità delle voci, la cui stesura non è mai conclusa.

Infine la crescita delle voci, essendo frutto di una collettività disomogenea, non è guidata e segue un'evoluzione praticamente casuale e dettata dagli interessi di masse di collaboratori. Frutto di questa evoluzione è per esempio il fenomeno per il quale la crescita una voce può essere sproporzionata rispetto alla sua importanza. Per esempio specifiche voci di attualità sono spesso più dettagliate di importanti voci che però alla maggior parte della comunità interessa poco.

2.1.5 Valutazione dei contenuti

Il progetto Wikipedia, più è diventato famoso, più ha ricevuto numerose critiche ed altrettanti riconoscimenti. Sull'analisi della qualità dei contenuti è presente una vasta letteratura che si è espressa con le opinioni più disparate; si rimanda alla pagina italiana di Wikipedia sull'argomento [24].

Le critiche mosse a Wikipedia in genere si basano sul fatto che non vi è alcuna assicurazione sull'affidabilità, sull'attendibilità e sull'autorevolezza delle voci, in particolare perché un utente qualsiasi può immettere informazioni errate ma plausibilmente vere senza che vengano rimosse. Questo è certamente vero ma tuttavia sorge il dubbio se sia più autorevole un testo potenzialmente verificato da milioni di persone piuttosto che uno redatto da una sola persona con presunta autorevolezza. Un'altra frequente critica riguarda la superficialità delle voci, poiché non sono per forza inseriti tutti gli elementi di descrizione ma quelli maggiormente noti ai più.

Sono inoltre stati effettuati dei test comparativi con enciclopedie tradizionali autorevoli, che hanno di solito verificato la buona qualità delle informazioni, inferiore a quelle

autorevoli come l'Enciclopedia Britannica ma superiore a quelle più commerciali quali Encarta. Da confronti con l'Enciclopedia Britannica a volte Wikipedia è risultata superiore per completezza per quanto riguarda gli argomenti tecnologici e di attualità.

In definitiva i contenuti di Wikipedia non sono per forza attendibili ed autorevoli, ma sono sicuramente un'ottima base per affrontare un argomento; per questo nella stesura dei testi delle voci le linee guida di Wikipedia consigliano di includere in fondo alla pagina il numero più possibile completo di referenze alle informazioni, che permette sia di verificare la veridicità delle informazioni sia di approndire l'argomento.

La qualità richiesta di Wikipedia è inoltre da rapportare all'ambito di utilizzo. È chiaro che l'accuratezza degli argomenti contenuti ha un livello che varia da voce a voce, ma è per esempio superiore alle richieste di un insegnamento elementare.

Nonostante le dispute sulla completezza di Wikipedia, l'affermazione e la diffusione di Wikipedia è giunta ad un punto tale che anche studi accademici [28], libri di vario genere [29], giornalisti [25] e casi giudiziari [26][27] propongono collegamenti alle sue voci, di solito allo scopo informativo e non referenziale.

2.2 Analisi delle tecnologie

Tutti i progetti Wikimedia sono attualmente basati su un software wiki di nome *MediaWiki* [30]. Il nome è un gioco di parole che arriva dall'inversione del nome della fondazione Wikimedia. Il software è rilasciato con licenza GNU General Public License (GPL) versione 2 [127]; è scalabile essendo pensato per essere utilizzato da un grande mole di utenti ed offre la possibilità di aggiungere estensioni. Offre inoltre numerose funzionalità non presenti sui tradizionali software wiki.

È scritto con il linguaggio PHP [108] che elabora e presenta le pagine HTML, e utilizza un database per contenere i dati, a scelta tra MySQL [95] e PostgreSQL [96].

MediaWiki presenta all'utente un insieme di pagine con collegamenti per la loro gestione. Le pagine sono memorizzate con un linguaggio appositamente concepito per facilitare la composizione di testo formattato. Questo linguaggio si chiama *wikitext*; esso permette ad un utente (che può anche non conoscere il linguaggio HTML) di modificare facilmente e velocemente il contenuto delle pagine. MediaWiki si occuperà quindi di interpretare il wikitext e tradurlo in testo HTML formattato.

Le pagine possono essere molto facilmente rimosse o create. Di ogni pagina esistita MediaWiki mantiene la storia completa delle modifiche, permettendo un facile ripristino del testo precendente oppure il confronto di due versioni della pagina.

MediaWiki permette inoltre la gestione di pagine con funzionalità particolari, fra le quali la possibilità di utilizzare una pagina che automaticamente redirige il browser ad un'altra pagina; la pagina che redirige prende il nome di "pagina di redirect". MediaWiki permette la gestione di contenuti grafici e multimediali, che sono memorizzati sul filesystem.

Gli utenti di MediaWiki possono essere registrati ed avere accesso a diversi livelli di permessi aggiuntivi. Inoltre ogni utente registrato è dotato di pagina con il proprio nome.

MediaWiki è pensato per contenere in ogni pagina una voce enciclopedica di Wikipedia.

È tuttavia utilizzato da molti enti ed individui per i propri scopi; è infatti un buon software che, oltre a permettere di facilmente inserire i contenuti, essendo la base di Wikipedia è continuamente rivisto alla ricerca di bug.

2.2.1 Struttura dei contenuti

Non tutte le pagine in MediaWiki hanno la stessa funzione. Esistono differenti gerarchie nell'organizzare le pagine.

Innanzitutto ci sono le pagine normali, tipicamente la stragrande maggioranza, che contengono le informazioni; su Wikipedia è qui che sono presenti le voci enciclopediche. Il loro percorso di accesso (URL) è composto direttamente dal titolo della voce. A questo tipo di pagine appartiene la *pagina principale* caricata di default. Sono anche pagine normali le pagine di redirect: esse infatti hanno al posto del testo un comando specifico per la redirezione alla pagina destinazione.

Ci sono poi le pagine di servizio o "pagine speciali", ovvero pagine proprie di MediaWiki obbligatorie per il funzionamento, per esempio per registrare un utente o per caricare i contenuti grafici. Queste pagine si riconoscono poiché il loro titolo contiene il prefisso `Speciale:`; queste sono le uniche che non possono essere modificate o aggiunte dagli utenti, amministratori inclusi.

Seguono le "categorie", che permettono di raggruppare i contenuti con delle gerarchie. Le categorie vengono automaticamente aggiornate impostando in una pagina normale l'appartenenza ad una di esse; in caso la categoria non esista essa verrà generata. Nelle pagine delle categorie è possibile aggiugere testo o impostare l'appartenenza di questa ad un'altra categoria; si genera così una gerarchia di categorie, utile per organizzare i contenuti su Wikipedia. Le categorie contengono il prefisso `Categoria:`; la pagina speciale `Speciale:Categorie` contiene l'elenco completo delle categorie presenti.

Esistono anche i "template", ovvero pagine contenenti testo che può essere incluso nelle pagine normali richiamandone soltanto il nome; in verità è anche possibile inviare comandi al template, che quindi si comporterà in modo differente, analogamente ad una funzione in un linguaggio di programmazione. Il titolo di queste è composto dal prefisso `Template:`. Hanno una grande utilità in Wikipedia poiché permettono di automatizzare l'immissione di elementi ricorrenti nelle voci, come per esempio un menù informativo per i comuni.

Come si è detto gli utenti registrati dispongono di una propria pagina. Queste sono chiamate "pagine utente" e hanno come prefisso `Utente:`.

A tutte le pagine (tranne le speciali) è associata una pagina gemella, detta "di discussione", che offre uno spazio dove discutere l'argomento della pagina associata. Le pagine di discussione delle pagine utenti vengono per esempio utilizzate come meccanismo di comunicazione fra gli utenti.

Ognuno dei tipi di pagine illustrate appartengono ad un cosiddetto "namespace"; ad ogni namespace è attribuito un numero, che per le pagine normali è `0`. La gestione dei namespace è particolarmente utile nel database per riconoscere l'appartenenza di una pagina al suo gruppo; non è però del tutto evidente all'utilizzatore di MediaWiki. Per

questo l'elenco dei namespace presenti su Wikipedia è riportato in sezione 4.2.1, dove si discute della selezione dei contenuti.

2.2.1.1 Schema di memorizzazione

MediaWiki utilizza un DBMS per memorizzare i contenuti del wiki. Tutti i progetti *wiki** sono basati su database MySQL, uno fra i più noti e affidabili DBMS opensource.

Nel database MediaWiki memorizza tutte le revisioni di tutte le pagine, con l'indicazione dell'autore e della data per ogni revisione. È così possibile facilmente identificare le parti di testo modificate o aggiunte da qualunque autore.

I contenuti multimediali sono invece memorizzati sul filesystem in un'apposita gerarchia di directory.

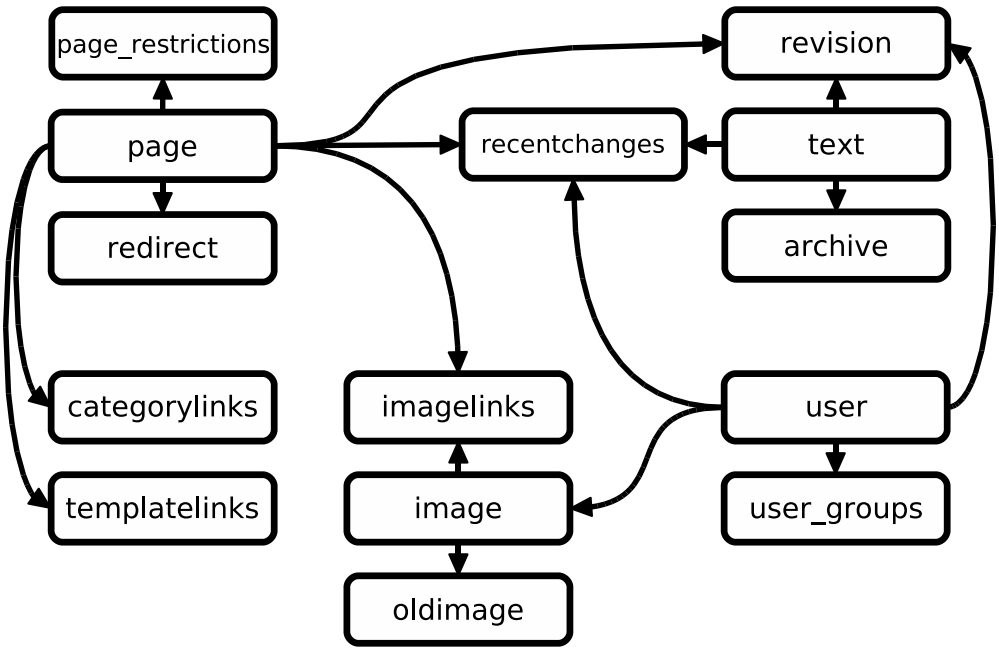

Figura 2.1. Schema semplificato delle tabelle del database in MediaWiki versione 1.5.

L'organizzazione delle tabelle del database è riportata in figura 2.1; è uno schema semplificato in quanto mancano alcune tabelle di supporto poco utili alla comprensione dello schema di memorizzazione. Le frecce indicano la relazione fra le tabelle, che sono di tipo uno a molti.

La tabella `page` è centrale al funzionamento di MediaWiki, poiché permette di mettere in relazione tutte le altre tabelle per la costruzione della pagina; la chiave della tabella è un numero intero univoco, chiamato *id* [3]. La tabella contiene informazioni di base per

[3] Nel resto del testo indicando *id* si intenderà sempre questo numero.

ogni voce; per ogni voce è possibile indentificare il testo della revisione corrente (grazie alla relazione con `revision`) e le immagini contenute nella pagina (grazie alla relazione con `imagelinks`).

Nel caso in cui la voce sia un redirect, al posto di ricercare il contenuto della pagina tramite `revision` si estrae la pagina destinazione da `redirect`.

Sono inoltre presenti due tabelle di supporto, `categorylinks` e `templatelinks`, che raccolgono l'elenco delle categorie e dei template ai quali le pagine possono appartenere.

La tabella `image` contiene l'elenco di tutti i file multimediali caricati, siano immagini che musiche oppure video. La tabella non contiene in verità i file stessi, ma soltanto il loro percorso sul filesystem; sono univocamente identificati dalla stringa del loro nome. La tabella `image` è messa in relazione con le pagine grazie alla tabella intermedia `imagelinks`, che contiene una corrispondenza tra id delle pagine e nome dei file.

Il testo delle voci è contenuto in `text`; la corrispondenza tra id della voce a cui fa riferimento e le differenti versioni di testo della voce è mantenuta da `revision`, che contiene anche l'indicazione dell'utente che ha generato tale testo. L'utente è presente nella tabella `User`, identificato univocamente dalla stringa del nome.

Lo schema del database di MediaWiki presentato è semplificato ed è valido per le ultime versioni di MediaWiki. Infatti durante la sua evoluzione lo schema di memorizzazione cambia di versione in versione [31]; vi è tuttavia stata una notevole rivisitazione nella versione 1.5, da cui non si è molto differenziata fino all'attuale versione 1.10.

2.2.1.2 Componenti di funzionamento

Il funzionamento di MediaWiki è basato interamente sul linguaggio PHP; il codice è gestito da varie classi, ognuna delle quali si occupa di un gruppo di operazioni logicamente affini [32]. Nel tempo il numero di classi e di metodi, per ampliare le funzionalità, è cresciuto notevolmente senza una guida precisa, per cui il codice è oggi abbastanza 'labirintico', avendo molte classi che si chiamano l'un l'altra circolarmente.

Uno schema estremamente semplificato del codice è presente in figura 2.2; anche qui il codice è ovviamente mutato nel tempo, anche se con meno variazioni che nella gestione del database. La versione di riferimento è ancora quella del periodo da MediaWiki 1.5 a 1.10.

La classe principale, che istanzia le altre e gestisce le operazioni da compiere, si chiama `MediaWiki`. È presente nel file `includes/Wiki.php`; è anche l'unica eccezione nella convenzione dei nomi dei file, che devono avere il nome della classe che contengono.

La richiesta di una pagina è quindi indirizzata alla classe `MediaWiki`, che parte con il caricare le impostazioni tramite la classe `Setup`. Procede poi ad ottenere il testo sotto forma di *wikitext* dalla classe `Article` ed effettua la formattazione della pagina (con relativa traduzione del wikitext) utilizzando la classe `OutputPage`, che provvede infine a stampare la pagina richiesta.

Le richieste dei contenuti sotto forma di wikitext vengono quindi effettuate da `Article`, che per le richieste al database utilizza `Title` e `Revision`. Si noti che l'interfaccia vera e propria al database è demandata a `Database`, che astrae il database effettivamente

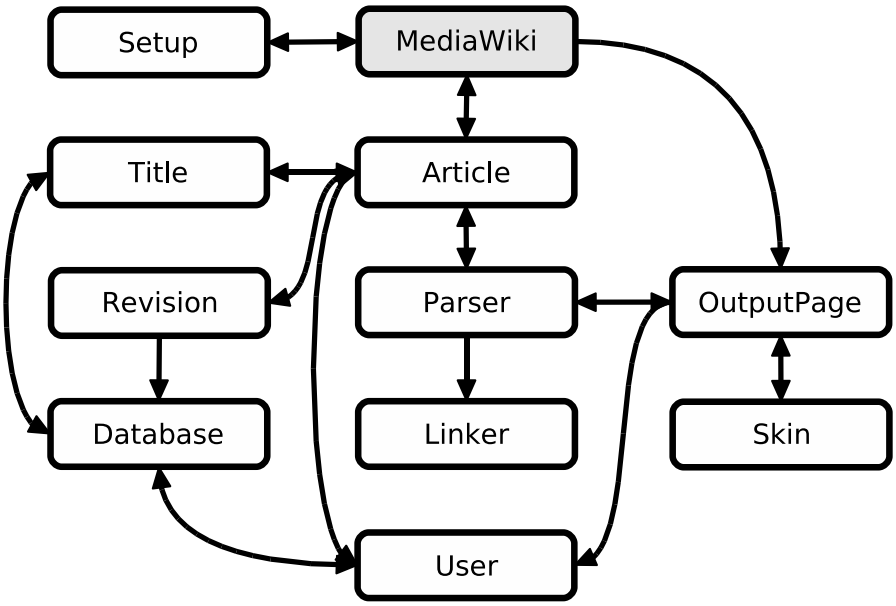

Figura 2.2. Diagramma semplificato delle classi di MediaWiki versione 1.5.

utilizzato; tuttavia in altre parti del codice vengono sporadicamente effettuate richieste direttamente al database.

La classe `OutputPage` si occupa sia della traduzione del wikitext passato da `MediaWiki` che della formattazione della pagina. Il wikitext viene tradotto grazie alla classe `Parser`, che demanda la gestione dei collegamenti a `Linker` e ottiene informazioni aggiuntive, per esempio il codice dei template, grazie ad `Article`. `OutputPage`, prima di rilasciare il testo, provvede alla resa grafica della pagina ed all'applicazione di un tema, utilizzando la classe `Skin`.

Oltre ai file PHP presenti in `includes/` che servono al funzionamento proprio di MediaWiki, ve ne sono di aggiuntivi, siti in una directory di nome `maintenance/`, che hanno gli utilizzi più disparati. Essi in genere sono utili all'amministratore per la gestione di MediaWiki, per esempio per installare MediaWiki senza utilizzare un server web oppure per effettuare dei backup del database, oppure anche per l'aggiornamento.

Tra i file ve ne sono due, di nome `dumpHTML.php` e `dumpHTML.inc`, che servono ad esportare le voci del database sotto forma di pagine HTML. In pratica esse istanziano l'oggetto `MediaWiki` ed effettuano l'esportazione del contenuto per ogni *id*. Il tema applicato da `OutputPage` è particolare ed adattato allo scopo. Questi file sono importanti per gli scopi del progetto WaNDA, poiché da essi è nato lo sviluppo del codice attuale.

2.2.1.3 La rete di Wikipedia

Infine diamo un'occhiata alla struttura della rete di Wikipedia, per identificare meglio l'impiego di MediaWiki e la sorgente delle informazioni per il progetto WaNDA.

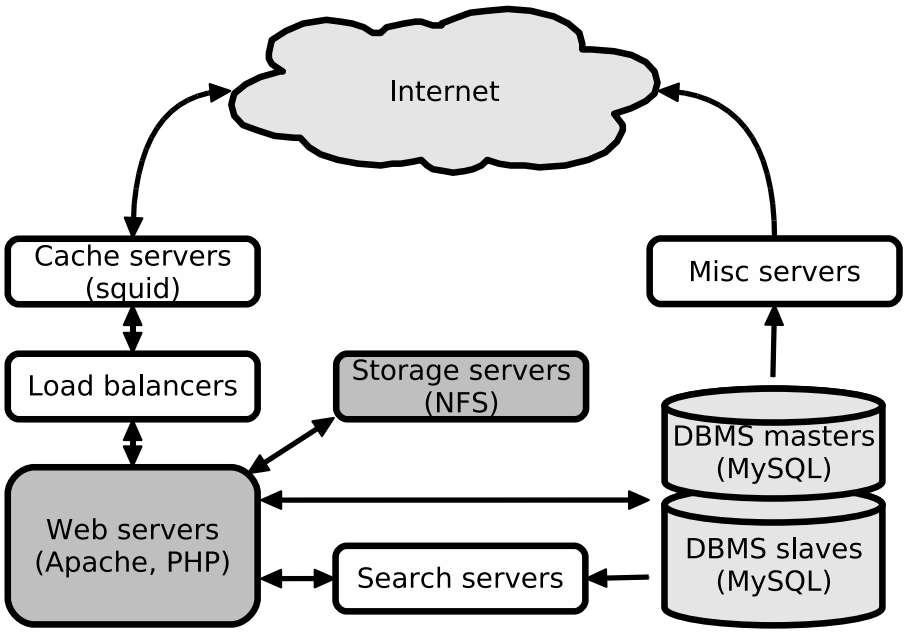

Figura 2.3. Organizzazione dei server di Wikipedia nel 2006.

La rete ha una conformazione abbastanza stabile nel tempo, anche se ovviamente di tanto in tanto vengono aggiunti nuovi server per servire meglio il crescente numero di accessi e di voci [33]. Essa è schematicamente rappresentata in figura 2.3.

La rete è composta da due blocchi fondamentali, i *web server* e i *database server*.

I web server si occupano di eseguire MediaWiki, rispondendo alle richieste di accesso a Wikipedia; per l'esecuzione sono dotati fondamentalmente del server web Apache e dell'interprete PHP, con un sistema operativo GNU/Linux. Nel 2006 questi server erano 105.

L'accesso da Internet ai web server non è diretto; per distribuire uniformemente il carico vengono utilizzati quattro *load balancer* che scelgono il web server adatto a servire la richiesta.

Oltre a ciò sono presenti una cinquantina di server, posti in luoghi geografici differenti, che si occupano di memorizzare le pagine delle voci maggiormente frequentate in modo da non doverle richiedere ogni volta ai web server. Ciò permette di non caricare i server nel caso in cui le richieste che arrivano da Internet siano di sola lettura di voci popolari, ovvero il tipo più comune di richiesta. Questi server sono indicati come *cache server* e

hanno una architettura mista, anche se il software che esegue la copia locale delle pagine richieste, Squid, è comune a tutti.

I *database server* contengono il database di MediaWiki organizzato su diversi server, *master* e *slave*. I database master sono fondamentalmente tre, divisi per lingua; uno si occupa di enwiki, gli altri due si ripartiscono gli altri *wiki. Ognuno di questi ha dietro di sè altri server database di tipo slave, per suddividere la mole di dati. Il software di riferimento per tutto il sistema di database è MySQL.

Per velocizzare i tempi, quando un utente effettua una ricerca in MediaWiki la ricerca non procede direttamente ai database server, ma viene compiuta da dei server intermedi, che periodicamente indicizzano i contenuti dei database.

Nell'analisi dello schema di memorizzazione di MediaWiki è stato visto che le immagini e gli altri contenuti aggiunti all'enciclopedia non sono presenti nel database ma su filesystem: nella rete di Wikipedia essi sono presenti su dei server a parte (i *storage server*) che esportano un filesystem di rete NFS.

Vi sono server aggiuntivi che offrono alcuni servizi misti, come per esempio il server di posta di Wikipedia; fra le funzionalità vi è quella di backup, che genera dei dump del database di ogni lingua. Il server che effettua quest'operazione è particolarmente utile al progetto WaNDA, poiché offre un dump XML della parte italiana di Wikipedia (itwiki), aggiornato periodicamente e liberamente accessibile tramite Internet.

I database server contengono tutte le versioni di Wikipedia nelle diverse lingue e la loro gestione è unificata; i server si situano a Tampa, in Florida (USA), tranne i cache server che sono presenti in Europa ed in Asia per rispondere in tempi più brevi.

2.2.2 Adattamenti ed utilizzi

In origine MediaWiki è stato concepito per il funzionamento di Wikipedia. Il software, rilasciato sotto GPL, si è diffuso per la gestione di altri siti in cui è comodo disporre di un wiki; non è l'unico software wiki con licenza opensource, nè il migliore dal punto di vista prestazionale per pochi contenuti, tuttavia ha il vantaggio di essere sicuramente scalabile, dato l'utilizzo che ne viene fatto in Wikipedia, e di essere costantemente aggiornato e controllato dalla comunità di sviluppatori, poiché la presenza di un bug serio potrebbe avere conseguenze disastrose in un progetto così ampio e noto.

MediaWiki è predisposto per l'ampliamento da parte di terzi: per primo vi sono le estensioni, che permettono di personalizzare in modo indefinito il linguaggio wikitext. Per aggiungere oggetti PHP che effettuino operazioni a più basso livello, nelle ultime versioni di MediaWiki è stata introdotta una classe di nome `Api`, che permette di accedere con meccanismi standard a diverse componenti della struttura interna di MediaWiki.

Inoltre, come già riportato, nella directory `maintenance` di MediaWiki sono presenti vari script e programmi per gestire funzionalità aggiuntive di MediaWiki che esulano dalla semplice interfaccia tramite browser. A volte queste funzionalità utilizzano il codice di MediaWiki vero e proprio; questo meccanismo è utilizzato dal progetto WaNDA per interpretare il più fedelmente possibile il wikitext.

La sintassi del wikitext di MediaWiki non è completamente definita, per due ragioni. La prima (e principale) è dovuta alle innumerevoli aggiunte apportate dalle estensioni; la

seconda è dovuta alla variabilità nel tempo, poiché sia il linguaggio wikitext che soprattutto la traduzione in testo HTML cambia di versione in versione. Questo ha portato all'implementazione di innumerevoli traduttori alternativi del wikitext per i motivi più disparati, nessuno di questi però completo e definitivo.

2.3 Lo sviluppo di Wikipedia

Il progetto Wikipedia fin dalla sua creazione ha sempre avuto notevole successo e si è sempre più diffuso, soppiantando dall'inizio il progetto di lancio, Nupedia. Il meccanismo di inserimento di informazioni è semplice ed accattivante, l'idea di fondo è ammirevole, l'utilizzo è estremamente comodo. La filosofia collaborativa con la quale è sviluppata ha innumerevoli vantaggi poiché riesce a raggiungere una quantità di autori molto più vasta di qualsiasi altra enciclopedia e ne permette idealmente la correzione progressiva delle inesattezze. Lo sviluppo collaborativo libero ha però alcune pecche, come la possibile presenza di informazioni non verificate e i casi di soggettività degli argomenti.

Prima di Wikipedia erano disponibili soltanto enciclopedie commerciali, con voci verificate da esperti e di solito affidabili; anche fra di esse vi sono diversi livelli di qualità delle voci. Alcune erano già liberamente consultabili poiché con diritti d'autore scaduti, ma di conseguenza anche non aggiornate.

Facendo un confronto con le enciclopedie classiche [34] [35], si nota immediamente quanto Wikipedia sia molto più fornita di voci, disponendo di un'ampiezza di argomenti invidiabile. Tuttavia la qualità degli articoli mediamente non raggiunge gli standard della più autorevole enciclopedia moderna, l'Enciclopedia Britannica. Wikipedia è quindi un ottimo strumento, anche se l'affidabilità delle voci non è garantita e i testi possono avere lacune; la crescita costante di Wikipedia lascia però ben sperare non solo in un futuro ampliamento degli argomenti, ma anche approfondimenti e miglioramenti della qualità.

Rispetto alle enciclopedie classiche Wikipedia porta una gestione più moderna dei contenuti. Le enciclopedie classiche nascono da un ambiente cartaceo e la loro trasposizione in forma multimediale ne risente, disponendo di una ricerca sui contenuti ed una gestione per categorie di solito poco strutturata. Wikipedia nasce invece in modo inverso, senza vincoli su un qualsiasi ordine e con una gestione delle categorie molto efficace, disponendo quindi di maggiore libertà nell'accesso ai contenuti.

La crescita di Wikipedia è inoltre un circolo senza fine, poiché l'aumento di notorietà porta con sè una maggiore utenza che consulta i contenuti e quindi, essendo modificabile da chiunque, probabilmente un aumento degli autori; questi accrescono i contenuti, aumentando la notorietà di Wikipedia sia come fonte o semplicemente come risultato nei motori di ricerca.

2.3.1 Dati sul progetto

La parte italiana di Wikipedia (itwiki) nell'estate del 2007 era caratterizzata da 120 milioni di parole, composte da 320 mila voci enciclopediche normali, definite come pagine che contengano almeno un collegamento ad un'altra voce e almeno un frase. Erano circa 600 mila le pagine di redirect, con un totale fra tutte le pagine di 7 milioni di collegamenti.

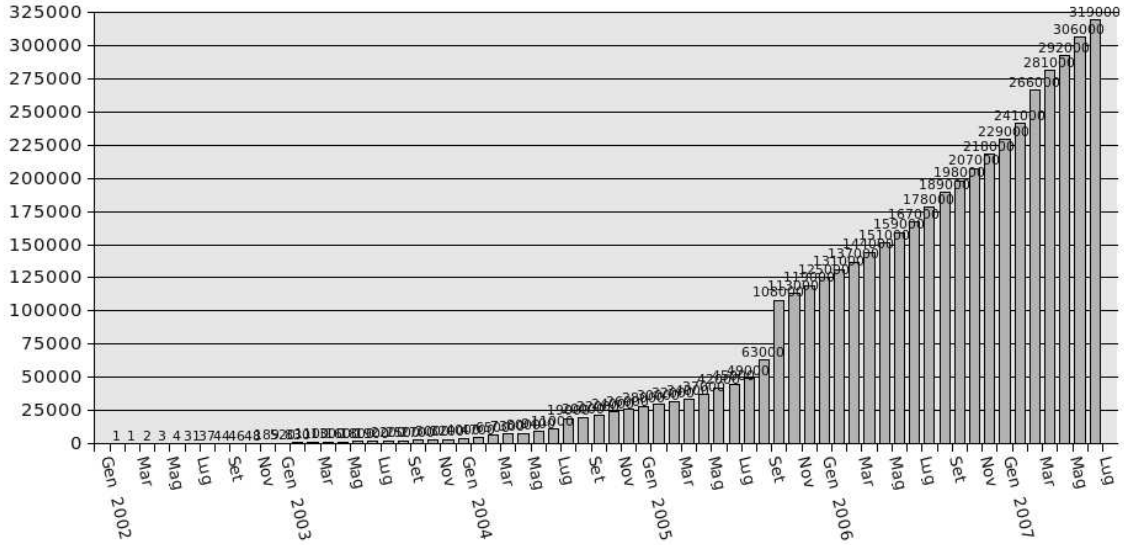

Figura 2.4. Crescita di *itwiki* dalla nascita ad oggi; la scala temporale è rappresentata dalle ascisse, sulle ordinate vi è il numero di pagine normali con almeno una frase ed un collegamento.

La dimensione del database della sola parte italiana si può stimare sugli 80 GB, di cui 1 GB di wikitext dei soli testi delle voci attuali, escludendo quindi cronologie e tutti i metadati.

Le statistiche ed i conteggi dei vari progetti sono minuziosamente riportati in un sito apposito automaticamente agggiornato [36].

A titolo di esempio della notevole crescita di Wikipedia si riporta il grafico (figura 2.4) della crescita delle voci della parte italiana a partire dalla sua nascita in gennaio 2002 fino ad oggi.

Per il confronto con le altre versioni linguistiche di Wikipedia si riporta in figura 2.5 l'andamento congiunto delle maggiori lingue. È possibile notare che l'evoluzione della parte inglese segue un progesso che accellera ad un ritmo maggiore rispetto alle altre lingue.

2.3.2 Progetti derivati

Data la licenza dell'intero progetto Wikipedia, è lecito effettuare una copia di tutti i contenuti e riutilizzarli nel rispetto della GNU FDL. È così possibile duplicare l'intero contenuto od una parte dell'enciclopedia e farne una versione parallela, che eventualmente evolve per conto suo. Su Wikipedia è presente un elenco [38] dei siti che replicano i propri contenuti, alcuni senza neanche rispettare la licenza GNU FDL con la quale sono inzialmente rilasciati i testi.

Tra i siti che utilizzano le voci di Wikipedia, è degno di nota il sopracitato progetto *Citizendium* [12], gestito da Sanger, uno dei fondatori di Wikipedia. Questo progetto

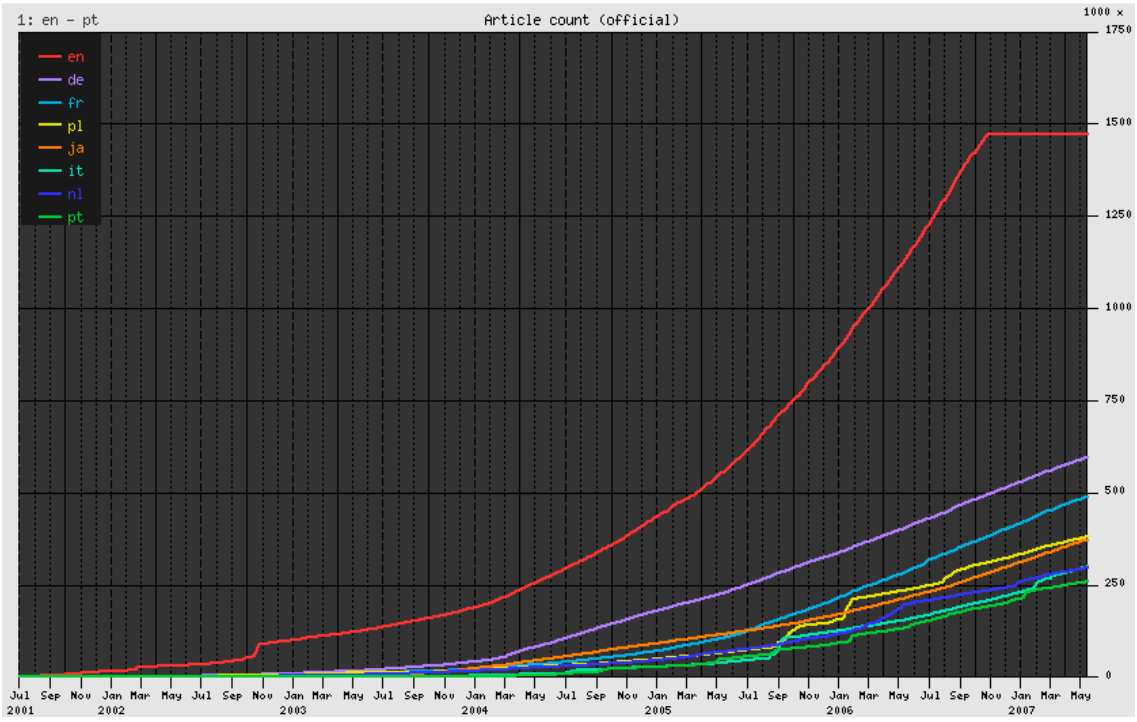

Figura 2.5. Crescita delle otto maggiori lingue di Wikipedia, dalla nascita di Wikipedia nel 2001 ad oggi. Il grafico è tratto da [37].

vuole costruire un'enciclopedia collaborativa con però maggiori controlli sui testi delle voci; le modifiche alle voci infatti devono essere approvate prima di passare nel testo definitivo. Vale nel progetto la stessa licenza GNU FDL di Wikipedia, oppure a volte ne viene applicata una compatibile di tipo Creative Commons Public License (queste licenze sono illustrate in sezione 6.4).

I contenuti iniziali dell'enciclopedia sono stati presi da Wikipedia nel 2005 e da allora Citizendium si è sviluppata in parte in modo indipendente e in parte attingendo da nuovi contenuti di Wikipedia.

Capitolo 3

Il progetto WaNDA

La presenza online di un'enciclopedia sempre più vasta e documentata ha portato ad una notevole notorietà Wikipedia, attirando l'attenzione di un numero sempre crescente di persone che navigano sul web. Ciò ha fatto diventare Wikipedia un punto di riferimento per una buona parte dell'informazione che si può trovare su Internet.

Poter disporre di quest'informazione senza dover dipendere da una connessione alla rete è un'idea molto interessante, che grazie alla qualità delle voci ed allo sviluppo tecnologico risulta sempre più fattibile. Si menziona lo sviluppo tecnologico in quanto nonostante la crescita di Wikipedia sia continua ed elevata, la crescita della tecnologia è superiore; questo fattore è importante data la mole di informazioni da elaborare e la presenza di dispositivi con spazio sufficiente a contenerle.

Il *progetto WaNDA* [39] si occupa di definire dei meccanismi per elaborare le informazioni enciclopediche di Wikipedia in modo da ottenere un formato memorizzabile su un supporto portabile; il supporto portatile permette quindi la consultazione offline dell'enciclopedia, che comprende testi e immagini. Sia i meccanismi di elaborazione, sia quelli di consultazione, sia il formato del supporto si basano su diffuse tecnologie opensource e nel limite del possibile sono indipendenti dalla piattaforma di utilizzo.

Il progetto WaNDA ha prodotto differenti meccanismi per l'elaborazione, di cui l'ultima versione è la più perfezionata ed è giunta ad un livello completo [40].

Il progetto si è anche occupato delle tematiche legali inerenti la diffusione offline di un'enciclopedia i cui autori siano misti ed il più delle volte ignoti, e in cui non vi è un controllo preventivo sui contenuti. Benché sia permesso dalle licenze di Wikipedia quest'utilizzo non è propriamente contemplato dall'attuale ordinamento giuridico italiano.

L'idea di poter consultare l'enciclopedia Wikipedia su calcolatori sconnessi dalla rete Internet utilizzando dispositivi portatili, primo fra i quali il DVD, nasce all'interno dell'ateneo nell'estate 2005 nell'ambito del gruppo *linux@studenti*. Tale gruppo, nella sfera delle attività del Centro di competenza per l'opensource e il software libero *open@polito*, si occupa della distribuzione di sistemi operativi e software opensource, e promuove alcuni progetti per la diffusione del software libero e dei contenuti aperti non soltanto all'interno dell'ateneo.

`linux@studenti` fa capo a Enrico Venuto ed Alessandro Ugo; quest'ultimo è stato il

primo promotore di questo progetto ed è colui che ne ha dato il nome *WaNDA*, acronimo ricorsivo di *WaNDA ancora Non Definitivamente Aggiornata*.

3.1 Scopo e motivazioni

Il progetto WaNDA [39] si è fin dall'inizio posto degli obiettivi ben chiari e precisi; questi sono stati definiti in modo tale da assecondare le motivazioni che hanno dato origine al progetto.

Le due ragioni principali dell'investimento nell'attività del progetto sono:

- Utilizzo nell'istruzione.
  Poter disporre di un'enciclopedia multimediale su supporto fisico con licenza libera permette l'utilizzo nelle scuole a scopo di apprendimento. Oltre ad essere un'ottima opportunità di diffusione della conoscenza, non vi sono spese di sorta nè per l'acquisto nè per la diffusione. D'altronde è un indispensabile strumento nella gran parte delle scuole soprattutto primarie e secondarie dove non è presente una connessione ad Internet.

  Come risvolto positivo alla diffusione delle scuole si aggiunge il miglioramento di Wikipedia stessa, poiché è maggiormente probabile che il docente o lo studente voglia correggere od ampliare una o più voci dell'argomento in esame. Inoltre si diffonde la cultura dell'*open content* e del sapere libero.

- Ridurre il *digital divide*.
  Nonostante gli sforzi per portare connettività Internet in tutta Italia sono ancora molte le zone non coperte; si aggiunge anche il fatto che la necessità dello strumento Internet è tuttora poco sentita tra la popolazione italiana rispetto alla media europea [41].

  Disporre di Wikipedia in formato offline, DVD, chiave USB, palmare o altro supporto che sia, permette la diffusione di un'ottima enciclopedia gratuita. Inoltre viene trasmessa agli utenti l'importanza dell'utilizzo di Internet.

Oltre a questi due punti fondamentali, vi sono sicuramente ulteriori vantaggi nello sviluppo del progetto WaNDA. Quelli che ci sono sembrati i più rilevanti sono:

- Vengono ridotti i tempi di accesso alla versione online. Essendo un progetto senza fini di lucro Wikipedia dispone di un sistema di server che riesce a soddisfare le richieste dei visitatori con poco scarto; i tempi di accesso alle voci sono quindi abbastanza lunghi.

- Si rende disponibile a chiunque una versione di Wikipedia memorizzabile su un generico supporto fisico compatibile con qualsiasi piattaforma. Al pari delle altre enciclopedia moderne si offre quindi a Wikipedia la possibilità di competere efficamente anche sul piano offline.

- È possibile utilizzare i contenuti offline in applicazioni particolari, per esempio dispositivi per diversamente abili.

- Per ultimo è psicologicamente differente poter accedere ai contenuti online e sapere di avere disponibile sullo scaffale in qualsiasi momento un'enciclopedia completa.

Inoltre WaNDA, essendo nato in un ambiente sensibile alle tematiche della libera conoscenza e dell'opensource, si prefigge anche lo scopo di sostenere e diffondere Wikipedia, poiché è un ottimo progetto dotato di grande valore.

Analizzando lo scopo ultimo di WaNDA, ovvero la consultazione offline dell'enciclopedia Wikipedia, e le motivazioni di fondo, si conclude che il progetto WaNDA deve dividersi in due componenti logiche fondamentali: uno strumento per ottenere un formato dell'enciclopedia ed un sistema per poter consultare questo formato; esse verranno descritte in sezione 3.3.

Il primo si compone di diversi programmi multipiattaforma che costruiscono il formato finale contenente l'enciclopedia. È stata la parte in cui si sono incontrate maggiormente difficoltà nello sviluppo, principalmente dovute alla mutabilità di organizzazione dei dati enciclopedici.

Il secondo è un misto di HTML e JavaScript che utilizzando un browser web, incluso per alcune piattaforme, permette l'accesso e la navigazione ai contenuti enciclopedici presenti nel formato finale.

Se si considerano i progetti noti paralleli a WaNDA, elencati alla fine del capitolo, si può notare che anch'essi sono dotati di queste due componenti, siano esse effettuate manualmente o tramite applicativi.

3.1.1 Linee guida del progetto

Le linee guida sono state definite durante la fase di progettazione e seguite durante lo svolgimento del progetto WaNDA e la risoluzione delle varie questioni incontrate. Avendo inizialmente collaborato con Wikimedia Italia, le linee guida del progetto sono state definite anche con il suo contributo. Il progetto parte da un ambiente accademico rivolto all'opensource e attento alle tecnologie multipiattaforma, per cui queste idee ne hanno pesantemente condizionato lo svolgimento.

Le linee guida seguite nello sviluppo del progetto WaNDA sono in parte differenti per le due componenti, poiché sebbene l'intento del progetto è unico differenti sono i requisiti da seguire.

I principi comuni alle due componenti sono:

- Sistema compatibile con il più alto numero possibile di piattaforme, in particolare quelle moderne più diffuse.

- Utilizzo di tecnologie opensource e di libero utilizzo sia per lo sviluppo che per la presentazione dei contenuti.

- Facilità di utilizzo. Questo vale in particolar modo per la presentazione dei contenuti, che dovrebbe essere fedele all'impostazione delle pagine di Wikipedia, per non sconvolgere l'utente e rendere transparente il passaggio dall'uno all'altro. Per quanto riguarda il meccanismo per ottenere il contenitore si deve tenere in mente che chi lo sviluppa non è detto che sia colui che poi lo utilizza.

- Si deve evitare l'obsolescenza futura. Il sistema per costruire il contenitore deve poter essere utilizzato per le future versioni di Wikipedia, che si tratti del formato del *wikitext* oppure dell'architettura di memorizzazione delle informazioni; essendo il futuro ignoto, eventuali mutamenti dovranno richiedere adattamenti minimi. La consultazione dei contenuti deve invece effettuarsi tramite tecnologie affermate e diffuse, e quindi disponibili in futuro.

- Le due componenti devono essere progettate in modo da permettere frequenti aggiornamenti del contenitore.

Le linee guida specifiche al sistema per la consultazione del contenuto enciclopedico sono più numerose poiché questa componente del progetto deve maggiormente raffrontarsi con l'utenza finale. Oltre quindi a quelle già elencate si aggiungono:

- Il contenitore deve poter essere memorizzato con un generico dispositivo hardware.

- La consultazione deve poter essere effettuata grazie ad una generica interconnessione con il dispositivo.

- Possibilmente è da evitare l'installazione di software o hardware aggiuntivo per poter effettuare la consultazione. Questo oltre a facilitare l'utente semplifica il requisito di genericità della piattaforma.

- La consultazione per essere usabile deve essere veloce.

- Il contenitore deve essere "hot plug", nel senso che il dispositivo deve poter essere facilmente connesso alla piattaforma dall'utente ed essere subito utilizzabile.

3.1.1.1 Supporti di memorizzazione

Dalle caratteristiche volute del progetto ne segue che il dispositivo con i contenuti enciclopedici deve essere dotato di accesso casuale o semi-casuale e non deve far parte della piattaforma utente ma sarebbe meglio un dispositivo portabile. Inoltre dall'analisi della quantità di dati da memorizzare la richiesta di spazio sul dispositivo deve all'incirca essere di 4 GB.

Allo stato attuale i dispositivi che possono essere utilizzati sono quindi:

- Supporti ottici: in questa casistica rientrano i DVD ed i DVD Double Layer.

- Memorie a stato solido: qui troviamo le chiavi USB, le schede di memoria Flash, i digital audio player e così via.

- Dischi magnetici: sono di questo tipo i Microdrive, i dischi esterni USB, i dischi IOmega, e così via.

L'utente ha quindi a disposizione un'ampia gamma di dispositivi sui quali può essere memorizzato il contenuto enciclopedico, permettendo quindi di essere collegati alla generica piattaforma utente; questa può andare dal personal computer fisso al notebook, ma anche dal palmare allo smartphone.

3.2 Strade perseguibili

Lo svolgimento del progetto WaNDA è stato molto articolato per vari motivi; inoltre le due componenti del progetto non hanno seguito lo stesso percorso.

Per quanto riguarda il sistema di memorizzazione e consultazione si è giunti abbastanza rapidamente alla definizione del formato quasi definitivo, che poi ha subito piccoli aggiornamenti successivi per correggere alcuni bug e per l'aggiornamento dei testi di presentazione.

Si è partiti con la valutazione delle tecnologie applicabili e quanto fossero multipiattaforma. Questi sono i metodi presi in considerazione con i quali memorizzare le informazioni:

1. Contenere tutti i dati enciclopedici in un database presente sul supporto, consultabile tramite un'applicazione da installare. Tra i vantaggi di questa soluzione vi sono la facilità di implementare un meccanismo di ricerca sui contenuti e lo spazio di memorizzazione sul supporto ridotto. Tuttavia è necessario implementare un'applicazione per la consultazione del database per le diverse piattaforme che si vogliono supportare; la consultazione non è quindi universale. Inoltre viene richiesta l'installazione di un'applicazione, andando in parte contro il requisito di consultazione "hot plug".

   Questa strategia è generalmente impiegata nei prodotti commerciali per l'accesso a contenuti multimediali, come per esempio le enciclopedie su DVD commerciali.

2. Contenere tutti i dati enciclopedici in un database a cui si accede tramite un'interfaccia presente sul supporto. È una soluzione che risolve alcuni limiti della precedente. Per esempio si può implementare l'interfaccia di navigazione con il linguaggio ad alto livello Java; le piattaforme di consultazione devono quindi essere dotate di JVM per eseguire l'applicativo di consultazione. Questa soluzione non richiede un'installazione per la consultazione, ma tuttavia richiede la presenta di un ambiente di esecuzione che, pur essendo disponibile per molti sistemi, non è veramente multipiattaforma.

3. Utilizzare un insieme organizzato di file, in cui ognuno contiene informazioni su una voce. Qui la piattaforma per effettuare la consultazione riesce immediatamente ad accedere ai file, non richiedendo un mezzo per accedere ai contenuti. Il formato dei file deve essere il più standard possibile ed offrire caratteristiche di formattazione avanzate; una buona scelta è l'HTML, sia poiché tutte le piattaforme sono dotate di

un'interfaccia per tale formato (detto comunemente *browser*), sia perché è il formato naturale del progetto Wikipedia.

Questa soluzione tra i vantaggi ha la massima compatibilità; gli svantaggi la staticità implicita del formato, per cui è più difficile implementare un meccanismo di ricerca sui contenuti.

La strada scelta nel progetto WaNDA è la terza. Per la ricerca sui contenuti anche qui è possibile scegliere tra differenti tecnologie per l'esecuzione di codice attivo; la più diffusa ed immediatamente disponibile assieme alla maggior parte dei browser è JavaScript.

In futuro potrebbe essere una buona alternativa la seconda soluzione, utilizzando un programma Java come interfaccia di accesso ai contenuti. Potrebbe anche essere un'evoluzione plausibile l'integrazione della terza soluzione con un programma Java che implementi un motore di ricerca. Inoltre come esempio è stato preso Java, ma potrebbe anche essere utilizzato un altro linguaggio interpretato disponibile su più piattaforme come Python, Ruby o Tcl; l'interprete Java è però maggiormente diffuso. La soluzione Java è stata scartata a causa della licenza d'uso della JVM storicamente non libera; tuttavia ultimamente è stata avanzata l'ipotesi di rilasciare l'implementazione della JVM con licenza libera, rendendo quindi conforme il Java ai requisiti di progetto.

Una volta che sono state vagamente definite le caratteristiche del supporto che contiene le informazioni, è necessario definire il meccanismo di traformazione dei dati per ottenerlo. L'approccio non è deliberatamente scelto, ma è dettato dalla tecnologia utilizzata da Wikipedia e dal formato delle informazioni enciclopediche che essa rilascia. In questo senso il percorso di sviluppo del programma è stato tortuoso dati i differenti formati disponibili. Inoltre la trasformazione della grande mole di dati enciclopedici è un'operazione lunga che ha sicuramente dilatato i tempi nella ricerca della soluzione ottimale.

A prima vista sembrerebbe possibile effettuare una banale copia locale delle pagine HTML tramite uno script apposito; tuttavia questa soluzione lascia irrisolti tutta una serie di problemi, per esempio il fatto che tutti i collegamenti in una pagina sono validi anche se la pagina di destinazione ancora non esiste. Inoltre questo comportamento è dissuaso dalle policy di utilizzo di Wikipedia che può bannare l'indirizzo IP, poiché per la struttura di Wikipedia a livello di server un'accesso consecutivo all'intero insieme delle voci è molto oneroso.

Per l'elaborazione sono quindi disponibili vari dump in formato XML, che oltre ad essere pratici come backup, sono semplici per uno sviluppatore di un progetto poiché offrono con un formato standard i contenuti: sono tuttavia disponibili dal 2006. Precedentemente erano disponibili i dump direttamente SQL di Wikipedia.

Inoltre sono fruibili da novembre 2005 le copie delle pagine di Wikipedia, sotto forma di pagine HTML raggruppate per directory e sottodirectory; questi file fanno parte del cosiddetto "static dump" [42], nato per adeguarsi maggiormente alla licenza GNU FDL che raccomanda la disponibilità intera dei contenuti, dato che il download diretto è scoraggiato. Inoltre il dump statico, per motivi di carico, era effettuato poco più di una volta l'anno.

Un'elaborazione che parte da pagine HTML per ottenere pagine HTML è molto onerosa e poco flessibile dato che si deve costruire uno strumento di parsing sul testo; un'elaborazione che parte dal *wikitext* è più diretta, dato che esso è già predisposto per una conversione

all'HTML. Inoltre un parser grammaticale sul formato HTML dove essere rivisto ed adattato ad ogni versione di Wikipedia poiché come si è sperimentalmente verificato esso varia di versione in versione di MediaWiki e delle sue estensioni.

3.2.1 Percorso di sviluppo del progetto

Inizialmente le prime versioni di WaNDA erano prodotte in stretta collaborazione con Wikimedia Italia.

Il programma di trasformazione dei contenuti enciclopedici partiva dagli "static dump"; questi dump sono l'insieme delle voci di Wikipedia sotto forma di pagine HTML simili a quelle accessibili online, raccolte in tre livelli di directory. Il programma era scritto in parte da G. Ceresa di Wikimedia Italia in Java e caricava uno ad uno questi file, per poi effettuare dei filtri sui testi e conversioni sui collegamenti HTML. Alla fine ogni pagina era scritta su una pagine HTML finale con un nome del file appropriato; l'insieme della pagine era formattato in modo tale da poter essere memorizzato su un CD oppure un DVD.

Quest'elaborazione, oltre ad essere molto lenta poiché doveva effettuare parsing di testo HTML, non elaborava alcuna immagine. I collegamenti alle immagini non erano elaborati e tutti i contenuti grafici con licenza opportuna venivano scaricati una tantum e aggiunti alla fine dell'elaborazione; questo provocava per esempio la presenza di didascalie per immagini mancanti.

È in questa versione che viene sviluppato il motore di ricerca in JavaScript ancora attualmente utilizzato. L'indice di ricerca e le pagine di navigazione (per dettagli sul funzionamento del motore di ricerca si veda la sezione 4.3.3) erano generati grazie a degli script shell a partire da una lista delle voci generata nella fase di conversione dal programma in Java.

Questa versione del progetto ha prodotto, seppur con alcuni difetti, alcune versioni accettabili dell'enciclopedia offline con il primo dump statico (fine 2005) e alcuni successivi. Il problema principale di questa versione del progetto era nella manutenzione problematica, poiché, per i dump statici successivi, le modifiche da apportare alla struttura del programma erano ogni volta notevoli.

Nel frattempo aumenta l'interesse per il progetto e viene effettuata una versione di WaNDA per sistemi embedded per non vedenti; il nome di questa versione è *WaNDA-nv*.

Il programma Java è leggermente modificato in modo da ottenere un formato delle pagine testuale e con una gestione particolare dei collegamenti. Il testo è quindi interpretato da un sintetizzatore vocale.

Con la diffusione sempre maggiore di Wikipedia vi è sempre più l'esigenza di un formato di Wikipedia standard; nasce il dump XML che è portabile contrariamente ai dump SQL. Il dump XML di Wikipedia verrà aggiornato inizialmente ogni due-tre mesi, per poi essere aggiornato continuamente in ciclo su tutti i progetti Wikimedia: ad oggi il dump di Wikipedia versione italiana è quindi aggiornato ogni mese, salvo imprevisti quali problemi di hardware.

Con l'arrivo di questo nuovo formato viene sviluppata una versione di transizione del processo di elaborazione, che utilizza il dump XML maggiormente aggiornato per

generare il "static dump" in locale utilizzando MediaWiki; a questo punto si lavora con il programma Java come al solito, che però non deve essere ritoccato di versione in versione dato che il MediaWiki utilizzato è sempre lo stesso.

In questa versione viene sviluppato il filtro sul dump XML per ridurne le dimensioni e velocizzare in processo. Inoltre gli indici per il motore di ricerca in JavaScript non vengono più generati alla fine dell'intero processo con lo script shell ma durante l'utilizzo di MediaWiki per la costruzione del "static dump". Nasce quindi uno script PHP da aggiungere a MediaWiki.

In questa versione manca ancora l'elaborazione delle immagini e mancano le categorie. Inoltre ottenere il formato finale è ancora un processo complesso e laborioso.

3.2.2 Progetto finale

La versione definitiva di WaNDA elimina la componente in Java del progetto: il programma di conversione del dump XML è interamente implementato in PHP con un programma che in parte si basa su MediaWiki. È stata quindi compiuta una semplificazione ed unificazione dei passi del processo di conversione. La soluzione della conversione sul testo HTML non era mantenibile poiché il formato variava da versione a versione; la soluzione attuale è invece ragionevolmente compatibile con le versioni future di Wikipedia ed eventuali incompatibilità sono facilmente superabili.

Il programma provvede alla conversione del *wikitext* utilizzando il traduttore di MediaWiki stesso, si occupa automaticamente della generazione di un indice per il motore di ricerca JavaScript e gestisce in modo intelligente le immagini da includere secondo la loro licenza d'uso.

Si noti che molte parti del programma PHP non sono state progettate da capo ma ci si è basati sul funzionamento di alcune delle parti scritte in Java e degli script shell.

Per il funzionamento il programma utilizza un database nel quale registra varie informazioni durante l'interpretazione del *wikitext*, utili alla fine della conversione delle pagine per la gestione degli indici e delle immagini.

Dopo l'esecuzione del programma, il formato finale è pronto; i tempi sono quindi notevolmente ridotti. Inoltre il modello di funzionamento permette di eseguire filtri sui contenuti enciclopedici come prima operazione del processo.

Il formato contenente le voci enciclopediche, già ben progettato dall'inizio, non è stato più modificato se non per alcune correzioni al testo di alcune pagine di presentazione.

Dopo la definizione di questa versione, che è stata chiamata *WaNDA-ng*, nel tempo sono seguite alcune rivisitazioni dei passi e del codice in modo da ottimizzare le elaborazioni e quindi i tempi di esecuzione.

3.2.3 Traduzione del wikitext

Il linguaggio wikitext è un linguaggio per la formattazione di testo che vuole semplificare la diretta composizione di testo HTML, al fine di essere facilmente imparato ed utilizzato da un vasto insieme di utenti nei software wiki. Esso permette una grande varietà di

elementi di formattazione; per esempio è agevole la composizione degli elenchi, oppure un collegamento interno al software viene automaticamente riconosciuto. Il software wiki provvede quindi ad interpretare (o meglio tradurre) il wikitext scritto dagli utenti e a generare il testo HTML corrispondente.

Non esiste uno standard sul formato del wikitext: ogni software, come in questo caso MediaWiki, utilizza una sua sintassi. Inoltre a volte in versioni successive del software tale sintassi viene ampliata o modificata; anche l'utilizzo di estensioni in MediaWiki può ampliare o modificare la sintassi del suo wikitext. Nell'ambito del progetto WaNDA con wikitext si indende genericamente il formato utilizzato da MediaWiki.

Nel database di MediaWiki i testi delle voci sono presenti come wikitext, poiché esso viene interpretato diversamente secondo lo stato attuale: per esempio la traduzione avviene diversamente se la destinazione di un collegamento è presente o no, oppure se l'utente è registrato, o anche a seconda delle estensioni presenti. Di conseguenza anche i dump XML contengono il wikitext dei contenuti enciclopedici.

I progetti come WaNDA, che devono adattare il testo a particolari esigenze, utilizzano quindi dei traduttori dal linguaggio wikitext verso il formato desiderato. Questo può essere HTML, secondo una sintassi particolare, oppure tutt'altro formato testuale. All'indirizzo [43] è presente un elenco dei software noti a Wikimedia che implementano un qualsiasi tipo di traduttore del wikitext (in inglese tecnico *parser*). Chiaramente questi software non sono ufficiali e sono alternativi al traduttore principale, che è MediaWiki stesso ed è il punto di riferimento. Questi traduttori possono avere vari utilizzi, anche se in genere lo scopo della conversione è ottenere un formato per la consultazione offline.

Pur essendo le specifiche libere e riportate all'indirizzo [44], il linguaggio wikitext subisce variazioni a causa di modifiche delle estensioni e di MediaWiki; questo costringe ad una regolare messa a punto dei traduttori alternativi. La conseguenza è che molti dei traduttori riportati sono incompleti o formattano volutamente un insieme specifico di elementi.

Per evitare di dover sia implementare che mantenere un traduttore, nel progetto WaNDA si è pensato fin dall'inizio di sorvolare il problema, utilizzando in qualche modo MediaWiki; esso infatti è il traduttore ufficiale, che oltre ad essere già presente, è sicuramente completo ed aggiornato.

Nasce però il problema di adattare la conversione effettuata da MediaWiki secondo le proprie necessità. All'inizio del progetto si era pensato di utilizzare i dump statici, che erano generati con MediaWiki; ci si è però successivamente accorti che l'HTML generato era molto variabile e la conversione non era mantenibile nel tempo. La soluzione è stata quindi successivamente scartata in favore di un adattamento diretto del traduttore, con la scrittura di una componente che si aggiunge a MediaWiki.

3.3 Funzionamento

Prima di proseguire è necessaria una chiarificazione sui nomi: il *progetto WaNDA* attualmente si occupa dell'ultima implementazione, indicata con il nome *WaNDA-ng*. Esso comprende alcuni programmi raccolti sotto il nome di *WaNDA-tools*: principalmente il

programma PHP da integrare con MediaWiki, ma anche un filtro scritto in AWK per ridurre la dimensione a priori del dump XML, alcuni programmi per una verifica finale e vari file di informazioni da seguire per configurare l'ambiente.

Come già ribadito, il progetto WaNDA si divide in due componenti: il formato contenente l'enciclopedia, da memorizzare su un dispositivo, e un sistema per ottenere il formato finale, composto dai programmi di WaNDA-tools che sostanzialmente permettono di estrarre i contenuti da un dump XML.

3.3.1 Il formato finale

Il formato finale contiene tutte le voci enciclopediche sotto forma di file HTML, memorizzate in sottodirectory secondo la lettera iniziale del titolo delle voci. Vi sono altri file di tipo HTML che presentano i contenuti, forniscono informazioni utili e contengono i disclaimer.

Oltre ai file HTML il formato contiene dei fogli di stile per la presentazione delle pagine HTML evitando ridondanza e dei codici JavaScript che implementano le funzionalità attive del supporto; la più importante di queste è il motore di ricerca, ma vi sono anche funzioni aggiunte come il cambio di tema ottimizzato per la stampa e l'estrazione casuale di una voce enciclopedica.

Inoltre il formato può includere alcuni browser eseguibili su certe piattaforme, dove magari di default non vi è un browser pienamente standard oppure per diffondere l'opensource. Inoltre può rendere più agevole per l'utente finale l'apertura automatica dell'enciclopedia.

Il formato finale nell'ambito del progetto viene chiamato *albero* [1]; esso si compone di due parti che vengono costruite in modo del tutto differente:

- Le pagine HTML di presentazione, dette *pagine di servizio*, i fogli di stile, i codici JavaScript e gli eventuali eseguibili sono precostruiti e non richiedono elaborazione. Fanno parte del pacchetto WaNDA-tools e possono eventualmente richiedere una modifica a mano, ma non vengono modificati durante il processo di estrazione. Questa componente viene chiamata *ossatura* [2].

- Le pagine HTML contenenti le voci enciclopediche, tutte le immagini, la base per il motore di ricerca e le pagine di indice che vengono generate durante il processo di estrazione.

3.3.2 Il processo di conversione

Il processo di estrazione, o di conversione, permette di compiere la trasformazione dal dump XML all'*albero*, utilizzando i programmi di WaNDA-tools. È presente in figura 3.1 un grafico che rappresenta a grandi linee i passi effettuati nel processo di conversione.

Il processo si compone di due fasi:

[1] D'ora in avanti quando si scriverà "albero" si intende sempre questa componente del progetto

[2] D'ora in avanti quando si scriverà "ossatura" si intende sempre questa componente del progetto

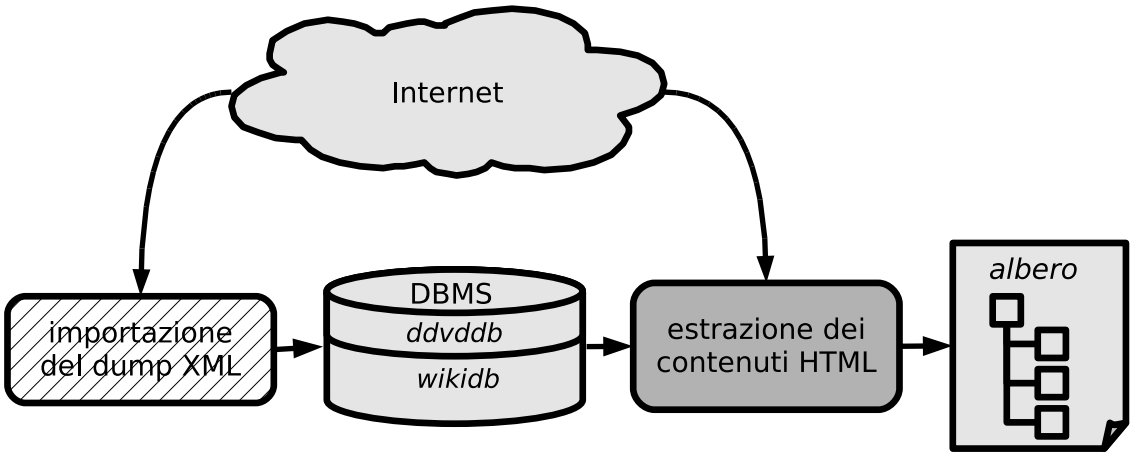

Figura 3.1. Schema generale delle componenti ed operazioni per la conversione dei contenuti enciclopedici dal dump XML di Wikimedia alla struttura HTML finale utilizzando WaNDA-tools

- *Importazione* del dump XML in un database locale. Questa fase prevede quindi il download del dump XML dai server di Wikipedia, l'esecuzione del filtro AWK (che fa parte di WaNDA-tools) su di esso per ridurne la dimensione, e l'inserimento dei dati del dump in una base di dati chiamato convenzionalmente *wikidb*.

- *Esportazione* dell'*albero* partendo dai dati del database locale wikidb. I contenuti enciclopedici vengono estratti ed interpretati dal programma PHP (il principale programma di WaNDA-tools) con l'ausilio di MediaWiki, per poi essere scritti al loro posto nell'albero. Il programma PHP si occupa anche di generare la base di ricerca in JavaScript e le pagine HTML di indice, utilizzando un database di supporto chiamato convenzionalmente *ddvddb*. Utilizzando questo database il programma si occupa anche di ottenere dal sito di Wikimedia le immagini da includere nell'albero.

Tutti i programmi di WaNDA-tools sono eseguiti da linea di comando ed utilizzano tecnologie opensource portabili su molti sistemi differenti. Il programma PHP è adattabile a molte esigenze ed è dotato di un completo file di configurazione. Inoltre ha tutta una serie di opzioni che possono essere utilizzate per effettuare a pezzi le diverse operazioni elencate, oppure per escluderne alcune. Questo può essere comodo per esempio nella fase di esportazione per escludere il download delle immagini ed effettuarlo in un secondo tempo, oppure per eseguire su due calcolatori diverse parti del processo totale.

3.4 Valutazione del progetto e progetti paralleli

Nel corso degli anni, soprattutto dopo la diffusione capillare di Wikipedia, sono nati vari progetti in qualche modo simili progetto WaNDA. Per progetti paralleli si intende

quei progetti software il cui scopo sia la consultazione parzialmente o totalmente offline dell'enciclopedia Wikipedia.

Il 2006 è stato l'anno in cui sono hanno iniziato a circolare sulla rete informazioni circa la maggior parte dei progetti paralleli; il motivo di tanta attesa, considerato il fatto che Wikipedia in inglese esiste dal 2001 (dal 2003 la Wikimedia Foundation), è probabilmente legato sia a fattori tecnologici che divulgativi.

Per quanto riguarda i fattori tecnologici, da una parte era vietato dalle politiche di Wikipedia il download consecutivo di tutte le pagine. Dall'altra la conversione dal formato SQL richiedeva un'installazione di MediaWiki e una conversione composta da vari passi (come effettivamente il progetto WaNDA). È stato possibile lavorare più facilmente dal momento in cui sono stati rilasciati i dump in formato XML come pratici backup di Wikipedia; è possibile infatti elaborare i contenuti direttamente dall'XML.

Un fattore inoltre da tenere sempre presente è la mole di dati che viene elaborata; benchè la crescita di Wikipedia sia continua e pare inarrestabile, la tecnologia informatica cresce a ritmi ancor maggiori: la grande quantità di dati un tempo difficile da gestire è sempre più agevolmente trattabile.

L'affermazione di Wikipedia nel mondo dell'informazione ha portato ad un interesse maggiore da parte di individui e di enti verso l'enciclopedia; per esempio addirittura enti attendibili come quelli governativi od accademici propongo sempre più spesso collegamenti all'enciclopedia. Da una parte questo ha portato ad una regolamentazione maggiore sui meccanismi interni di Wikipedia, rafforzandone la struttura, e dall'altra a procedure chiare di interfacciamento con gli altri enti.

A questo punto è stato quindi più facile per un ente commerciale investire in progetti che si basino sui contenuti di Wikipedia; la diffusione stessa di Wikipedia aiuta un ente che vuole distribuire a pagamento l'enciclopedia poiché gli assicura un profitto tale da poter coprire eventuali spese legali riguardanti eventuali contenuti discutibili.

Infine l'utilizzo offline di contenuti disponibili sulla rete spesso porta al cambiamento degli ambiti legislativi di competenza. In molti Stati, quali ad esempio l'Italia, le tematiche legislative su questo tipo di conversione non sono chiare e definite. Questo porta inevitabilmente ulteriore ritardo nello sviluppo dei progetti quali WaNDA e quelli paralleli. Gli aspetti legali ed economici di questi progetti verranno approfonditi nel capitolo 6, dove viene preso in esame WaNDA.

Segue una analisi dei progetti paralleli noti a metà del 2007. Non vi sono progetti identici a WaNDA, poiché differenziano per approccio e risultato; il progetto WaNDA ha quindi tuttora ragione di continuità poiché presenta caratteristiche uniche che gli attribuiscono un elevato prestigio.

I progetti sono divisi a seconda del loro principale ambito di collocamento, se opensource o comunque unicamente gratuiti oppure principalmente commerciali.

3.4.1 Progetti gratuiti

- *MoulinWiki* [45] è un progetto opensource che ha come obiettivo primario la diffusione dell'enciclopedia in francese nella regione dell'Africa occidentale, dove è molto

scarsa la diffusione della connettività ad Internet; il progetto riceve finanziamenti da diversi enti governativi.

Il contenuto dell'enciclopedia (per ora esiste solo la versione francese) è presente su un CDROM a cui si accede tramite un'interfaccia ivi contenuta. Le pagine sono soltanto testuali e sono quindi prive di grafica ed immagini (mancano le formule matematiche). Sul CDROM è contenuto il dump XML di Wikipedia compresso, un database minimale per una ricerca rapida dei contenuti sul dump ed un web server minimale per accedere ai contenuti. I binari necessari all'esecuzione del database e del web server sono disponibili per tre sistemi operativi: Windows XP, GNU/Linux e Mac OSX, tutti su piattaforma Intel x86. Oltre a questi è incluso un browser minimale per l'accesso al web server eseguito dal CDROM. Conseguentemente al meccanismo di funzionamento sono necessari privilegi di amministratore.

Il progetto è ancora in fase di sviluppo, l'immagine ISO delle versioni di prova della versione francese sono disponibili dall'homepage del progetto.

- *Wikipedia CD Selection* [46] è una selezione annuale delle migliori voci enciclopediche. La prima versione è stata rilasciata nel 2006 ed era originariamente sviluppata dall'organizzazione *SOS Children* [47] per la diffusione dell'enciclopedia su CD nei paesi del terzo mondo. Il progetto è stato appoggiato da Wikimedia stessa che ne ha preso in parte l'incarico ed ha rilasciato la versione del 2007 con le pagine di maggio.

  La selezione comprende sia il testo che la grafica: le voci vengono selezionati in modo da essere adatte ad un utilizzo scolastico di base (l'età considerata va da 8 a 15 anni). I testi delle voci sono rivisti da un gruppo di volontari; le immagini sono scelte secondo la licenza d'uso ed il loro contenuto. Le voci vengono inoltre categorizzate secondo raggruppamenti che aiutano un percorso di apprendimento.

  Le voci sono presenti sotto forma di singole pagine HTML con un layout rivisitato, che permette la navigazione su qualsiasi piattaforma. La navigazione procede soltanto tramite i numeri indici e categorie creati dai volontari.

  La selezione comprende 4655 voci e circa 8000 immagini. È possibile visitare online la selezione del 2007 al sito [48]. Il download dell'immagine ISO per DVD è consentito solo per una versione da 800MB che contiene le immagini ridotte; la versione intera viene distribuita gratuitamente sotto forma di DVD dagli uffici di *SOS Children*.

- *Encyclopodia* [49] è un software che permette l'installazione dell'enciclopedia sugli Apple iPod [50] dotati di schermo.

  Per la consultazione è richiesta l'installazione sul dispositivo di un progamma rilasciato sotto licenza GPL che permette la consultazione di *eBook* in un formato documentato; l'enciclopedia viene quindi distribuita sotto forma di *eBook*. Il lavoro di conversione dal dump XML al formato eBook viene effettuato ogni 6 mesi dal manutentore del progetto.

  L'enciclopedia è rilasciata in differenti lingue; non sono presenti immagini e la conversione del *wikitext* lascia testo spurio. L'accesso ai contenuti avviene tramite una

ricerca eseguita dal lettore installato sull'IPod. Assieme al lettore di eBook per l'IPod ed agli eBook viene anche rilasciato un programma di installazione per la facile installazione del lettore per Windows e GNU/Linux.

- Un semplice progetto opensource nato nel 2007 [51] di nome "Building a (fast) Wikipedia offline reader" permette l'accesso al dump XML compresso utilizzando un computer.

  L'accesso ai contenuti enciclopedici avviene da riga di comando tramite un motore di ricerca scritto in Python; l'utente sceglie la voce tra quelle trovate ed essa si apre nel browser predefinito. Il motore di ricerca richiede una fase di preparazione composta dall'esecuzione di un programma scritto in C che costruisce un indice dei titoli per poter rapidamente orientarsi nel dump XML che rimane compresso. Per la traduzione del *wikitext* esso utilizza Mediawiki con una patch [52] sviluppato da un'associazione tedesca di nome Free Software Lab; non sono presenti le immagini, ma tuttavia le formule matematiche sono generate al volo da LaTeX .

  Si noti che tutti i programmi utilizzati dal progetto sono opensource e disponibili per ogni piattaforma. L'accesso è basilare, macchinoso e non è distribuito su supporto rimovibile, tuttavia è quello che ha più punti di contatto con il progetto WaNDA.

3.4.2 Progetti commerciali

- *TomeRaider* [53] è probabilmente il primo progetto per la consultazione offline di Wikipedia versione inglese. Esso è prodotto da Yadabyte, che si occupa di vendere sia un software di consultazione che le versioni di Wikipedia da installarvi. Il formato utilizzato viene indicato come "TomeRaider database" [54]; l'enciclopedia è presente in versioni sia con i contenuti grafici che senza; sono inoltre presenti delle versioni gratuite prodotte da Wikipedia http://download.wikimedia.org/tomeraider/.

  Il software proprietario di consultazione è stato sviluppato specificatamente per i palmari ed permette la consultazione di moltissimi ebook e contenuti multimediali. Le prime versioni di Wikipedia inglese sono state prodotte nel periodo in cui necessitava poco spazio; con lo spazio a disposizione oggi sui dispositivi portatili le versioni complete di immagini di Wikipedia sono sempre meno problematiche.

- *Wikipediaondvd* [55] è un progetto portato avanti da LinterWeb [56] per la generazione di un CD-ROM che contiene voci scelte della versione inglese.

  Si accede alle voci, circa 2000 con i contenuti grafici, grazie ad un programma opensource di nome Kiwix [57]; il programma è molto flessibile e oltre ad offrire un motore di ricerca permette l'accesso alla cronologia delle voci e la configurazione da parte dell'utente dell'interfaccia. Nel CD-ROM viene fornito compilato per piattaforme Intel x86 e sistemi operativi Windows, Mac OSX e GNU/Linux.

  Le pagine sono selezionate e controllate da un gruppo di volontari, in parte gli stessi del progetto *Wikipedia CD Selection*.

Il CD-ROM è sia in vendita da LinterWeb, sia navigabile online, che scaricabile sotto forma di immagine ISO compressa.

- *Directmedia Publishing* si occupa della vendita di una versione tedesca di Wikipedia su DVD-ROM Double Layer. L'azienda ha sviluppato un meccanismo proprietario per la generazione del contenuto; ogni anno viene generata una nuova versione. Si noti che l'enciclopedia viene fornita in toto senza eseguire selezioni.

  La consultazione delle voci nel formato generato avviene invece tramite un software da installare sul PC rilasciato sotto licenza GPL di nome *digibib* [58]; tale software era già impiegato per la consultazione di altri libri in formato digitale distribuiti dall'azienda (è il suo core business).

  Il software *digibi* da installare per la consultazione del DVD DL è rilasciato dall'azienda pronto per essere utilizzato su molte piattaforme tra le quali Windows, GNU/Linux, Mac OSX e BSD.

  L'azienda ha rilasciato gratuitamente per il download l'immagine ISO del DVD delle edizioni vecchie; l'immagine è presente in due versioni, l'una da oltre 8 GB per DVD Double Layer che comprende le immagini e l'altra per DVD normale senza immagini.

  Questo progetto è stato il primo a porre i contenuti in modo offline su un supporto ottico e il suo sviluppo è iniziato alla fine del 2005.

- *EXA* è un'azienda italiana che sviluppa e vende contenuti multimediali; nel 2007 ha prodotto la versione italiana di Wikipedia [59] utilizzando tecnologie proprietarie.

  Anch'essa utilizza quindi un DVD-ROM Double Layer occupando 7 GB di spazio tra testi ed immagini; la consultazione non richiede l'installazione di un applicativo ma è effettuata da binario presente sul supporto.

Capitolo 4

Architettura

La principale applicazione di WaNDA-tools è il programma PHP, in quanto le altre componenti gli fanno da corredo. In questo capitolo vengono illustrate le componenti del programma principale, definendone nel dettaglio la struttura. Sono anche descritte approfonditamente le caratteristiche dell'*albero*, sia delle parti statiche (l'ossatura) che in particolar modo delle pagine esportate.

Il capitolo è quindi diviso in tre parti: la prima descrive le classi e i metodi del programma PHP con le indicazioni delle correlazioni fra di loro e con MediaWiki; sono inoltre descritte le componenti dell'ossatura poiché sono incluse sotto forma di archivio nel programma PHP.

La seconda parte descrive la gestione dei contenuti da parte del programma, ovvero i meccanismi ed i formati utilizzati nel programma PHP per poter effettuare l'elaborazione. In particolare si descrive il funzionamento del filtro sui contenuti, la gestione delle immagini e degli autori per ogni voce, la codifica necessaria per adattare il nome dei file e le modifiche che vengono apportate al testo HTML.

L'ultima parte presenta il formato dell'albero, descrivendo il formato delle pagine HTML, il loro foglio di stile ed il funzionamento dei codici attivi in JavaScript; in particolare si illusta il funzionamento del motore di ricerca. Si analizzano inoltre due formati di memorizzazione dell'albero sul supporto, l'uno per dischi ottici e l'altro per memorie solide.

4.1 Componenti del programma

L'applicativo principale di WaNDA-tools richiede, per essere utilizzato, di una installazione completa del software MediaWiki; è quindi forse meglio definirlo un'estensione di MediaWiki che ne aggiunge nuove funzionalità. Tuttavia data la natura logica del progetto, in cui WaNDA-tools sfrutta MediaWiki per effettuare l'estrazione ed il parsing delle voci, si considera MediaWiki una dipendenza di WaNDA-tools.

WaNDA-tools è scritto prevalentemente utilizzando il linguaggio PHP, scelta non dettata da considerazioni prestazionali o architetturali ma piuttosto per facilitare l'integrazione di MediaWiki e non spaziare su troppe tecnologie differenti fra di loro. Inoltre il PHP

ha una vasta base di utenti essendo basato sullo stile del diffuso linguaggio C/C++, con modifiche tratte dal linguaggio Perl per dotarlo di capacità di scripting.

Per le funzionalità dinamiche presenti sul DVD o sul filesystem per la navigazione offline, il linguaggio scelto è JavaScript, in quanto essendo stato introdotto oltre dieci anni fà ha acquisto una diffusione ed una standardizzazione sufficiente. Infatti praticamente ogni browser grafico moderno riesce ad interpretare il codice di WaNDA, scritto utilizzando soltanto funzioni standard per ogni implementazione.

WaNDA-tools viene rilasciato sotto forma di archivio compresso, che contiene alcuni file organizzati in directory. I file che devono essere integrati con MediaWiki sono presenti in una directory di nome *maintenance*, poiché per l'installazione devono essere copiati nella directory di MediaWiki con lo stesso nome.

Questi file sono nove codici in linguaggio PHP e due archivi:

- `dumpDVD.php` - file di accesso a WaNDA-tools, l'unico con estensione `.php`;
- `dumpDVD.inc` - contiene la classe principale;
- `dumpDVDdb.inc` - è il punto di accesso alla base dati di servizio;
- `dumpDVDtext.inc` - contiene le funzioni per esportare il testo HTML delle voci;
- `dumpDVDhistory.inc` - gestisce la lista degli autori delle voci;
- `dumpDVDindex.inc` - elabora la lista delle voci per creare gli indici di ricerca offline;
- `dumpDVDlog.inc` - funzioni di base per presentare all'utente informazioni utili;
- `dumpDVDfilter.inc` - include i filtri da eseguire sul database locale di Wikipedia, è modificabile dall'utente;
- `dumpDVD.ini` - file di configurazione modificabile dall'utente;
- `dumpDVD/wnd_html.tar.bz2` - archivio contenente le pagine di servizio, come la pagina iniziale ed gli script JavaScript;
- `dumpDVD/wnd_bin.tar.bz2` - archivio contenente gli eseguibili per Windows per la navigazione offline con strumenti opensource.

4.1.1 Componenti PHP di WaNDA-tools

Il file di accesso `dumpDVD.php` non contiene una classe con dei metodi, ma è impostato come successione di comandi. Questo implica che il file PHP è gestito come fosse uno script od un programma batch.

Per prima cosa si effettua il caricamento delle impostazioni dal file di configurazione `dumpDVD.ini`.

Il file di configurazione è diviso in tre sezioni:

- `release infos` - impostazioni da modificare ogni volta che si effettua un'estrazione;
- `path infos` - definizione dei percorsi;
- `advanced infos` - opzioni avanzate, da modificare solo per esigenze particolari.

Il file principale continua quindi istanziando le varie classi che saranno utilizzate nel programma; esse sono elencate nelle sottosezioni successive.

A riga di comando `dumpDVD.php` può essere invocato con o senza argomenti; senza esso effettua cinque operazioni per l'estrazione completa dei contenuti enciclopedici. Gli argomenti si dividono in due gruppi: i comandi esclusivi, che permettono di eseguire alcune specifiche operazioni in modo indipendente; le opzioni di estrazione, che modificano lo svolgersi delle cinque operazioni.

I comandi esclusivi si escludono l'un l'altro: `--makeiso`, crea la ISO del DVD; `--filterdb`, esegue i filtri sul database locale; `--getimg`, scarica le immagini; `--makeimglist`, aggiorna la lista delle immagini date le categorie impostate; `--copyimg`, copia soltanto le immagini già scaricate.

Le opzioni di estrazione possono essere abbinate: `--nomakeimg`, salta l'elaborazione delle immagini; `--nomergedb`, forza la creazione di un nuovo database di servizio; `--nofilterdb`, non esegue i filtri sul database locale; `--idstart=NUM`, imposta l'id della pagina da cui iniziare l'estrazione (implica `--nomergedb` e `--nofilterdb`); `--idstop=NUM`, imposta l'id della pagina con cui terminare l'estrazione ed esce (implica `--nomergedb` e `--nofilterdb`).

I dettagli sull'utilizzo delle opzioni sono descritti nel capitolo 5.2.3. A seconda dei comandi esclusivi specificati si procede ad eseguire i metodi oppure le chiamate a sistema necessarie a svolgere i compiti.

Il file PHP procede quindi all'esecuzione normale delle cinque fasi per l'estrazione completa delle informazioni, tenendo eventualmente conto delle sopracitate opzioni. Si noti che prima di iniziare si verifica che il percorso di destinazione per i file sia vuoto. Brevemente le fasi sono: esecuzione dei metodi di `FilterDB` per ridurre le informazioni futili; creazione o merging della base di dati di supporto (*ddvddb*) utilizzando la classe `DumpDVDDB`; elaborazione del *wikitext* ed estrazione delle pagine HTML, utilizzando la classe `DumpDVD` e gli archivi contenenti i file di servizio; generazione degli indici per la navigazione con la classe `DumpDVDindex`; download e gestione delle immagini utilizzando la classe `DumpDVDDB`.

I dettagli sulle fasi ed il loro utilizzo delle risorse di sistema sono riportati nel capitolo 5. Il file di accesso `dumpDVD.php` non effettua ulteriori operazioni e conclude.

Le classi da esso utilizzate sono ognuna inclusa in un file PHP differente, il cui nome ricorda quello della classe e ha estensione `.inc`.

4.1.1.1 Classe DumpDVD

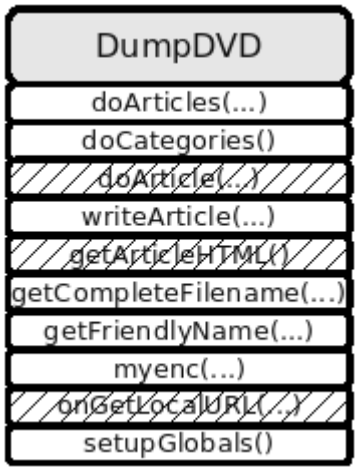

Figura 4.1. L'oggetto DumpDVD. Le classi che si interfacciano con MediaWiki sono indicate con il tratteggio.

Questa classe contiene i metodi utilizzati per eseguire l'estrazione del contenuto delle pagine, il parsing del *wikitext*, la raccolta degli autori, la scrittura dei file HTML.

È inoltre l'unico punto di accesso all'installazione locale di MediaWiki; non è stato scritto da capo, ma inizialmente partendo dalla classe `DumpHTML` presente nel file `maintenance/dumpHTML.php` di MediaWiki, il cui scopo è l'estrazione delle pagine presenti in MySQL. Questo garantisce con ogni probabilità la compatibiltà futura tra WaNDA-tools e MediaWiki, in quanto le funzioni di accesso a MediaWiki sono le stesse.

Vengono di seguito riportati i metodi presenti nella classe, assieme alla descrizione per ognuno.

`DumpDVD.doArticles(start, end)`

Questo metodo provvede a guidare l'estrazione delle pagine, estraendo i titoli delle voci dalla base di dati ed eseguendo il metodo `DumpDVD.doArticle(title)` sul titolo.

Le voci da elaborare possono essere selezionate secondo il loro *id*, impostando l'intervallo dato da `start` e `end`. Vengono scelte soltanto le pagine con il titolo appartenente al namespace principale (ovvero il numero 0, per l'elenco dei namespace si veda 4.2.1). Il metodo riporta inoltre un'indicazione della pagina attualmente in elaborazione e la percentuale di avanzamento sull'intervallo impostato.

Inoltre la funzione provvede ad aggiungere il titolo della voce al database di servizio tramite il metodo `DumpDVDDB.page_add(title)`.

`DumpDVD.doCategories()`

Il metodo è simile a quello precedente, ma si occupa di elaborare tutte le pagine delle categorie. Le pagine vengono infatti estratte utilizzando `DumpDVD.doArticle(title)` sul titolo.

Il titolo di ogni categoria viene estratto dalla tabella *wikidb.categorylinks* del database locale di Wikipedia.

Le categorie vengono utilizzate come mezzo di accesso al contenuto enciclopedico, procendendo per definizione successiva dell'ambito di interesse (i meccanismi di accesso ai contenuti sono riportati in sezione 4.3.2).

`DumpDVD.doArticle(title)`

Questo metodo scrive la pagina della voce indicata dalla stringa del titolo passata per argomento.

Le operazioni che effettua sono semplicemente l'estrazione del codice HTML della pagina con `DumpDVD.getArticleHTML(title)` e la sua scrittura con `DumpDVD.writeArticle(title, text)`.

`DumpDVD.writeArticle(title, text)`

Chiamata soltanto dal metodo precedente, questa funzione provvede a costruire e verificare il percorso del file dato da `title`, nel quale scrivere `text`.

La costruzione del nome del file è data da `DumpDVD.getCompleteFilename(title)`.

`DumpDVD.getArticleHTML(title)`

Qui avviene la costruzione del testo HTML della pagina indicata da `title`; la funzione ritorna infatti una stringa in UTF-8 contentente tutto il testo.

La pagina può essere un redirect oppure una voce vera e propria. Nel primo caso si interpella `DumpDVDtext.redirectPage(path)`, che provvede a fornire il testo per redirigere il browser al percorso `path` indicato.

Nel secondo caso invece si ottiene il testo istanziando un oggetto di tipo `Article` (che fa parte del codice di MediaWiki), con cui si costruisce la pagina dal *wikitext* date alcune opzioni, tra le quali la soppressione dei collegamenti per editare il testo (le "editsection").

Il testo HTML così ottenuto è solo il corpo della voce: esso va incluso in una cornice HTML, che oltre ai menù di navigazione comprende le parti in JavaScript per offrire offline le funzionalità dinamiche. Tale operazione viene effettuata dal metodo `DumpDVDtext.articlePage(title, path, url, text, history)`; per la composizione della pagina HTML finale, oltre al titolo, al collegamento alla versione online ed al testo stesso della voce, serve anche la parte di testo che elenca gli autori della voce (passata nell'argomento `history`). Questa si ottiene con la funzione `DumpDVDhistory.getHistory(article)`, che utilizza l'oggetto `Article` istanziato nel metodo corrente.

I tre metodi che seguono hanno lo scopo di formattare la stringa del titolo delle voci in modo da essere comodamente ed universalmente gestito dai browser oppure utilizzato come nome del file della voce.

`DumpDVD.getCompleteFilename(title)`

Questo metodo ritorna una stringa che rappresenta il percorso completo della pagina il cui titolo è indicato da `title`. Esso utilizza `DumpDVD.myenc(text)` per generare parte del percorso; un esempio di percorso è: `p/prova.html`, in cui "prova" è data da questa funzione.

`DumpDVD.getFriendlyName(name)`

Il metodo formatta il titolo della voce secondo l'RFC 3986, in modo tale da essere una stringa composta da soli caratteri US-ASCII 7-bit e utilizzabile come nei collegamenti HTML.

La stringa, che si dice essere encodata in *percent encoding*, è il valore di ritorno del metodo; il formato è ampiamente descritto in sezione 4.2.4.

`DumpDVD.myenc(text)`

Questo metodo, simile al precendente, formatta il titolo in modo simile ma adatto ad essere utilizzato come nome del file sul filesyetem lowcase, come la FAT o l'ISO-9660 del DVD.

Le differenze sono l'aggiunta di due operazioni: la sostituzione del punto (".") con il corrispettivo *escape* HTTP ("%2e") e la conversione della stringa in soli caratteri minuscoli.

`DumpDVD.onGetLocalURL(title, url, query)`

Questo metodo viene chiamato dalla classe *Hooks*, che fa parte di MediaWiki. Esso serve durante l'elaborazione del *wikitext* per decidere il formato dei collegamenti HTML tra le voci (i cosidetti "link interni"); il collegamento è ottenuto dalla funzione `DumpDVD.getCompleteFilename(title)`.

`DumpDVD.setupGlobals()`

Questo metodo imposta le variabili globali di MediaWiki ai loro valori di default. Questo metodo è indispensabile poiché, essendo questa classe `DumpDVD` il punto d'ingresso agli oggetti di MediaWiki, è necessaria un'impostazione di base ad ogni variabile globale. Eventuali messe a punto vanno fatte nel file `LocalSettings.php` di MediaWiki.

Come indicato, questa classe è l'unica che presenta punti di contatto con l'installazione di MediaWiki utilizzata. In particolare i metodi interessati sono:

- `DumpDVD.doArticles(start, end)`, dove viene istanziato un oggetto di tipo `Title` per ottenere le informazioni salienti sulla voce data dall'*id*: titolo, namespace e se è un redirect.
- `DumpDVD.getArticleHTML(title)`, che istanzia un oggetto `Article` sia per ottenere il *wikitext* che la destinazione di una pagina di redirect. Inoltre si utilizzano le classi `OutputPage` e `ParserOptions` di MediaWiki per riuscire con `Article` ad effettuare la conversione dal *wikitext* al testo HTML.

- `DumpDVD.onGetLocalURL(title, url, query)`, che viene chiamato dall'oggetto `Hooks` di MediaWiki per decidere, durante il parsing del *wikitext*, come gestire i collegamenti presenti nel testo delle voci.

4.1.1.2 Classe FilterDB

Figura 4.2. L'oggetto FilterDB.

Questa classe raccoglie i filtri da eseguire sulla base di dati locale contenente l'enciclopedia. La classe è impostata in modo tale che il costruttore richiami i metodi per effettuare i diversi filtri; è quindi facile per un utente modificare i filtri o aggiungerne oppure rimuoverne: questa classe è presente nel file `dumpDVDfilter.inc`.

Attualmente i filtri sono di due tipi: il filtro sui *namespace*, ovvero gli ambiti ed i tipi delle pagine presenti sul dump ma che, non essendo propriamente enciclopediche, non devono essere presenti nel formato finale (i namespace sono presentati in sezione 4.2.1); il filtro sui doppioni delle pagine e dei redirect alle pagine che si vengono a formare passando ad un contesto *case insensitive*.

I motivi che hanno portato all'esecuzione di questi filtri con queste modalità sono descritti in sezione 4.2.1: è molto più comodo ed è più performante eseguire questi filtri sul database prima di passare all'estrazione delle pagine e dopo aver importato il dump XML ridotto.

`FilterDB.FilterDB()`

Il costruttore della classe provvede ad inizializzare la connessione con la base di dati locale contenente il dump di Wikipedia; inoltre vengono chiamati i due metodi (`FilterDB.filterNS()`, `FilterDB.filterRedirect()`) corrispondenti a due filtri SQL descritti in sezione 5.2.3.1.

`FilterDB.deleteNS(num)`

Questo metodo contiene semplicemente la query da eseguire su MySQL per rimuovere le pagine dato il namespace `num`. Viene utilizzato da `FilterDB.filterNS()`, da cui viene chiamato per migliorare la leggibilità del codice.

`FilterDB.filterNS()`

Il metodo esegue la query definita in `FilterDB.deleteNS(num)` più volte, in modo eliminare le pagine appartenenti ai namespace 1, 2, 3, 4, 5, 7, 9, 11, 12. 13 e maggiori di 15 compreso; i numeri dei namespace corrispondono alle pagine di discussione, alle pagine utente e le loro discussioni, le pagine di servizio di Wikipedia e di MediaWiki e le loro discussioni, le pagine di discussione alle immagini e ai template, le pagine di aiuto ed i portali e le loro discussioni. Una descrizione esaustiva dei namespace è presente in sezione 4.2.1, mentre il loro ruolo nell'organizzazione di MediaWiki è presente nel capitolo 2.2.1.

In aggiunta vengono eliminate alcune pagine spurie che non appartengono al namespace di Wikipedia ma che trattano argomenti simili: sono identificate dal prefisso `Wikipedia:`, come di solito quelle del namespace Wikipedia.

`FilterDB.filterRedirect()`

Questo metodo deve rimuovere le pagine di redirect che hanno in un contesto *case insensitive* lo stesso titolo di voci normali; deve anche poter gestire il caso di due pagine normali con lo stesso nome *case insensitive*, situazione che accade quando un utente crea una nuova voce senza verificare che ne esista già una con il titolo scritto con caratteri diversi.

Questa operazione è indispensabile poiché generalmente per ogni voce normale esistono diverse pagine di redirect con il titolo scritto nei modi più disparati; un'eventuale pagina di redirect che sul filesystem FAT oppure ISO-9660 è presente in concomitanza ad una pagina di voce normale porta ad un redirect a sè stesso infinito, situazione che oltre ad essere sgradevole fa perdere molto spazio del formato finale essendoci molti file fantasma.

Il filtro è abbastanza laborioso e sono state testate differenti soluzioni, siccome MySQL non permette la modifica della tabella su cui si effettua una query. Piuttosto che l'utilizzo di una tabella temporanea di supporto, la soluzione più rapida è composta da un'elaborazione per metà effettuata in PHP.

Per prima cosa si effettua una selezione sul titolo (preso come stringa minuscola) e la lunghezza del testo delle voci normali; la selezione è ordinata in modo decrescente secondo la lunghezza.

Si passa quindi alla cancellazione delle pagine con lo stesso titolo; l'ordinamento effettuato porta alla cancellazione delle pagine con meno testo. Durante questo ciclo il titolo delle pagine non cancellate vengono memorizzate in un vettore PHP.

Successivamente si effettua la selezione del titolo, convertito in minuscolo, delle sole pagine di redirect; un secondo ciclo provvede ad eliminare tutte le pagine di redirect che hanno lo stesso titolo e a memorizzare in un altro vettore PHP man mano i titoli rimasti.

Infine un terzo ciclo, questa volta effettuato sul vettore PHP dei redirect, effettua la cancellazione delle pagine di redirect confrontandosi con il vettore delle voci normali. Si noti che questo ciclo è abbastanza veloce essendo il vettore dei redirect già scremato. Inoltre per ogni ciclo è effettivamente effettuata una cancellazione, contrariamente al modo di procedere in cui si cancellano pagine di redirect con lo stesso nome di voci normali senza sapere se esse davvero siano presenti.

`FilterDB.close()`

È un semplice metodo per terminare correttamente la connessione alla base di dati.

4.1.1.3 Classe DumpDVDDB

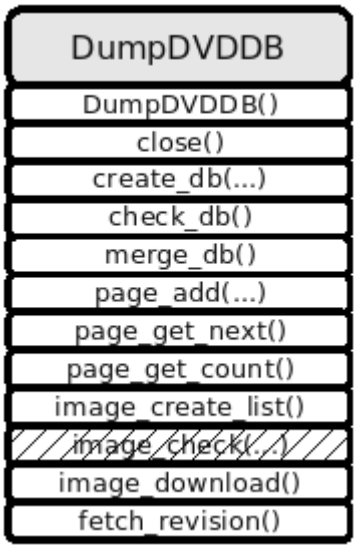

Figura 4.3. L'oggetto DumpDVDDB. Le classi che si interfacciano con MediaWiki sono indicate con il tratteggio.

Questa classe presenta tutte le funzioni per accedere sia al database di servizio di WaNDA-tools che al database locale contenente il dump enciclopedico.

I metodi sono suddivisi in vari gruppi, poiché sono relativi a diverse fasi dell'elaborazione:

- creazione e merging del database di servizio;
- aggiunta ed estrazione dei titoli delle pagine estratte, per la generazione degli indici;
- gestione delle immagini da includere nel formato finale;
- estrazione della lista di autori di una pagina.

`DumpDVDDB.DumpDVDDB()`

Il costruttore si occupa di inizializzare le due connessioni al database di servizio e a quello enciclopedico. Di default in WaNDA-tools il primo si chiama *ddvddb*, mentre il database locale di Wikipedia *wikidb*.

`DumpDVDDB.close()`

È un semplice metodo per terminare correttamente le due connessioni ai database.

`DumpDVDDB.create_db(populateimages)`

Questo metodo provvede a create le tabelle del database di servizio; esse sono due, *ddvddb.pages* e *ddvddb.images*. Prima di essere creato, un eventuale database con questo nome viene rimosso.

Se l'argomento è posto a `true`, dopo aver creato il database è chiamata la funzione `DumpDVDDB.image_create_list()`, che provvede a riempire la tabella *ddvddb.images* con il titolo delle immagini appartenenti alle categorie desiderate. L'argomento è opzionale e se assente vale `false`.

`DumpDVDDB.check_db()`

È un semplice metodo che verifica se il database di servizio è già presente e se contiene tabelle corrette; ritorna quindi una variabile booleana.

È utilizzato dal file PHP di accesso per sapere in automatico se creare oppure eseguire il merging del database di servizio *ddvddb*, invocando quindi `DumpDVDDB.create_db()` oppure `DumpDVDDB.merge_db()`. Passando da riga di comando l'argomento opportuno è possibile forzare il ripristino del database anche se essendo già presente dovrebbe essere effettuato il merging.

`DumpDVDDB.merge_db()`

Questo metodo viene utilizzato nel caso in cui venga elaborato più di un dump sulla stessa piattaforma. Permette durante una successiva elaborazione di evitare il download di immagini già presenti in locale; può quindi essere utilizzato soltanto quando `DumpDVDDB.check_db()` ci assicura che il database di servizio già esiste.

Siccome il merging interessa soltante la tabella *ddvddb.images*, la tabella di indice *ddvddb.pages* viene svuotata dai suoi valori.

L'operazione di merge procede nel modo seguente: si ottiene dal database enciclopedico *wikidb* la lista delle immagini attualmente richieste, selezionate secondo le categorie di appartenenza impostate.

Per ogni immagine, il cui titolo è unico, si verifica se già esiste nel database di servizio *ddvddb.images*; in caso positivo si confrontano i timestamp, poiché un'immagine potrebbe essere stata aggiornata; se i timestamp sono differenti si aggiorna la riga contrassegnando l'immagine. Nel caso il titolo non esista in *ddvddb.images* lo si aggiunge. Nulla viene fatto se il titolo esiste e i timestamp sono uguali, il che significa che l'immagine richiesta è già localmente presente.

`DumpDVDDB.page_add(pagetitle)`

Il metodo aggiunge il titolo di una pagina alla tabella *ddvddb.pages*. È chiamato durante la fase di scrittura delle pagine HTML, in modo da avere alla fine un indice delle pagine presenti.

`DumpDVDDB.page_get_next()`

Questa funzione estrae in modo ordinato i titoli delle pagine presenti in *ddvddb.pages*. Deve essere chiamata una prima volta per eseguire la query, e ritorna `true`; dalla seconda volta in poi ritorna una stringa che è il titolo della pagina. Si invoca il metodo per ottenere ogni titolo, finchè non ritorna una stringa ma `false`, il che significa che tutti i titoli sono stati estratti.

A questo punto è possibile effettuare di nuovo l'operazione dall'inizio. La funzione è utilizzata dopo l'estrazione delle voci per generare sia il contenuto delle pagine di indice di tutte di pagine, che la base di ricerca per il motore JavaScript.

`DumpDVDDB.page_get_count()`

Questo semplice metodo ritorna il numero di titoli presenti in *ddvddb.pages*.

`DumpDVDDB.image_create_list()`

È il primo dei metodi relativi alla gestione delle immagini. Viene sempre chiamato prima dell'estrazione delle pagine poiché si occupa di generare l'elenco dei titoli delle immagini che hanno le licenze opportune. La selezione è fatta scegliendo le categorie impostate nel file di configurazione di WaNDA-tools (`dumpDVD.ini`).

In dettaglio, per ogni categoria si estrae da *wikidb.categorylinks* il titolo ed il timestamp dell'immagine, per poi inserirli nel database di supporto *ddvddb.images*.

`DumpDVDDB.image_check(img_name, img_width)`

Il metodo è chiamato dalla classe `Linker` di MediaWiki durante l'operazione di parsing del *wikitext* ogni qualvolta incontra un'immagine da includere.

Essa si occupa di contrassegnare nel database di servizio *ddvddb.images* l'immagine con il nome `img_name` se presente; inoltre se è indicata la dimensione dell'immagine nel *wikitext*, è anche passato come argomento l'altezza e la larghezza in pixel.

Se il titolo dell'immagine non è presente nel database di servizio la funzione ritorna `false`; in questo modo il codice in `Linker` sà che può ignorare l'immagine con il collegamento alla pagina immagine e l'eventuale didascalia.

`DumpDVDDB.image_download()`

Questa funzione è sempre chiamata dopo la fase di estrazione delle voci, quando è completo l'elenco delle immagini che devono essere incluse nel formato finale; la chiamata è effettuata da `DumpDVD.php`.

Si effettua una query sul database di servizio *ddvddb.images* da cui si ottiene la lista dei titoli e della dimensione delle immagini che sono state contrassegnate e non sono localmente presenti.

Nel ciclo eseguito sugli elementi si costruisce dal titolo il percorso locale che dovrà avere l'immagine e l'URL online per ottenerla (il server che le contiene è `http://upload.wikimedia.org/wikipedia/it/`). Si procede quindi al download ed al ridimensionamento dell'immagine, che viene salvata nella directory `images/` di MediaWiki.

Come ultimo passo del ciclo ogni immagine scaricata e quindi presente localmente è segnata nel database di servizio.

`DumpDVDDB.fetch_revisions(mtitle, limit)`

L'unico metodo di questa classe che lavora soltanto con il database enciclopedico *wikidb* è questo. Esso ritorna un vettore di stringhe contenente la lista degli autori di una pagina dato il titolo `mtitle`. L'elenco è ordinato per numero decrescente delle revisioni, in modo che il primo autore della lista sia l'ultimo che abbia effettuato modifiche.

4.1.1.4 Classe DumpDVDtext

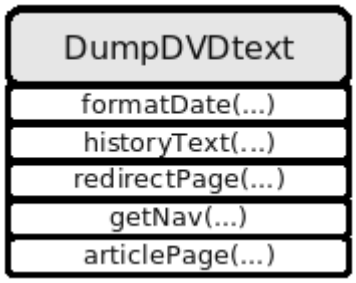

Figura 4.4. L'oggetto DumpDVDtext.

Tutti i testi che vengono introdotti nelle pagine HTML presenti nel formato finale sono definiti in questa classe; tra questi vi è la "cornice" HTML che contiene il testo formattato delle voci, la formattazione delle date, la presentazione degli autori, il testo delle pagine di redirect ed il testo delle pagine di navigazione che contengono l'indice alfabetico delle pagine.

Il testo di tutte le pagine nel formato finale, come l'ISO-9660 per DVD, è codificato in UTF-8; infatti è il modo più comodo per rappresentare il contenuto delle voci, che spesso contengono caratteri estesi. Ciò significa che il testo qui presente deve già essere encodato in UTF-8: tuttavia PHP di base gestisce soltanto testo encodato in US-ASCII a 7 o 8 bit. Ecco quindi che serve l'estensione PHP per il supporto alle stringhe *multibyte*, che abilita la comprensione delle diverse codifiche di tipo UTF e Unicode [60].

Il file stesso `dumpDVDtext.inc` per contenere le stringhe di testo in UTF-8 è presente in questo formato.

`DumpDVDtext.formatDate(dateString)`

Questo metodo converte una stringa numerica di 12 cifre, che rappresenta una data, in una stringa di testo che la descrive in modo più immediatamente comprensibile. Stringhe di testo predefinite come i nomi dei mesi sono impostati nel file di configurazione di WaNDA-tools.

La funzione ritorna quindi una stringa di tipo "ora:minuti, giorno mese anno"; eventuali modifiche al formato del testo possono facilmente essere impostate in questa funzione. È chiamata principalmente per formattare la data dell'ultima revisione ad una voce.

`DumpDVDtext.historyText(lasttimestamp, userlist)`

Il metodo è utilizzato da `DumpDVDhistory.getHistory(article)` per presentare la lista degli autori e la data dell'ultima modifica. `userlist` è il vettore contenente l'elenco degli autori. La stringa ritornata viene inserita nella parte inferiore di ogni pagina enciclopedica.

`DumpDVDtext.redirectPage(url)`

Questa funzione ritorna il testo di una pagina HTML di redirect che deve redirigere al percorso indicato in `URL`. Inizialmente comprendeva tre tipi meccanismi di redirect, per massimizzare la compatibilità con i browser web: tuttavia si è preferito rendere la pagina è la più concisa possibile, in modo da occupare poco spazio. La soluzione adottata presenta nell'`head` HTML un campo `meta` che effettua un refresh immediato della pagina, impostando l'URL di destinazione per il rinfresco. Questo meccanismo è nello standard HTML W3C [80] ed è supportato da praticamente ogni browser moderno; se ciò non dovesse funzionare (magari perché disabilitato) il `body` presenta un collegamento HTML classico per passare alla pagina di destinazione.

`DumpDVDtext.getNav(list, pnprefix, cpage, tpage)`

L'indice navigabile di tutte le voci enciclopediche è presente all'interno di un gruppo di pagine HTML. Il metodo ritorna il testo di una delle pagine: il numero della pagina da generare è impostato da *cpage*, mentre *tpage* è il numero dell'ultima pagina. Il numero di pagine su cui spezzettare l'indice è definito nel file di configurazione di WaNDA-tools `dumpDVD.ini`.

Questa funzione provvede soltanto a formattare il testo: *list* contiene la serie di collegamenti che fanno parte del troncone della lista indicato dal numero di pagina. È `DumpDVDindex.makeHTMLindex()` che si occupa, dopo aver generato i segmenti della lista secondo il numero delle pagine, di chiamare questa funzione.

Viene utilizzata la stessa "cornice" HTML delle voci normali, per cui il testo che presenta i link per viaggiare tra le diverse pagine di indice e i collegamenti alle voci è passato come l'argomento `content` di `DumpDVDtext.articlePage(utftitle, filepath, onlineversion, content, history, path)`; la funzione ritorna quindi il testo completo della pagina.

`DumpDVDtext.articlePage(utftitle, filepath, onlineversion, content, hist ory, path)`

Questo metodo contiene il testo HTML "cornice" delle pagine: la parte di `head` con i collegamenti ai file JavaScript utilizzati ed ai documenti di stile CSS, e la parte di `body` che presenta con l'aiuto di JavaScript il testo delle voci enciclopediche. La descrizione accurata del formato delle pagine è illustrata in sezione 4.2.

4.1.1.5 Classe DumpDVDhistory

Questa classe presenta una sola funzione, che genera l'elenco degli autori di una voce.

Figura 4.5. L'oggetto DumpDVDhistory.

`DumpDVDhistory.getHistory(article)`

Il metodo è sempre chiamato da `DumpDVD.getArticleHTML(title)`, da cui prende un'istanza dell'oggetto `Article` di MediaWiki; utilizzando l'istanza ed il metodo `DumpDVDDB.fetch_revisions(title, limit)`, si estrae un vettore contenente l'elenco degli autori per la voce indicata.

Tale vettore viene quindi verificato per nascondere gli indirizzi IP degli utenti non registrati ed eliminare i doppioni, e formattato in una stringa testuale da `DumpDVDtext.historyText(lasttimestamp, userlist)`.

4.1.1.6 Classe DumpDVDindex

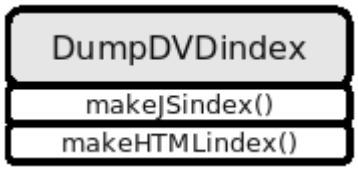

Figura 4.6. L'oggetto DumpDVDindex.

Per accedere ai contenuti enciclipedici, presenti come pagine HTML, sono presenti meccanismi di diverso tipo descritti in sezione 4.3.2; la pagine che contengono la lista di tutte le voci e la base di ricerca per il motore JavaScript vengono create dalle funzioni di questa classe.

L'elenco di tutte le voci è disponibile dalla fine della fase di esportazione delle loro pagine; esso è disponibile nel database *ddvddb.pages*.

`DumpDVDindex.makeJSindex()`

Questo metodo crea la base per il motore di ricerca JavaScript; esso si compone di un vettore associativo, composto da celle quante sono le voci enciclopediche normali e di redirect. Ogni cella ha come chiave la stringa del titolo della voce in formato *percent encoding* e come valore il percorso della pagina; nel formato finale attualmente utilizzato il percorso della pagina di una voce si basa sulla prima lettera del titolo: esso è composto dalla stringa in *percent encoding* della prima lettera.

Questo sistema è ottimo ai fini dei brevi tempi di risposta del motore di ricerca; tuttavia è laborioso per il browser, in quanto esso deve caricare nella memoria riservata all'interprete JavaScript. Ecco perché il vettore associativo viene smembrato secondo un ordine alfabetico prestabilito in 10 segmenti, presenti in 10 file diversi.

Per coordinare questi 10 file di indice è necessario averne un ulteriore che contenga la corrispondenza tra file di indice e lettere in esso presente. In questo modo il motore di ricerca interpella questo sommario e carica soltanto la parte di indice che serve alla ricerca. Una ricerca estesa su tutto l'indice richiede quindi il caricamento progressivo delle 10 parti.

Il metodo `DumpDVDindex.makeJSindex()` accede quindi alla lista delle voci tramite `DumpDVDDB.page_get_next()`; per ogni voce si costruisce il percorso e la stringa *percent encoding*, e si aggiunge un elemento al vettore della parte di indice corretta.

Dopo aver completato i vettori presenti nei 10 file, si scrive il contenuto del file di sommario. La distribuzione alfabetica delle voci nei vettori è impostata in modo tale da equilibrare al massimo il numero di elementi in ogni vettore; tale ordine predefinito è modificabile nelle opzioni avanzate del file di configurazione `dumpDVD.ini`.

Per il funzionamento dettagliato del motore di ricerca JavaScript si rimanda alla sezione 4.3.3.

`DumpDVDindex.makeHTMLindex()`

Questo metodo costruisce e scrive il contenuto delle pagine di navigazione; esse contengono l'elenco delle voci, con i collegamenti alle loro pagine, diviso in 50 pagine HTML. Il numero delle pagine di navigazione è modificabile a piacere dal file di configurazione: un valore troppo basso porta però ad avere pagine molto lunghe.

Ognuna delle pagine di navigazione, per permettere all'utente di muoversi, è collegata con la prima, la precendente, la successiva e l'ultima pagina.

L'elenco è ordinato alfabeticamente: i titoli delle voci sono anch'essi estratti con la funzione `DumpDVDDB.page_get_next()`. Ognuno dei 50 segmenti della lista viene formattato in HTML come elenco non ordinato (il tag HTML `UL`) di collegamenti alle voci; tale elenco viene passato a `DumpDVDtext.getNav(list, pnprefix, cpage, pnnum)` che ritorna il testo HTML della pagina completa. Ogni pagina di navigazione viene quindi salvata nella radice dell'*albero*, dove sono presenti le pagine di servizio di WaNDA-tools.

4.1.1.7 Classe DumpDVDlog

Per presentare all'utente di WaNDA-tools le informazioni più critiche si utilizza questa classe; questi metodi sono utili sopratutto per eventuali utilizzi che richiedano la stampa delle informazioni non solo su *standard out*.

`DumpDVDlog.info(text)`

Metodo che stampa sul terminale messaggi informativi.

`DumpDVDlog.warn(text)`

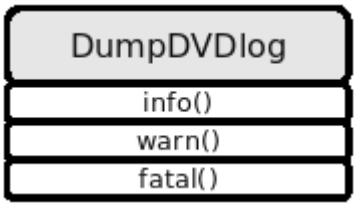

Figura 4.7. L'oggetto DumpDVDlog

Metodo che stampa sul terminale messaggi di avviso per errori non critici.

`DumpDVDlog.fatal(text)`

Metodo che stampa sul terminale messaggi critici dopo i quali avviene l'interruzione della normale esecuzione del programma.

4.1.1.8 Classi e chiamate agli oggetti

Come riepilogo in figura 4.8 vengono riportati i vari oggetti PHP con l'indicazione delle chiamate ai loro metodi. Sono anche indicate le cinque fasi (i cinque blocchi al centro della figura) che fanno uso del codice PHP per la generazione del formato finale; il percorso completo da effettuare per la conversione dal dump XML all'*albero* è presentato nel capitolo seguente. Le cinque fasi compongono la seconda ed ultima parte del processo: si parte dal database locale già contenente il dump enciclopedico e si ottiene l'*albero* finale completo.

I metodi barrati indicano quelli che si interfacciano con le classi di MediaWiki:

- `DumpDVD.doArticles(...)` che istanzia l'oggetto `Title` di MediaWiki;
- `DumpDVD.getArticleHTML(...)` che utilizzando l'oggetto `Article` interpreta il *wikitext*
- `DumpDVD.onGetLocalURL(...)` che viene chiamato da MediaWiki per gestire i collegamenti HTML tra le voci
- `DumpDVDDB.image_check(...)` che è utilizzato dall'oggetto `Linker` di MediaWiki per discernere le immagini.

4.1.2 Componenti per la presentazione

Le parti in PHP finora presentate possono generare le seguenti componenti del formato finale (l'*albero*): le pagine HTML delle voci, le immagini da includere e il vettore per la ricerca in linguaggio JavaScript.

Per ottenere l'*albero* completo mancano ancora:

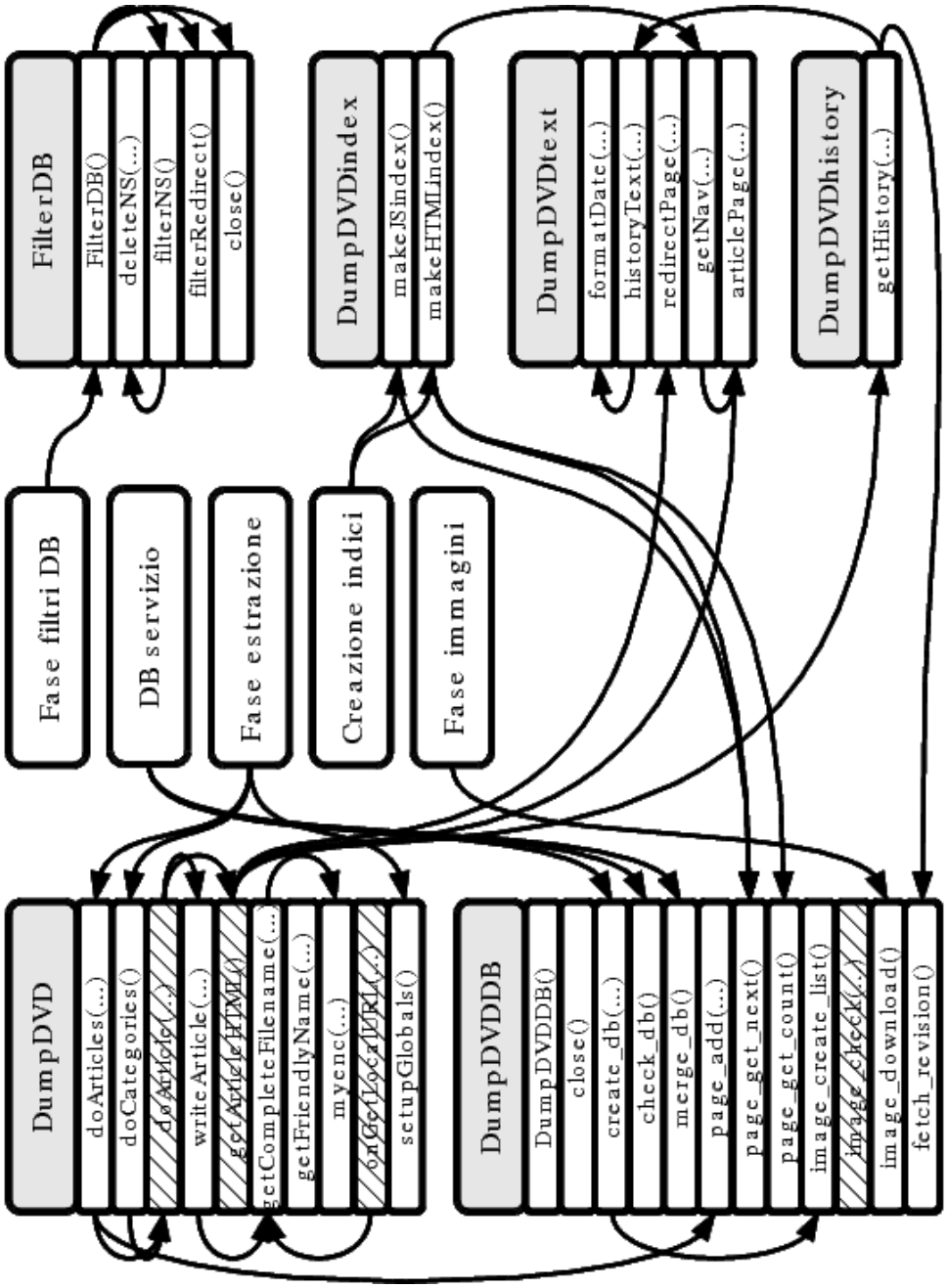

Figura 4.8. Classi PHP di WaNDA-tools, con indicazione del loro utilizzo nelle cinque fasi dell'elaborazione e le reciproche chiamate dei metodi. I metodi barrati indicano quelli che si interfacciano con le classi di MediaWiki.

- Le pagine HTML di servizio: per presentare inizialmente i contenuti enciclopedici, per raccogliere le indicazioni di aiuto alla consultazione, per riportare i disclaimer e le licenze, e per eventuali ringraziamenti.
- Il foglio di stile CSS per tutte le pagine HTML.
- I codici attivi in JavaScript necessari alla presentazione delle pagine HTML ed al funzionamento del motore di ricerca.
- I binari per poter eseguire sulle piattaforme proprietarie un browser OpenSource.

Queste sono componenti statiche, in q¡uanto sono inseribili senza necessitare elaborazioni nel contesto dell'*albero* ricavato da una qualsiasi versione del dump XML. Per questa loro caratteristica statica, sono piuttosto le componenti elaborate con le funzioni PHP viste (pagine HTML delle voci, immagini e vettore di ricerca) che vanno ad integrarsi a questi statici componenti di base; per questo le quattro componenti elencate fanno parte della cosidetta *ossatura*.

L'*ossatura* è presente nei due archivi `dumpDVD/wnd_html.tar.bz2` e `dumpDVD/wnd_bin.tar.bz2`. Il primo archivo contiene le pagine HTML di servizio, il foglio di stile CSS ed i file JavaScript, il secondo contiene i file eseguibili.

4.1.2.1 Pagine HTML di servizio

L'*ossatura* è divisa nei due archivi per poter comodamente modificare il contenuto delle pagine HTML di servizio senza dover toccare il pacchetto dei binari. Esse infatti possono dover subire modifiche *una tantum* per aggiornare informazioni sul progetto o sugli autori, oppure modificare caratteristiche stilistiche di presentazione delle pagine HTML.

Le pagine HTML di servizio attualmente sono:

- `pagina_aiuto.html`: illustra le caratteristiche di presentazione delle pagine, la navigazione dei contenuti ed il funzionamento del motore di ricerca.
- `pagina_benvenuto.html`: è una pagina introduttiva che presenta Wikipedia e l'enciclopedia offline.
- `pagina_credits.html`: elenca gli autori del progetto e i ringraziamenti dovuti.
- `pagina_disclaimer.html`: contiene i dovuti chiarimenti sulle resposabilità degli enti in gioco, la licenza FDL e le licenze delle immagini, ed i disclaimer di Wikipedia. Queste tematiche sono illustrati nel capitolo 6.
- `pagina_donazioni.html`: recapito per eventuali donazioni al progetto.
- `pagina_principale.html`: la prima pagina che si apre quando si accede all'*albero*; essa deve almeno contenere collegamenti alle altre pagine di servizio, alle principali categorie ed al motore di ricerca, per permettere all'utente finale di raggiungere facilmente i contenuti enciclopedici. Inoltre una breve introduzione e collegamenti online ai progetti Wikimedia sono consigliati.

- `pagina_ricerca.html`: è una pagina opzionale che raccoglie le tre funzionalità del motore di ricerca JavaScript, descritte in sezione 4.3.3.

Per la formattazione di tutte le pagine HTML, enciclopediche e di servizio, si utilizza la tecnologia CSS, *Cascading Style Sheet* [61]; esso provvede anche a caricare i contenuti grafici delle pagine. Il CSS permette di ridurre la dimensione di ogni pagina, diminuendo considerevolmente lo spazio totale occupato dell'*albero*. Inoltre per modificare in qualche modo elementi grafici di presentazione è necessario apportare modifiche ad un solo file. Il file CSS e gli altri elementi di stile, cioè le immagini in formato PNG e le icone, sono presenti nella direcory di nome `data/style/`.

4.1.2.2 Componenti dinamiche

I contenuti attivi del formato finale si situano nella directory `data/js/`: sono tutti implementati in linguaggio JavaScript 1.0. Essi permettono sia di presentare le pagine HTML sia di effettuare ricerche all'interno dell'*albero*.

La presentazione delle pagine sfrutta il JavaScript poiché permette di ridurre ulteriormente la dimensione totale. Infatti sia il menù di navigazione che le componenti fisse della pagina non vengono ripetute come testo HTML fisso in ogni pagina, ma è presente in una funzione JavaScript. La pagina HTML di ogni voce è quindi composta da un HEAD scarno e un BODY che chiama la funzione JavaScript che provvede a stampare il testo HTML. La funzione JavaScript ha come argomento il contenuto della voce enciclopedia, ovvero il solo testo HTML che varia da pagina in pagina; la funzione aggiunge il testo fisso del menù di navigazione, l'intestazione e la parte inferiore,

Il JavaScript serve inoltre a far funzionare il motore di ricerca del progetto WaNDA. Si può quindi comodamente effettuare una ricerca sui titoli delle voci, e come funzionalità aggiunta è possibile l'individuazione di una pagina a caso. Si accede al motore di ricerca Javascript tramite un `form` HTML che si interfaccia ad un gruppo di pagine HTML contenenti diversi script JavaScript; tali pagine hanno il nome del tipo `pagina_risultati_*.html`.

È stato verificato che i vari metodi JavaScript utilizzati siano disponibili nelle varie implementazioni degli interpreti, quali SpiderMonkey [62], JScript [63], KJS [64], JavaScriptCore [65]. Ciò assicura la massima compatibilità con tutti i browser, oltre ad seguire lo standard implementativo ECMAscript [66] e le linee guida dell'RFC 4329 [67].

Infine assieme all'*albero* sono stati aggiunti degli eseguibili per offrire funzionalità specifiche per l'una o l'altra piattaforma. In particolare è comodo disporre di un browser opensource ufficialmente testato sui sistemi Microsoft, dove il browser ufficiale potrebbe non essere totatalmente compatibile con gli standard.

Uno studio approfondito sulla presentazione grazie al CSS, sul funzionamento del motore di ricerca e sul browser incluso è presente nella sezione 4.3.

4.2 Gestione contenuti

Vengono ora raccolti ed approfonditi vari aspetti sui meccanismi di gestione dei contenuti enciclopedici, ovvero la descrizione di specifiche operazioni utili alla comprensione del resto del capitolo sulla presentazione dei contenuti e del capitolo successivo sulla descrizione del processo di estrazione.

Le operazioni sui contenuti illustrate sono:

- Motivazioni e campo di applicazione dei filtri sui contenuti; vengono descritti i *namespace* di MediaWiki.
- Licenze e formato dei contenuti multimediali.
- Indicazione degli autori delle voci.
- Formato di codifica dei testi, dei collegamenti e del nome dei file con UTF-8 e *percent encoding*.
- Modifiche apportate al testo delle voci, che quindi richiedono alterazioni di alcuni file di MediaWiki.

4.2.1 Epurazione dei contenuti non enciclopedici

Il dump fornito da Wikimedia contenente l'enciclopedia Wikipedia deve essere filtrato principalmente per tre motivi:

- Primo, è necessario ridurre al massimo lo spazio occupato dall'*albero*, poiché esso dovrà essere facilmente maneggevole all'utente. Il supporto ottico a basso costo, DVD single layer, ha una dimensione limitata. Le memorie flash o simili (il pendrive USB) hanno costi esponenziali a seconda della capienza. Inoltre un file di dimensioni ridotte ne facilita la diffusione e lo scambio di rete.

  È quindi inutile includere le informazioni che esulano dal contesto enciclopedico: pagine di servizio, pagine di discussione, pagine di guida alla versione online, pagina di iscrizione e così via.

- Secondo, alcune pagine devono essere filtrate in quanto sono dal contenuto discutibile e potenzialmente diffamatorio. Esse sono segnate come nNPOV (non - *Neutral Point of View*); si noti che la selezione viene effettuata online dal team di Wikimedia Italia ed di conseguenza nel database locale.

- Per ultimo ci sono i motivi tecnologici. Il formato finale dell'albero deve poter essere copiato su un filesystem *case-insensitive*, ovvero che interpreti un carattere maiuscolo e minuscolo come se fosse il medesimo. È di questo tipo sia il formato ISO-9660 che il filesystem FAT.

  Su Wikipedia si presenta il problema inverso, in quanto MediaWiki è *case sensitive* poiché permette la creazione di più voci con lo stesso nome ma scritto con combinazioni di *case* differente. La Wikimedia Foundation ha quindi consigliato a tutti i

progetti wiki di utilizzare una sola pagina di riferimento che contiene la voce completa e convertire le altre in rimandi alla voce completa; il *case* della voce principale dovrebbe avere come convenzione la prima lettera del titolo e dei nomi propri in esso contenuti maiuscolo. Spesso accade che un utente di Wikipedia crei una nuova pagina senza controllare se ne esista già una il cui titolo è scritto con un *case* differente, per cui gli amministratori di Wikipedia regolarmente provvedono all'unione dei contenuti.

Nell'ambito *case insensitive* è quindi necessario tenere la voce principale, rimuovendo i redirect con il titolo simile. Inoltre può capitare che gli amministratori non abbiano fatto in tempo ed unire il doppione, nel qual caso il software WaNDA decide di mantenere la voce contenente più informazioni; si noti che questo evento accade di rado.

Le prime due selezioni avvengono rimuovendo nella base di dati tutte le voci appartenenti a determinati *namespace*: sono dei tipi di pagine che estendono i tipi predefiniti di MediaWiki (voci normali, discussioni, ...), servono come organizzazione logica per gestire comodamente i diversi contenuti su Wikipedia. I namespace sono identificati univocamente su tutti i progetti Wikimedia da un numero.

La lista completa dei namespace della parte italiana di Wikipedia è riportata in tabella 4.1.

Questa lista dei namespace può essere mutabile nel tempo e possono esserne presenti altri oppure mancarne secondo i diversi **wiki* locali; l'elenco aggiornato per `itwiki` è disponibile all'indirizzo [68]. I namespace da -1 a 15 sono presenti di default in MediaWiki.

Si può facilmente intuire che tutti i *namespace* che raggruppano le discussioni non interessano ai fini enciclopedici, essendo utilizzato come meccanismo di comunicazione fra gli utenti; così anche le pagine utente esulano dagli scopi del progetto. Le pagine di aiuto non servono in quanto presentano informazioni relative alla modifica on-line delle pagine. Inoltre le pagine interne di Wikipedia, che offrono i servizi interattivi quali ricerca, indici e autenticazione, non sono applicabili all'ambito offline.

Si noti che i pseudo-namespace -2 e -1 non generano pagine vere e proprie, e all'interno del database agli oggetti di questi namespace non viene assegnato un `id`; ne consegue che esse non vengono filtrate (nè sarebbe possibile per le pagine di servizio) poiché non vengono selezionate ed esportate nel processo di scrittura delle pagine.

Sono quindi rimosse tutte le pagine che fanno parte dei namespace:

- 1 (*Discussioni*);
- 2 (*Utente*);
- 3 (*Discussioni_utente*);
- 4 (*Wikipedia*);
- 5 (*Discussioni_Wikipedia*);

numero	nome	pagine contenute
-2	*Media*	pseudo-namespace per file non grafici
-1	*Speciale*	pseudo-namespace per pagine speciali
0		namespace principale che contiene le voci normali
1	*Discussioni*	discussioni alle voci normali
2	*Utente*	pagine personali degli utenti registrati
3	*Discussioni_utente*	discussioni alle pagine personali
4	*Wikipedia*	pagine di servizio di Wikipedia
5	*Discussioni_Wikipedia*	discussioni alle pagine di servizio
6	*Immagine*	pagine delle immagini caricate su MediaWiki
7	*Discussioni_immagine*	discussioni alle immagini
8	*MediaWiki*	pagine protette contenenti testi di MediaWiki
9	*Discussioni_MediaWiki*	discussioni alle pagine protette di MediaWiki
10	*Template*	raccoglie i *template*
11	*Discussioni_template*	discussioni ai *template*
12	*Aiuto*	pagine di aiuto
13	*Discussioni_aiuto*	discussioni alle pagine di aiuto
14	*Categoria*	pagine di raccolta dei collegamenti fra categorie
15	*Discussioni_categoria*	discussioni alle categorie
100	*Portale*	pagine dei portali di Wikipedia
101	*Discussioni_portale*	discussioni ai portali
102	*Progetto*	pagine dei progetti di Wikipedia (i *WikiProject*)
103	*Discussioni_progetto*	discussioni ai progetti

Tabella 4.1. Elenco dei namespace in Wikipedia versione italiana.

- 7 (*Discussioni_immagini*);
- 9 (*Discussioni_MediaWiki*);
- 11 (*Discussioni_template*);
- 12 (*Aiuto*);
- 13 (*Discussioni_aiuto*);
- 15 e maggiori (*Discussioni_categoria*, *Portali* e così via).

La descrizione del meccanismo con il quale vengono rimossi i contenuti indesiderati e la funzione delle pagine che appartengono ai namespace rimanenti sono riportate nel capitolo successivo, illustrando il filtro implementato in PHP per questo scopo.

4.2.2 Contenuti grafici

I contenuti multimediali dell'enciclopedia sono tipicamente di tipo grafico; sono immagini, grafici e formule matematiche. Si cerca più possibile di utilizzare il formato PNG [69] [70] che oltre ad essere uno standard aperto offre ottime prestazioni nel caso di immagini con pochi colori [71] quali sono i grafici e le formule matematiche.

Eventuali contenuti di tipo audio o video vengono automaticamente scartati dal meccanismo di esportazione delle pagine poiché non vengono installate le estensioni MediaWiki per gestirli.

Tutti i contenuti grafici si possono differenziare in quattro tipi secondo il meccanismo con cui vengono ottenuti:

- Immagini di carattere misto presenti nelle pagine HTML. Queste immagini sono selezionate secondo la licenza con la quale sono state rilasciate dagli autori; esse sono tipicamente contestuali alla voce nella cui pagina sono collegate. A questo tipo appartengono anche le immagini di Wikipedia per indicare graficamente informazioni di servizio o avvisi, oppure anche bandiere dei Paesi o altro ancora.

  Le licenze selezionate devono permettere la redistribuzione successiva, sia di tipo commerciale che non; quelle di Wikipedia sono state per convenzione inserite con licenza FDL.

  Queste immagini possono essere nei formati JPEG o PNG.

- Immagini che raffigurano formule matematiche: vengono generate con un estensione di MediaWiki che traduce del testo TEX presente nel *wikitext*. Queste immagini sono tutte in formato PNG, che considerando il tipo di figure con due colori (bianco e nero) genera file di piccole dimensioni.

- Immagini che raffigurano gli istogrammi, principalmente gli andamenti demografici dei comuni. Questi grafici e schemi sono immagini di tipo PNG generati con un'estensione apposita.

- Immagini sempre presenti nell'*albero* che provengono dall'*ossatura*. Questi file sono statici ed introdotti da WaNDA-tools. Queste immagini comprendono i dettagli grafici utilizzati dal tema delle pagine HTML, i loghi e le immagini utilizzate dalle pagine di servizio.

I contenuti grafici occupano parecchio spazio rispetto ai contenuti testuali per la loro stessa natura. Può quindi essere necessario ridurre la dimensione occupata da quesi contenuti.

Gli schemi e le formule matematiche, poiché sfruttano bene il formato PNG, sono una parte minima di questi contenuti; il grosso dei contenuti grafici è rapprensentato dalle immagini del primo tipo. Per ridurre lo spazio occupato è quindi meglio lavorare sulle immagini degli utenti. Per esempio si riduce la dimensione delle immagini in modo da tenere soltanto una thumbnail della grandezza del collegamento nelle pagine in cui viene incontrata; il funzionamento verrà illustrato nel capitolo seguente.

Inoltre è possibile modificare l'elenco delle licenze permesse, in modo da tagliare gruppi di immagini; per esempio si potrebbe decidere di includere soltanto le immagini FDL e Public Domain.

4.2.3 Autori

Tutte le pagine delle voci enciclopediche, secondo la licenza FDL con la quale sono rilasciati i contenuti, devono riportare l'elenco degli autori, ovvero di chiunque abbia contribuito al testo.

Per cercare di ridurre nell'*albero* sia il numero di file sia la dimensione si è optato per includere un elenco degli autori di una pagina in fondo alla stessa. Questo elenco è ricavato dalla lista delle modifiche, tenendo per ognuna l'autore della modifica; questa lista è ottenuta utilizzando una chiamata a MediaWiki. Gli autori non devono per forza registrarsi, onde per cui può capitare di avere come identificativo dell'autore un indirizzo IP; come policy si è scelto di indicare gli autori non registrati con il testo "Utente anonimo".

Per avere maggiori informazioni sulla porzione di testo modificata da un determinato utente è necessario recarsi sulla versione online della voce, subito accessibile con un collegamento presente in testa alla pagina; è però necessario disporre di un collegamento ad Internet.

4.2.4 Percent encoding

Per poter rappresentare i caratteri estesi presenti nell'*albero* si utilizza la codifica UTF-8 (*8 bit Unicode Tranformation Format*). Tale formato permette la rappresentazione di qualsiasi carattere con uno o più gruppi di 8 bit; se il bit più significativo di un gruppo è posto a 0, allora quel byte è da prendere come tale; altrimenti il byte che lo precede fa parte della sequenza di bit identificativa del carattere. La corrispondenza tra sequenza di bit e carattere è definita dal consorzio Unicode [60] e definita in RFC 3629 [72]; la corripondenza tra sequenza di bit e carattere è anche definita dallo standard ISO 10646 [73]. Le sequenze di un byte con il bit più significativo posto a 0 sono esattamente uguali alla loro codifica US-ASCII a 7 bit. I caratteri estesi del formato ASCII a 8 bit non sono quindi compatibili.

Tutti i documenti HTML dell'*albero* sono formattati con la codifica UTF-8 e sono impostati in modo da essere correttamente interpretati dal browser poiché contengono nell'HEAD un campo META che specifica questa codifica.

Per ottenere una corrispondenza fra il titolo delle voci e il nome del file HTML che le contengono senza dover scrivere sul filesystem direttamente in UTF-8 si è cercato un modo di codificare le informazioni utilizzando soli caratteri US-ASCII.

La codifica cercata è stata immediatamente trovata dato che la soluzione ideale era già utilizzata dal browser per trasmettere caratteri estesi nel nome del file. Infatti, come ribadito dal recente RFC 3986 [74], ogni URI deve essere contenere soltanto caratteri alfanumerici e pochi altri caratteri speciali; per poter rappresentare caratteri aggiuntivi è necessario utilizzare una codifica chiamata *percent encoding* o meno genericamente *url*

car.	ASCII	hex Unicode	bit Unicode	UTF-8	percent enc.
A	0x41	0041	01000001	01000001	A
#	0x23	0023	00100011	00100011	%23
è	0x8A	00E8	11101000	*11*000011 *10*101000	%C3%A8
ξ	non esiste	03BE	0011 10111110	*11*001110 *10*111110	%CE%BE

Tabella 4.2. Esempi di caratteri codificati Unicode, UTF-8 e secondo *percent encoding*. Il carattere "`A`" è un normale carattere alfanumerico (rappresentato su 7 bit) che quindi resta come tale; il carattere "`#`" è riservato e pur essendo rappresentato su 7 bit viene tradotto con un byte; il carattere "`è`" è rappresentato su 8 bit e viene tradotto con due byte, si noti che il valore ASCII esteso non corrisponde a quello Unicode; il carattere "ξ" supera gli 8 bit e viene tradotto con due byte.

encoding [75]. Questa codifica prevede di rappresentare i byte UTF-8 come triplette di caratteri:

- Il primo è il carattere percentuale ("`%`"), che avverte dell'utilizzo di questa rappresentazione.
- Il secondo ed il terzo carattere sono la codifica esadecimale del byte UTF-8; i caratteri possibili sono quindi `0-9` e `A-F`.

Il *percent encoding* permette quindi di rappresentare non solo i caratteri estesi, ma qualsiasi carattere con la codifica UTF-8; tuttavia ciò è sconsigliato dagli standard. I caratteri che non devono essere tradotti, oltre a quelli alfanumerici, sono "`-`", "`_`", "`.`", "`~`" e ovviamente "`%`" con il significato particolare descritto; inoltre quando devono essere interpretati possono essere presenti i caratteri "`?`", "`+`", "`/`", "`:`", "`=`", "`@`" e "`&`" che in un URI hanno significati ben noti. Si noti che la codifica è *case insensitive* per cui un carattere minuscolo equivale ad uno maiuscolo, come negli URI.

In tabella 4.2 sono presenti degli esempi di codifiche di alcuni caratteri secondo le codifiche esaminate. Per la piena comprensione dello schema si precisa che un carattere il cui valore Unicode è espresso su 8 bit viene per definizione scomposto in due byte UTF-8 secondo la maschera `xxyyyyyy -> 110000xx 10yyyyyy`. Le coppie di bit più significative dei due byte (11 e 10) indicano che il carattere UTF-8 si presenta su due byte.

La codifica deve essere utilizzata anche per la trasmissione di dati del tipo MIME `application/x-www-form-urlencoded`, ovvero i dati trasmessi tramite un FORM HTML.

Il *percent encoding* viene quindi utilizzato nell'*albero* per le seguenti stringhe:

- I nomi dei file delle voci sul filesystem, utilizzando solo caratteri minuscoli; questo vuol dire che i caratteri utilizzati per la memorizzazione dei file fanno parte dell'insieme composto da "`a-z`", "`0-9`", "`%`", "`_`" e "`.`". Un esempio di file è `data/t/torino_%28disambigua%29.html` in cui si può notare la traduzione delle parentesi con il loro equivalente in *percent encoding*.

- I collegamenti HTML fra le voci; questi devono obbligatoriamente seguire il *percent encoding* per essere interpretati dal browser. Utilizzando la codifica sul filesystem sorge però un problema di corrispondenza poiché il brower interpreta la sequenza che inizia con %, nell'esempio precendente cercando di caricare il file di nome `data/t/torino_(disambigua).html`. La soluzione sta nel rappresentare il carattere `%` con il suo equivalente in *percent encoding* `%25`; la stringa del collegamento sarà quindi `data/t/torino_%2528disambigua%2529.html` che tradotta dal browser caricherà il file voluto di nome `data/t/torino_%28disambigua%29.html`.

- I nomi dei file nella base di ricerca JavaScript; questi hanno la sintassi dei file sul filesystem per evitare inutili tempi di elaborazione per effettuare la traduzione nel momento in cui devono essere caricati.

- Le chiavi di ricerca, poiché il browser invia le stringhe del FORM per immettere i dati di ricerca con questa codifica.

4.2.5 Elaborazione del testo

Durante la fase di esportazione MediaWiki si occupa di interpretare il *wikitext* delle pagine e costruisce il contenuto delle pagine HTML. In questa fase può essere necessario effettuare delle alterazioni al testo HTML; per fare ciò è quindi necessario agire su MediaWiki.

Le alterazioni al testo HTML sono disivibili in due gruppi:

- Impostazioni particolari di MediaWiki che permettono le modifiche volute. La modifica della classica pagina HTML si ottiene impostando una configurazione di MediaWiki particolare; ciò viene fatto automaticamente nel file PHP di WaNDA-tools `maintenance/dumpDVD.inc` prima di procede in qualsiasi operazione.

  A questa categoria appartiene per esempio la rimozione del collegamento "Modifica" che permette di editare il contenuto di una parte della pagina.

  Possono essere intese in questa categoria anche le estensioni di MediaWiki, che in effetti permettono di interpretare particolari campi del *wikitext* e generare codice HTML e immagini.

- Modifiche "ad-hoc" al codice di MediaWiki: ciò vuol dire modificare alcune parti del codice interno di MediaWiki, presente nella directory `includes`. Si è chiaramente cercato di limitare al massimo tali interventi, ma sono tuttavia inevitabili per una corretta generazione dell'*albero*.

  A questo tipo di modifiche appartengono principalmente quelle che modificano il formato dei collegamenti alle voci in modo da utilizzare il *percent encoding*.

4.2.5.1 Aggiustamento dei collegamenti

Le modifiche da apportare a MediaWiki per formattare con la nostra convenzione i collegamenti HTML si situano tutte nel file `includes/Linker.php`. Assieme a WaNDA-tools è presente un file `Linker.php` con indicazione chiara delle modifiche apportate.

Le modifiche apportate al codice PHP eseguono queste operazioni:

- I collegamenti HTML alle pagine non esistenti, anziché essere di colore rosso, devono essere rimossi; ovviamente deve restare la stringa del collegamento in modo da non sconvolgere il testo della voce.
- I collegamenti HTML ai progetti esterni, che appaiono come collegamenti a pagine non esistenti dato che effettivamente sul database locale non sono presenti, devono essere rimossi "in toto" compreso il testo del collegamento, per non lasciare in fondo alle pagine delle parole poco comprensibili all'utente.
- Ogni immagine viene inclusa in un collegamento HTML alla pagina dell'immagine; non avendo nel database locale le pagine delle immagini, questi collegamenti devono essere rimpiazzati con i collegamenti alla pagina online dell'immagine. È anche qui che viene effettuata la scelta se tenere o meno l'immagine richiamando il metodo di WaNDA-tools `dumpDVDDB.image_check()` descritto in sezione 4.1.1.3. Si noti che se l'immagine non è da includere tutta la cornice con la disdascalia non deve essere inclusa.
- I collegamenti alle pagine esistenti devono essere formattati con utilizzando il *percent encoding* e devono utilizzare caratteri minuscoli.

A dispetto delle apparenze le modifiche sono poco invadenti e tipicamente si risolvono nell'inserire dopo la dichiarazione di un metodo della classe `Linker` una chiamata ad un funzione definita in WaNDA-tools.

4.2.5.2 Altro

Essendo MediaWiki un progetto software in continua evoluzione possono e potranno sorgere esigenze di limitazione delle funzionalità di traduzione del *wikitext*.

Ad esempio l'unica attuale è la rimozione delle gallerie di immagini, che non sono attualmente gestibili poiché richiedono la presenza nel database locale delle pagine immagini appartenenti ad una galleria. Questa modifica si effettua commentando una riga del file di MediaWiki `includes/Parser.php`.

4.3 Presentazione dei contenuti

I contenuti dell'*albero* devono essere resi disponibili ad un'utenza il più generico possibile.

Questo pone pochi problemi per la formattazione corretta delle pagine HTML, essendo una tecnologia estremamente accolta e diffusa; le tecnologie più ricercate quali CSS e JavaScript, comunque diffuse, sono impiegate nell'ottica della compatibilità con diverse piattaforme. In quest'ottica eventuali eseguibili permettono di allargare la compatiblità con alcune piattaforme scelte.

Il discorso dell'utenza generica vale anche per il supporto di memorizzazione dell'*albero*, che deve utilizzare un filesystem diffuso ed affermato.

4.3.1 Il foglio di stile

Il foglio di stile di default, presente nel formato finale nel file `data/style/main.css`, contiene tutti gli stili, la posizione dei blocchi di testo e gli elementi grafici per la presentazione di tutte le pagine HTML, sia delle voci enciclopediche sia di quelle di servizio.

Il documento di stile è scritto con il linguaggio *Cascading Style Sheet* (*CSS*) versione 2 [76], una raccomandazione del 1998 proposta dal consorzio W3C; inoltre la sintassi dei selettori segue la versione 3, raccomandazione del 2001 [77].

Esso carica gli elementi grafici (quali sfondo, bordi, elenchi puntati), presenti come immagini PNG nella stessa directory. Il tema di partenza è *monobook*, il tema ufficiale dei progetti di Wikimedia, come appunto Wikipedia. La scelta dello stesso stile di presentazione, rilasciato assieme a MediaWiki con licenza GPL, è stata fatta in modo da rendere trasparente la transizione fra supporto offline e Wikipedia online; sia gli utento finali non vengono disorientati e confusi da differenti formati, sia l'utilizzo dell'*albero*, avendo a disposizione l'accesso ad Internet, si integra coerentemente con le pagine online.

Ogni voce è infatti dotata di un collegamento alla "versione attuale", che permette se si è dotati di accesso ad Internet di visualizzare la versione online; per questo motivo sono stati scelti per la formattazione delle voci dei colori nei collegamenti che non confondessero l'utente. I collegamenti esterni che richiedono una connessione ad Internet, vengono presentati in verde scuro, colore assente in Wikipedia online; i collegamenti interni hanno il solito colore blu, più scuro se visitato.

Il foglio di stile è stato modificato in modo da semplicarlo e renderlo compatibile al massimo con i vari browser HTML. Inoltre il CSS è stato sfruttato al massimo, trasferendovi alcuni elementi di presentazione che nel tema d'origine erano inseriti utilizzando i tag HTML; questo permette di ridurre il più possibile la dimensione occupata da ogni singola pagina HTML, come le voci enciclopediche, in modo da contenere la dimensione dell'*albero*.

Il motivo dell'utilizzo dei selettori per i CSS della versione 3 risiede unicamente nella sua capacità di formattare i collegamenti (in HTML i tag *anchor*) secondo un'espressione regolare sulla stringa del collegamento. Questo permette di decidere nel CSS quale formattazione attribuire ai diversi tipi di collegamenti, anziché indrodurre attributi specifici nel testo HTML. Si noti che il comportamento del browser che non segue questa versione di CSS ma una precedente non è definito in quanto ogni browser può aver implementato o meno questa caratteristica; nel peggiore dei casi il collegamento verrà presentato con un colore errato. I browser Gecko, tra cui quello ufficiale, sono pienamente compatibili con queste caratteristiche.

Considerando, per esempio, 300 mila voci enciclopediche le cui pagine HTML contengono in media 10 link, la presenza di un attributo HTML per formattare il collegamento richiede la presenza di almeno 12 caratteri (non estesi) nell'HTML; questo comporta un aumento di $300000 * 12 * 10$ byte, ovvero di almeno 34 MB. Piccoli accorgimenti come questo permettono di ridurre la dimensione del formato finale senza modificare l'aspetto delle pagine HTML formattate dal browser.

È inoltre presente un secondo foglio CSS di nome `data/style/commonprint.css` che

provvede a formattare la pagina HTML in caso la voglia stampare. Esso rimuove gli elementi grafici, nasconde le barre di navigazione, descrive i collegamenti esterni con l'URL fra parentesi e presenta i testi in modo da migliorarne la leggibilità sullo stampato.

Infine ma non meno rilevante l'utilizzo del foglio di stile CSS è utilissimo nel caso si voglia cambiare qualche accorgimento grafico oppure l'intero tema: il file da modificare è uno soltanto.

4.3.2 Accesso ai contenuti

L'*albero* complote deve essere fornito di almeno un meccanismo di navigazione del contenuto, in modo da poter comodamente accedere alle informazioni contenute nelle più di 300 mila voci enciclopediche. Finora il progetto WaNDA ha previsto almeno tre meccanismi principali per accedere alle voci, ovviamente collegati nella pagina principale di accesso all'albero:

- Una gerarchia di categorie e sottocategorie, che permettono di accedere alle voci enciclopediche procedendo per argomenti. Sulla pagina di accesso sono raggruppate le principali 50 categorie nei 5 temi principali: *scienze matematiche, fisiche e naturali*; *arte, letteratura, lingue, musica*; *scienze sociali, storia, geografia, religione*; *hobby e società*; *tecnologia e scienze applicate*. Si procede per categorie progressivamente più precise dell'argomento di interesse, fino a trovare il collegamento alla voce cercata.

  Rientrano in questo tipo di accesso alle informazioni anche la lista delle biografie, essendo anch'essa una categoria, e la radice di tutte le categorie, ovvero la categoria "categorie", da cui si diramano tutte le altre.

  Tutte le pagine che compongono queste gerarchie di categorie sono scritte durante la fase di esportazione delle pagine, essendo importate nel database locale.

- Un indice di tutte le voci presenti nell'enciclopedia, redirect e pagine delle categorie escluse; l'indice delle voci è disposto per ordine alfabetico crescente.

  Questo indice è presente sotto forma di varie pagine HTML, ognuna contenente un listato sottoinsieme dell'indice totale; in ogni pagina sono presenti collegamenti alla pagina iniziale, finale, precedente e successiva quella attuale. Il numero di queste pagine è definito nel file di configurazione di WaNDA-tools; più tale numero è piccolo più sono lunghe le pagine HTML. Di default sono generate 50 pagine, ed essendo circa 300 mila le voci totali in ognuna delle pagine HTML sono presenti circa 6000 collegamenti.

- Un motore di ricerca in tecnologia JavaScript. La base di ricerca è composta dai titoli delle voci, inclusi i redirect, in modo da massimizzare la probabilità di trovare la chiave cercata; la ricerca effettuata è di tipo *case-insensitive*.

- Un generatore casuale di voci. Esso sfrutta il motore di ricerca, permettendo di estrarre velocemente una pagina casualmente scelta dell'albero.

Si noti che essendoci 600 mila redirect e 300 mila voci, spesso la pagina estratta è un redirect, per cui più probabilmente verrà estratta una pagina che ha molti modi di essere scritta.

Al motore di ricerca ed al generatore casuale di voci è dedicata la sezione seguente, che ne illustra il funzionamento.

4.3.3 Il motore di ricerca

Si accede al motore di ricerca Javascript tramite un `form` HTML che si interfaccia ad un gruppo di pagine HTML contenenti diversi script JavaScript; tali pagine hanno il nome del tipo `pagina_risultati_*.html`.

Questo meccanismo composto da differenti pagine HTML è necessario poiché il JavaScript a livello implementativo non ha accesso al filesystem locale; si può però richiamare una pagina HTML che ha il nome del file noto tramite il metodo JavaScript 1.0 `location.replace()`. Questa funzione è pensata per redirigere il browser ad un URI (*Uniform Resource Identifier* [75]), in modo da poter automaticamente accedere ad una nuova pagina HTML. Così facendo si può navigare automaticamente attraverso diverse pagine HTML, accedendo di volta in volta al contenuto dei JavaScript ivi contenuti.

4.3.3.1 Trasmissione di informazioni tramite solo browser

Il passaggio da un file HTML all'altro durante la raccolta di informazioni, richiede però la memorizzazione dello stato corrente. Per trasmettere variabili ed informazioni di stato da un file HTML all'altro si utilizza in tutti i passaggi del motore di ricerca il meccanismo di GET HTTP dei FORM, definito dall'RFC 2616 sull'HTTP/1.1 e dalla raccomandazione del W3C sull'HTML 4.01 [80]. Le variabili sono quindi trasmesse come parte dell'URI da accedere, dopo il nome della pagina HTML.

Il nome del file e le variabili sono separati dal carattere `?`, mentre le variabili sono separate fra di loro dal carattere `&`. Una variabile è composta dal nome ed il suo valore, separati dal carattere `=`. Si noti che le variabili sono tutte di tipo stringa.

Il formato del GET HTTP prevede che il testo dell'URI sia codificato in ASCII. Inoltre sia il nome delle variabili che il valore non possono contenere i caratteri utilizzati come delimitatori; per questo le stringhe sono presenti nel formato *percent encoding*, la stessa codifica utilizzata per il nome delle pagine HTML, descritta in sezione 4.2.4.

Come un esempio pratico si consideri l'URI completo di una trasmissione tramite GET HTTP il cui obiettivo sia la pagina `index.html` a cui si vuole passare una variabile di nome `domanda` contenente la stringa "`che ora è?`". L'URI sarà questo:

```
index.html?domanda=che+ora+%c3%a8%3f
```

La comunicazione tra le varie componenti avviene quindi tramite il passaggio di metodi GET HTTP; tuttavia la ricerca può dover trasmettere molti dati da una componente all'altra, il che potrebbe superare i limiti del browser. Infatti sia l'RFC 2616 sull'HTTP/1.1 [78] e precedenti come l'obsoleto RFC 2068 [79], che le raccomandazioni del W3C sull'HTML

non accennano minimamente al un limite per la lunghezza di un URI; tuttavia per motivi tecnici è necessario impostare una lunghezza massima. Per poter inviare una stringa di lunghezza arbitraria infatti si può utilizzare l'altro meccanismo per la trasmissione di variabili tramite form, il metodo POST. Non è però possibile sfruttarlo con il meccanismo di redirezione del browser descritto, poiché le variabili di un POST vengono inviate ad un entità server e non possono essere catturate lavorando soltanto dal lato client, come si riesce a fare con la GET e JavaScript.

Ogni browser implementato in modo differente il limite per la lunghezza dell'URI; per esempio da prove sperimentali i browser recenti basati su Gecko [81] (ovvero Mozilla e affini) hanno un limite oltre vari miliardi di caratteri ASCII, quelli basati sulle ultime versioni di Presto [82] (Opera 7, 8 e 9) hanno un limite a 2^{16} caratteri, i browser basati su KHTML [83] (Konqueror e Safari) sono limitati a 2^{15} caratteri, mentre quelli basati su Trident [84] (Internet Explorer 4, 5.5 e 6) si fermano a soli 2^{11} caratteri.

La prima pagina di ricerca a cui si accede è `pagina_risultati_main.html`: essa è l'arbitro che interpreta i comandi e decide il da farsi e quale pagina caricare. Nel caso della ricerca estesa essa sarà anche l'ultima pagina richiamata dato che si occupa di presentare i risultati.

Si accede a questa pagina tramite un form HTML di tipo GET; l'attributo di tipo `action`, per indicare l'obiettivo del form, è quindi la pagina offline `pagina_risultati_main.html`. Le operazioni del motore di ricerca, come accennato precedentemente, sono tre: ricerca *esatta*, ricerca *estesa* ed estrazione di una pagina a caso. Il comando che seleziona con quale modalità richiamare l'arbitro della ricerca è una variabile di nome `cmd`. Nelle prime due modalità è anche necessario impostare una chiave di ricerca.

La pagina di arbitro, d'ora in poi brevemente indicata come *main*, elabora i comandi ricevuti e decide le operazioni da compiere.

4.3.3.2 Ricerca esatta

Il primo tipo di ricerca che il motore JavaScript può effettuare è il più veloce ed immediato. Esso è stato definito *ricerca "secca"* poiché illustra efficacemente l'operazione compiuta. La ricerca verifica se esiste una pagina che ha il titolo esattamente uguale alla stringa di ricerca immessa dall'utente; il confronto è *case insensitive* e quindi variazioni nei caratteri maiuscoli e minuscoli non influenzano il funzionamento della ricerca.

Dato che nei vettori di ricerca non esistono due stringhe identiche, questo è un tipo di ricerca esatta che determina in modo booleano l'esito della ricerca. Nel caso in cui il titolo venga trovato, si carica automaticamente la pagina corrispondente senza interazioni da parte dell'utente; nel caso in cui non venga trovato un titolo, la ricerca termina qui.

Per il dump di maggio 2007, il numero delle pagine corrispondenti ad una voce normale è circa 300 mila; il numero totale delle pagine, quindi contando anche le pagine di redirect, è invece circa 900 mila. Ciò significa che per ogni voce enciclopedica esistono in media tre titoli, quello della pagina vera e propria ed i due delle pagine di redirect. Inoltre nel conteggio delle voci normale ci sono anche le *voci disambigue*, che dato un termine generico offrono collegamenti a diverse voci differenti. La probabilità di trovare una voce cercata

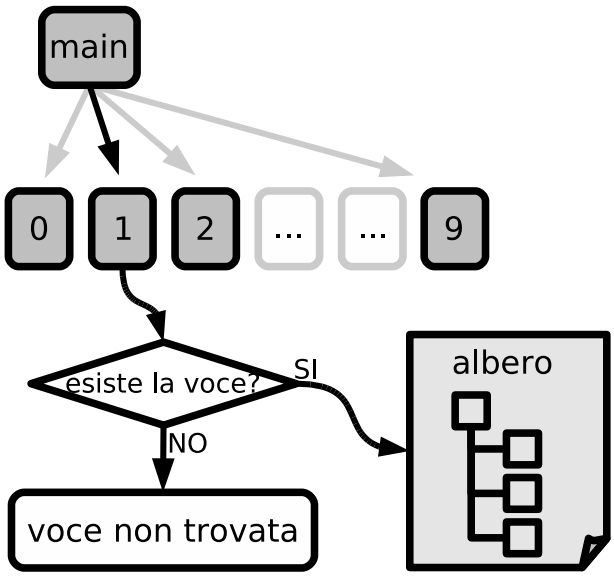

Figura 4.9. Motore di ricerca offline: ricerca esatta o "secca"

data una stringa di ricerca è quindi maggiore di quanto possa sembrare. Inoltre spesso le pagine di redirect aggiuntive ad una pagina normale sono presenti per quei titoli che sono in qualche modo complicati da scrivere oppure di solito scritti con errori ortografici; la loro presenza è utile su Wikipedia per trovare facilmente una voce ed evitare così di iniziare una voce che già esiste.

Come illustrato in figura 4.9, la ricerca parte dal *main* che determina a quale sotto-pagina passare la ricerca. Ogni sotto-pagina si occupa infatti di una parte dell'indice totale; se le sotto-pagine sono dieci, allora ognuna si occuperà di un vettore approssimativamente 1/10 dell'indice totale.

La sotto-pagina a cui passare la ricerca viene determinata secondo il primo carattere della stringa ricercata. In figura esso corrisponde alla numero 1. La pagina 1 gestisce il vettore dei titoli che iniziano con il carattere voluto; essa verifica in brevissimo tempo se la stringa esiste, poiché il vettore è una array associativo la cui chiave è il titolo. Il tempo di elaborazione si concentra quindi principalmente nel tempo impiegato dal browser nel caricare il vettore associativo.

Se il titolo viene trovato, la sotto-pagina compone il rapidamente il percorso della pagina HTML della voce utilizzando il valore del vettore associativo, che contiene il nome della directory, e il titolo stesso. Tale percorso viene quindi caricato nel browser tramite la chiamata JavaScript `location.replace()`.

Se il titolo non viene trovato la ricerca termina; tuttavia come descritto successivamente di solito si utilizza in cascata la ricerca estesa, combinando i pregi delle due strategie.

4.3.3.3 Ricerca estesa

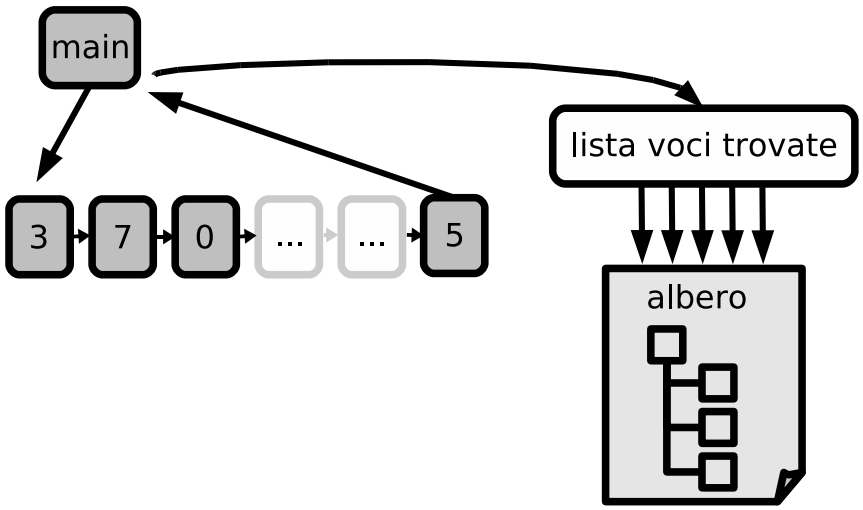

Figura 4.10. Motore di ricerca offline: ricerca estesa

La ricerca maggiormente esaustiva viene effettuata sempre sui titoli delle voci; tuttavia non si controlla se esiste un titolo che sia uguale, ma si genera una lista dei titoli che contengono la stringa cercata. Chiaramente anche questa ricerca è *case insensitive*. Questa ricerca utilizza gli stessi componenti e vettori della ricerca esatta descritta nella sezione precedente; questa volta però tutte le sotto-pagine vengono coinvolte nell'operazione di ricerca.

Il funzionamento è illustrato in figura 4.10; la pagina di partenza è sempre *main*. Essa decide per prima cosa un ordine di attraversamento delle sotto-pagine in modo casuale: questo è utile per due motivi, l'uno tecnologico l'altro di presentazione.

Nel caso il numero di risultati trovati sia troppo grande, la lunghezza dell'URI potrebbe potrebbe superare le capacità del browser. I risultati non sono quindi tutti presenti ma l'elenco appare troncato; ecco quindi che effettuare nuovamente la ricerca presenta i risultati in ordine differente. Questa soluzione cerca di risolvere un problema non definito, poiché ogni browser è dotato di una lunghezza differente per la stringa massima di un URL; si noti che il fenomeno di troncamento avviene soltanto quando sono presenti centinaia di risultati.

Il secondo motivo è rappresentato dall'impossibilità di sapere quanto un risultato sia più o meno pertinente rispetto agli altri. Non vi è quindi un ordine preferenziale per la dispozione dell'elenco dei risultati, per cui essi sono riportati secondo un ordine casuale.

La pagina *main* passa quindi il controllo delle operazioni alla prima sotto-pagina, in figura la numero 3. Essa con un ciclo cerca la stringa di ricerca in ogni chiave del vettore associato che gestisce. Siccome quest'operazione non è immediata e si tratta di elaborazioni invisibili all'utente, esso potrebbe pensare che il browser sia bloccato. Per questo prima di

partire con la ricerca nel vettore viene stampata una pagina HTML di base che presenta una barra di avanzamento. Ovviamente essendo questa la prima sotto-pagina, la barra dovrà scorrere per 1/10 della lunghezza totale.

Una volta che la prima sotto-pagina ha finito la ricerca, essa chiama la seconda sotto-pagina (numero 7 in figura), passandole con il GET HTTP i risultati trovati. La seconda sotto-pagina effettua le stesse operazioni della prima ma sul suo vettore; la raccolta di risultati continua così fino all'ultima sotto-pagina, che passa il controllo all'arbitro *main*. Questo riceve così l'elenco del risultati trovati nelle varie sotto-pagine e provvede a generare i percorsi del collegamenti alle voci, in modo da formattare una lista di collegamenti HTML da mostrare all'utente. L'utente potrà quindi scegliere quale voce visitare da questo elenco.

Questa ricerca è abbastanza lenta (su un PC di medie prestazioni nel 2006 impiega approssimativamente 30 secondi) ma fornisce un maggior numero di riscontri rispetto alla ricerca esatta. Anche su computer più lenti la presenza della barra di avanzamento assicura comunque la progressione della ricerca. Nel caso che l'*albero* sia presente su supporto ottico, il tempo di ricerca dipende in gran parte dalla qualità del lettore DVD; le memorie solide, per esempio di tipo NAND, hanno tempi che dipendono sia dalla qualità della memoria stessa che dalla banda offerta dal bus di interconnessione con il sistema, per esempio la porta USB.

Si noti che anche qui la presenza di molte pagine di redirect può migliorare l'efficacia della ricerca, specialmente nel caso di una voce il cui titolo ha diverse ortografie.

4.3.3.4 Ricerca completa

I due tipi di ricerca *esatta* e *estesa* sono complementari, poiché la prima è molto veloce e restituisce uno o nessun risultato, mentre la seconda è più lenta e restituisce un elenco di risultati. è quindi conveniente combinare i due algoritmi, effettuando per primo la ricerca esatta, e nel caso non trovi una corrispondenza effettuare la ricerca estesa: questa ricerca combinata, illustrata in figura 4.11, prende il nome di ricerca *completa*.

Nel menù di navigazione di ogni pagina dell'*albero* è presente un form per ricercare una stringa: l'operazione effettuata di default è la ricerca *completa* ma è possibile cliccare su un pulsante apposito per effettuare la sola ricerca *estesa*. Questo permette ad un utente di dirigersi semplicemente verso una voce, oppure di verificare quali voci possano riguardare la stringa cercata; questo doppio tipo di ricerca è ispirato a quello originale di MediaWiki, che offre infatti due tipi di ricerca che all'utente appaiono simili a quelli implementati.

4.3.3.5 Estrazione casuale

Il software MediaWiki dispone di un collegamento per estrarre una voce a caso; utilizzando il meccanismo della ricerca esatta è possibile disporre della stessa funzionalità.

La pagina *main* decide casualmente a quale sotto-pagina passare il controllo; essa invece sorteggierà casualmente un titolo dal vettore associativo. Dalla chiave e dal valore della cella individuata si costruisce il percorso della pagina da caricare.

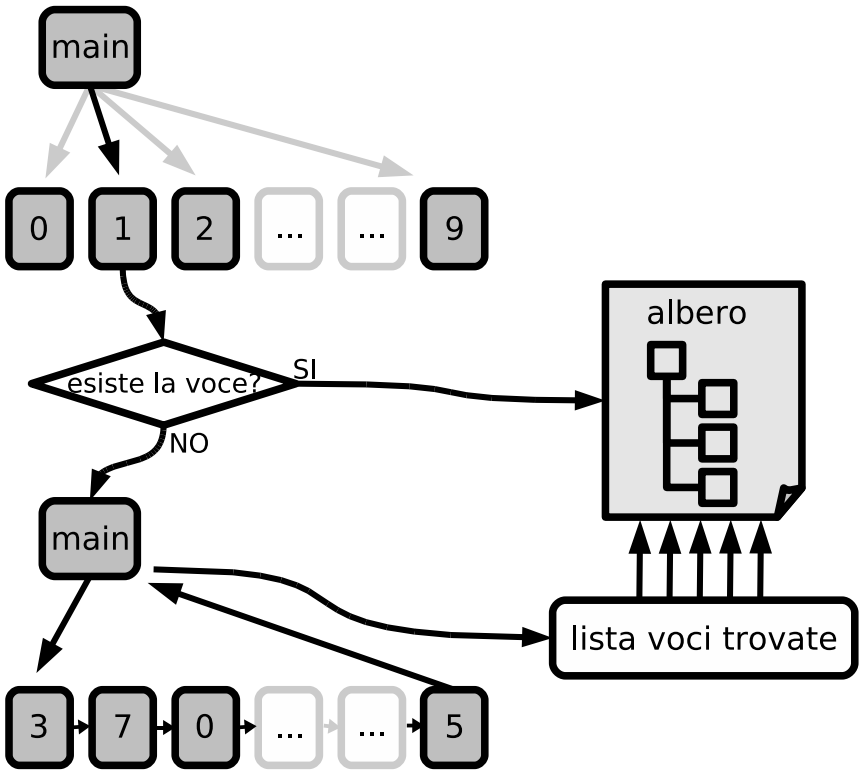

Figura 4.11. Motore di ricerca offline: ricerca completa

L'operazione è veloce, anche se leggermente più lenta dell'operazione della ricerca esatta, a causa dell'elaborazione aggiuntiva sul vettore associativo. Le estrazioni casuali sono guidate dal generatore di numeri pseudocasuali fornito dal JavaScipt 1.0, `Math.random()`, che è implementato in modo molto approssimativo.

Siccome le pagine di redirect sono in rapporto 3:1 con le voci normali, è molto frequente caricare una fra le pagine di redirect.

4.3.4 Integrazione del browser

Assieme all'*albero* sì è pensato introdurre un browser opensource basato sul motore di rendering *Gecko*; questo permette sia da una parte di fornire un meccanismo di navigazione ufficialmente testato, sia di disporre di un browser moderno dotato di linguette di navigazione, che implementi il meccanismo dei bookmark e che sia sicuro, sia di diffondere le tecnologie opensource.

È importante notare che la presenza di questo browser non è assolutamente necessaria, poiché l'*albero* è completo e funzionante anche senza di esso. Ciononostante è un valore aggiunto offerto all'utente.

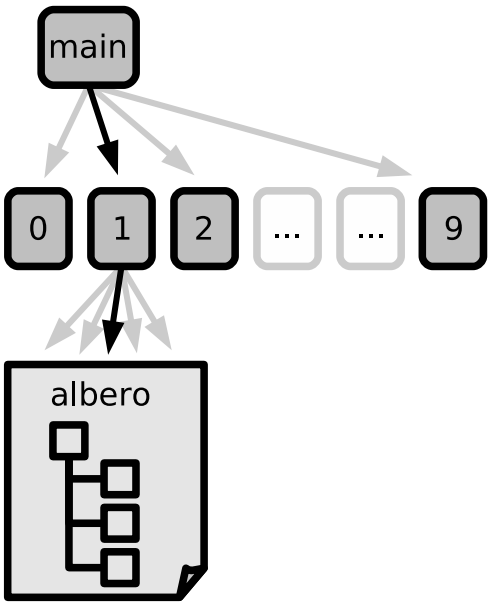

Figura 4.12. Motore di ricerca offline: estrazione casuale

Il browser deve occupare poco spazio e deve poter essere utilizzato non solo per navigare sui contenuti offline ma anche a quelli online, dato che per ogni voce è presente il collegamento alla versione online.

La presenza di un browser ufficiale basato su Gecko assicura la compatibilità con le tecnologie utilizzate, ovvero HTML 4.01, JavaScript 1.0 e CSS v2. Sui sistemi di derivazione UNIX, opensource e non, sono di solito presenti di default browser di derivazione Gecko o KHTML, per cui è già disponibile un browser dalle caratteristiche simile a quello ufficiale. Il problema si pone sui diffusi sistemi commerciali a sorgente chiuso, come Microsoft Windows e in parte minore Apple MacOS.

Per questo si è scelto di introdurre il browser *K-Meleon* [85], che già dal 2005 rispondeva alle necessità. È di derivazione Gecko, è totalmente personalizzabile, viene fornito precompilato per Windows e ha la caratteristica di poter essere eseguito direttamente da un CDROM senza richiedere l'installazione. Nel caso in questione, poiché si distribuisce una versione immagine ISO per DVD, è presente un sistema di "autorun" per Windows [86] che provvede ad avviare automaticamente K-Meleon dal DVD con la pagina iniziale `pagina_principale.html`; tutto questo è presente nell'*albero*.

In futuro, secondo la diffusione del progetto, si potrà pensare di fornire un browser anche per MacOSX, e si potrebbe scegliere un browser più diffuso e testato come Mozilla Firefox, che oggi esiste nella versione che non richiede installazione di nome *Firefox-Portable* [87].

4.3.5 Il formato del supporto

Il formato finale, che comprende tutte le voci enciclopediche, le pagine di servizio e i binari, per la consultazione deve essere posto su un supporto. Per essere comodo per l'utente il supporto deve potersi trasportare e collegato "a caldo" al sistema operativo in uso; i supporti che hanno le caratteristiche volute possono essere dischi ottici quali DVD, dischi magnetici come i microdrive e memorie solide di tipo flash come i pendrive USB e memory card.

In tutti questi dispositivi i dati sono organizzati su un filesystem tipicamente semplice e compatibile con la maggior parte delle piattaforme; i filesystem presi in considerazione, data la loro diffusione capillare e le specifiche libere, sono l'ISO-9660:1999 e il FAT32.

L'organizzazione dei dati dell'*albero*, ovvero la struttura delle directory, la lunghezza ed i caratteri dei nomi dei file, deve essere tale da adattarsi a questi filesystem.

È infine da considerare che i filesystem utilizzati dai sistemi operativi moderni sono meno restrittivi del formato ISO-9660 e FAT; onde per cui l'*albero* è memorizzabile su un insieme molto più vasto di filesystem. Per sapere se un filesystem preso in esame può custodire l'*albero* è sufficiente che risponda ai requisiti minimi illustrati nella sezione del formato ISO-9660. Per esempio può essere memorizzato sui filesystem: UDF, NTFS, UFS, FFS, EXT2, ReiserFS, ZFS, XFS, HFS, HPFS, JFS, QFS, BFS.

4.3.5.1 Il formato ISO-9660

Il filesystem ISO-9660 [88] è un formato dati standardizzato la prima volta nel 1987 per introdurre un sistema multipiattaforma per contenere dati sui dischi ottici CD-ROM. Dovendo funzionare su molte piattaforme, è abbastanza restrittivo per quanto riguarda il formato dei file ivi memorizzati; inoltre è uno standard datato e non vi era la necessità di supportare funzionalità oggi importanti per un filesystem.

Il fatto di offrire funzionalità minimali ha portato alla nascita di vari estensioni più o meno proprietarie per ogni sistema operativo. Inoltre nel tempo è stato rivisto con varie modifiche dette "livelli".

La versione più aggiornata che vuole normalizzare le funzionalità aggiunte è l'ISO-9660:1999 [89]; questa versione, indicata anche come ISO-9660 version 2, è quella più indicata per il progetto WaNDA.

I limiti del filesystem ISO-9660:1988, la prima versione, che hanno rilevanza per accogliere l'*albero* sono:

- La lunghezza massima del nome dei file è 12 caratteri; il nome del file deve obbligatoriamente seguire il formato "*nomefile*`.`*estensione*", in cui *nomefile* è lungo massimo 8 caratteri e *estensione* massimo 3. Il *nomefile* non è opzionale e nel caso l'*estensione* si mancante deve comunque essere presente il "`.`".
- La lunghezza massima del nome delle directory è 8 caratteri; non è presente un'estensione.
- I caratteri del nome dei file, delle estensioni dei file e delle directory devono far parte dell'insieme delle lettere maiuscole `A-Z`, `0-9` ed il carattere "_".

- Il livello di profondità per i percorsi non deve superare le 8 directory.
- A causa del cosiddetto *path table*, il massimo numero di sottodirectory per ogni directory è 65535.
- La dimensione massima per un file è di 2 GiB.

Il formato ISO-9660 livello 2 e livello 3 rilassa in parte i limiti imposti, in particolare:

- Il nome dei file non deve necessariamente avere un "." ed esso può essere presente ovunque tranne come primo ed ultimo carattere del nome del file; inoltre può essercene soltanto uno.
- La lunghezza masssima per il nome dei file e delle directory dipende dagli attributi impostati, e può così raggiungere i 180 caratteri.
- Il nome delle directory continua a non permettere un ".".

Invece il filesystem ISO-9660:1999 o versione 2 permette una gestione più flessibile e moderna del filesystem:

- La lunghezza massima di un nome di file o directory è 207 byte Unicode.
- Non vi è limite sul livello di profondità delle directory.
- Il carattere "." non ha un significato particolare e viene trattato come gli altri caratteri.
- I caratteri disponibili possono seguire una codifica Unicode con distinzione fra caratteri maiuscoli e minuscoli (*case sensitive*).

Questo formato permette quindi di ospitare l'*albero* senza ricorrere all'utilizzo delle estensioni al formato ISO-9660:1988. Inoltre essendo uno standard dovrebbe comportarsi in modo simile sulle diverse piattaforme. In particolare le caratteristiche utili per la memorizzazione corretta dell'*albero* sono:

- L'elevata lunghezza del nome dei file.
- Il supporto ai caratteri del formato di codifica *percent encoding* descritto in sezione 4.2.4; quindi è possibile avere nel nome dei file e directory tutti i caratteri alfanumerici, "_", "%" e ".".
- Essendo *case sensitive* ogni piattaforma deve interpretare i nomi dei file minuscoli come tali; questo è importante dato che sui diversi sistemi il formato ISO-9660:1988 con estensioni non sempre viene caricato con lo stesso *case*. La scelta di memorizzare tutti i file e directory con testo minuscolo è puramente convenzionale e sarebbe possibile utilizzare caratteri maiuscoli; l'importante è che il filesystem sia letto nello stesso modo su più piattaforme possibili.

Un'alternativa alla versione 2 che è stata testata consiste nel generare un'immagine con il primo standard ISO-9660:1988 con *level 2* dotata almeno delle seguenti caratteristiche:

- indici delle directory con Joliet (l'estensione non standard di Microsoft per stringhe Unicode), che aggiunge il file TRANS.TBL;
- indici delle directory con Rock Ridge (un'estensione non standard per sistemi operativi che hanno link a file e che siano *case sensitive*);
- rilassare in modo non convenzionale i limiti sulla lunghezza dei nomi dei file, arrivando a 178 caratteri;
- avere tutti i nomi dei file con caratteri minuscoli sugli indici delle estensioni sopracitate.

Così facendo si vìola lo standard, anche se, come verificato sperimentalmente durante fasi di testing, nessuno sistema operativo incontrato segue rigidamente l'ISO-9660; il problema è quindi generare un'immagine che si comporti in modo il più possibilmente simile sul massimo numero di piattaforme.

Si noti inoltre che non è assicurata la compatibilità del nuovo standard ISO-9660-1999 con i vecchi sistemi; tuttavia la scelta di questo formato è stata effettuata vista la perdita di notevole spazio sul DVD a causa degli grossi indici delle directory (si consideri le migliaia di file in ogni directory alfabetica). Inoltre essendo un formato comunque affermato si suppone che perduri a lungo.

4.3.5.2 Il formato FAT

Il formato FAT (*File Allocation Table*) è molto semplice e come il formato ISO-9660 ha origini lontane. Esso è stato esteso progressivamente nel tempo in modo tale da supportare le dimensioni crescenti dei supporti. L'ultima versione, rilasciata da Microsoft nel 1995, prende il nome di FAT32; tale tecnologia è stata fra l'altro oggetto di discussioni, dato che è stata registrata presso l'ufficio brevetti statunitense dopo che era già largamente diffusa e implementata [90] [91] [92]; il suo uso è pertanto esente da royalities.

Per quanto riguarda il progetto WaNDA, il suo sviluppo si è concentrato tenendo in mente il formato per dischi ISO-9660, dato che le sue specifiche sono più rigide di quelle del formato FAT32. È necessario utilizzare quest'ultima versione poiché è l'unica che supporta un volume maggiore di 2 GB, dimensione limitativa per contenere l'intero *albero*.

Analizzando punti focali per poter memorizzarvi l'*albero*, il filesystem FAT32 non offre ostacoli di sorta. Infatti:

- La lunghezza massima dei file e delle directory è data da 255 caratteri.
- I caratteri utilizzabili sono tutti gli Unicode tranne i particolari " ", "/", ":", "*", "?", "", "<", ">" e "|".
- Non vi è limite sulla struttura delle directory.

- Il filesystem è *case insensitive* ma è anche *case preserving*: ciò significa che non possono esistere due file con lo stesso nome formato da combinazioni diverse di maiuscole e minuscole, anche se la combinazione è mantenuta. Un nome di file composto da soli caratteri minuscoli verrà letto come tale su ogni piattaforma.

Esso è quindi l'ideale per la memorizzazione sui dispositivi di memorizzazione NAND come i pendrive o schede di memoria, ma anche per i palmari e smartphone dato che praticamente qualsiasi dispositivo è capace di leggere questo semplice filesystem.

Capitolo 5

Utilizzo

Questo capitolo raccoglie informazioni sull'utilizzo di WaNDA-tools. Oltre a descrivere il funzionamento del programma PHP principale, si presenta anche quello del filtro sul dump XML e l'integrazione con il software MediaWiki.

Esso si articola in tre sezioni: la prima presenta una tipica piattaforma di esecuzione con indicazione dei requisiti software ed hardware.

La seconda approfondisce tutte le fasi del processo di conversione, diviso nelle due parti di importazione del dump ed esportazione delle voci enciclopediche.

La terza sezione offre un'analisi dei tempi e delle risorse occupate durante le diverse fasi del processo. Si continua quindi con la descrizione di alcune tecniche software, hardware ed architetturali per ridurre le risorse occupate dal processo di estrazione, siano esse spazio di memorizzazione o tempi di elaborazione.

5.1 Piattaforma di sviluppo

Prima di poter generare l'immagine ISO del DVD o di un altro supporto fisico per la distribuzione dei contenuti enciclopedici, è ovviamente necessario configurare un ambiente adatto ad eseguire le varie fasi del lavoro. Esse si suddividono sostanzialmente in un processo di preparazione iniziale, il cui risultato è la costruzione di una base di dati locale simile a quella dei server online ma ridotta di dimensioni, e l'estrazione vera e propria dei contenuti, che genera la struttura finale di file e directory (l'*albero*); questo può quindi essere impacchettato nel formato del supporto scelto e distribuito.

I requisiti hardware e software della piattaforma su cui eseguire le varie fasi del processo di generazione dell'ISO sono abbastanza generici.

Il progetto WaNDA si basa su applicativi opensource e gratuitamente disponibili, quali PHP e MySQL; inoltre tali applicativi sono disponibili per la maggior parte dei sistemi operativi moderni. Ciò assicura di conseguenza la possibilità di esecuzione dei software di WaNDA-tools su molti ambienti software differenti (*multi piattaforma*).

Dal punto di vista hardware qualsiasi tipo di calcolatore con moderate risorse può essere sufficiente; è però richiesto abbastanza spazio sulla memoria secondaria, sulla quale memorizzare i dati con cui si ha a che fare. L'elaborazione è quasi totalmente automatica

per cui poche sono le interazioni con l'utilizzatore. Ne consegue che il tempo di elaborazione totale può essere relativamente poco importante; attualmente con la piattaforma di sviluppo utilizzata (un calcolatore dalle buone prestazioni) si aggira sui cinque giorni. Un'analisi sui tempi di elaborazione e la dimensione occupata nelle varie fasi è condotta in sezione 5.3 alla fine della presentazione di ogni fase. Data la mole di informazioni da elaborare un hardware scelto specificatamente per quest'utilizzo ed un sistema operativo adeguato permettono di ridurre i tempi di generazione e di assicurare la stabilità del sistema.

Verranno di seguito illustrati i punti su cui focalizzare l'attenzione per la scelta dell'hardware e degli applicativi richiesti.

5.1.1 Requisiti software

Gli applicativi richiesti sulla piattaforma di sviluppo per la fase di estrazione dei contenuti sono sostanzialmente gli stessi del software MediaWiki. Essi si riducono essenzialmente a:

- MediaWiki [93] stesso, essendo WaNDA-tools un'espansione di questo software wiki. Le versioni compatibili testate sono MediaWiki 1.5, 1.6, 1.7, 1.8, 1.9 e 1.10; future versioni di MediaWiki sono supportate finchè non cambia radicalmente la struttura principale, con la conseguente modifica sostanziale dello script `dumpHTML.php`, presente in forma immutata nelle versioni citate. Maggiori dettagli riguardanti il punto di raccordo fra MediaWiki e WaNDA-tools sono presenti in sezione 4.1.1.8.

- Interprete per linguaggio *PHP*.
  La versione richiesta per il progetto WaNDA è almeno PHP 4. Tuttavia dalla versione 1.7 di MediaWiki la versione di PHP richiesta è la 5 [108]. Si noti che a maggio 2007 l'ultima versione stabile di MediaWiki è la 1.10, per cui è desiderabile disporre almeno di PHP 5.2.

  Oltre al linguaggio PHP di base, per l'utilizzo di MediaWiki e di WaNDA-tools sono necessarie alcune estensioni per ampliarne le funzionalità implementando funzioni di sistema aggiuntive. Esse sono:

  - `php-mysql`, per consentire l'utilizzo delle funzioni di accesso al DBMS MySQL;
  - `php-mbstring`, per consentire di maneggiare stringhe *multibyte*, ovvero contenenti caratteri Unicode; dalla versione 5 di PHP è anche possibile attivare quest'abilità direttamente nel core PHP, impostandolo in fase di compilazione.
  - `php-pcre`, fornisce funzioni per l'identificazione di stringhe utilizzando le espressioni regolari *Perl-Compatible Regular Expression* (PCRE).

  Si noti inoltre che per l'utilizzo di WaNDA-tools il linguaggio PHP deve disporre dell'interfaccia a riga di comando, detta *PHP-CLI*. L'utilizzo di PHP è infatti generalmente associato ad un applicativo web server, per esempio Apache [94]. Tuttavia la generazione dell'albero a partire dalla base di dati non necessita di un web server, in quanto prevede soltanto operazioni da linea di comando. L'installazione di

questi è però consigliata in quanto principalmente rende più agevole l'installazione di MediaWiki, la cui procedura più semplice è effettuata tramite la compilazione di una pagina Web locale. Inoltre la presenza di un Web server locale permette di verificare, una volta che la base di dati è stata caricata, la corretta formattazione delle pagine presenti sul sistema prima della fase di esportazione.

- La base di dati *MySQL*.
  Il software MediaWiki può utilizzare come DBMS (*Data Base Management System*) sia MySQL [95] che PostGRESQL [96]; tuttavia tutti i progetti di Wikimedia sono basati su MySQL. Sarebbe possibile utilizzare PostGRESQL per la generazione del DVD, in quanto il database dei contenuti è presente con un formato XML pubblicamente definito. Questo formato di rappresentazione è ad alto livello rispetto ad un dump SQL e permette la successiva importazione in una qualsiasi base di dati.

  Ciononostante alcune parti del database, illustrate successivamente, sono presenti solo come dump SQL. Inoltre la scelta di utilizzare MySQL permette di minimizzare il numero di possibili incongruenze con la piattaforma Wikipedia online.

  Sia la versione 4 che la versione 5 possono essere soddisfacenti. Si noti che MySQL può utilizzare diversi motori per la gestione della base di dati: i due più importanti sono *MyISAM* e *InnoDB*. Il primo è il vecchio standard, è molto veloce nel caso di query di tipo "`SELECT`", ma presenta diversi limiti sia a livello di query (in query complicate spesso è necessario l'impiego di tabelle temporanee) che di struttura (per esempio limiti nel caso di query innestate), ed è poco resiliente in caso di guasti [97].

  InnoDB viene introdotto come default nella versione 5; ha però una carenza per quanto riguarda la ricerca estesa di tipo "fulltext" negli indici [98]. Avendo MySQL 5 a disposizione, MediaWiki permette di utilizzare entrambi i motori citati; si utilizza ovunque InnoDB, tranne per la ricerca dei titoli sulla tabella indicizzata dove si utilizza MyISAM con una conversione ad Unicode. Inoltre MySQL 5 ha una gestione più semplice delle stringhe Unicode, che permette quindi di avere tabelle indicizzate con stringhe che contengono caratteri estesi. In definitiva la versione raccomandata di MySQL è quindi la 5.

5.1.1.1 Estensioni

Oltre all'installazione di MediaWiki, sui server ufficiali dei progetti *wiki** sono presenti innumerevoli estensioni (*extentions*); alcune sono fondamentali alla comprensione del testo, quale la resa grafica delle formule matematiche, altre sono inutili al fine del progetto, quali la gestione di contenuti interattivi. Inoltre uno dei principi base del progetto è la resa delle pagine possibilmente più simile fra albero offline e pagine online.

È così necessario installare almeno quattro di queste estensioni, le più importanti:

- *ParserFunctions* [99], che espande le funzionalità del wikitext, permettendo di riconoscere particolari tag per la gestione di meccanismi di logica nel testo; per esempio

è possibile che il wikitext presenti un contenuto differente a seconda di particolari condizioni. Esso inoltre è molto utilizzato nei collegamenti ai progetti wiki non enciclopedici.

Una sua assenza provoca la presenza di sgradevole testo spurio e collegamenti privi di significato in innumerevoli pagine HTML.

- *Cite* [100], che interpreta i tag `<ref>` del wikitext per la costruzione automatica dei riferimenti nelle citazioni. Siccome una delle recenti linee guida di Wikimedia è quella di inserire i riferimenti per ogni affermazione presentata, molte pagine utilizzato questa estensione per formattarli automaticamente in una sezione finale delle pagine indicata con il nome "Note".

  La mancanza di quest'estensione semplicemente non popola la sezione indicata.

- *texvc* permette di convertire principalmente le formule matematiche ed i caratteri delle citazioni in una rappresentazione grafica piacevole. Il linguaggio utilizzato è il TEX , che viene convertito grazie ad un eseguibile il cui codice sorgente è distribuito assieme a MediaWiki.

  Per l'esecuzione richiede la presenza di *OCaml* versione 3 [101] e di *dvipng* [102].

  Le immagini prodotte sono in formato `PNG` [70]; un'immagine mancante provoca la visualizzazione sulla pagina HTML del testo matematico in linguaggio TEX .

- *Timeline* per rappresentare molti dei grafici presenti nell'enciclopedia; per esempio nella pagine di ogni comune è presente un'istogramma che rappresenta l'andamento della popolazione. Le immagini sono presenti in formato `PNG`, e vengono generate a partire di dati testuali con il programma open source *Ploticus* [103]; solo la versione 2.32 è funzionante con l'estensione Timeline. Inoltre quest'estensione di MediaWiki utilizza per interfacciarsi a Ploticus uno script in linguaggio Perl [104], sviluppato dal team di Wikimedia, compatibile con qualsiasi versione.

5.1.1.2 Software aggiuntivi

Altri software che possono essere utili al progetto WaNDA nel processo di generazione sono:

- Un'implementazione compatibile del processore sintattico *AWK*; questo è comunemente presente sui sistemi UNIX o derivati in quanto è un programma del corredo base utile in moltissime situazioni.

- Un applicativo per il download di file tramite HTTP da linea di comando; è consigliabile che permetta il ri-trasferimento automatico in caso di fallimenti, date le caratteristiche dei download da effettuare. Anche questo di solito è già presente nei sistemi UNIX compatibili, essendo parte dei comandi base; nel caso pratico è stato utilizzata l'implementazione BSD chiamata *fetch*.

- Il classico comando `tar` abbinato con la libreria di compressione *bzlib* per la decompressione degli archivi bzip2. Fa parte dei comandi base di tutti gli UNIX moderni.

- Un eseguibile per calcolare l'hash MD5 di un file; è utile per poter distribuire le immagini dei supporti assieme ad un checksum che ne verifichi la correttezza. Un comando di questo tipo è presente di base su tutti gli UNIX; è possibile cambiare il nome dell'eseguibile nel file di configurazione di WaNDA-tools in modo da scegliere un algoritmo differente, come per esempio il più lento SHA-1.

- Un applicativo per la conversione di immagini da linea di comando. È meglio ridimensionare le immagini ottenute dal download dalla raccolta multimediale online di Wikipedia detta *commons*, per limitare lo spazio utilizzato dall'enciclopedia; l'operazione è effettuata automaticamente. Un applicativo comodo a questo scopo è *convert*, fornito con la suite *ImageMagick* [105].

- Un applicativo per la generazione di un'immagine ISO. Quest'applicativo serve solo nel caso si voglia tenere e distribuire il contenuto dell'albero sotto forma di immagine DVD. Il software deve poter generare il formato ISO-9660:1999 detto anche ISO-9660 versione 2. Questo permette la massima compatibilità con tutti i sistemi moderni, permettendo di gestire file con nomi lunghi fino a 207 caratteri e distinzione fra maiuscole e minuscole.

  Il software di riferimento per l'operazione di impacchettamento dell'albero nell'immagine ISO è *mkisofs*, parte del pacchetto *cdrtools* [106]. Esso supporta tutti i formati ISO e le varianti non standard, con la possibilità di una selezione fine ed accurata di ogni parametro. Per impostare la creazione di un'immagine ISO-9660:1999 basta utilizzare il comando `-iso-level v4`.

  Il pacchetto cdrtools è disponibile gratuitamente ed è opensource; viene distribuito su licenza mista GPL e CDDL dal 2006, anno in cui è nato un fork totalmente GPL di nome *cdrkit*, di cui fa parte *genisoimage* che ne mantiene le stesse caratteristiche e opzioni da linea di comando del corrispondente mkisofs [107]. I due comandi *mkisofs* e *genisoimage* possono essere indifferentemente utilizzati.

5.1.1.3 Sistema operativo

Come già evidenziato, gli applicativi richiesti dal progetto sono disponibili per tutte le piattaforme moderne; nella scelta del sistema operativo è però maggiormente comodo lavorare con sistemi UNIX derivati, per differenti motivi:

- offrono ottime garanzie di robustezza e performance;

- appena installati dispongono di un ambiente pronto all'uso e di facile utilizzo;

- la piattaforma di Wikipedia è di questo tipo, con la conseguenza che la configurazione e le estensioni di MediaWiki sono sicuramente applicabili;

- sono congeniali al contesto del progetto;
- ne esistono alcuni open source e senza costi di utilizzo.

È stata testata un'installazione sia di *GNU/Linux* (distribuzione *Gentoo* versione 06 e 07) che di *FreeBSD* (versione 6.2-STABLE). Il secondo sistema è stato definitivamente scelto, soprattutto a causa di maggiore stabilità, ma anche per comodità d'uso personale; si consideri che il sistema deve sopportare un notevole carico, essendo impegnato per diversi giorni ad un continuo troughput tra la memoria principale e la secondaria, ed avendo il processore sempre impegnato in elaborazioni. Inoltre su FreeBSD sia l'ambiente sia il programma di *nawk* (l'implementazione BSD dell'interprete AWK) sono risultati maggiormente idonei all'iniziale lavoro di scrematura del dump XML.

Infine ma non meno importante, il sistema operativo scelto deve supportare l'hardware scelto; in particolare nel caso di un hardware SMP deve supportare differenti processori e disporre di librerie *multi-threaded*. A questo proposito si noti che le versioni della base di dati MySQL citate sono predisposte al multi-thread.

5.1.1.4 Spazio richiesto

Per la configurazione del sistema operativo l'unica raccomandazione è lasciare uno spazio sufficiente per la partizione `/var` per i sistemi UNIX derivati o equivalente altrove, che dovrà contenere sia il contenuto della base di dati che l'albero di esportazione. Per stimare lo spazio richiesto è possibile elencare tutte le componenti che occupano maggiormente spazio durante l'elaborazione:

- i dump compressi ottenuti da Wikipedia (almeno 4 GB);
- i dump ridotti da importare nella base di dati (almeno 4 GB);
- il database (almeno 7 GB);
- la directory che contiene le immagini generate da MediaWiki (circa 1 GB);
- la directory che contiene il futuro contenuto dell'albero (4 GB);
- l'immagine del DVD e l'archivio offline di un albero (9 GB).

Lo spazio totale richiesto dal progetto WaNDA è quindi di almeno 30 GB. Le dimensioni sono approssimative, ma non possono che aumentare vista la continua crescita di Wikipedia Italia. In aggiunta il sistema tipicamente deve mantenere come precauzione un dump di backup della precedente basi di dati elaborata, le immagini già ottenute dalla rete in modo da non doverle richiedere nelle esportazioni successive, e gli archivi dell'albero delle versioni precedenti.

5.1.2 Analisi hardware

È evidente che il primo requisito hardware è una memoria secondaria con spazio sufficiente in grado di contenere i file sopra elencati. Inoltre l'accesso ai dischi rigidi deve essere caratterizzato da un'elevata banda di accesso, in quanto ne giova grandemente le prestazioni della base di dati. Per questo nel sistema adottato ci si è avvalsi di un bus SCSI Ultra 320 con un disco da 145 GB a 10000 RPM con basso seek time.

La fase di estrazione esegue il bytecode PHP, eseguito dall'interprete Zend, che effettua innumerevoli query SQL al DBMS locale; per migliorare le prestazioni è consigliabile l'impiego di un sistema con più processori. Il codice PHP viene eseguito su un unico processore, essendo il bytecode PHP eseguito come unico thread, mentre il DBMS MySQL può utilizzare molti processori disponibili, secondo le sue necessità.

Inoltre l'affiancamento, oltre che di cache di primo livello, di più una larga cache di secondo livello migliora notevolmente la fase di filtro decompressione bzip2 e l'esecuzione del codice PHP. È un fattore spesso sottovalutato ma molto importante; tipicamente i processori per server hanno una buona cache di secondo livello.

Oggi giorno a causa del costi contenuti i calcolatori di fascia medio-bassa sono sempre più spesso dotati di processori con due o quattro unità elaborative (dual-core); abbinati con adeguata memoria principale e abbastanza cache velocizzano notevolmente il processo di estrazione. L'hardware impiegato era composto da un processore Xeon dual-core con 2MB di cache di secondo livello e 2 GB di RAM. Rispetto ad un precedente sistema dotato di processore Pentium 4 con maggiore frequenza di clock e quindi con poca cache è stato possibile stimare che la sola introduzione del nuovo processore abbia più che dimezzato i tempi di elaborazione preventiva del dump XML.

Per il download del dump da Wikipedia e l'accesso alle immagini di `upload.wikimedia.org` è necessario una qualche sorta di collegamento a Internet. Essendo le dimensioni dei file in gioco considerevoli può essere opportuno considerare un accesso veloce.

Per il resto non ci sono particolari requisiti. Bisogna inoltre tenere in conto che tale sistema deve restare in funzione a massimo carico per alcuni giorni, per cui un locale adeguatamente protetto e termoregolato è consigliabile.

5.2 Il processo di estrazione

Il processo di costruzione dell'albero a partire dai dati disponibili su Internet si compone di varie fasi, che vengono ora illustrate passo per passo; tutte le fasi, illustrate con i blocchi di figura 5.1, si possono raggruppare in tre sezioni:

- creazione di una base di dati locale contenente una versione ridotta della parte italiana di Wikipedia;
- estrazione delle pagine (l'*albero*), inclusi eventuali filtri iniziali sulla base di dati ed la gestione delle immagini;
- impacchettamento e verifica dell'albero da mettere sul supporto di distribuzione.

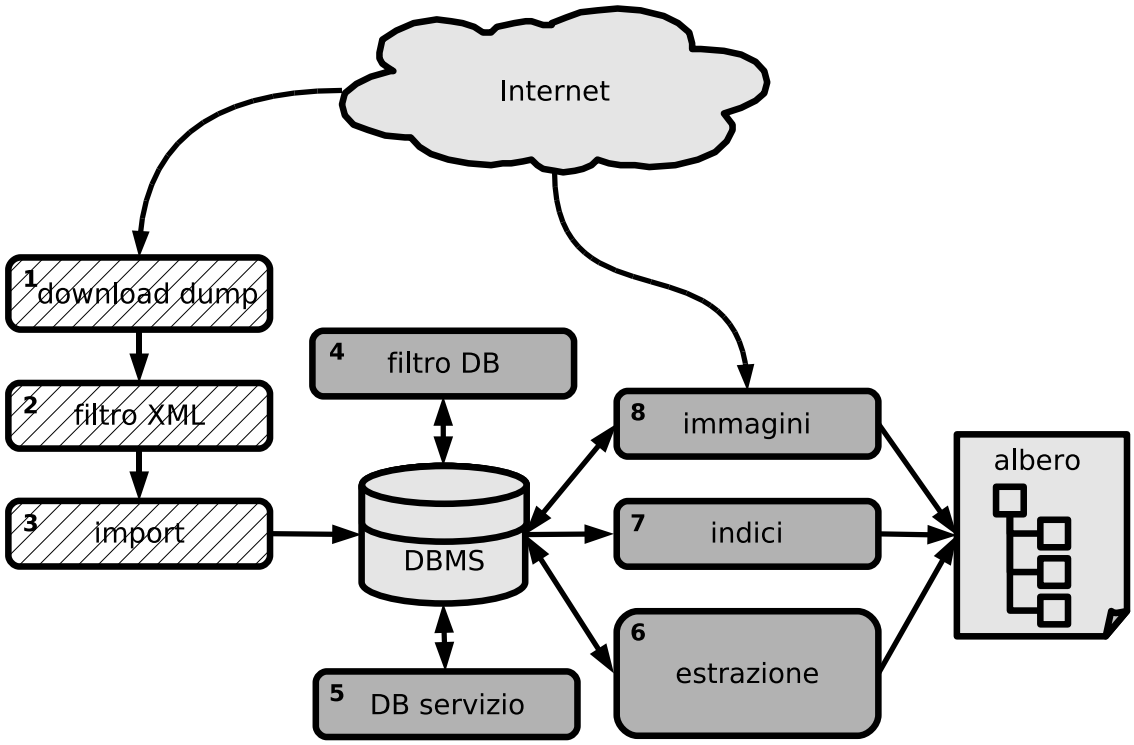

Figura 5.1. Le fasi del processo totale. In ordine sono download (1), filtro (2) e import (3) del dump XML nella base di dati; filtro sul database (4), creazione del database di servizio (5), fase di estrazione delle pagine HTML (6), generazione degli indici per la ricerca (7) e gestione delle immagini (8). Le fasi sono divise in due sezioni, l'importazione del dump e l'estrazione delle pagine. Le fasi di estrazione (segnate con il colore scuro), dal filtro sul database alla gestione delle immagini, sono implementate dalla parte PHP di WaNDA-tools.

Le prime due sezioni sono le più importanti; corrispondono alle due fasi descritte in sezione 3.3.2.

L'ultima sezione è in verità molto versatile in quanto sia il supporto di distribuzione può essere variegato, andando dall'immagine ISO per un DVD ad un filesystem FAT32 per un'unità di memoria solida, sia la verifica può essere effettuata con differenti approcci, dalla *community* ai programmi automatici. Per questi motivi questa fase verrà tratta nel capitolo seguente, concentrandosi ora all'analisi dei primi due punti, esaminando anche i tempi e lo spazio occorrente; a questo riguardo si noti che tutte le misure riportate sono tratte da un'elaborazione effettuata sull'hardware descritto all'inizio del capitolo con il materiale disponibile in maggio 2007, che ha permesso di generare l'immagine ISO versione `0.21-01Jul07`.

5.2.1 Installazione e configurazione dell'ambiente

Prima di procedere all'esecuzione delle fasi proprie del processo di generazione dell'albero raggruppate nelle tre sezioni sopraelencate, è necessario mettere a punto il sistema ed i requisiti software descritti precedentemente.

Si installa quindi il DBMS MySQL e l'interprete del linguaggio PHP con i relativi moduli. L'installazione del server web Apache è opzionale, in quanto non strettamente necessario ma utile per agevolare l'installazione di MediaWiki; inoltre permette la verifica a campione della corretta configurazione di MediaWiki e della basi di dati dopo l'importazione, prima di passare alla lunga operazione di esportazione delle pagine.

L'installazione di MediaWiki, nel caso Apache sia presente, consiste nello decomprimere un archivio contenente i vari file PHP in un direttorio e nell'accedervi tramite un programma di navigazione web (browser). Compare quindi all'indirizzo della macchina in questione una pagina di configurazione, in cui vengono specificati i parametri di accesso ed il tipo di DBMS; per la codifica di testi memorizzati nel database è opportuno utilizzare la stessa configurazione di wikipedia, per cui si deve scegliere l'impostazione classica (non sperimentale) per mysql 4; la codifica è binaria e non UTF-8.

A questo punto sarà necessario copiare il file `LocalSettings.php` secondo quanto indicato sul browser; questo è il file di configurazione generale di MediaWiki.

Nel caso in cui il server web Apache non sia presente, è possibile installare MediaWiki con un pò più di lavoro decomprimendo l'archivio in una directory a piacimento e copiare il file di configurazione `LocalSettings.php` fornito assieme a WaNDA-tools; il file di configurazione va ancora personalizzato con i parametri di accesso alla base di dati, in modo simile alla procedura tramite browser web descritta prima. Inoltre per la creazione delle tabelle nel database è necessario eseguire i due script in linguaggio SQL `maintenance/tables.sql` e `maintenance/database.sql`.

Il pacchetto `WaNDA-tools` contiene, oltre alcuni file di corredo (fra i quali il `LocalSettings.php` preconfigurato), una patch a MediaWiki, sotto forma di una directory di nome `maintenance`; questa contiene tutti i file PHP utili all'estrazione (contenenti le classi descritte nel capitolo 4) e deve essere copiato sopra la directory `maintenance` presente in MediaWiki, in modo da fare un merge dei file. A questo proposito si noti che tutti i file di WaNDA-tools hanno come prefisso nel nome del file `dumpDVD`, per cui le componenti aggiunte sono facilmente identificabili e non sovrascrivono parti di MediaWiki.

Si procede a questo punto all'installazione delle estensioni di MediaWiki, almeno quelle elencate in sezione 5.1.1.1. Si noti che la loro attivazione richiede la modifica del sopracitato file di configurazione `LocalSettings.php`; può anche essere utilizzata direttamente il file `LocalSettings.php` di riferimento, fornito assieme a WaNDA-tools.

Per il funzionamento delle estensioni di MediaWiki è possibile che sia necessario configurare alcuni parametri aggiuntivi.

Per esempio l'estensione *texvc* può richiedere la compilazione del binario; i sorgenti sono siti in `math/` e per compilarli basta eseguire il `Makefile` utilizzando *gmake*.

Inoltre per quanto riguarda l'estensione *timeline* può essere necessario personalizzare

nel file `extensions/timeline/EasyTimeline.php` il percorso di *Ploticus* e di una sua directory temporanea.

Oltre al file di configurazione generale è necessario modificare secondo le proprie necessità il file di configurazione di WaNDA, sito in `maintenance/dumpDVD.ini`; il formato è quello standard PHP. Esso si suddivide in tre parti: la prima raccoglie le opzioni della release da esportare, come la versione e le categorie delle immagini con licenza desiderata da includere nell'albero; la seconda permette di configurare i percorsi, come la destinazione dell'immagine ISO; la terza contiene molte opzioni avanzate, fra le quali potrebbe essere necessario adeguare i percorsi completi dei comandi di base utilizzati e del programma per costruire il filesystem ISO-9660.

Infine si riporta l'indirizzo del sito presso il quale è possibile trovare tutte le configurazioni delle piattaforme di Wikipedia, qui raggiungibile [109].

5.2.2 Importazione della base di dati

Il primo passo di figura 5.1 è il download dei dump della base di dati di Wikipedia da Internet. Essi sono liberamente disponibili all'indirizzo `http://download.wikipedia.org/`; possono essere utili come backup in caso di guasti. Inoltre, considerato che un download pagina per pagina dell'intera Wikipedia è vietato dalle regole di utilizzo per problemi di carico (la questione è stata descritta nel capitolo 3), il dump fornisce tutti i contenuti testuali di Wikipedia, che sono appunto coperti dalla licenza GNU FDL. Le immagini non sono incluse in questi dump, sia per motivi logistici (le immagini sono presenti in gran parte su un server apposito di nome `commons`) sia per questioni di licenza, poichè come visto non tutte le immagini sono poste con medesima licenza e solo alcune sono liberamente ridistribuibili.

Sono presenti differenti dump con diverse selezioni delle voci: ai fini del progetto WaNDA serve quello contenente sia le voci che ogni revisione delle stesse (è quello più corposo), ma ci sono anche i dump con solo gli articoli o con le pagine di discussione. I formati dei dump presenti sono sostanzialmente due:

- l dump dei testi, sotto forma di documento XML (*eXtended Markup Language*), i cui tag sono definiti da un XSD (*XML Schema Definition* [110]) presente all'indirizzo [111]. Di questo tipo è il dump di tutte le voci con ogni revisione. I dump sono compressi con bzip2, un algoritmo di compressione molto efficace sul testo, poichè utilizza nella codifica di Huffman dei dizionari relativamente grandi, 900 kB; come rovescio della medaglia ha dei tempi relativamente lunghi di compressione e decompressione.

- I dump di tipo SQL con sintassi MySQL, che contengono i dati non testuali, come le tabelle di collegamenti, presenti nel database di Wikipedia e non inclusi nei dump testuali. I dump sono compressi con gzip, un algoritmo veloce che meglio si adatta vista la limitata dimensione di questi dump.

Nel caso italiano il repository comune è `http://download.wikipedia.org/itwiki/`; qui è sempre disponibile la versione più aggiornata dei dump. Si noti che durante l'esecuzione del backup da parte di Wikipedia i dump non sono disponibili; eventuali note sono riportate in testa al sito. Per ricreare in locale la base di dati contenente le informazioni utili al progetto è necessario ottenere il dump XML `itwiki-latest-pages-meta-history.xml.bz2`, che contiene tutte le voci di tutti i namespace con ogni revisione di queste: bisogna avere tutte le revisioni di un articolo poichè assieme ad ognuna è indicato l'autore, e secondo la licenza GNU FDL nel redistribuire i contenuti è obbligatorio indicare *tutti* gli autori che vi hanno contribuito. Il file è grande circa 4 GB compresso, decompresso sfiora i 100 GB, il che chiaramente giustifica l'utilizzo del lento ma efficace bzip2 per comodità di immagazzinamento e di trasferimento. Inoltre per i collegamenti alle immagini con le categorie di appartenenza è necessario procurarsi il dump SQL con i collegamenti delle categorie `itwiki-latest-categorylinks.sql.gz`; compresso è meno di 20 MB, decompresso sui 100 MB; il dump SQL è più grande dello spazio occupato dalle stesse informazioni nella base di dati.

5.2.2.1 Filtro sul dump XML

I due dump così come sono possono essere importati nel DBMS MySQL locale. Tuttavia per quanto riguarda il dump XML degli articoli ci si rende conto che la dimensione di tale base di dati è elevata; di più il dump XML occupa meno spazio che le stesse informazioni presenti in MySQL. Ed il tempo richiesto all'operazione di inserimento nella base di dati locale è tanto più lenta quanti sono i dati da inserire.

Da queste considerazioni si è valutato, esaminandone il formato, che alcune informazioni inutili potevano essere rimosse. Nella sequenza di revisioni di una voce, oltre all'autore ed alla data di ogni revisione, è riportato integralmente tutto il testo dell'articolo di questa specifica revisione; per cui l'ultima revisione contiene la il contenuto completo della pagina più recente. Dato che sull'albero interessa avere solo la pagina più recente, con indicati gli autori di tutte le revisioni, si è valutato di azzerrare il contenuto di tutte le revisioni tranne la più recente per tutti gli articoli.

Considerato che nel dump XML le revisioni per un articolo sono anche centinaia, questa rimozione alleggerisce parecchio il dump. È stato quindi sviluppato uno script con il processore grammaticale AWK, chiamato `xmldumpskimmer.sh`, che esamina un file XML che segue il formato del file XSD sopracitato e provvede alla rimozione del contenuto delle revisioni tranne della più recente, lasciando intatte le informazioni sugli autori e sul timestamp; esso accetta in *standard in* il dump XML e in *standard out* stampa l'XML modificato.

I risultati sono ottimi, in quanto il dump XML decompresso e filtrato per la versione di maggio 2007 è di soli 3.8 GB, con un ratio di 1:20.

Si noti che il dump XML così filtrato contiene ancora molte informazioni non fondamentali ai fini del progetto, in quanto prima dell'esportazione si effettua una serie di query MySQL per filtrare contenuti indesiderati, come per esempio le voci appartenenti a namespace indesiderati (4.2.1). Tuttavia in questa fase si preferisce effettuare un primo

livello di sgranatura del dump, senza entrare nei dettagli delle selezioni più delicate effettuate successivamente; inoltre la selezione su MySQL è, a causa delle tabelle indicizzate del DBMS, più veloce che la sequenziale elaborazione dei testi utilizzando AWK.

Lo script AWK è stato scritto in due versioni, nel tentativo di ridurre i tempi del processo. La sintassi di lancio è:

```
bzip2 -dc itwiki-latest-pages-meta-history.xml.bz2 | \
./xmldumpskimmer.sh > itwiki-latest-pages-meta-history-RIDOTTO.xml
```

Il lancio dello script è come si vede in pipe con bzip2, che decomprime il dump XML nel suo standard in. Questo comando con la prima versione dello script impiega 72 ore. La seconda versione è stata riscritta in modo da migliorare l'esecuzione del codice e probabilmente lo script non è ulteriormente ottimizzabile; esso impiega 68 ore.

Nelle figure 5.2, 5.3, 5.4 e 5.5 sono riportate l'occupazione delle risorse sulla piattaforma FreeBSD durante l'esecuzione del filtro.

```
last pid: 89405;  load averages:  1.00,  1.06, 1.35    up 12+05:00:25  23:12:48
43 processes:  2 running, 41 sleeping
CPU states: 49.6% user,  0.0% nice,  0.0% system,  0.0% interrupt, 50.4% idle
Mem: 41M Active, 729M Inact, 193M Wired, 37M Cache, 112M Buf, 1660K Free
Swap: 2023M Total, 104K Used, 2023M Free


  PID USERNAME  THR PRI NICE   SIZE    RES STATE  C   TIME   WCPU COMMAND
95948 err         1 119    0  2796K  1952K RUN    0 255:29 84.24% awk
33216 err         1  -8    0  5004K  4280K pipewr 0  12:14  8.22% bzip2
93335 root        9  20  -76  3716K  1552K kserel 0  51:42  0.00% hpasmd
55121 root        1  96    0 14204K  7476K select 0   0:16  0.00% httpsd
...
```

Figura 5.2. Decompressione del dump XML e filtro AWK: output del comando `top`.

5.2.2.2 Costruire la base di dati

A questo punto il dump XML ridotto contenente le voci ed il dump SQL con i collegamenti delle categorie devono essere importati nel DBMS locale; l'ordine di importazione è ininfluente. Per il secondo, essendo già nel formato SQL di MySQL, è semplicemente eseguito dal client di accesso al database:

```
gunzip itwiki-latest-categorylinks.sql.gz | mysql -u user -p wikidb
```

Il tempo per eseguire questo comando, che decomprime l'archivio ed importa i dati, è di circa 30 minuti. Le categorie presenti sono 29589, che formano un totale di 1701941 collegamenti tra categorie e voci.

```
    1 users    Load  1.01  1.07  1.36                  Jun  4 23:12

Mem:KB    REAL            VIRTUAL                       VN PAGER  SWAP PAGER
        Tot   Share      Tot    Share    Free           in  out     in  out
Act   31552    6096    67844    12796   39848 count
All 1022204    7900152977308    17520         pages
                                                        Interrupts
Proc:r  p  d  s  w    Csw  Trp  Sys  Int  Sof  Flt        cow    4014 total
     1     7 34       447  185  358  192  192    7 197328 wire         1: atkb
                                                  41400 act          6: fdc0
 0.4%Sys   0.0%Intr 49.6%User  0.0%Nice 50.0%Idl   746104 inact        15: ata
|    |    |    |    |    |    |    |    |    |      38188 cache      6 16: uhc
>>>>>>>>>>>>>>>>>>>>>>>                             1660 free       5 48: bge
                                                        daefr      1 72: cis
Namei         Name-cache    Dir-cache                   prcfr   2001 cpu0: time
    Calls     hits    %     hits    %                   react   2001 cpu1: time
                                                        pdwake
                                       7 zfod           pdpgs
Disks   da0 pass0                      3 ozfod          intrn
KB/t  72.00  0.00                     44 %slo-z   114304 buf
tps       1     0                     24 tfree        11 dirtybuf
MB/s   0.06  0.00                                  70236 desiredvnodes
% busy    1     0                                 58505 numvnodes
                                                  17559 freevnodes
```

Figura 5.3. Decompressione del dump XML e filtro AWK: output del comando `systat -vmstat 1`.

```
                   /0   /1   /2   /3   /4   /5   /6   /7   /8   /9   /10
     Load Average   >>>>>

                   /0   /10  /20  /30  /40  /50  /60  /70  /80  /90  /100
root     idle: cpu1 XXXXXXXXXXXXXXXXXXXXXXXX
err              awk XXXXXXXXXXXXXXXXXXXXXX
err           bzip2 X
```

Figura 5.4. Decompressione del dump XML e filtro AWK: output del comando `systat -pigs 1`.

Altra questione è invece l'importazione delle voci contenute nel dump XML [113]. È possibile importarlo nel DBMS locale utilizzando uno fra i seguenti diversi programmi:

- *maintenance/importDump.php*
  Questo script PHP è fornito assieme a MediaWiki e permette sempre l'inserimento, in quanto è l'implementazione ufficiale; si noti che il codice è speculare a exportDump.php, il programma utilizzato invece per generare il dump XML. Il suo grave

```
                  /0   /1   /2   /3   /4   /5   /6   /7   /8   /9   /10
     Load Average   |||||

          /0   /10  /20  /30  /40  /50  /60  /70  /80  /90  /100
cpu  user|XXXXXXXXXXXXXXXXXXXXXXX
     nice|
   system|
interrupt|
     idle|XXXXXXXXXXXXXXXXXXXXXXX

          /0   /10  /20  /30  /40  /50  /60  /70  /80  /90  /100
da0   MB/s
      tps|X
pass0 MB/s
      tps|
```

Figura 5.5. Decompressione del dump XML e filtro AWK: output del comando `systat -iostat 1`.

limite è l'inefficienza, in quanto ha tempi elevati e necessita di esagerata memoria nel caso di un dump XML come quello in questione; il suo uso è quindi sconsigliato dal team stesso di Wikipedia. Sulla piattaforma di estrazione descritta esso utilizza tutta la memoria virtuale dopo poco tempo, obbligando il sistema a terminare il processo.

- *mwdumper* [114]
  Si tratta di un parser XML scritto in Java che riconosce il formato XSD del dump, e lo trasforma in comandi SQL per riempire la base di dati di MediaWiki. Le tabelle della base di dati devono avere il formato di Mediawki 1.5, lo stesso formato con cui è testato Wanda. È multipiattaforma essendo scritto in java e viene distribuito già precompilato; essendo un parser completo del formato XSD occupa abbastanza risorse, sia come tempi che come memoria. Per l'esecuzione è richiesta una JVM (*Java Virtual Machine*) superiore a Java 1.4, meglio se versione 1.5 (detta J2SE [112]).

  Sono stati rilevati miglioramenti in termine di velocità eseguendo l'ambiente Java con almeno 200 MB riservati per la memoria heap e scegliendo la JVM in modalità ottimizzata per server.

  Il comando utilizzato per eseguirlo è quindi:

  ```
  cat itwiki-latest-pages-meta-history-RIDOTTO.xml | \
  java -Xmx200M -server -jar mwdumper.jar --format=sql:1.5 | \
  mysql -u user -p wikidb
  ```

  Questo sistema è quello preferito in quanto è il giusto compromesso tra compatibilità

con il formato dei dump e tempi. L'importazione del dump XML filtrato con lo script AWK precendente impiega 7 ore.

Nelle figure 5.6, 5.7, 5.8 e 5.9 sono riportate le statistiche sul carico del sistema che effettua l'importazione.

```
last pid: 8810; load averages: 0.37, 0.25, 0.18        up 24+03:06:55  21:19:18
59 processes:  1 running, 58 sleeping
CPU states:  1.7% user,  0.0% nice,  5.6% system,  0.0% interrupt, 92.7% idle
Mem: 107M Active, 667M Inact, 184M Wired, 41M Cache, 112M Buf, 1660K Free
Swap: 2023M Total, 176K Used, 2023M Free

  PID USERNAME  THR PRI NICE   SIZE    RES STATE  C   TIME   WCPU COMMAND
98947 mysql       8  20    0 61272K 39120K kserel 0  49:20  3.22% mysqld
93335 root        9  20  -76  3716K  1552K kserel 0 101:57  0.00% hpasmd
39333 err         4  20    0   364M 32892K kserel 0   1:38  0.00% java
55121 root        1  96    0 14204K  7436K select 0   0:37  0.00% httpsd
21250 root        1  96    0  3704K  2196K select 0   0:30  0.00% sendmail
92330 root        1   8    0  1464K   852K nanslp 0   0:05  0.00% cron
 5417 err         1   4    0  5492K  3636K sbwait 0   0:04  0.00% mysql
86549 err         1  96    0  3008K  2280K select 0   0:04  0.00% screen
25479 root        1  96    0  1428K   828K select 0   0:03  0.00% syslogd
37590 root        1  96    0  1340K   644K select 0   0:02  0.00% usbd
48624 err         1  -8    0  1316K   588K pipewr 0   0:01  0.00% cat
```

Figura 5.6. Importazione del dump XML: output del comando `top`.

- *mwimport.pl* [115]
  Dati i tempi lunghi di mwdumper, la comunità Wikimedia ha deciso di riscriverlo nel linguaggio di scripting Perl (per certi versi più veloce di Java [116]) semplificando la sintassi riconosciuta del dump al fine di ridurre i tempi ma soprattutto la memoria utilizzata. Il programma, essendo scritto per poter lavorare con il dump di `enwiki`, ha delle difficoltà ad elaborare i dump delle altre lingue, e non importa correttamente il testo delle voci del dump `itwiki` filtrato.

È possibile trovare una pagina ufficiale di Wikimedia sulle varie procedure per l'importazione dei dump all'indirizzo web [117].

Soltanto *importDump.php* costruisce direttamente la base di dati partendo dal dump XML; esso effettua internamente in PHP l'interfacciamento con il server MySQL. Gli altri programmi elencati lavorano accettando in ingresso il dump XML (come flusso nello *standard in* o dandogli il nome del file) e producono in *standard out* i comandi SQL compatibili con MySQL per inserire i dati.

I file utilizzati da MySQL per mantenere i contenuti della base di dati occupano più spazio della dimensione del dump XML, a causa delle informazioni di servizio proprie di un DBMS, quali per esempio gli indici sulle tabelle. Dal dump XML di maggio 2007 filtrato, si ottiene un database che occupa all'incirca 6 GB.

```
    4 users    Load  0.61  0.36  0.22                  Jun 16 21:21

Mem:KB    REAL            VIRTUAL                       VN PAGER  SWAP PAGER
        Tot   Share      Tot    Share    Free           in  out     in  out
Act   95912    7072   532292    14404   39024 count
All 1021884   14112153501336    27312         pages
                                                         Interrupts
Proc:r  p  d  s  w    Csw  Trp  Sys  Int  Sof  Flt        cow    4614 total
          10 48      2115   8910867 1232   77   13 188800 wire         1: atkb
                                                 109536 act          6: fdc0
 4.5%Sys   0.4%Intr  1.9%User  0.0%Nice 93.3%Idl   687104 inact        15: ata
|    |    |    |    |    |    |    |    |    |      37312 cache   304 16: uhc
==>                                                 1712 free      8 48: bge
                                                         daefr   296 72: cis
Namei         Name-cache    Dir-cache                    prcfr  2003 cpu0: time
    Calls     hits    %     hits    %                    react  2003 cpu1: time
                                                         pdwake
                                       13 zfod           pdpgs
Disks   da0 pass0                      12 ozfod          intrn
KB/t  18.56  0.00                      92 %slo-z   114304 buf
tps     433     0                      67 tfree         6 dirtybuf
MB/s   7.85  0.00                                  70236 desiredvnodes
% busy   92     0                                  19500 numvnodes
                                                   17521 freevnodes
```

Figura 5.7. Importazione del dump XML: output del comando `systat -vmstat 1`.

```
                    /0   /1   /2   /3   /4   /5   /6   /7   /8   /9   /10
     Load Average   ||

                    /0   /10  /20  /30  /40  /50  /60  /70  /80  /90  /100
root     idle: cpu1 XXXXXXXXXXXXXXXXX
             <idle> XXXXXXXXXXXXXXXX
root     idle: cpu0 XXXXXXXXXXXXXXX
mysql        mysqld X
```

Figura 5.8. Importazione del dump XML: output del comando `systat -pigs 1`.

5.2.3 Esportazione delle voci

Con il database disponibile localmente è possibile utilizzare il programma PHP di WaNDA-tools aggiunto a MediaWiki. Prima però è consigliato controllare il corretto funzionamento di MediaWiki, e quindi la correttezza delle pagine esportate, se è disponibile il server Apache sulla piattaforma: basta collegarsi ad esso con un browser ed accedere alle pagine come fosse la versione online di Wikipedia. Quest'operazione è molto utile per verificare

```
                    /0   /1   /2   /3   /4   /5   /6   /7   /8   /9   /10
     Load Average   |||

          /0   /10  /20  /30  /40  /50  /60  /70  /80  /90  /100
cpu  user|X
     nice|
   system|XX
interrupt|
     idle|XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

          /0   /10  /20  /30  /40  /50  /60  /70  /80  /90  /100
da0   MB/sXXXX
      tps|XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX439.99
pass0 MB/s
      tps|
```

Figura 5.9. Importazione del dump XML: output del comando `systat -iostat 1`.

la corretta installazione delle estensioni, descritta in sezione 5.2.1, ed il funzionamento dei template.

Tutti i comandi della seconda fase di WaNDA vengono lanciati eseguendo lo script PHP sito in `maintenance/dumpDVD.php`. È importante eseguirlo nella directory radice di MediaWiki in quanto deve poter accedere alle sue differenti componenti.

Senza predisporre l'ambiente, la semplice esecuzione di `php maintenance/dumpDVD.php` dovrebbe fallire con un messaggio di errore, che indica il direttorio dove andrà effettuata l'operazione di esportazione e la costruzione dell'albero. Questo percorso è indicato nel file di configurazione `maintenance/dumpDVD.ini` alla voce $defdest$; esso può essere sia assoluto che relativo alla directory di lavoro corrente. Per essere disponibile all'esportazione esso deve essere scrivibile dall'utente che esegue WaNDA-tools e deve essere vuoto.

Oltre a scrivere sul percorso indicato di destinazione dell'albero, WaNDA-tools inoltre scrive informazioni di servizio su un database che viene creato nello stesso server MySQL utilizzato per contenere le voci di Wikipedia. Questi dati di servizio contengono fondamentalmente la lista delle voci esportate, utili alla generazione degli indici di ricerca, e la lista delle immagini desiderate, necessaria per la corretta formattazione dei loro collegamenti nel testo.

Eseguito come indicato, senza opzioni, il comando effettua completamente tutti i cinque passi necessari alla costruzione dell'albero:

1. applicazione dei filtri sul database delle voci (il *wikidb*);
2. creazione del database di servizio (il *ddvddb*);
3. elaborazione delle voci, in ordine crescente secondo il loro *id* nel *wikidb*;
4. generazione degli indici di ricerca;

5. download ed integrazione delle immagini.

I vari passi verranno illustrati in dettaglio nelle sezioni successive. Il percorso descritto può essere modificato dalle seguenti opzioni:

- `--nofilterdb`: non esegue i filtri sul *wikidb* (salta il passo 1);
- `--nomergedb`: al passo 2, in caso che il database di servizio esista già, forza la creazione di uno nuovo;
- `--idstart=`: imposta l'*id* della voce da cui iniziare l'elaborazione delle voci del passo 3 (presuppone che il database di servizio esista già);
- `--idstop=`: imposta l'*id* dove terminare l'elaborazione delle voci, ed esce senza effettuare i passi successivi al 3;
- `--nomakeimg`: non effettua l'integrazione delle immagini (salta il passo 5).

Inoltre possono essere impostate opzioni che effettuano delle operazioni specifiche:

- `--filterdb`: esegue i filtri sul *wikidb*;
- `--makeimglist`: aggiorna la lista nel database di servizio delle immagini da includere;
- `--getimg`: scarica le immagini secondo il database di servizio;
- `--copyimg`: integra le immagini scaricate nell'albero finale;
- `--makeiso`: genera il file immagine ISO-9660 per il DVD contenente l'albero.

Tranne nel caso in cui vengano specificate le operazioni specifiche, la prima operazione effettuata da WaNDA è la costruzione della struttura di base dell'*albero*, quella che viene chiamata *ossatura*. Essa include una gerarchia di directory per contenere le pagine HTML di servizio e le loro immagini, i file JavaScript, il browser *K-Meleon* e l'*Autorun*, il testo della licenza GNU FDL ed un testo di istruzioni all'uso.

L'*ossatura* è fornita con WaNDA-tools e comprende file e directory che non saranno successivamente modificati. Essa è presente sotto forma di due archivi compressi: `maintenance/dumpDVD/wnd_html.tar.bz2` contenente le pagine HTML, immagini e JavaScript; `maintenance/dumpDVD/wnd_bin.tar.bz2` contenente i binari e l'autorun per Windows. Un'eventuale modifica alle pagine di servizio comporta la decompressione del primo file e la modifica del testo nel file voluto.

Un'*ossatura* già scompattata nel percorso di destinazione per l'*albero* impedisce il successivo riutilizzo dello stesso percorso per una nuova elaborazione.

5.2.3.1 Filtro sulla base di dati

Lo scopo dello script AWK eseguito sul dump XML serve unicamente a ridimensionarlo per poter quindi ridurre i tempi e lo spazio necessari a gestirlo.

È però ancora necessario rimuovere dati principalmente per tre motivi:

- è necessario ridurre il più possibile il numero di file e quindi lo spazio occupato dall'*albero*, per cui vengono rimossi i *namespace* non appartenenti ai contenuti enciclopedici;
- è consigliabile filtrare le pagine dal contenuto discutibile secondo una selezione applicata dal team di Wikimedia Italia;
- bisogna permettere all'*albero* di essere contenuto su un filesystem *case insensitive*, risolvendo ambiguità nel caso di pagine di redirect e pagine normali che abbiano lo stesso titolo ma con un *case* differente.

Una volta che il dump XML è importato nel database MySQL, prima di passare alla fase di esportazione, vengono quindi applicati vari filtri; tali filtri si devono porre in questa fase poichè l'esecuzione di query SQL che eseguono operazioni di rimozione è il modo migliore e più veloce di operare.

Un'analisi dettagliata delle finalità e dell'ambito di applicazione dei filtri è stato descritto nel capitolo 4.2.1.

Il filtro lavora utilizzando semplici query SQL all'interno del codice PHP; il file che contiene il codice dei filtri è `maintenance/dumpDVDfilter.php`. Il file PHP è organizzato in modo tale da essere facilmente configurabile secondo future necessità.

Le pagine che rimangono fanno parte dei namespace elencati di seguito (anche per l'elenco completo dei namespace si rimanda alla sezione 4.2.1):

- 0 (namespace principale): qui si trovano le pagine, da esportare, delle voci normali che interessano ai fini enciclopedici, pagine di redirect incluse.
- 6 (*Immagine*): sono le pagine che descrivono la licenza e riportano la thumbnail per ogni immagine; queste pagine non saranno esportate, ma sono utili poichè permettono di avere nella base di dati locale le informazioni complete sulle immagini e permettono di avere i collegamenti nel testo delle voci normali automaticamente formattati correttamente.
- 8 (*MediaWiki*): queste pagine protette non sono cancellabili, essendo parte di MediaWiki, e non verranno comunque esportate.
- 10 (*Template*): i template sono fondamentali alla corretta esportazione delle pagine delle voci normali, poichè essi contengono parti di testo e informazioni che sono incluse in quasi tutte le voci; essi non sono esportati come tali, però inclusi nel testo delle pagine delle voci normali.

- 14 (*Categoria*): le categorie sono utili sia all'elaborazione delle immagini (descritta in sezione 4.2.2), sia sotto forma di pagine esportate per disporre di un'interfaccia di navigazione delle voci normali secondo il loro argomento (descritta in sezione 4.3.2).

Il problema dei redirect e delle doppie pagine con lo stesso titolo *case insensitive* richiede una notevole elaborazione sul DBMS. Si lavora in due passi successivi: si raccoglie la lista degli articoli non redirect ordinati per dimensione decrescente del contenuto e si rimuovono dal database i doppioni *case-insensitive*. Si ottiene così la lista delle voci contenente gli articoli che saranno presenti sul tree finale, con la cancellazione della pagina più scarna nel caso di doppie con il medesimo titolo.

Il secondo passo prevede la rimozione dal database delle pagine di redirect con il titolo *case-insensitive* uguale a quello di una pagine propria.

La lista di tutte le pagine, ordinate per `id` univoco, è presente del database di nome `page`; qui sono già indicati *namespace*, se è un redirect oppure no, e la lunghezza della pagina. L'elaborazione è pesante in quanto, oltre al fatto che tabella contiene centinaia di migliaia di tuple quante sono le voci compresi i redirect (circa 900 mila per il dump di maggio 2007), è necessario andare a modificare la tabella stessa sulla quale si esegue la *select*; quest'operazione non è consentita su MySQL, e viene consigliato l'utilizzo di una tabella temporanea di supporto.

Dopo vari tentativi utilizzando la base di dati si è giunti ad una soluzione abbastanza rapida che utilizza in modo misto il MySQL e PHP. Prevede di costruire la lista delle pagine proprie come *array* in PHP durante il primo passo e successivamente per il secondo passo ciclare con un'unico loop su tale *array*, cancellando nel database i redirect con lo stesso titolo di ogni elemento.

Il carico del sistema sul quale si esegue il codice è illustrato dalle figure 5.10, 5.11 e 5.12. Il processore è poco carico; viene invece utilizzato al massimo la banda disponibile per l'accesso al disco. Il tempo totale impiegato per eseguire tutti i filtri è di 3 ore e mezza; disponendo di un IO con maggior throughput i tempi si riducono.

5.2.3.2 Database di servizio

La parte vera e propria di esportazione delle pagine prevede l'utilizzo di un piccolo database di supporto, utile per due motivi:

- Permette di mantenere l'elenco delle pagine esportate, utile per creare alla fine dell'esportazione un motore di ricerca e la lista di tutte le pagine.
- È necessario per la gestione delle immagini; alla creazione del database si inserisce la lista di tutte le immagini che si vuole permettere nell'albero finale, tipicamente secondo la loro licenza di utilizzo (tale scelta è effettuata nel file di configurazione `maintenance/dumpDVD.ini`.

  Si noti come premessa che le immagini sono identificate univocamente dal loro nome. Durante l'esportazione delle pagine si controlla per ogni immagine se è presente in questa lista; in caso positivo, si inserisce il tag per l'inclusione dell'immagine ed il

```
    4 users    Load  0.18  0.05  0.02                  Jun 18 17:13

Mem:KB    REAL            VIRTUAL                       VN PAGER  SWAP PAGER
        Tot   Share      Tot    Share    Free           in  out     in  out
Act  111936    7516   176496    13880   39024 count
All 1021924    9292154592456    18428         pages
                                                         Interrupts
Proc:r  p  d  s  w    Csw  Trp  Sys  Int  Sof  Flt        cow    4554 total
           7 50      1810   4012083 1154   41      188240 wire         1: atkb
                                                 125552 act          6: fdc0
 5.2%Sys   0.0%Intr  2.2%User  0.0%Nice 92.6%Idl  671696 inact        15: ata
|    |    |    |    |    |    |    |    |    |    37364 cache   278 16: uhc
===>                                               1660 free      5 48: bge
                                                        daefr   273 72: cis
Namei         Name-cache    Dir-cache                   prcfr  1999 cpu0: time
    Calls     hits    %     hits    %                   react  1999 cpu1: time
                                                        pdwake
                                        zfod            pdpgs
Disks   da0 pass0                       ozfod           intrn
KB/t  19.47  0.00                       %slo-z   114304 buf
tps     427     0                    72 tfree         5 dirtybuf
MB/s   8.11  0.00                                  70236 desiredvnodes
% busy   87     0                                  20323 numvnodes
                                                   17555 freevnodes
```

Figura 5.10. Filtro sul database: output del comando `systat -vmstat 1`.

```
                    /0   /1   /2   /3   /4   /5   /6   /7   /8   /9   /10
     Load Average

          /0   /10  /20  /30  /40  /50  /60  /70  /80  /90  /100
cpu  user|X
     nice|
   system|XX
interrupt|
     idle|XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

          /0   /10  /20  /30  /40  /50  /60  /70  /80  /90  /100
da0   MB/sXXXX
      tps|XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX436.68
pass0 MB/s
      tps|
```

Figura 5.11. Filtro sul database: output del comando `systat -iostat 1`.

```
                    /0   /1   /2   /3   /4   /5   /6   /7   /8   /9   /10
     Load Average   |

                    /0   /10  /20  /30  /40  /50  /60  /70  /80  /90  /100
root     idle: cpu1 XXXXXXXXXXXXXXXX
             <idle> XXXXXXXXXXXXXXX
root     idle: cpu0 XXXXXXXXXXXXX
mysql        mysqld X
```

Figura 5.12. Filtro sul database: output del comando `systat -pigs 1`.

collegamento HTML, oltre a segnalare nel database il riscontro dell'immagine; in caso negativo nel testo della pagina non si fà riferimento all'immagine, omettendo l'eventuale didascalia.

Alla fine dell'esportazione, l'elaborazione delle immagini verterà su quelle segnate nel database.

Il database viene creato nello stesso MySQL utilizzato per contenere il dump di Wikipedia; di default il suo nome è *ddvddb*.

Siccome le immagini devono poi essere ottenute scaricandole dal sito `commons.wikimedia.org`, nel caso di differenti elaborazioni successive con dump di Wikipedia differenti è possibile ridurre il traffico di rete riscaricando solo le immagini nuove. Infatti il database di supporto per le immagini contiene anche un campo per indicare se l'immagine è già presente sulla piattaforma locale.

Di default in caso WaNDA-tools trovi già un database di nome *ddvddb* ne verifica la struttura e provvede a cancellare solo l'indice delle pagine. Per quanto riguarda l'indice delle immagini genera la lista delle immagini da includere e modifica il database di servizio secondo questa nuova lista: un'immagine presente nel *ddvddb* e non nella lista viene rimossa, un'immagine presente nella lista e non nel *ddvddb* viene aggiunta; le immagini già presenti nel *ddvddb* vengono controllate secondo il loro timestamp, per verificare non siano state aggiornate.

Tenendo conto che la maggior parte delle immagini della selezione impostata secondo la loro licenza d'uso nel tempo restano invariate e la loro presenza non varia grandemente, poche sono le nuove immagini da riscaricare tra due elaborazioni di due dump.

In ogni caso è sempre possibile forzare la creazione del database di supporto (tramite l'opzione `--nomergedb`), che provvede a pulire *ddvddb* e a reinizializzarlo secondo le immagini contenute del dump di Wikipedia importato.

La modifica delle categorie delle immagini secondo le licenze richieste nel file di configurazione non costringe al ripristino del *ddvddb*, in quanto è compatibile con l'operazione di *merging* del database di supporto.

5.2.3.3 Scrittura delle pagine

L'esportazione procede secondo l'`id` univoco di ogni voce presente nel database di Wikipedia locale (il *wikidb*). Sull'output è presente lo stato attuale dell'elaborazione: il numero dell'ultimo `id` multiplo di 50 esportato e lo stato percentuale sul totale dell'esportazione.

Ogni voce enciclopedia viene estratta dal database selezionando l'`id`; si ottiene quindi, utilizzando le funzioni di MediaWiki, il testo in formato *wikitext* della voce (vedi il Percent Encoding in sezione 4.2.4). Sempre utilizzando il codice di MediaWiki, il *wikitext* viene convertito in normale testo HTML. L'utilizzo di MediaWiki stesso per effettuare quest'operazione evita la necessità di modifiche successive al progetto WaNDA per poter interpretare le nuove varianti del *wikitext*; MediaWiki deve per forza di cose interpretarlo completamente, essendo il software di riferimento per la sua interpretazione.

Inoltre vengono agevolmente gestiti i frammenti di codice elaborati dalle estensioni, dato che basta installarle senza effettuare complicate messe a punto. Le immagini generate dalle estensioni *timeline* e *texvc* vengono scritte nella normale directory `images/`, apposita per le immagini di MediaWiki; nell'ultima fase di costruzione dell'albero essere saranno copiate assieme alle fotografie ottenute da *commons* nella posizione a loro riservata nell'albero.

Oltre che per la formattazione del testo HTML, il codice di MediaWiki viene utilizzato per accedere agevolmente alla base *wikidb* per estrarre l'elenco degli autori che hanno effettuato modifiche a ciascuna voce. Tale elenco verrà incluso a piè di pagina di ogni voce, in modo da seguire le indicazioni della licenza GNU FDL, discussa in sezione 2.1.2.

Il testo HTML e la lista di autori per ogni voce vengono quindi inglobati da WaNDA in un contenitore HTML e JavaScript, che ne permette la presentazione con qualsiasi browser web che supporti almeno JavaScript 1.0. La pagina HTML è composta da un header ridotto all'osso che include il foglio di stile CSS ed i file JavaScript utilizzati, e dal body che contiene la chiamata ad una funzione JavaScript che presenta il contenuto della pagina. In questo modo si riduce al massimo la quantità di spazio richiesta da ogni pagina HTML, essendo le barre di navigazione e le altre informazioni di servizio presenti, che sono le stesse per tutte le pagine, scritte in un solo file JavaScript. Le tecnologie e le modalità impiegate per presentare le pagine HTML all'utente sono state descritte approfonditamente nel capitolo 4.

Le pagine di redirect sono esportate come semplici pagine HTML il cui header contiene un campo `meta` che redirige alla pagina completa. Esse sono comode in quanto permettono di aumentare le probabilità di trovare una chiave cercata con il semplice motore di ricerca sul titolo delle voci.

Questo semplice motore di ricerca, esaminato in sezione 4.3.3, richiede la lista di tutte le pagine presenti nell'albero su cui effettuare la ricerca. In questa fase di elaborazione il titolo della voce di ogni pagina scritta è aggiunta al database di servizio *ddvddb*, in modo da avere un elenco completo delle pagine presenti.

Le pagine da esportare sono di due tipi. Il primo tipo (la maggior parte) sono le voci enciclopediche, redirect inclusi; il secondo tipo sono le pagine delle categorie. Queste ultime sono delle liste di collegamenti a voci enciclopediche, gestiti internamente da MediaWiki

impostando l'appartenenza di una pagina ad una determintata categoria. A volte alla categoria è associato un breve testo iniziale; inoltre le categorie possono raccogliere altre categorie, formando una gerarchia. Le categorie con i riferimenti sono contenute nel dump SQL importato all'inizio del processo; le stesse categorie servono ad identificare la licenza delle immagini.

Le categorie devono essere esportate come pagine html di raccolta dei collegamenti alle voci, in modo tale che un utente che acceda all'*albero* possa navigare nella gerarchia delle categorie. Le principali categorie saranno accessibili dalla prima pagina introduttiva del DVD.

L'estrazione e la scrittura delle pagine può essere interrotta in qualunque momento inviando il classico segnale di interrupt SIGINT tramite `ctrl-C`. L'elaborazione può essere ripresa impostando l'opzione `--startid=` seguita dall'*id* della pagina con la quale continuare.

Abbinando l'opzione `--stopid=` che termina l'elaborazione all'*id* indicato, è possibile effettuare spezzoni del processo totale di estrazione. Si noti che senza `--stopid=` vengono eseguite anche le fasi successive a quella di scrittura delle pagine, mentre in assenza di `--startid=` vengono eseguite le fasi precedenti.

Il numero di voci normali esportate per il dump di maggio 2007 è di circa 300 mila, compresi i redirect si arriva a 900 mila; ad esse si aggiungono il numero delle pagine di categorie, che sono quasi 30 mila. L'elaborazione di tutte le voci impiega il 70% del processore per il PHP ed il restante per il DBMS; l'accesso al disco si concentra sul DBMS, mentre il tempo di scrittura delle pagine è trascurabile. L'esportazione completa impiega circa 43 ore; nelle figure 5.13, 5.14, 5.15 e 5.16 sono presenti le statistiche di utilizzo delle risorse per questa fase.

```
last pid: 44355; load averages: 0.72, 0.23, 0.09       up 27+18:50:15  13:02:38
59 processes:  2 running, 57 sleeping
CPU states: 47.4% user,  0.0% nice,  0.8% system,  0.0% interrupt, 51.9% idle
Mem: 135M Active, 614M Inact, 193M Wired, 51M Cache, 112M Buf, 8172K Free
Swap: 2023M Total, 136K Used, 2023M Free

  PID USERNAME  THR PRI NICE   SIZE    RES STATE  C   TIME   WCPU COMMAND
52928 err         1 121    0 29724K 23644K RUN    0   0:52 76.07% php
79785 mysql       9  20    0 56716K 34832K kserel 0   0:04  2.15% mysqld
93335 root        9  20  -76  3716K  1552K kserel 0 117:21  0.00% hpasmd
55121 root        1  96    0 14204K  7384K select 0   0:44  0.00% httpsd
...
```

Figura 5.13. Scrittura delle pagine: output del comando `top`.

```
    4 users    Load  0.83  0.30  0.12                  Jun 20 13:03

Mem:KB    REAL            VIRTUAL                       VN PAGER  SWAP PAGER
        Tot   Share      Tot    Share    Free           in  out     in  out
Act  124848    7388   188932    13880   57164 count
All 1017932    9252150656460    18428         pages
                                                         Interrupts
Proc:r  p  d  s  w    Csw  Trp  Sys  Int  Sof  Flt    248 cow    4038 total
     1     6 51      1257 499126043  230  519 4190 197044 wire         1: atkb
                                                 138348 act          6: fdc0
 3.7%Sys   0.0%Intr 41.0%User  0.0%Nice 55.4%Idl   632044 inact        15: ata
|    |    |    |    |    |    |    |    |    |      51528 cache    23 16: uhc
==>>>>>>>>>>>>>>>>>>>                               5636 free       6 48: bge
                                                         daefr    17 72: cis
Namei         Name-cache    Dir-cache                356 prcfr  1996 cpu0: time
    Calls     hits    %     hits    %                    react  1996 cpu1: time
      943      933   99                                 pdwake
                                     281 zfod            pdpgs
Disks   da0 pass0                    242 ozfod           intrn
KB/t  16.00  0.00                     86 %slo-z   114304 buf
tps      17     0                    585 tfree       50 dirtybuf
MB/s   0.26  0.00                                  70236 desiredvnodes
% busy   11     0                                  60374 numvnodes
                                                  17556 freevnodes
```

Figura 5.14. Scrittura delle pagine: output del comando `systat -vmstat 1`.

```
                    /0   /1   /2   /3   /4   /5   /6   /7   /8   /9   /10
     Load Average   ||||

                    /0   /10  /20  /30  /40  /50  /60  /70  /80  /90  /100
root     idle: cpu1 XXXXXXXXXXXXXXXXXX
err             php XXXXXXXXXXXXXX
            <idle> XXXXXXXXXX
root     idle: cpu0 XXX
mysql        mysqld
```

Figura 5.15. Scrittura delle pagine: output del comando `systat -pigs`.

5.2.3.4 Indici per la ricerca

Uno dei meccanismi di accesso ai contenuti, probabilmente il più efficace, è il meccanismo di ricerca JavaScript descritto nel capitolo precendente.

Il funzionamento è garantito dalla presenza di un indice delle voci presenti sul supporto, la *base di ricerca*. Tecnicamente è composta da vettori associativi in JavaScript, in cui

```
                  /0   /1   /2   /3   /4   /5   /6   /7   /8   /9   /10
     Load Average   |||||

          /0   /10  /20  /30  /40  /50  /60  /70  /80  /90  /100
cpu  user|XXXXXXXXXXXXXXXXXXXXXXXX
     nice|
   system|
interrupt|
     idle|XXXXXXXXXXXXXXXXXXXXXXXX

          /0   /10  /20  /30  /40  /50  /60  /70  /80  /90  /100
da0   MB/s
       tps|X
pass0 MB/s
       tps|
```

Figura 5.16. Scrittura delle pagine: output del comando `systat -iostat 1`.

per ogni cella la chiave è il titolo della voce ed il contenuto è il percorso sull'albero in cui trovare la pagina. La ricerca di una chiave esatta sul vettore associativo è praticamente istantanea. Non tutte le voci sono nello stesso vettore per motivi di risorse, in quanto i browser hanno difficoltà a gestire array javascript più grandi di 1 MB; l'elenco è quindi suddiviso in 10 vettori, il che genera la presenza di 10 file JavaScript contenenti i vettori, grandi circa mezzo MB.

La ricerca si può effettuare con due livelli di profondità:

- Un primo di tipo di ricerca, chiamata *ricerca esatta*, verifica velocemente se esiste una voce uguale alla stringa cercata. Essendo le voci divise nei 10 vettori secondo un raggruppamento specifico della prima lettera delle voci, tale ricerca procede caricando solamente il vettore JavaScript contenente le voci che iniziano con la prima lettera della stringa cercata. Con il vettore corretto si verifica immediatamente se la stringa è presente; è quindi una ricerca che si effettua in brevissimo tempo.

  Quando una voce viene trovata, la ricerca conclude essendo possibile soltanto una soluzione, caricando quindi nel browser la corretta pagina HTML dell'*albero*.

- Il secondo tipo di ricerca, detto *ricerca estesa*, è più completo e più lento. Esso verifica se la stringa cercata è contenuta fra le celle di tutti i vettori: la ricerca della stringa in ogni chiave del vettore si effettua tramite un ciclo `for`. L'ordine dei vettori associativi sui quali eseguire la ricerca completa è scelta casualmente, per cui l'ordine dei risultati cambia di volta in volta. La stringa di ricerca deve essere più lunga 3 caratteri.

  Durante la ricerca, la maggior parte del tempo di attesa è composta dai tempi per il caricamento dei file JavaScript contenenti i vettori associativi; per non far pensare all'utente che il sistema sia in stallo viene raffigurata una barra di caricamento che

avanza ad ogni file caricato. Il tempo totale di ricerca, siccome dipende dai tempi per caricare i file JavaScript, dipende solo in parte dalla stringa di ricerca o dal numero di risultati, e quindi il tempo totale per la ricerca è quasi sempre lo stesso; esso si aggira su qualche decina di secondi per un sistema medio.

I risultati della ricerca sono presentati come lista di collegamenti alle voci trovate, in una pagina creata dinamicamente con codice JavaScript.

- Un generatore casuale di voci. Esso sfrutta il motore di ricerca di primo tipo, permettendo di estrarre velocemente una pagina casualmente scelta dell'albero; si procede selezionando a caso un vettore, e quindi sempre a caso una chiave all'interno di esso.

Si noti che entrambi i tipi di ricerca effettuano confronti di tipo *case insensitive*, consentono la ricerca di stringhe composte da più parole e composte da qualsiasi carattere unicode.

I due algoritmi di ricerca se combinati assieme prende il nome di *ricerca completa*; prima si effettua la ricerca esatta, e se nessuna voce viene trovata si passa automaticamente alla ricerca estesa. Di default la ricerca effettuata dal *form* HTML presente in ogni pagina dell'*albero* effettua la ricerca completa. È anche possibile eseguire soltanto il secondo tipo di ricerca, cliccando l'apposito tasto nel *form*. Si noti che tale comportamento emula a grandi linee il funzionamento dei due tipi di ricerca presenti sulla versione on-line di Wikipedia.

In questa fase viene anche generato l'elenco delle voci presenti nell'albero. Non fanno parte di questo elenco le categorie ed i redirect.

L'indice è organizzato secondo l'ordine alfabetico delle voci ed è suddiviso in varie pagine HTML, dal nome `pagina_indice_[0-9][0-9].html`; il numero di pagine è impostato nel file di configurazione di WaNDA (di default è 50). In ogni pagina sono presenti collegamenti alla pagina iniziale, finale, precedente e successiva.

Lo stile di queste pagine è esattamente lo stesso delle altre voci di WaNDA e la lista appare come il testo di una voce normale; è tuttavia mancante il collegamento alla voce online come nelle pagine di servizio.

L'indice di tutte le voci presenti ed i motori di ricerca richiedono quindi la gestione di una lista nel database delle pagine scritte nell'*albero*.

L'indice delle voci viene direttamente generato in questa fase di elaborazione, scrivendo nel percorso `data/` dell'*albero* le 50 pagine HTML descritte, estraendo dal database di supporto le voci ordinate alfabeticamente.

La base dei motori di ricerca offline, ovvero i 10 file JavaScript contenenti i vettori associativi, sono scritti nel percorso `data/js/`. Ognuno di essi non dev'essere più grande di 1 MB per essere caricato in tempi ragionevoli dal browser. Essendo tutte le chiavi ordinate secondo la loro prima lettera, ed essendo i casi possibili composti da tutti i caratteri (anche non italiani) ed i numeri, è necessaria una mappatura che permetta di identificare il file JavaScript contenente il carettere cercato. Tale mappatura è impostata nel file di configurazione di WaNDA. In questa fase si scrive quindi anche un file JavaScript che presenta tale mappatura.

La creazione dell'indice delle voci e della base dei motori di ricerca impiega 1.5 ore di elaborazione, tempo dovuto in gran parte all'esecuzione delle query SQL per l'estrazione ordinata dei titoli delle voci. Questa fase di elaborazione è strettamente legata alla fase precedente di scrittura delle pagine, per cui queste due fasi sono sempre eseguite di seguito.

Inoltre alla fine di questa fase viene scritta la versione dell'*albero* e la data nel file `leggimi.txt` presente in radice, in modo da avere l'indicazione univoca sulla versione corrente sempre disponibile.

5.2.3.5 Gestione immagini

La gestione delle immagini deve sempre svolgersi come ultima operazione dell'estrazione; questo poichè tutte le immagini che devono essere incluse nell'*albero* dipendono dalla generazione delle pagine HTML delle voci. Le immagini vengono poco a poco scritte nella directory `images/` di MediaWiki, per essere quindi copiate nell'*albero*.

Le immagini presenti, come analizzato in sezione 4.2.2, sono di quattro tipi e si differenziano per il meccanismo con il quale sono ottenute:

- Immagini di carattere misto scaricate dal repository `upload.wikimedia.org`, che sono state incontrate durante la fase di scrittura delle pagine HTML e che corrispondono alle licenze indicate; sono quelle elencate nella tabella *ddvddb.images* del database di supporto *ddvddb*.

  I nomi di tutte le immagini sono presenti dalla fase di creazione del *ddvddb*; esse vengono contrassegnate durante l'interpretazione del *wikitext* e indicata la risoluzione con la quale sono presenti. Nell'evento in cui l'immagine incontrata è già segnata, si controlla se la risoluzione attuale è inferiore a quella già indicata, nel qual caso si segna la risoluzione minore: ciò permette di compattare la dimensione delle immagini.

  L'immagine è scaricata dal repository costruendo l'URL univoco dato dal nome (univoco) dell'immagine nel *wikitext*; essa è quindi ridimensionata secondo la dimensione scritta nel database di servizio. Il nome del file con la quale ogni immagine è salvata nell'*albero* corrisponde al nome della sua pagina immagine e quindi del collegamento nel *wikitext*; i file sono salvati in directory alfabetiche divise per la prima lettera del nome.

- Immagini che raffigurano formule matematiche: queste sono generate con il binario *textvc* dai frammenti di testo tex presente nel *wikitext*.

  Esse sono presenti, come nella versione di Wikipedia online, in una gerarchia di 3 livelli di directory e hanno il nome del file lungo 32 caratteri, dato dall'hash MD5 della stringa tex che rappresentano. Questa gerarchia si situa nella directory `images/math/` di Mediawiki.

- Immagini che raffigurano gli istogrammi, principalmente gli andamenti demografici dei comuni. Questi grafici vengono generati con l'estensione *timeline*; anche qui il nome del file è dato dall'hash MD5.

Tutti i file sono costruiti nella directory `images/timeline/` di Mediawiki.

- Immagini sempre presenti nell'*albero* che provengono dall'*ossatura*. Questi file sono statici ed introdotti dal programma WaNDA; prima della dell'inclusione nell'*albero* sono presenti nell'archivio `maintenance/dumpDVD/wnd_html.tar.bz2`. Queste immagini comprendono i dettagli grafici utilizzati dal tema delle pagine HTML, i loghi e le immagini utilizzate dalle pagine di servizio.

Il secondo ed il terzo tipo di immagini sono create in `images/` man mano che vengono incontrate durante l'interpretazione del *wikitext*. Il quarto tipo essendo incluso nell'*ossatura* non richiedono elaborazione preventiva e sono installate nella fase iniziale dell'elaborazione.

Il primo tipo di immagine viene gestito in questa fase, che ne effettua il download, la ridimensione ed il salvataggio in `images/`, dove sono già presenti *math* e *timeline*. Questa operazione può essere eseguita con il comando `--getimg`, per esempio per effettuare di colpo il download quando l'accesso ad Internet è disponibile.

Questa fase comporta anche l'integrazione di tutte le immagini con l'*albero*; esse vengono in pratica copiate o spostate da *images/* al percorso sull'*albero* adatto. Quest'operazione si può effettuare con il comando `--copyimg`.

Solo quelle del primo tipo sono copiate, in quanto posso essere riutilizzate quando si effettua il *merging*. Le altre immagini generate dalle estensioni sono spostate, poichè dipendono dal dump attualmente presente.

Questo consente di riscaricare da `upload.wikimedia.org` solo le immagini che sono in qualche modo cambiate o sono state aggiunte. Il controllo sulla mutabilità dell'immagine viene effettuate mantenendo nella tabella *ddvddb.images* il timestamp dell'ultima modifica apportata su Wikipedia.

5.2.4 Impacchettamento

Una volta che l'operazione di integrazione dei contenuti grafici è finita, il ciclo di esportazione dell'*albero* è terminato ed esso è quindi completo.

L'albero può quindi essere archiviato e compresso per essere distribuito.

Per generare l'immagine ISO-9660 automaticamente è presente un comando aggiuntivo, `--makeiso`, che provvede a costruire l'ISO ed il suo hash MD5 nella directory indicata nel file di configurazione di WaNDA; il nome del file è dato dalla data e versione dell'*albero*.

Effettuare l'immagine ISO-9660:1999 con *mkisofs* è un'operazione che sulla piattaforma di riferimento impiega un'ora e mezzo, un tempo lungo dovuto alla presenza di migliaia di piccoli file.

L'immagine ISO può essere compressa con *bzip2* che ne riduce la dimensione di circa 1/6; per la diffusione in rete ciò è comodo poiché la trasmissione richiede la trasmissione di soli 500-600 MB invece di un DVD.

Per la distribuzione su memoria solida l'*albero* deve soltanto essere impacchettato con un programma di archiviazione, come per esempio *tar* o *zip*. Sarà l'utente finale, infatti, ad estrarre i file dell'*albero* su una partizione contenente un filesystem FAT32. Tale partizione

è quella presente sul dispositivo scelto dall'utente, che sia un pendrive, una scheda di memoria, un microdrive od un palmare.

Oltre ad essere archiviato, conviene anche comprimere l'archivio in modo da ridurne la dimensione come per l'immagine ISO. Per archiviare e comprimere, dato l'ottimo rapporto di compressione, la licenza e la presenza di applicativi per ogni piattaforma si consiglia l'utilizzo dell'applicativo *7-Zip* [118].

5.3 Analisi risorse: tempo e spazio

Riepilogando tutte le operazioni effettuate durante l'esecuzione completa di un'estrazione sono:

- Download del dump SQL e XML compressi - i tempi dipendono dalla connessione ad Internet e dal carico del server `download.wikipedia.org`.
- Decompressione del dump XML ed esecuzione del filtro AWK - l'operazione impiega 68 ore.
- Importazione dei dump nel DBMS locale - 7.5 ore.
- Installazione *ossatura* - pochi minuti.
- Filtro SQL sul DBMS - 3.5 ore.
- Creazione o merging del database di servizio - 30 minuti.
- Interpretazione del *wikitext* e scrittura delle pagine HTML - 45 ore.
- Generazione degli indici per la ricerca - 1.5 ore.
- Download delle immagini dei testi enciclopedici - i tempi dipendono dalla quantità di immagini richieste, dalla connessione ad Internet e dal carico del server `upload.wikimedia.org`.
- Copia delle immagini - pochi minuti.
- Creazione del file immagine ISO-9660 - 1.5 ore.

I tempi riportati si riferiscono all'esecuzione delle operazioni sulla piattaforma di riferimento. È stato osservato che i tempi delle due fasi che implicano il download di dati dai server di Wikipedia non dipendono tanto dall'accesso ad Internet utilizzato (nella fattispecie la rete GARR) quanto dal carico dei server di Wikipedia.

Come si è potuto notare dall'analisi dei tempi le due fasi particolarmente lunghe dell'intero processo sono:

- Decompressione del dump XML ed esecuzione del filtro AWK per ridurne la dimensione.

- Interpretazione del *wikitext* ed esportazione delle pagine dalla base di dati, con implicitamente il progressivo aggiornamento del database di servizio e l'esecuzione delle estensioni.

Queste due fasi utilizzano per la loro esecuzione differenti risorse del sistema.

5.3.1 Fase di elaborazione del dump XML

La prima fase è composta da due operazioni, ovvero l'esecuzione dell filtro e la decompressione; vengono eseguiti in pipe per ridurre i tempi e lo spazio richiesto, poichè eseguirli separamente implica dover scrivere sulla memoria secondaria il dump XML decompresso e comporta sia un passo in piu', sia la neccessità di quasi 100 GB si spazio su disco.

`bzip2` è un compressore lossless che lavora con l'algoritmo di ordinamento di blocchi Burrows-Wheeler e un codifica di Huffman [119]. La compressione è buona e superiore agli algoritmi classici degli `zip` (compressori di tipo LZ77); le prestazioni sono però inferiori, però rientra nei tempi adeguati per essere utilizzato normalmente. Le prestazioni dell'operazione di decompressione sono tipicamente superiori a quella di compressione, in quanto in questa fase non è necessario ordinare i blocchi di dati. Inoltre la struttura di dati di tipo testuale, quale è il dump XML, sono compressi con ottimi risultati: il dump compresso per esempio risulta grande 4 GB.

`AWK` è un linguaggio di programmazione per realizzare scanner di schemi testuali, nato con i primi UNIX nel 1977. La versione utilizzata sul sistema FreeBSD in esame è `nawk`, la nuova versione del 1985, disponibile sotto licenza BSD. Esso è particolarmente idoneo all'operazione da effettuare sul dump, composto da solo testo. Esistono varie implementazioni, dotate di caratteristiche differenti; nella sezione successiva si esamina un'implementazione particolarmente ottimizzata che riesce a ridurre i tempi di `nawk`.

Le risorse del sistema maggiormente cariche durante questa fase sono (vedere le figure 5.2, 5.4 e 5.3) il processore e la memoria principale; nel sistema multiprocessore utilizzato soltanto uno dei due è carico, essendo sia bzip2 che AWK non strutturati con diversi thread. Anche l'esecuzione in pipe costringe i due processi all'esecuzione su un solo processore; la coda FIFO che collega lo standard out dell'uno e lo standard in dell'altro è implementata inevitabilmente in questo modo.

Uno dei core è quindi dedicato per il 10% del tempo a bzip2, mentre per il 85% ad AWK. È ragionevole approssimare che su 68 ore di esecuzione, circa 8 ore sono utilizzate dall'operazione di decompressione, mentre 60 dal filtro; il processo più lento è quindi l'esecuzione del filtro AWK.

5.3.2 Fase di esportazione delle pagine HTML

La fase di esportazione delle pagine comporta l'utilizzo intensivo del DBMS, elaborazione di codice PHP e in parte minore esecuzione delle estensioni e scrittura di file. Il codice PHP, che gestisce le varie operazioni da effettuare, non è eseguibile su molteplici processori; le estensioni sono eseguite all'interno di MediaWiki con chiamate di tipo exec, per cui vengono eseguite sullo stesso core. Ciò è comprensibile se si considera l'ambito con il quale

MediaWiki è stato concepito, ovvero un software wiki modulare che, dopo aver elaborato la richiesta, fornisce all'utente il test HTML e le immagini prodotte dalle estensioni.

Cionostante la parte più impegnativa di questa fase resta l'esecuzione delle query per estrarre i contenuti dal database. Il DBMS utilizzato, MySQL, sfrutta la presenza di diversi core per eseguire le query. I maggiori limiti all'esecuzione di questa fase si pongono quindi sull'accesso alla memoria secondaria, ove risiedono i dati del database, e sul processore, che sia performante sia fornito di molta cache accellera l'interpretazione del bytecode PHP.

5.3.3 Ridurre lo spazio occupato

Per quanto riguarda lo spazio per memorizzare i dati necessari all'estrazione (basi di dati, *albero*, immagini, etc), non è possibile ulteriormente ridurre la dimensione occupata dalle componenti, essendo il meccanismo già concepito per occupare meno spazio possibile. D'altra parte la dimensione dei supporti di memorizzazione cresce ad un ritmo più elevato di quello del progetto *itwiki*.

Avendo poco spazio su disco disponibile per l'estrazione è possibile cancellare i dump XML ed SQL dopo il loro inserimento nel DBMS locale; questo libera spazio per la fase di costruzione dell'*albero*. Inoltre è possibile effettuare l'importazione del dump XML filtrato nel database consecutivamente alle operazioni di decompressione e di filtraggio. Si può procedere eseguendo le fasi in cascata con:

```
bzip2 -dc itwiki-latest-pages-meta-history.xml.bz2 | \
./xmldumpskimmer.sh | java -Xmx200M -server -jar mwdumper.jar \
--format=sql:1.5 | mysql -u user -p wikidb
```

Effettuare quest'operazione evita di dover memorizzare il dump XML filtrato in locale, e accellera l'esecuzione dei passi effettuati singolarmente dato che non viene scritto sulla memoria secondaria il dump XML filtrato. Tuttavia questa ottimizzazione è sconsigliata poichè un eventuale errore o qualsiasi problema costringe ad effettuare di nuovo tutte le operazioni. Nell'esempio un fallimento nell'importazione in MySQL (operazione che eseguita separamente impiega 7 ore) obbliga a dover eseguire di nuovo la lunga fase di decompressione e applicazione del filtro (68 ore).

5.3.4 Ottimizzazioni software

È possibile cercare di ridurre i tempi di elaborazione con differenti approcci. Questi possono essere divisi tra ottimizzazioni a livello software e miglioramenti hardware; quest'ultimo può essere un semplice potenziamento delle componenti della piattaforma più usate ma anche un ristrutturamento architetturale, come l'impiego di più sistemi.

Delle due operazioni della fase di elaborazione del dump XML quella di decompressione è già ottimizzata; bzip2 è un ottimo software che offre il miglior compromesso fra rapporto di compressione e tempo di decompressione. L'esecuzione del filtro con AWK, che è anche l'operazione più impegnativa, è invece ottimizzabile.

Di implementazioni di AWK ne esistono molteplici: ci soffermiamo soltanto su quelle opensource che puntano al miglioramento delle prestazioni.

`awka` [120] è una versione di AWK che cerca di ottimizzare le istruzioni compilando il programma in un binario. Il file contenente le operazioni da far eseguire ad AWK viene quindi compilato in un eseguibile contenente anche l'implementazione stessa di AWK. Tale meccanismo funziona bene quando le istruzioni sono complesse; non è però il nostro caso, in cui i tempi lunghi sono dovuti alla massiccia quantità di dati da macinare. Ed in effetti tale approccio non migliora notevolemente la durata del processo.

`mawk` [121] è una versione che, partendo dal sorgente di `nawk`, implementa in modo differente ed ottimizzato la gestione delle strutture dati interne; questo a scapito di un maggior utilizzo di memoria principale del sistema. L'esecuzione del filtro sul dump XML con questa implementazione è in effetti molto più veloce: il processo di decompressione con bzip2 ed applicazione del filtro con `mawk` impiega 17 ore, contro le 68 di `nawk`. L'utilizzo della memoria è di poco superiore; varia invece molto la percentuale di occupazione del processore da parte del filtro AWK. Il 70% del tempo il processore elabora istruzioni di `mawk`, mentre il 25% è dedicato a bzip2: circa 13 ore sono dedicate al filtro, mentre solo più 4 ore a bzip2. Il minor carico ed i minor tempi di attesa da parte di AWK permette quindi anche all'operazione di decompressione di essere eseguita più velocemente.

In figura 5.17 è riportato un estratto del carico del sistema mentre si esegua la fase di decompressione ed esecuzione di `mawk`. Questa implementazione di AWK è quella attualmente utilizzata; si riporta in bibliografia [122] una serie di benchmark sulle diverse implementazioni di AWK prodotta dallo sviluppatore di awka.

```
last pid: 43595;  load averages: 0.41, 0.10, 0.03      up 34+09:55:15  22:43:50
9 processes:   2 running, 7 sleeping
CPU states: 48.9% user,  0.0% nice,  1.1% system,  0.0% interrupt, 50.0% idle
Mem: 108M Active, 158M Inact, 86M Wired, 46M Cache, 112M Buf, 602M Free
Swap: 2023M Total, 1648K Used, 2021M Free

  PID USERNAME  THR PRI NICE   SIZE    RES STATE  C   TIME   WCPU COMMAND
 4449 err         1 103    0  1868K  1204K RUN    0   0:16 70.58% mawk
 1408 err         1  -8    0  5004K  4420K pipewr 0   0:06 25.31% bzip2
76992 err         1  20    0  5080K  2944K pause  0   0:00  0.00% tcsh
29160 err         1  96    0  2204K  1832K select 0   0:00  0.00% screen
...
```

Figura 5.17. Filtro AWK sul dump XML: output del comando `top`.

La fase di esportazione delle pagine HTML non ha possibili miglioramenti a livello software, in quanto le varie operazioni qui effettuate sfruttano i meccanismi già ottimizzati del PHP, di MySQL e di MediaWiki.

Per quano riguarda MediaWiki si potrebbe progettare un codice più veloce per gestire il parsing del *wikitext*, però verrebbe meno la compatibilità futura garantita dal riutilizzo

della stessa piattaforma di Wikipedia. Infatti esistono parser del *wikitext* di MediaWiki alternativi, però nessuno garantisce il completo supporto del linguaggio; in sezione 3.2.3 è presente un discussione su questo punto.

5.3.5 Ottimizzazioni hardware

La piattaforma di riferimento è un intel x86 compatibile poichè per costi e diffusione è il sistema più facilmente disponibile. Inoltre i software utilizzati (FreeBSD, PHP, MySQL, etc), seppur disponibili per altre piattaforme, sono comunque ottimizzati e maggiomente testati per questi ambienti.

Un potenziamento dell'hardware deve porsi come obiettivi, per ridurre i tempi di entrambe le fasi più lente indicate, le caratteristiche del sistema elencate nella tabella 5.1; sono riportate le operazioni salienti delle due fasi più lunghe e per ognuna quale aspetto del sistema sfrutta maggiormente.

Caratteristica	bzip2	AWK	MySQL	PHP
throughput della memoria primaria		X		X
quantità di memoria primaria		X		
velocità del processore	X	X		X
cache del processore	X	X		X
throughput della memoria secondaria			X	
sistema multi processore			X	

Tabella 5.1. Utilizzo delle componenti hardware durante le maggiori operazioni.

5.3.6 Ottimizzazioni con più sistemi

Tra le due fasi maggiormente lunghe indicate precedentemente, quella della decompressione del dump è un'operazione, per come è gestita, poco scalabile; la fase di esportazione delle pagine è invece maggiormente scalabile, ed è scomponibile in processi da eseguire in parallelo. Una gestione strutturata è quindi più utile alla fase di esportazione delle pagine.

5.3.6.1 Fase di elaborazione del dump XML

Senza modificare lo schema di funzionamento della decompressione e del filtro AWK, l'unico accorgimento (impiegando differenti piattaforme) che potrebbe ridurre ulteriormente i tempi è l'esecuzione del processo di decompressione e l'esecuzione di AWK su due sistemi connessi tramite la scheda di rete. Il cavo di rete che interconnette i due sistemi deve essere dedicato: sul primo sistema, bzip2 provvede a leggere il dump XML, decomprimerlo e vengono trasmessi i dati decompressi; sul secondo, i dati in ingresso sull'interfaccia di rete dedicata vengono dati in pasto al filtro AWK, e quindi salvati sulla memoria secondaria.

Facendo un'analisi teorica dei tempi, il secondo sistema è sempre più carico del primo, essendo il filtro un'operazione quasi tre volte più lenta di bzip2; l'esecuzione del filtro,

avendo a disposizione molti più cicli del processiore, migliorerebbe circa del 133%. Il tempo totale, utilizzando due sistemi come quello utilizzato come riferimento dell'utilizzo delle risorse riportate in questo capitolo, raggiungerebbe circa le 12 ore.

Nel calcolo si è tenuto conto della banda passante delle interfacce di rete, che per due schede Gigabit dedicate hanno un throughput massimo tipicamente di 90 MB/s (il 75% di 1024 Mbps); il trasferimento è quindi trasparente rispetto alla fase di decompressione.

Un meccanismo di questo tipo può facilmente essere costruito utilizzando il programma di rete universale *NetCat* [123] (sono disponibili varie implementazioni su licenza BSD o GPL e per ogni piattaforma). Sul primo sistema si esegue:

```
bzip2 -dc itwiki-latest-pages-meta-history.xml.bz2 | \
nc IP_secondo_sistema 1025
```

mentre sul secondo:

```
nc -l 1025 | ./xmldumpskimmer.sh > \
itwiki-latest-pages-meta-history-RIDOTTO.xml
```

Si noti infine che una soluzione di questo tipo è probabilmente sconveniente essendo più costosa che la semplice sostituzione del processore per aumentare la frequenza di clock del sistema.

5.3.6.2 Fase di esportazione delle pagine HTML

La fase di esportazione delle pagine, essendo più scalabile, ha un maggior numero di possibili configurazioni.

Utilizzando due differenti sistemi, uno può essere utilizzato come server per eseguire il DBMS contenente il database di Wikipedia e l'altro effettua l'esportazione delle pagine, accedendo al server DBMS tramite un'interfaccia di rete. In questa configurazione è utile che il server DBMS sia un sistema multi-processore, in quanto i vari core vengono efficacemente utilizzati dal software MySQL. In tale modo il carico, inteso come accesso alla memoria secondaria (tipica di un DBMS), si concentra su un server che deve efettuare prevalentemente quest'operazione.

Il sistema che esporta le pagine deve soprattutto elaborare il *wikitext*, per cui utilizza soprattutto il processore e la memoria principale. In questo senso tale sistema non necessita di elevate banda di I/O, nè il PHP sfrutta l'eventuale presenza di diversi processori.

Tuttavia nel caso questo sistema sia dotato di più processori, è possibile sfruttare le opzioni di WaNDA `--idstart` e `--idstop`, al fine di elaborare tronconi di voci in parallelo. Questa tecnica è anche applicabile utilizzando più di due sistemi, per esempio un server DBMS e due o più piattaforme dotate dei medesimi componenti software per l'esportazione in parallelo di due o più parti dell'*albero*; esso deve quindi essere presente su un filesystem di rete condiviso, per esempio NFS.

Capitolo 6

Risultati

Il processo di generazione del progetto WaNDA ha diversi elementi variabili che dipendono da molti fattori, dato che si basa sull'utilizzo di componenti software variegate. Di più, la variabilità del processo è anche dato da elementi che variano di versione in versione dei dump, di MediaWiki e delle sue estensioni.

Ne consegue che le versioni dell'*albero* prodotte con il progetto WaNDA devono essere verificate prima di essere diffuse. La prima parte del capitolo riporta i meccanismi di verifica, tra cui un programma automatico presente in WaNDA-tools.

Il capitolo offre inoltre un'analisi a posteriori del progetto, riportando possibili miglioramenti futuri al contenuto offline e al processo di conversione. I problemi attuali vanno dai limiti del supporto di memorizzazione a quelli architetturali della conversione; dove possibile è quindi indicata una plausibile soluzione.

Inoltre il progetto tale e quale si applica a qualsiasi installazione di MediaWiki e va quindi ben oltre la sola enciclopedia Wikipedia, ma può essere utilizzato per generare una versione offline in primis di tutti gli altri progetti di Wikimedia.

Infine è presente una sezione sugli aspetti giuridici riguardanti il progetto. Vengono analizzate sia gli aspetti di licenza sui contenuti, essendo essi il frutto di un lavoro non coordinato di un gruppo di autori, sia quello che riguarda la diffusione del prodotto, con qualsiasi mezzo avvenga.

L'ambito giuridico preso in considerazione è quello dove è stato sviluppato il progetto WaNDA, ovvero l'attuale legislazione italiana. Le tematiche proposte sono state analizzate anche in collaborazione con legali del settore.

La mancanza di una giurisdizione adeguata che contempli l'ambito open content del progetto ne ha fortemente rallentato lo sviluppo, poiché la responsabilità sia civile che penale (in caso di diffamazione) dovrebbe probabilmente essere assunta da un soggetto editore. A tutt'oggi non è ancora stata rilasciata pubblicamente alcuna versione su supporto ottico; sarà probabilmente più semplice riuscire nel rilascio online dell'archivio contenente l'albero.

6.1 Verifica

Il processo di generazione dell'albero è composto dalle varie fasi descritte; esse sono compiute utilizzando un'insieme numeroso di software differenti, per cui è possibile che dopo l'aggiornamento di alcuni componenti il prodotto finale presenti alcuni problemi.

L'aggiornamento e l'aggiunta di estensioni a MediaWiki è un'operazione obbligata, poiché è necessario mantenersi al passo con il dump XML di Wikipedia. L'enciclopedia online, ed il software che la fà funzionare, hanno un'evoluzione costante: MediaWiki viene ampliato o modificato, e lo stesso vale per le estensioni. Anche il formato del database utilizzato da MediaWiki e quindi il formato del dump XML subiscono modifiche a seconda della versione di MediaWiki.

Tutto questo tende a portare alcuni errori nel formato finale, il che richiede alcune messe a punto alle componenti utilizzate per la generazione. Gli errori possono essere di diverso tipo e sostanzialmente due sono i metodi per verificare il formato finale.

6.1.1 Tipologie problemi

I problemi riscontrati nel formato finale possono essere problemi dal punto di vista formale, del contenuto o problemi nei collegamenti.

Nel caso di errori formali vi possono essere problemi di presentazione della pagina o di formattazione del testo. Tipicamente questi errori o sono già presenti nel dump XML oppure vengono introdotti da discrepanze tra la versione locale di MediaWiki rispetto a quella online. A questo proposito si noti che la versione online, prima di inviare il testo HTML della pagina all'utente, effettua una correzione automatica di eventuali errori di formato; questo controllo sulle pagine è tuttavia abbastanza pesante. È stato notato che questi errori sono molto sporadici e di solito sono tag HTML errati presenti nelle pagine di *template*, descritte in sezione 5.2.3.1; dopo una verifica manuale dell'albero è quindi facile andare a correggere la sintassi del template nel MediaWiki locale.

I problemi sul contenuto possono riguardare il testo o le immagini. Gli errori riscontrati del primo tipo erano riconducibili ad errori formali dell'HTML; un eventuale problema nel testo dovrebbe comunque essere dovuto alle modifiche introdotte nel codice di MediaWiki (presentate in sezione 4.2.5). Questi problemi inoltre sono difficilmente identificabili, poiché per la verifica è necessario confrontare la pagina dell'albero con il corrispondente online alla data del dump XML.

Gli errori sulle immagini sono più frequenti e possono avere molte forme. La prima è la mancanza di tutte le immagini matematiche o dei grafici generati da MediaWiki, dovuta ad una configurazione errata di MediaWiki. Poi possono essere mancanti alcune immagini caricate dali utenti; le cause possono essere la loro rimozione dal repository di Wikimedia, per cui il loro nome è presente nel dump XML ma non sono più accessibili online. Anche gli spostamenti od il cambio di nome online provoca la mancanza delle stesse. Inoltre nel tempo il repository dal quale ottenere le immagini è cambiato, per cui possono essere problemi siccome WaNDA ne utilizza soltanto uno. Un'immagine mancante con il nome atteso sul repository lascia la cornice e la didascalia della stessa pur essendo assente.

I problemi sui collegamenti, i più frequenti, possono riferirsi a collegamenti verso voci

locali, voci esterne oppure immagini. Tutti questi possono essere ricondotti alle modifiche introdotte nel codice di MediaWiki ai fini del progetto WaNDA. I collegamenti interni, funzionanti nella versione online, che sono presentati come testo semplice nell'albero implicano che la voce locale di destinazione sia mancante; potrebbe accadere per pagine rimosse per qualche motivo di selezione, oppure per problemi nel dump XML iniziale.

I collegamenti alle voci esterne ed alle immagini potrebbero presentare problemi nella misura in cui la voce online sia stata spostata o cancellata.

6.1.2 Controllo automatico

Il meccanismi per controllare la validità della versione dell'albero prodotta sono due: un programma PHP sviluppato appositamente per lo scopo, e una comunità di persone che verificano il funzionamento.

Il programma è stato scritto in PHP per non utilizzare un'ulteriore tecnologia differente; fa parte del pacchetto WaNDA-tools. Esso si compone di varie funzioni, contenute e richiamate all'inizio di un file: la principale è una funzione ricorsiva che individua i collegamenti interni nel testo di una pagina HTML di WaNDA e procede alla verifica di esistenza delle pagine collegate; un'altra funzione è pensata per scorrere tutti i file HTML dell'albero e verificare la presenza delle immagini incluse.

È quindi possibile effettuare tre controlli, selezionabili individualmente; oltre al controllo da effettuare, deve essere indicato il percorso dell'albero.

- Verifica in ampiezza dei collegamenti interni delle pagine dell'albero: il programma apre gli indici HTML di tutte le pagine (descritti in sezione 5.2.3.4) ed effettua un controllo con profondità due, effettuando quindi il controllo di validità di tutte le pagine indicizzate e dei loro collegamenti.

- Verifica in profondità dei collegamenti interni: il programma apre la pagina principale dell'albero e ne verifica la validità dei collegamenti, per poi passare al controllo delle pagine collegate; questo controllo potrebbe teoricamente non concludersi mai, ma poiché analizza in modo più casuale le pagine offre rapidamente un'analisi statistica sulla qualità dei collegamenti.

- Verifica delle immagini, incluse le formule matematiche e i grafici: il programma effettuata una scansione di tutte le pagine dell'albero e controlla la presenza dei file delle immagini incluse.

6.1.3 Community

Il controllo della validità dell'albero è maggiormente efficace se effettuato da parte di una comunità di persone, principalmente perché le tipologie di problemi verificabili sono maggiori.

La comunità ideale è quella di Wikipedia stessa, che conosce già l'enciclopedia e molti dei tipi di errore presenti anche online; in verità anche persone non affini a Wikipedia sono molto utili per notare elementi di usabilità che possono sfuggire alle persone maggiormente addette ai lavori.

Per questi motivi il progetto WaNDA è supportato da una mailing list privata di Wikimedia Italia in cui vengono comunicate le versioni dell'albero e i resoconti sui problemi.

Essendo un progetto opensource l'idea di fondo è che la comunità si ingrandisca con altri sviluppatori che migliorano il software di WaNDA e pareri sulla presentazione dei contenuti. Per questo per il futuro, se il progetto prende piede, sono già disponibili canali di comunicazione pubblici.

6.2 Miglioramenti futuri

Vari miglioramenti possono essere apportati a WaNDA-tools, principalmente per aggiungere funzionalità all'albero, ma anche per superare alcuni dei limiti attuali del meccanismo di conversione.

6.2.1 Limiti progettuali

Il dump XML di Wikipedia è una fotografia istantanea di un sistema dinamico; avviene quindi implicitamente una sorta di trasformazione a monte del progetto. Alcune pagine possono quindi presentare errori sul contenuto o sui collegamenti, dovuto all'operazione di dump del database compiuto quando Wikipedia è in funzione. Per esempio possono contenere testo spurio comparso per soltanto un breve lasso di tempo, oppure possono non essere raggiungibili pagine perché in quel momento erano state spostate.

Questo tipo di limite è implicito a qualsiasi sistema offline che utilizzi programmi automatici. Il progetto WaNDA cerca di contenere questo problema generando per ogni versione del dump XML il corrispettivo formato finale, che migliora le incongruenze passate, ma può anche introdurne di nuove; l'aggiornamento costante permette di disporre di una versione possibilmente recente, evitando di tenere a galla inesattezze per troppo tempo.

La conversione da sistema dinamico a statico è effettuata in un'altra fase, più propria di WaNDA, intendendo come dinamico l'installazione di MediaWiki locale e come sistema prevalentemente statico l'albero. MediaWiki infatti permette per determinate funzionalità non solo di richiamare la pagina da URL, ma anche di aggiungere comandi specifici; per esempio questo è utilizzato per visionare versioni vecchie delle voci, oppure per la ricerca di MediaWiki. Una di queste condiziona il formato finale: le categorie. Esse infatti se contengono più di un determinato numero di voci (di default 200), vengono divise in sotto pagine con lo stesso nome, ma distinte da parametri aggiuntivi nell'URL.

Il risultato nell'albero è che di default queste categorie corpose contengono collegamenti, anziché alle altre pagine della categoria, alla stessa pagina. Un modo provvisorio per risolvere questo problema è stato di incrementare in MediaWiki il numero di voci per ogni pagina, rischiando però di generare pagine molto lunghe. Si noti però che una prassi recente di Wikipedia è quella di raggruppare le categorie in sotto categorie, in modo da evitare anche online queste lunghe categorie.

Un'altra operazione effettuata in questa fase è quella di adattamento al supporto finale, che prevede l'eliminazione della differenza tra titoli delle voci in maiuscolo e minuscolo.

Se quest'operazione è pressochè invisibile per le voci che hanno un titolo lungo, non è così evidente per le stringhe con pochi caratteri (per esempio tre). Una voce con due o tre caratteri è statisticamente più probabile che possa avere molteplici significati e quindi diverse voci, magari scritte con differenti combinazioni di maiuscolo e minuscolo. Questo problema è anche presente su Wikipedia, poiché la distinzione non è evidente per l'utente finale; a questo proposito le linee guida recenti di Wikipedia consigliano di scrivere una pagina di disambiguità con il titolo minuscolo e le altre voci che dovrebbero avere il titolo simile dovrebbero essere meglio identificate posponendo l'ambito della voce tra parentesi tonde.

Un modo per risolvere entrambi i limiti descritti potrebbe essere quello di posporre nel nome dei file dell'albero un identificativo univoco della voce, per esempio basato su un hash del titolo che tenga conto della differenza tra caratteri maiuscoli e minuscoli.

Infine un limite implicito è dato dall'elaborazione automatica del processo partendo da informazioni presenti online, che può portare ad una selezione errata dei contenuti. Questo tipo di errore può quindi dipendere dall'appartenenza errata di una pagina ad una categoria, oppure ad una sua non appartenza. Essendo i contenuti di Wikipedia non affidabili, neppure queste informazioni di corredo lo sono, anche se in misura molto minore dato il numero molto più ristretto di utenti che se ne occupano. Un primo esempio può essere la categoria delle pagine non neutrali, non per forza affidabile; un altro pagine proprie della gestione di Wikipedia che non fanno parte del namespace corretto.

Questo limite seppur presente può essere ridotto andando a correggere la versione online di Wikipedia, in modo tale che il dump XML successivo sia dotato di informazioni maggiormente corrette; questo vale chiaramente anche nel caso degli problemi elencati precedentemente che sono frutto di errori di Wikipedia. Per esempio i template che hanno una sintassi errata possono facilmente essere notati nell'albero e corretti online, ma difficilmente sono individuabili online.

6.2.2 Il limite del supporto

Un limite di altra natura è quello che riguarda il supporto che deve contenere l'albero. Le ultime versioni di WaNDA, ad ogni versione di dump, si stanno progressivamente avvicinando al limite dei 4.4 GB dei DVD; siamo ancora sui 3.5 GB. Sarà poi possibile ridurre l'elenco delle categorie di immagini da includere, ma può che essere una soluzione definitiva. Per i supporti su FAT32 il problema non dovrebbe porsi, dato che ad oggi i palmari, i pendrive, le schede di memoria e così via sono disponibili con spazio di memoria da 8 GB ed il trend di crescita è costante.

Il problema di spazio si porrà quindi soltanto per i dischi ottici; vi sono alternative future, come i DVD Dual Layer, i Blu-Ray Disc e i HD DVD. Ad oggi esistono nel commercio di largo consumo alcuni masterizzatori e supporti DVD Dual Layer, che offrono una capacità di 8 GB; il problema è che sono ancora poco diffusi e i supporti sono costosi. Oltretutto male di adatta all'idea di libera diffusione dei contenuti, se poi l'albero non può facilmente essere copiato e ridistribuito. Gli altri due formati sono ancora agli inizi della loro diffusione, la loro capienza è nell'ordine delle decine di GB, e per ora non sono ancora utilizzati per contenere qualsiasi tipo di dati ma soltanto filmati.

In futuro i supporti DVD DL scrivibili potranno quindi probabilmente essere utilizzati; nel frattempo se la dimensione dell'albero dovesse superare un DVD e questi supporti Dual Layer non essere abbastanza diffusi, è comunque disponibile la versione alternativa per supporti su FAT32, che comporta però problemi nella distribuzione fisica poichè il supporto è più costoso.

6.2.2.1 Supporti ottici multipli

Un'alternativa presa in seria considerazione per superare i limiti di spazio del DVD, è quella di dividere i contenuti su due dischi. Tecnologicamente è possibile farlo e renderlo compatibile per più piattaforme. Il browser di sistema infatti una volta che ha caricato la pagina HTML non tiene aperto il file; questo dovrebbe essere valido per tutti i browser poiché è un comportamento simile alla natura *stateless* del protocollo HTTP normalmente utilizzato. Si dovrebbe quindi sviluppare un ulteriore componente in JavaScript che gestisce il cambio di disco: è necessaria una lista delle voci presenti sull'uno e sull'altro disco, oltre che ad un meccanismo per la verifica del disco presente ed il passaggio di informazioni di stato.

Più complessa è invece la gestione dei contenuti stessi. In primo luogo le immagini, che devono essere presenti sul disco dove sono presenti le pagine che le includono: richiede un minimo di elaborazione maggiore ma non è un problema per i grafici e le formule matematiche che hanno una corrispondenza uno ad uno con le pagine che le includono. Potrebbe essere più delicata la gestione delle immagini caricate dagli utenti, che possono essere incluse da più di una pagina e devono quindi essere sdoppiate.

Il più grande problema a questa soluzione resta la praticità per l'utente: è molto fastidioso dover cambiare alternativamente il disco passando di argomento in argomento. Sarebbe possibile cercare di raggruppare le pagine secondo ipotetici pattern di navigazione dell'enciclopedia, cercando di ridurre i cambi; tuttavia richiederebbe un'analisi molto approfondita della struttura dei collegamenti che probabilmente non varrebbe la pena intraprendere, dato che comunque non risolverebbe completamente il problema.

6.2.3 Motore di ricerca avanzato

Una delle prime ipotesi sulle tecnologie da utilizzarsi per i contenuti attivi prevedeva un programma Java presente sul supporto, da eseguire grazie ad una Java Virtual Machine opensource. La JVM potrebbe essere sia già presente sulla piattaforma utente (opensource oppure proprietaria SUN che sia), sia essere installata dal supporto.

Questa soluzione, pensata inizialmente per effettuare la ricerca e scartata poiché non è multi piattaforma e richiede probabilmente un'installazione, potrebbe affiancarsi alla ricerca attuale effettuata con tecnologia più diffusa. Il vantaggio di un motore di ricerca Java risiede nella maggiore libertà di implementazione, che permetterebbe di effettuare una ricerca non solo sul titolo delle voci ma anche sui contenuti delle pagine; il supporto disporebbe quindi di due motori di ricerca, uno di base già implementato e qui descritto, ed uno complementare più completo, magari non funzionante su tutte le piattaforme.

Il motore di ricerca Java dovrebbe essere dotato di una sorta di database che contiene

un'indicizzazione dei contenuti. È possibile per esempio generare un elenco di termini con indicato le pagine in cui il termine è presente, con un'indicazione della percentuale di rilevanza. Tale specie di indicizzazione deve essere effettuata in fase di elaborazione. La ricerca potrebbe anche implementare un meccanismo di ricerca non esatta sui termini, ma approssimativo utilizzando una funzione per il calcolo delle distanze fra due stringhe.

6.3 Estensione del progetto

Lo scopo del progetto WaNDA è quello di generare una copia offline dell'enciclopedia Wikipedia; il software prodotto per lo scopo permette tuttavia di generare le copie offline di molti altri contenuti.

La prima estensione del progetto può essere la generazione delle copie offline delle altre lingue oltre a quella italiana. L'applicabilità a tutti i *wiki è garantita dal fatto che la piattaforma dei server è la medesima per tutte le lingue; gli applicativi di WaNDA-tools sono generici e funzionano con tutti i formati dei dump, siano anche tutti i testi codificati con caratteri non occidentali.

L'unico limite di utilizzo dell'albero generato per altre lingue riguarda la dimensione che esso occupa; per esempio applicare WaNDA a enwiki vuol dire, facendo una prima stima molto approssimativa da un confronto con itwiki (da circa 320 mila voci si ottiene un albero da 3.5 GB), che da 1.4 milioni di voci si ottiene un albero da 15 GB. Tale dimensione potrebbe essere memorizzata su un disco o su un palmare con abbastanza spazio, ma non certo su un supporto ottico di oggi. Il problema non si presenterebbe comunque per la maggior parte delle altre lingue.

Il processo di conversione dei contenuti potrebbe essere applicato anche agli altri progetti di Wikimedia (i wiki*), elencati in sezione 2.1.1. Essi non differiscono molto da Wikipedia, le uniche possibili differenze riguardano le estensioni di MediaWiki, da installare come descritto nel capitolo 5. Anche qui vale il limite della dimensione dell'albero, anche se gli altri progetti non sono così sviluppati quanto Wikipedia e quindi non dovrebbero avere questo problema.

Come ambito di applicazione potrebbe per esempio avere un senso, utilizzando Wiktionary, disporre di un dizionario su CD; volendo si possono ottenere vari CD uno per lingua.

Infine è possibile adattare il progetto a qualsiasi MediaWiki. WaNDA-tools è ottimizzata e pensata per le installazioni di Wikimedia, ma il software funziona con poche modifiche agli altri siti. Le differenze nel processo possono riguardare il formato del dump da importare nel database locale, la versione di MediaWiki e le estensioni. Potrebbe per esempio essere utilizzato in ambito accademico per generare un formato offline di contenuti elaborati collettivamente su un'installazione locale di MediaWiki.

6.4 Licenze e giurisdizione

Nei vari capitoli precedenti sono stati presentati sostanzialmente vari programmi per la gestione di contenuti testuali e grafici. Ci si sofferma ora sulle implicazioni legali che ciò può avere, in particolare al caso in questione dove i contenuti hanno come autori diversi soggetti e in rapporto all'attuale diritto italiano; sono descritti sia gli aspetti del software per la generazione WaNDA che quelli dell'opera collettiva Wikipedia.

La tecnologia digitale nella società dell'informazione ha presentato sempre più spesso nuove implicazioni giuridiche non previste. Se questo da una parte ha causato l'introduzione di regolamentazioni che contemplano il prodotto software, dall'altra le licenze d'uso permissive del software sono state estese ad altri campi della creatività non software. È questo il caso delle licenze opencontent del testo di Wikipedia descritta in sezione 2.1.2, che deriva dalle licenze opensource; queste licenze creative sono infatti nate con la documentazione tecnica allegata al software, poi estesa al altri testi e ad altre opere figurative o musicali.

La tutela giuridica del software è relativamente recente, nata negli anni '70 e '80 dall'interesse nel creare dei diritti esclusivi per garantire profitti al mercato del software. Tale tutela è nata inizialmente negli USA, dove può essere attuata tramite brevetti e diritti d'autore; in Europa il software viene tutelato dal diritto d'autore, evitando il brevetto che nel campo software crea molti problemi. Una normativa europea del 1991 [134] sancisce una regolamentazione al livello comunitario, imponendo l'applicazione del codice sul diritto d'autore per la protezione del software; esso viene equiparato ad un'opera letteraria di carattere tecnico.

Il diritto d'autore difende in modo esclusivo la sua opera, software, testuale, figurativa. musicale o altro che sia; inoltre la forma esclusiva protegge l'integrità dell'opera, impedendone la modifica. Per concedere diritti, secondo la legislazione classica, l'autore deve stipulare un contratto apposito con il destinatario della concessione.

L'autore può però voler rilasciare l'opera con dei diritti particolarmente laschi, senza dover stipulare contratti ad hoc per ogni ente richiedente. A questo scopo sono nate forme di licenze, ovvero concessioni di diritti, di tipo non esclusivi; queste licenze libere (siano opensource o opencontent) hanno tutte come regola almeno l'indicazione dell'autore, che continua comunque a detenere i diritti sull'opera.

Si noti che le licenze d'uso non sono leggi ma sono regole con le quali l'autore decide di rilasciare un'opera; il diritto tutela quindi l'autore, non la licenza. Essa è uno strumento per elencare le concessioni dei diritti da parte dell'autore, che detiene comunque la paternità e può stipulare contratto ad hoc sull'opera.

Il diritto d'autore in forma esclusiva si applica automaticamente ad ogni opera; l'autore che intende rilasciarla con una licenza d'uso libera deve quindi fare una dichiarazione esplicita dei termini della licenza.

Negli anni sono sorte diverse licenze libere, le prime riguardanti solo il software, le ultime per ogni opera creativa. È possibile fare una classificazione per le licenze software, che in modo analogo può essere applicata alle opere non software; si dividono principalmente in due gruppi qui riassunti:

- software proprietari, in cui è permesso soltanto l'utilizzo del software (pertanto diffuso soltanto come binario), dietro compensazione economica o meno;

- software libero o opensource, in cui è permessa la redistribuzione e la modifica, eventualmente sotto certe condizioni.

Il software libero ha molti gradi di concessione dei diritti: esso copre infatti le combinazioni di diritti che vanno dal software proprietario al pubblico dominio. Con pubblico dominio si intende un'opera priva di diritti e quindi senza licenze, come sono per esempio le opere letterarie con diritti scaduti.

Alcune licenze libere molto efficaci sono quelle utilizzate nel progetto WaNDA, le licenze GNU GPL e FDL, che permettono l'utilizzo, la copia e la modifica dell'opera a condizione che sia riconosciuta la paternità all'autore e che un'opera derivata sia rilasciata sotto la medesima licenza (questa sorta di 'viralità' è descritta successivamente). Una delle licenze d'uso più libere è per esempio la licenza BSD (Berkeley System Distribution) [128], che prescrive la sola attribuzione dell'autore e qualsiasi utilizzo, copia o modifica dell'opera; il software può quindi essere riutilizzato, con indicazione dell'autore, in un prodotto proprietario. Una licenza molto elastica è quella Creative Commons (CCPL), che permette di concedere varie combinazioni di diritti a seconda delle esigenze dell'autore, ed è definita in rapporto alla legislazione locale di molti Stati.

Con la diffusione di molte licenze libere è nata la necessità di confrontarle fra di loro allo scopo di identificare le loro equivalenze. Per esempio una licenza Creative Commons che esclude l'ambito commerciale e obbliga alla redistribuzione con la stessa licenza è simile ad una licenza GNU GPL o FDL. Inoltre sono possibili dei passaggi di licenza, per esempio ad un'opera con licenza BSD può essere inclusa in un'opera GNU GPL, ma non viceversa.

Questi cambi di licenza sono molto interessanti, poiché permettono al progetto WaNDA di includere opere con licenze più lasche della GNU FDL.

6.4.1 Licenza GNU GPL

La licenza d'uso per il software GNU General Public License [127] nasce nel 1991 grazie alla figura di R. Stallman per contrastare il diffondersi del software rilasciato con licenze esclusive. La versione 2 è la più diffusa, oltre ad essere quella utilizzata per il software MediaWiki e WaNDA-tools. Il documento stesso della GNU GPL è rilasciato con la permissione di distribuzione e copia, ma non di modifica.

La licenza, pensata per il diritto statunitense, sancisce la libera distribuzione e la copia del software; la possibilità di modificare l'opera; nel caso di modifica o copia in forma binaria è necessario rilasciare anche il codice sorgente completo; assenza di garanzia, per scaricare giuridicamente qualsiasi responsabilità degli autori.

Infine un punto molto importante della licenza riporta che la redistribuzione successiva ad una modifica deve avvenire secondo i termini della licenza stessa. Questa condizione connota la licenza di un aspetto 'virale', poiché costringe ogni software derivato anche solo in minima parte da un software GNU GPL deve essere rilasciato dall'autore con la stessa

licenza GPL. Vuole anche dire che i programmi coperti o derivati da questa licenza non possono essere incorporati in un prodotto proprietario.

Questa licenza è quella che ha dato il via al filone di licenze e pensiero indicato come 'copyleft', di cui si è parlato in sezione 2.1.2.

6.4.2 Licenza GNU FDL

La licenza d'uso GNU GPL iniziò ad essere utilizzata anche per la documentazione tecnica in corredo ai software GNU GPL; per rispondere a questa lacuna nel 2000 il gruppo di Stallman redatte una licenza d'uso per i documenti sulla falsariga della GPL, che prende il nome di GNU Free Documentation License [126]. È molto simile alla sua equivalente per il software (anch'essa è di tipo copyleft) e si applica non solo alle opere testuali, ma è abbastanza generica da adattarsi ad altre opere artistiche, su qualsiasi supporto esse siano, cartaceo, digitale o altro.

È sancita la libertà di copiare e distribuire con o senza modifiche l'opera, con attestazione della paternità dell'opera originaria. Inoltre anche qui un'opera derivata deve essere rilasciata con la stessa licenza FDL; deve quindi essere allegata all'opera una copia del testo della licenza FDL. Nelle opere che hanno subito rielaborazioni o inclusioni di lavori di vari autori, devono essere dichiarati tutti gli autori delle singole modifiche.

Nel caso di riproduzione commerciale deve essere indicato un indirizzo web dover reperire l'opera stessa gratuitamente; inoltre deve essere previsto un modo per ottenere le versioni gratuite delle modifiche di ogni singolo autore.

La licenza GNU FDL è applicabile non solo ai testi, ma anche ad opere figurative o musicali; il suo ambito di applicazione di solito prevede formati digitali non proprietari facilmente modificabili e compatibili con diversi sistemi. È perciò perfetta per l'ambito del progetto WaNDA.

La licenza FDL è inoltre molto interessante per i progetti quali WaNDA, poiché dedica sia una sezione al caso di raccolta di testi rilasciati sotto licenza FDL, che una al caso di raccolta di testi FDL e non.

L'opera di raccolta enciclopedica è un'opera ottimamente protetta dalla licenza FDL, poiché si presenta come un'opera composta da molti testi con permessi di copia e modifica; non a caso FDL e Wikipedia, assieme al progetto GNUPedia, sono nate nello stesso periodo. In Wikipedia è perfettamente tutelato l'aggiornamento dei testi, con precisa attribuzione dell'autore ai frammenti di testo; inoltre le modifiche sono attuabili da chiunque.

Tutto il materiale di Wikipedia è rilasciato sotto licenza GNU FDL e include, come previsto dalla licenza, anche opere non FDL ma con licenze compatibili o più lasche; per esempio le immagini, che sono rilasciate sotto Creative Commons o sono di pubblico dominio. Sono poi state introdotte alcune cautele, ma non viene intaccato l'aspetto funzionale.

La licenza d'uso FDL, come la GPL, è una licenza 'virale', poiché l'opera derivata deve anch'essa essere distribuita con questa licenza; ne consegue che i contenuti del progetto WaNDA sono presenti con licenza FDL, poiché deriva da Wikipedia in quanto i contenuti vengono redistribuiti (non avviene una modifica sui testi).

Inoltre, tutto l'ambito copyleft è congeniale alle opere di informazione scientifica, poiché il suo scopo è principalmente il progresso e la diffusione della conoscenza. Le esigenze di libera diffusione e meno quelle di mercato fanno di questa licenza, e più in generale quelle open content, l'ideale negli ambiti didattici e nella ricerca. Inoltre la scuola pubblica, essendo basata sul finanziamento pubblico, non ha uno scopo commerciale e può necessitare di documentazione libera per la formazione degli studenti. Si noti che questo ragionamento può valere anche per i documenti e il software nella pubblica amministrazione, che essendo un ente pubblico richiederebbe trasparenza, adattabilità e indipendenza dal mercato.

6.4.3 Licenze Creative Commons

Le licenze Creative Commons Public Licenses (CCPL o brevemente CC) [124] nascono verso il 2002 sotto la direzione di L. Lessig; esse sono molto flessibili in quanto sono diverse licenze, caratterizzate da alcuni termini di base, con diverse concessioni di diritti. Sono pensate per essere applicate a qualsiasi ambito della creatività, che siano opere figurative, musicali, multimediali, software o altro. Inoltre il testo delle differenti licenze è stato adattato in molti paesi per essere congruente con la legislazione locale.

Riassumendo, le licenze CC come diritto riservato hanno l'"attribuzione", che prevede la paternità dell'autore originario. I diritti che possono essere concessi sono indicati con: "non opere derivate", che vieta la modifica o l'inclusione dell'opera in altri prodotti; "condividi allo stesso modo", prevede che le opere derivate o la distribuzione avvenga secondo la stessa licenza dell'opera originale (questo concessione è incompatibile con la precedente; si noti inoltre che questa concessione rende la licenza 'virale' come quella GNU FDL); "non commerciale", che esclude il profitto ottenuto con la distribuzione dell'opera. Dalla combinazione di queste tre concessioni è possibile ottenere un gruppo di sei licenze d'uso.

I contenuti grafici di Wikipedia, se non GNU FDL, sono spesso rilasciati con queste licenze, poiché più semplici e maggiormente applicabili alla legislazione italiana. Nel progetto WaNDA non è chiaro quale sarà la destinazione del supporto, ma probabilmente ricadrà nella definizione di profitto, che comprende anche il guadagno di immagine; ai fini del progetto WaNDA devono quindi essere esclusi i contenuti grafici rilasciati con una licenza CC di tipo "non commerciale".

6.4.4 Licenza d'uso di WaNDA-tools

Tutte le componenti software e non del progetto WaNDA sono rilasciate con licenze opensource. Per analizzare le ragioni e le licenze utilizzate è necessario dividere le componenti in due gruppi funzionalmente diversi: gli applicativi utilizzati per la generazione dell'albero, e le opere incluse nell'albero. In ogni caso si è cercato di utilizzare licenze libere e di utilizzare tecnologie opensource, poiché confacenti sia all'ambito di Wikimedia, sia a quello del gruppo linux@studenti, presso il quale è nato il progetto.

Le parti software del pacchetto WaNDA-tools utili all'esportazione dell'albero sono scritte con linguaggi di programmazione interpretati, come per esempio il linguaggio PHP. La licenza dei codici PHP, poiché originariamente uno dei file (`maintenance/dumpDVD.inc`)

deriva da MediaWiki, prodotto software rilasciato sotto licenza GNU GPL, devono seguire la stessa licenza. L'indicazione della licenza, e dell'autore originario per quel file in particolare, sono indicati in testa ad ogni file; assieme al pacchetto è inoltre presente il testo della licenza completa.

Le altre parti software per l'esportazione, come lo script AWK, sono anch'essi rilasciati sotto licenza GNU GPL.

Le opere presenti nell'albero sono di varia natura: sono presenti codici software per i contenuti attivi, testi redatti dal progetto WaNDA, i testi enciclopedici tratti da Wikipedia, immagini tratte da Wikipedia ed eventualmente binari per la consultazione.

Il codice software presente nell'albero, principalmente composto dal motore di ricerca JavaScript, è rilasciato con licenza GNU GPL; anche il tema utilizzato in tutte le pagine è sotto GPL, poiché è di derivazione MediaWiki. La documentazione del progetto WaNDA è rilasciata sotto licenza GNU FDL. Questo vale anche per i testi enciclopedici tratti da Wikipedia; in ogni voce enciclopedica è riportato l'elenco degli autori ed un collegamento alla voce attuale presente online.

Sono poi presenti le immagini di Wikipedia; per l'inclusione devono essere rispettate le licenze compatibili con la GNU FDL. Oltre alla licenza FDL, sono incluse le immagini di pubblico dominio e tutte le Creative Commons tranne quelle di tipo "non commerciale".

L'eventuale browser per la compilazione, dato l'ambito del progetto, deve essere di tipo opensource. Attualmente viene utilizzato K-Meleon, un browser di derivazione Gecko rilasciato sotto licenza GNU GPL; esso include quindi una parte di software rilasciato sotto Mozilla Public License (MPL) [129]. Nell'albero il software è presente in forma binaria, e data la sola distribuzione del prodotto è indicata la paternità dell'opera.

Nell'albero è presente il testo completo delle varie licenze citate.

Il pacchetto WaNDA-tools è rilasciato su un sito pubblicamente accessibile, ed essendo composto da software GPL e testi FDL, ogni sua modifica dovrà essere rilasciata con le stesse licenze indicate.

6.4.5 Responsabilità sui contenuti

È stato effettuato, nel corso dello sviluppo del progetto, un'analisi per la limitazione di responsabilità sui contenuti dell'albero. Il problema si pone per i contenuti tratti da Wikipedia, siano essi testuali o grafici, ottenuti tramite una selezione ed elaborazione automatica, che non richiede quindi intervento umano nell'elaborazione.

Wikipedia dal punto di vista giuridico è un servizio che rende disponibile agli utenti degli strumenti web per la raccolta e la memorizzazione di informazioni, composti da MediaWiki e dai server di memorizzazione siti in Florida. Per cui cui Wikimedia potrebbe essere considerata fornitrice di un servizio di "hosting" definito nell'articolo 16 del D.Lgs n. 70/03 [133], che quindi gode di una certa limitazione di responsabilità sui contenuti forniti dagli utenti. Ovviamente non appena le autorità competenti abbiano dato notizia di un contenuto illecito, Wikimedia deve provvedere alla rimozione degli stessi. Quindi secondo il punto di vista giuridico gli autori delle voci sono i destinatari del servizio.

La redistribuzione dei contenuti, resa molto facile dal progetto WaNDA, può avvenire in due modi: con il rilascio online del formato ISO-9660 o dell'archivio compresso, oppure con la distribuzione del supporto ottico. Questi due canali di distribuzione hanno grandi differenze in quanto a responsabilità civile a causa di un cambio di legislazione: nel primo caso si applica la legge sul commercio elettronico, nel secondo caso la legge sulla stampa. È un'anomalia legislativa, dovuta alle tecnologie in gioco che non sono propriamente contemplate dalla giurisdizione.

6.4.5.1 Distribuzione online

Il rilascio dell'albero sotto forma di immagine ISO-9660 e di archivio compresso su un sito online viene considerato come riproposta in formato diverso ma uguale come sostanza di opere gratuitamente presenti online. Si ricade quindi nell'ambito del D.Lgs 70/03 [133], che prevede la redistribuzione di contenuti nell'articolo 1 ("libera circolazione dei servizi della società dell'informazione").

La limitazione di responsabilità dell'articolo 16 prevede l'estraneità di chi fornisce il servizio, in questo caso Wikimedia, alle informazioni immesse dagli utenti, destinatari del servizio. Questo decreto non si applica quindi ai destinari del servizio, nel caso in questione un utente che preleva i contenuti da Wikipedia e li ripropone dopo averli elaborati con WaNDA-tools; questo vale anche per chi decide di riproporre sul proprio sito i contenuti di Wikipedia. Il decreto non può avere un'interpretazione estensiva, e quindi applicarsi a terzi che ripropongono il servizio, poiché è una legge con carattere eccezionale.

Un altro ente che volesse rilasciare online i contenuti di Wikipedia, come nel caso di WaNDA, per poter ricadere nella norma sulla limitazione di responsabilità dovrebbe disporre di un accordo in cui Wikimedia incarica l'ente della redistribuzione. La limitazione di responsabilità prevede che non appena il fornitore del servizio viene a conoscenza di un illecito da parte delle autorità, deve provvedere in tempi brevi alla rimozione del contenuto; per il formato digitale dell'albero si tratterebbe della rimozione del testo o dell'immagine che causa l'illecito, una facile operazione.

Ne consegue che il ruolo attivo del progetto WaNDA, ovvero chi effettua la fase di elaborazione utilizzando WaNDA-tools, per ricadere nella norma dovrebbe essere Wikimedia stessa oppure un altro ente da essa incaricata; questo vale anche per la distribuzione online del prodotto, che dovrebbe essere Wikimedia o un altro ente da essa incaricata.

6.4.5.2 Distribuzione su supporto fisico

La distribuzione del supporto fisico, come il DVD, contenente le informazioni di Wikipedia difficilmente può ricadere nel decreto sul commercio elettronico. Per la legislazione un prodotto enciclopedico digitale ricade piuttosto nel caso di pubblicazione stampata ed è meglio classificato come un prodotto editoriale. La legge nella quale si ricade è quella relativa alla stampa n. 47/1948 [135]. L'attività di stampa, nel caso vi siano contenuti lesivi dei diritti di terzi o diffamatori, prevede sia responsabilità civile che penale. Inoltre vi è la necessità di identificare la figura dell'editore, che raccoglie i contenuti enciclopedici e incarica un stampatore.

Da queste considerazioni ne consegue che per la diffusione dell'albero tramite supporto ottico è necessario che un editore si incarichi di responsabilità civile e penale dell'enciclopedia. Potrebbe essere il caso di un giornale, che decide di pubblicare in modo occasionale o sporadico il DVD, valutando i possibili rischi in relazione al profitto guadagnato.

6.4.6 Considerazioni sulle licenze e sulla distribuzione di contenuti web

Il rilascio del contenuto enciclopedico, venga esso diffuso online oppure pubblicato su supporto ottico, pone quindi molte problematiche. La diffusione di informazioni effettuata nel progetto WaNDA si svolge in un ambito senza scopo di lucro, in cui non vi è un ritorno economico tale da giustificare l'operazione stessa di diffusione. Questo fenomeno può essere considerato una "barriera economica" poichè la normativa impone un modello economico in questo caso eccessivamente restrittivo.

In primo luogo la legge sulla stampa, pensata per essere applicata ai testi in formato cartaceo, viene oggi spesso considerata come la regolamentazione sotto la quale avviene la distribuzione fisica di contenuti testuali su supporti diversi dalla carta, come è nel caso del progetto WaNDA; seppur con posizioni contrastanti in passato è anche stata a volte addirittura estesa alle pubblicazioni online.

Per quanto riguarda l'ordinamento penale, in cui si può ricadere nel caso della diffusione di contenuti diffamatori, è bene notare che l'ordinamento prevede principi quali la tassatività e sufficiente determinatezza. È quindi stato più volte fatto notare che la norma penale, riguardante beni primari dell'individuo, dovrebbe risultare certa e univocamente interpretabile, ponendo inoltre l'eventuale trasgressore nella piena coscienza della previsione del proprio comportamento come reato. L'applicazione della norma sulla stampa, che prevede "metodi meccanici o fisico-chimici", risulterebbe un'applicazione analogica ai mezzi elettronici e ottici del progetto, negando quindi l'applicazione del diritto penale.

Alcune novità legislative, quale la legge 62/2001 [136], presentano regolamentazioni in fatto di supporti ottici; infatti nell'articolo 2 viene considerato "prodotto editoriale" anche quello su "supporto informatico" se "destinato alla pubblicazione", con esclusione di quelli che contengono esclusivamente software o suoni e voci. Tuttavia tale esclusione non si applica al supporto del progetto WaNDA, che contiene prevalentemente testi ed immagini, e che quindi è da considerarsi prodotto editoriale.

L'ordinamento giuridico italiano è oggi conformato al modello classico dell'industria editoriale, che è però ormai desueto. L'attuale legislazione fornisce regolamentazioni nella distribuzione di contenuti attraverso canali telematici, con la legge sul commercio elettronico che deriva da una direttiva della comunità europea. Tale norma risponde all'esigenza di esonerare a livello europeo da qualsiasi responsabilità gli intermediari che hanno un ruolo passivo nel trasporto di informazioni; rischia però di avere un carattere marginale nell'ordinamento italiano. Infatti tale norma viene a volte interpretata come un esonero di responsabilità soltanto quando il prestatore ed il destinatario del servizio siano collegati da un rapporto diretto; secondo questa interpretazione, il progetto WaNDA richiederebbe quindi un esplicito accordo con Wikimedia, non essendo la protezione data dalla norma applicabile benché la licenza dei contenuti lo consente.

È difficile reputare sensato che un editore occasionale mosso da ragioni non di lucro,

si accolli gli oneri imposti dalla normativa per la pubblicazione (quali l'indicazione sui supporti dell'editore, dello stampatore, del luogo e della data) e dalle eventuali spese tributarie all'organo SIAE, oltre a possibili rischi derivati da sanzioni civili o penali per violazione di diritti altrui. È sicuramente gravemente lesivo dover sostenere spese legali per chi si trovi a pubblicare materiale gratuitamente senza finanziamenti, poiché la diffusione gratuita di contenuti "liberi" difficilmente rientra nell'ambito del rischio d'impresa. La legge italiana infatti non offre alcuna particolare tutela per le opere pubblicate senza scopo di lucro o sotto licenze libere. Oltre a non prevedere norme a riguardo delle licenze non esclusive, non risultano ad oggi pronunce giurisprudenziali utilizzabili come precedente.

Infine non giova l'assegnamento in via esclusiva della tutela dei diritti d'autore alla Società Italiana degli Autori ed Editori (SIAE), che conduce al rischio di subire richieste tributarie la cui resistenza, anche se legittima, potrebbe risultare eccessivamente costosa a chi non svolge il ruolo professionale di editore. A questo proposito è bene notare che recentemente [137] è stata ipotizzata una futura apertura da parte della SIAE in merito alle licenze libere, sotto forma di eccezioni alle regole in tema di pagamento dei diritti d'autore.

Capitolo 7

Note conclusive

Il progetto WaNDA si è articolato su due percorsi paralleli: lo sviluppo di una tecnologia per effettuare la conversione, WaNDA-tools, e lo studio di meccanismi per la diffusione del prodotto.

Il primo percorso, dopo aver svolto diverse soluzioni, è infine giunto a buoni risultati: il meccanismo di conversione attuale ha raggiunto gli obiettivi prefissati in fase di progettazione, i risultati del processo sono buoni, ma soprattutto il meccanismo utilizzato è mantenibile nel tempo.

Il secondo percorso, che è giunto alla definizione di due formati finali per la distribuzione, non ha ancora avuto la sua realizzazione, che consiste nella diffusione del prodotto. Ciò è dovuto a molti fattori esterni al progetto, tra i quali la legislazione italiana riguardo i canali di distribuzione descritti nell'ultimo capitolo, che non dà garanzie per quanto riguarda il supporto fisico; per quanto riguarda la distribuzione online, è stata fatta la richiesta di gestione a Wikimedia Foundation, che tuttavia per come è improntata necessita di tempi lunghi. Il progetto in questo senso ha però buone speranze, essendo il software fatto e applicabile ad altri contesti.

Dal punto di vista tecnico sono stati incontrati molti problemi con la procedura di conversione del wikitext, dovuti alla sua elasticità ed alla struttura del codice di MediaWiki. In questo senso è auspicabile che in futuro al codice di MediaWiki venga effettuato un *refactoring* in modo da semplificarne la struttura.

Inoltre è stato inizialmente un problema la mole di dati dell'enciclopedia Wikipedia, superati grazie al filtro descritto e all'utilizzo di una piattaforma dotata di un'elevata banda di accesso al disco, che ha rimpiazzato il sistema di riferimento descritto nel capitolo 5.

Il lavoro effettuato, oltre ad essere stato un modo per aiutare il progetto Wikipedia in cui credo molto, è stato utile alla definizione di soluzioni multipiattaforma appronfondendo temi quali JavaScript e codifiche di comunicazione tra pagine HTML, filesystem dei supporti e utilizzo di schemi XML. Infine il progetto è stato illuminante per la comprensione delle tematiche giuridiche del software opensource e delle licenze open content, in ambito locale ed internazionale.

Bibliografia

[1] Voce “Wiki” su Wikipedia. http://it.wikipedia.org/wiki/Wiki.

[2] Traffic analysis from Alexa.com. http://www.alexa.com/data/details/traffic_details?&range=6m&size=large&compare_sites=&y=t&url=http://it.wikipedia.org

[3] List of Wikipedias. http://meta.wikimedia.org/wiki/List_of_Wikipedias

[4] F. B. Viégas, M. Wattenberg, K. Dave, Studying Cooperation and Conflict between Authors with history flow Visualizations. CHI 2004. http://alumni.media.mit.edu/~fviegas/papers/history_flow.pdf

[5] Portland Pattern Repository. http://c2.com/ppr

[6] WikiWikiWeb. http://c2.com/cgi/wiki

[7] Voce “Bomis” su Wikipedia. http://en.wikipedia.org/wiki/Bomis

[8] Nupedia.com al 30 gennaio 2003. http://web.archive.org/web/20030130105253/http://nupedia.com/

[9] The Free Encyclopedia Project (GNUPedia). http://www.gnu.org/encyclopedia/

[10] UseModWiki. http://www.usemod.com/cgi-bin/wiki.pl

[11] Wikimedia Foundation. http://wikimediafoundation.org/wiki/Pagina_Principale

[12] Citizendium, a citizens’ compendium of everything. http://en.citizendium.org/wiki/CZ:About

[13] L. Grossman, Time’s Person of the Year: You. 13 December 2006. http://www.time.com/time/magazine/article/0,9171,1569514,00.html

[14] Exemption Doctrine Policy: criteri di ammissibilità dei file multimediali su it.wiki. http://it.wikipedia.org/wiki/Wikipedia:EDP_per_it.wiki

[15] Wikizionario. http://it.wiktionary.org/

[16] Wikibooks. http://it.wikibooks.org/

[17] Wikiquote. http://it.wikiquote.org/

[18] Wikisource. http://it.wikisource.org/

[19] Wikispecies. http://it.wikispecies.org/

[20] Wikinotizie. http://it.wikinews.org/

[21] Wikiversità. http://it.wikiversity.org/

[22] Commons.http://commons.wikimedia.org/

[23] Meta-Wiki. http://meta.wikimedia.org/

[24] Voce “Wikipedia” su Wikipedia. http://it.wikipedia.org/wiki/Wikipedia#Valutazioni

[25] Wikipedia as a press source. http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_as_a_press_source
[26] Wikipedia as a court source. http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_as_a_court_source
[27] Internet Cases: Wikipedia and the courts. 13 December 2005. http://www.internetcases.com/archives/2005/12/wikipedia_and_t_1.html
[28] Wikipedia as an academic source. http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_as_an_academic_source
[29] Wikipedia as a book source. http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_as_a_book_source
[30] MediaWiki. http://www.mediawiki.org/wiki/MediaWiki
[31] MediaWiki database layout (versione 1.10 1.9 1.8 1.7 1.6 1.5 1.4). http://www.mediawiki.org/wiki/Manual:Database_layout
[32] MediaWiki classes documentation. http://svn.wikimedia.org/doc/, http://www.mediawiki.org/wiki/Manual:Code
[33] Wikipedia servers. http://meta.wikimedia.org/wiki/Server
[34] University of Winconsin - Stout. Enciclopedias. http://www.uwstout.edu/lib/reference/encycl.htm
[35] Librarians' Internet Index. http://www.lii.org/pub/htdocs/search?action=show;search=encyclopedia;searchtype=keywords
[36] Statistiche di Wikipedia Italiano. http://stats.wikimedia.org/IT/TablesWikipediaIT.htm
[37] Wikipedia Statistics. Immagini sotto pubblico dominio. http://stats.wikimedia.org/EN/PlotsPngArticlesTotal.htm
[38] Wikipedia:Mirrors and forks. http://en.wikipedia.org/wiki/Wikipedia:Mirrors_and_forks
[39] Homepage del progetto WaNDA. http://linux.studenti.polito.it/wanda.php
[40] Release del progetto WaNDA presenti su SourceForge. http://souceforge.net/projects/wanda-tools
[41] ASSINFORM, La penetrazione di Internet e dell'e-Commerce. 2005. http://www.rapportoassinform.it/interna.asp?sez=110&ln=3
[42] Wikipedia static dumps. Pagina di dicembre 2005. http://web.archive.org/web/20051220112333/http://static.wikipedia.org/
[43] Alternative MediaWiki parsers. http://meta.wikimedia.org/wiki/Alternative_parsers
[44] Draft of Wikipedia DTD. http://meta.wikimedia.org/wiki/Wikipedia_DTD
[45] Moulin: la wikipédia offline. http://www.moulinwiki.org/l/it/
[46] Wikipedia: Wikipedia CD Selection.http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_CD_Selection
[47] SOS Children launches Wikipedia for Schools. 29 May 2007. http://www.soschildrensvillages.org.uk/charity-news/wikipedia-for-schools.htm
[48] 2007 Wikipedia Selection for schools. http://schools-wikipedia.org/
[49] Encyclopodia free software project. http://encyclopodia.sourceforge.net/en/index.html

[50] Apple IPod. http://www.apple.com/it/ipod
[51] T. Tsiodras, Building a (fast) Wikipedia offline reader. August 2007. http://www.softlab.ntua.gr/~ttsiod/buildWikipediaOffline.html
[52] Free Software Lab MediaWiki patch. http://fslab.de/svn/wpofflineclient/trunk/mediawiki_sa/
[53] TomeRaider. http://www.tomeraider.com/
[54] TomeRader 3 database format. http://en.wikipedia.org/wiki/Wikipedia:TomeRaider_database
[55] Wikipedia OnDVD. http://www.wikipediaondvd.com/
[56] LinterWeb. http://www.linterweb.fr
[57] Kiwix: Wikipedia offline reader. http://www.kiwix.org/index.php/Main_Page
[58] Digitale Bibliothek software. http://www.digitale-bibliothek.de/software
[59] EXA Media: Wikipedia, l'enciclopedia libera in DVD-Rom. http://www.exaspa.it/product_detail.asp?id=586
[60] The Unicode Standard, Version 5.0. http://www.unicode.org/standard/standard.html
[61] W3C Cascading Style Sheets. 1996, 1999, 2004.http://www.w3.org/Style/CSS/
[62] SpiderMonkey (JavaScript-C) Engine. http://www.mozilla.org/js/spidermonkey/
[63] Microsoft JScript (Windows Script Technologies). http://msdn2.microsoft.com/en-us/library/hbxc2t98.aspx
[64] KJS JavaScript (ECMAScript). http://developer.kde.org/language-bindings/js/
[65] Apple WebKit open source web browser engine. http://developer.apple.com/opensource/internet/webkit.html
[66] ECMA-262: ECMAScript Language Specification, 3rd edition. December 1999. http://www.ecma-international.org/publications/files/ecma-st/ECMA-262.pdf
[67] RFC 4329: B. Hoehrmann, Scripting Media Types. April 2006. http://www.ietf.org/rfc/rfc4329.txt
[68] Wikipedia: pagina Aiuto:Namespace. http://it.wikipedia.org/wiki/Aiuto:Namespace
[69] W3C Portable Network Graphics (PNG) Specification (Second Edition), ISO/IEC 15948:2003 (E). 10 November 2003. http://www.w3.org/TR/2003/REC-PNG-20031110/
[70] RFC 2083: T. Boutell, et. al., PNG (Portable Network Graphics) Specification Version 1.0. March 1997. http://www.ietf.org/rfc/rfc2083.txt
[71] G. Roelofs, PNG Frequently Asked Questions. August 2006. http://www.libpng.org/pub/png/pngfaq.html
[72] RFC 3629: F. Yergeau, UTF-8, a transformation format of ISO 10646. November 2003.http://www.ietf.org/rfc/rfc3629.txt
[73] ISO/IEC 10646:2003. Universal Multiple-Octet Coded Character Set (UCS). 2003. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39921

[74] RFC 3986: T. Berners-Lee, R. Fielding, L. Masinter. Uniform Resource Identifier (URI): Generic Syntax. January 2005. http://www.ietf.org/rfc/rfc3986.txt

[75] W3C Web Naming and Addressing Overview (URIs, URLs, ...). http://www.w3.org/Addressing/

[76] W3C Cascading Style Sheets level 2. 12 May 1998. http://www.w3.org/TR/REC-CSS2/

[77] W3C CSS3 module: Selectors Candidate Recommendation. 13 November 2001. http://www.w3.org/TR/2001/CR-css3-selectors-20011113/

[78] RFC 2616: R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, T. Berners-Lee. Hypertext Transfer Protocol – HTTP/1.1. June 1999. http://www.ietf.org/rfc/rfc2616.txt

[79] RFC 2068: R. Fielding, J. Gettys, J. Mogul, H. Frystyk, T. Berners-Lee, Hypertext Transfer Protocol – HTTP/1.1 (Obsoleted by 2616). January 1997. http://www.ietf.org/rfc/rfc2068.txt

[80] W3C HTML 4.01 Specification. 24 December 1999. http://www.w3.org/TR/REC-html40/

[81] Mozilla Gecko rendering engine. http://wiki.mozilla.org/Gecko:Home_Page

[82] Opera Browser. http://www.opera.com/products/desktop/

[83] Konqueror Web Browser. http://www.konqueror.org/features/browser.php

[84] Microsoft MSHTML Reference. http://msdn2.microsoft.com/en-us/library/aa741317.aspx

[85] K-Meleon Il Browser Sotto Controllo! http://kmeleon.sourceforge.net/

[86] Microsoft AutoRun: Creating an AutoRun-Enabled Application. http://msdn2.microsoft.com/en-us/library/aa969330.aspx

[87] PortableFirefox: your browser, your way... in your pocket. http://portablefirefox.mozdev.org/

[88] ISO 9660:1988. Volume and file structure of CD-ROM for information interchange. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=17505

[89] ISO/IEC DIS 9660:1999(E). Volume and file structure of CD-ROM for Information Interchange. http://www.y-adagio.com/public/standards/iso_cdromr/tocont.htm

[90] Public Patent Foundation, Microsoft FAT Patent. http://www.pubpat.org/microsoftfat.htm

[91] A. Orlowski, Microsoft's war on GPL dealt patent setback. 14 June 2004. http://www.theregister.co.uk/2004/06/14/ms_fat_patent_reexamined/

[92] Anne Broache, Microsoft's file system patent upheld. 10 January 2006. http://www.news.com/Microsofts+file+system+patent+upheld/2100-1012_3-6025447.html

[93] MediaWiki installation manual, http://www.mediawiki.org/wiki/Manual:Installation

[94] Apache HTTP server project: The Number One HTTP Server On The Internet, http://httpd.apache.org/

[95] MySQL: the world's most popular open source database, http://www.mysql.org/doc/refman/5.0/en/index.html

[96] PostgreSQL: the world's most advanced open source database, http://www.postgresql.org/

[97] Sistemi di gestione dei dati in MySQL: MyISAM e InnoDB, [http://dev.mysql.com/tech-resources/articles/storage-engine/part_1.html, http://dev.mysql.com/tech-resources/articles/storage-engine/part_2.html, http://dev.mysql.com/tech-resources/articles/storage-engine/part_3.html

[98] Restrizioni nella ricerca di tipo "fulltext" in MySQL utilizzando InnoDB, http://dev.mysql.com/doc/refman/5.0/en/fulltext-search.html, http://dev.mysql.com/doc/refman/5.0/en/fulltext-restrictions.html

[99] MediaWiki Extension: ParserFunctions. http://www.mediawiki.org/wiki/Extension:ParserFunctions_(extended)

[100] MediaWiki Extension: Cite. http://www.mediawiki.org/wiki/Extension:Cite/Cite.php/it

[101] Objective Caml. http://caml.inria.fr/ocaml/

[102] Dvipng: A DVI-to-PNG converter. http://dvipng.sf.net/

[103] Ploticus. http://ploticus.sourceforge.net/

[104] EasyTimeline perl script for Ploticus. http://svn.wikimedia.org/viewvc/mediawiki/trunk/extensions/timeline/EasyTimeline.pl

[105] ImageMagick: Convert, Edit, and Compose Images. http://www.imagemagick.org/

[106] Cdrtools - Highly portable CD/DVD/BluRay command line recording software. http://cdrecord.berlios.de/old/private/cdrecord.html

[107] Cdrkit - portable command-line CD/DVD recorder software. http://www.cdrkit.org/

[108] PHP: Hypertext Preprocessor, http://it.php.net/

[109] File di configurazione di MediaWiki, Apache e PHP utilizzati dai server Wikipedia, http://noc.wikimedia.org/conf

[110] XML Schema Part 0: Primer Second Edition. W3C Recommendation 28 October 2004, http://www.w3.org/TR/xmlschema-0/

[111] XML Schema description of MediaWiki's Special:Export system, http://www.mediawiki.org/xml/export-0.3.xsd

[112] Java JDK 5.0, http://java.sun.com/javase/downloads/index_jdk5.jsp

[113] Importing a Wikipedia database dump into MediaWiki, http://meta.wikimedia.org/wiki/Importing_a_Wikipedia_database_dump_into_MediaWiki

[114] MWDumper, http://www.mediawiki.org/wiki/MWDumper

[115] Mwimport.pl, [http://meta.wikimedia.org/wiki/Data_dumps/mwimport

[116] Confronto di prestazioni tra Java e Perl. http://www.ipd.uka.de/~prechelt/Biblio/jccpprtTR.pdf, http://kreiger.linuxgods.com/kiki/?Java+vs+Perl, http://use.perl.org/articles/02/09/16/1448246.shtml.

[117] Wikimedia meta-wiki: data dumps, http://meta.wikimedia.org/wiki/Data_dumps

[118] 7-Zip: a file archiver with a high compression ratio, http://www.7-zip.org/, http://p7zip.sourceforge.net/
[119] J. Seward, Bzip2 and libbzip2. Version 1.0.3, 15 February 2005, http://www.bzip.org/1.0.3/bzip2-manual-1.0.3.html
[120] Awka - Open Source AWK to C Conversion Tool, http://awka.sourceforge.net/index.html
[121] MAWK - Mike's AWK implementation. http://www.math.fu-berlin.de/~leitner/mawk
[122] Awka performance comparisons, http://awka.sourceforge.net/compare.html
[123] NetCat. http://www.vulnwatch.org/netcat/
[124] Elenco delle licenze Creative Commons Public Licenses (CCPL) per l'Italia, versione 2.5. Aprile 2005. http://www.creativecommons.it/Licenze/
[125] Cos'è il permesso d'autore (copyleft)? 03 August 2006. http://www.gnu.org/copyleft/copyleft.it.html
[126] GNU Free Documentation License, version 1.2. November 2002. http://www.gnu.org/copyleft/fdl.html
[127] GNU General Public License, version 2. June 1991. http://www.gnu.org/licenses/old-licenses/gpl-2.0.html
[128] The BSD License. Versione del 1998. http://opensource.org/licenses/bsd-license.php
[129] Mozilla Public License. Versione 1.1. http://www.mozilla.org/MPL/MPL-1.1.txt
[130] S. Aliprandi, copyleft & opencontent l'altra faccia del copyright. PrimaOra (Lodi), marzo 2005. http://www.copyleft-italia.it/libro/
[131] Legge n. 633 del 22 aprile 1941, aggiornato con il D.Lgs n.72 del 22 marzo 2004. Protezione del diritto d'autore e di altri diritti connessi al suo esercizio. http://www.interlex.it/Testi/l41_633.htm
[132] Direttiva 2000/31/CE. Direttiva sul commercio elettronico. 8 giugno 2000. http://www.interlex.it/testi/00_31ce.htm
[133] Decreto legislativo n. 70 del 9 aprile 2003. Attuazione della direttiva 2000/31/CE europea sul commercio elettronico. http://www.interlex.it/testi/dlg0370.htm
[134] Direttiva 91/250/CEE. Tutela giuridica dei programmi per elaboratore. 14 maggio 1991. http://www.interlex.it/testi/91_250ce.htm
[135] Legge n. 47 del 8 febbraio 1948. Disposizioni sulla stampa. http://www.interlex.it/Testi/l48_47.htm
[136] Legge n. 62 del 7 marzo 2001. Nuove norme sull'editoria e sui prodotti editoriali. http://www.interlex.it/Testi/l01_62.htm
[137] La Siae si apre al copyleft. 26 ottobre 2007. http://frontieredigitali.net/index.php/La_Siae_si_apre_al_copyleft

Il presente testo è rilasciato sotto licenza Creative Commons Public License Attribuzione - Non Commerciale 2.5. È qui riportato il riassunto della licenza; per la versione integrale si faccia riferimento alla pagina web:

`http://creativecommons.org/licenses/by-nc/2.5/it/legalcode`

È possibile contattarmi all'indirizzo email `e@richiardone.eu`