GLI ARCHIVI INFORMATICI: PROBLEMATICHE DI INTEGRITA' NEL TEMPO

Fondazione dell'Ordine degli Ingegneri della Provincia di Milano

Commissione per l'Ingegneria dell'Informazione

Ing. Gianluca Sironi

I FORMATI DI SALVATAGGIO E LE NUOVE VERSIONI

agenda

- tipologie di documenti
- formati proprietari e formati open
- nuovo formato standard per i documenti

tipologie di documenti / 1

Possiamo suddividere i documenti in tipologie:

- documenti di Office Automation
  (testi, fogli di calcolo, presentazioni, ...)
- disegni (mondo CAD)
- altri tipi di documenti (software, immagini, ...)

ci occuperemo soprattuto della prima tipologia

tipologie di documenti / 2

Le applicazioni ed i formati corrispondenti:

- Office Automation
  Microsoft Office è lo standard de facto
  file **DOC** per i testi, file **XLS** per i fogli di calcolo,
  file **PPT** per le presentazioni, ...

- Disegni
  Autocad di Autodesk è lo standard de facto
  file **DWG** per i disegni

differenze tra versioni di Microsoft Office

Pur essendo lo standard de facto bisogna tenere in considerazione che vi sono differenze sostanziali tra i formati di memorizzazione (e di conseguenza problemi di interscambio) dei documenti tra le varie versioni di Microsoft Office:

- nelle precedenti versioni di Microsoft Office (Office XP/2000/97, Office 95, ...) i documenti (file DOC, XLS, PPT, ...) sono creati in un formato proprietario binary-based
- a partire dalla versione di Microsoft Office 2003 i documenti sono creati in un formato proprietario XML-based

formato dei documenti / 1

il formato di un documento può essere **proprietario**, utilizzabile solo dalla applicazione che lo ha generato (o una sua evoluzione) o utilizzabile attraverso filtri in un'altra applicazione

il formato di un documento può essere **open**, con specifiche pubbliche, quindi utilizzabile da tutte le applicazioni che aderiscono alle specifiche

formato dei documenti / 2

esempi di formati propietari:

- il formato DWG di Autocad
- i formati DOC, XLS, PPT, ... di Microsoft Office

esempi di formati open:

- il formato PDF di Adobe (*)
- il formato HTML e XML del consorzio W3C
- i formati SX di OpenOffice 1

(*) il formato PDF è un formato realizzato per la memorizzazione e la distribuzione delle informazioni, non è un formato realizzato con lo scopo di condividere documenti modificabili

requisiti di un formato strandard

Vari enti governativi (tra cui la Commissione Europea) hanno da tempo manifestato l'esigenza di avere un **formato standard open** per l'archiviazione e lo scambio di documenti modificabili

Il formato standard per i documenti deve avere i seguenti requisiti e caratteristiche:

- essere open, XML-based, non binario, modificabile
- preservare il layout originario (fidelity)
- garantire accesso a lungo termine
- interoperabilità tra piattaforme diverse
- essere utilizzabile senza vincoli legali
- poter essere implementato senza restrizioni

documenti XML-based / 1

La scelta comune tra suite di office automation commerciali o suite open source commerciali di utilizzare un formato XML-based è dettata da molti fattori e vantaggi

Con questa scelta di formato i documenti utilizzano uno schema XML (dove viene definita la struttura logica del documento e indicato il tipo di dati contenuto)

Lo schema XML è facilmente accessibile (è un file plain text), estensibile e personalizzabile.

documenti XML-based / 2

Un documento in formato XML-based può essere visto come il passaggio tra documenti non strutturati (quali un file di testo Word) ed un database che ha una rigida schematizzazione

I vantaggi di questo approcio sono molti tra cui:

- semplicità di collaborazione sui documenti
- facilità di interscambio di informazioni
- dimensioni di memorizzazione dei documenti (aspetto determinante in documenti complessi)
- le strutture dati riducono gli errori nel riutilizzo
- le ricerche sono molto più efficaci (aspetto determinante in documenti complessi)

OASIS / 1

OASIS (Organization for the Advancement of Structured Information Standards) is a not-for-profit, international consortium that drives the development, convergence, and adoption of e-business standards"

Il consorzio OASIS ha numerosi membri in molti settori (industria, agenzie governative, finanza, università, biblioteche, tlc, ...):
Sun Microsystems, SAP, Visa, Microsoft, Adobe, Intel, IBM, Oracle, RSA, Boeing, MIT, Ford, Accenture, France Telecom, NTT, ...

OASIS / 2

Il consorzio OASIS ha prodotto la maggior parte degli standard relativi ai Web Services, oltre a standard su sicurezza, e-commerce e per il settore pubblico

Ad esempio il consorzio OASIS ha definito insieme all 'ONU lo standard ebXML per l'e-business (global framework for e-business data exchange)

Il consorzio OASIS è una delle poche organizzazioni che può sottoporre standard direttamente a ISO (International Organization for Standard)

standard OpenDocument / 1

OASIS Open Document Format for Office Applications

OpenDocument è un formato aperto e libero da royalty per la memorizzazione e lo scambio di documenti digitali modificabili (documenti di testo, fogli di calcolo, grafici e presentazioni)

Il formato OpenDocument è stato rilasciato dal consorzio OASIS; la versione standard OpenDocument 1.0 è stata approvata il 1° Maggio 2005

Il formato OpenDocument è basato su XML

standard OpenDocument / 2

Il formato standard OpenDocument è stato sviluppato sul formato XML-based originariamente creato da OpenOffice.org (OpenOffice versione 1)

Il principio con cui è stato realizzato il formato OpenDocument permette a chiunque l'elaborazione delle informazioni contenute utizzando XSLT (*) oppure attraverso qualsiasi strumento in grado di manipolare XML.

XML e XSLT sono standard realizzati dal W3C

(*) XSLT è un linguaggio di trasformazione tra documenti XML

standard OpenDocument / 3

il formato OpenDocument consente a chiunque, liberamente, di estendere il formato stesso per contenere elementi e attributi non specificati dallo schema XML originale

Per la gestione degli oggetti binari (immagini, oggetti OLE o altri tipi di media) che non sono supportati in modo nativo da XML, il formato OpenDocument utilizza un "package file" (in formato ZIP standard) dove vengono memorizzati i contenuti XML insieme agli oggetti binari associati

standard OpenDocument / 4

Diverse suite di Office Automation già supportano in modo nativo documenti nel formato standard OpenDocument **OD (*)**:

- OpenOffice (di OpenOffice.org)
- StarOffice (di Sun Microsystems)
- KOffice (di KDE.org)
- Workplace (di IBM/Lotus)

A breve sarà supportato da altre applicazioni:

- Lotus 1-2-3 (di IBM/Lotus)
- WordPerfect (di Corel)

(*) .odt text, .ods spreadsheet, .odp presentation, ...

standard OpenDocument / 5

Il formato "Microsoft Office Open XML format" di Office 2003 è un formato open e basato su XML

Presenta però alcune differenze rispetto allo standard:

- lo schema XML di Microsoft può contenere oggetti propietari che possono essere eseguiti solo in un ambiente Microsoft
- la licenza preclude modifiche o estensioni degli schemi XML; in sostanza presenta diversi vincoli che non lo rendono implementabile da altri vendor

formato OpenDocument / 1

Un (file) documento in formato OpenDocument è memorizzato un file JAR

Un file JAR (Java Archive) è un file compresso (ZIP) che contiene vari file XML (in plain text) tra cui:

- manifest.xml che contiene informazioni sull'archivio compresso JAR (in sostanza contiene le informazione del documento in formato OpenDocument)
- content.xml dove c'è il contenuto del documento
- settings.xml che contiene impostazioni ed informazioni specifiche per l'applicazione (molte impostazioni sono comuni)

confronto tra formati e applicazioni

confronto tra applicazioni e formati

format	open	non-binary	modify	fidelity	cross-platform	advanced features	wide-adoption
MS Office 2003	**Y**	**Y**	**Y**	**M**	**N**	**Y**	**Y**
MS Office XP	**N**	**N**	**Y**	**M**	**N**	**Y**	**Y**
MS Office 2000	**N**	**N**	**Y**	**M**	**N**	**Y**	**Y**
CorelWordPerfect	**N**	**N**	**Y**	**M**	**N**	**N**	**Y**
Adobe - PDF	**Y**	**Y**	**N**	**H**	**Y**	**N**	**Y**
OpenOffice 1	**Y**	**Y**	**Y**	**M**	**Y**	**Y**	**Y**
OpenOffice 2	**Y**	**Y**	**Y**	**M**	**Y**	**Y**	**-**
KDE KOffice	**Y**	**Y**	**Y**	**M**	**N**	**N**	**N**
Sun StarOffice 7	**Y**	**Y**	**Y**	**M**	**Y**	**Y**	**N**
Sun StarOffice 8	**Y**	**Y**	**Y**	**M**	**Y**	**Y**	**-**

Note:
- MS Office 2003 vincoli di licenza per implementazione
- OD formato nativo Koffice 1.4, StarOffice 8, OpenOffice 2
- OD solo lettura per OpenOffice 1.1.5

riferimenti

- consorzio OASIS:
  http://www.oasis-open.org
- W3C World Wide Web Consortium (HTML, XML, ...):
  http://www.w3.org
- ISO (International Organization for Standard)
  http://www.iso.org
- suite OpenOffice di OpenOffice.org
  http://www.openoffice.org
- OpenDocument XML Essentials
  http://books.evc-cit.info

GNU Free Documentation License

Via Stradella, 7 – 20129 Milano MI

gianluca.sironi @ gmail.com