Un archivio virtuale in XML: il progetto COVAX

di Luciana Bordoni e Fabrizio Poggi

Introduzione

Le arti e gli studi umanistici sono scienze basate principalmente sull'interpretazione di oggetti culturali quali testi, dipinti, reperti storico-etnologici, monumenti e altre opere d'arte. Tali oggetti sono spesso unici, molto preziosi, fragili, insostituibili e conservati presso musei, archivi, in aree urbane e di interesse storico. Archivi, musei e altre istituzioni culturali non si limitano semplicemente a conservare tali oggetti, ma gestiscono anche una cospicua documentazione su di essi sotto forma di raccolte fotografiche, record, analisi e studi scientifici. Sia gli oggetti che la documentazione supplementare a essi relativa sono spesso accessibili solo attraverso il contatto fisico con gli utenti. Le riproduzioni su carta di documenti di testo (ad esempio, edizioni critiche) o di immagini (facsimili, fotografie) sono estremamente costose in termini di manodopera, *know-how* e spese di stampa, e spesso non possono essere giustificate per un'utenza di ricercatori numericamente poco consistente. I formati elettronici per la documentazione di oggetti in biblioteche digitali riducono tali difficoltà di accesso. La nuova sfida, allora, è quella di fornire agli utenti strumenti aggiornati in grado di facilitare la fruizione e la ricerca del patrimonio culturale affinché anche le comunità non costituite da esperti o ricercatori possano trarne giovamento per le proprie finalità professionali sia individuali che collettive.

Dal punto di vista della tecnologia, il World Wide Web può rappresentare, al contempo, una piattaforma comunicativa standard per queste comunità e un punto di accesso per le applicazioni di biblioteche digitali basate su documenti.

Dunque, il miglioramento dei sistemi informatici per archivi, biblioteche e musei andrebbe considerato da un punto di vista globale. Occorre implementare le innovazioni necessarie a garantire l'accesso comune e combinato alle informazioni indipendentemente dalla loro ubicazione, dal tipo di trattamento (bibliografico, archivistico o museale) o dalla tipologia del documento, facilitando in tal modo l'accesso a tutte le informazioni memorizzate in tali sistemi.

Questo contributo presenta il lavoro svolto nel corso del progetto COVAX (Contemporary Culture Virtual Archives in XML, Archivi virtuali di cultura contemporanea in XML) finanziato dalla Commissione Europea nell'ambito del Programma IST (Information Society Technologies, Tecnologie della società dell'informazione) [1]. Nove partner appartenenti a settori di competenza diver-

LUCIANA BORDONI – FABRIZIO POGGI, ENEA/UDA-Advisor, via Anguillarese 301, 00060 S. Maria di Galeria (Roma), e-mail bordoni@casaccia.enea.it; poggif@casaccia.enea.it.

Bollettino AIB, ISSN 1121-1490, vol. 44 n. 1 (marzo 2004), p. 19-28.

si, e a cinque nazioni europee quali l'Austria, l'Italia, la Spagna, la Svezia ed il Regno Unito, hanno costituito il Consorzio in cui hanno espletato i seguenti ruoli:
- proprietari di contenuti: istituzioni culturali e di ricerca che detengono propri sistemi informatici e che, nel corso del progetto, hanno definito, implementato e convalidato il prototipo (Residencia de Estudiantes, Universitat Oberta de Catalunya, University of Karlskrona/Ronneby, Biblioteca Menéndez y Pelayo, London and South Eastern Library Region - sostituita dalla South Bank University -, ENEA);
- sviluppatori: aziende che hanno definito il progetto e lo sviluppo del prototipo COVAX (Software AG España, S.A, Angewandte Informationstechnik mbH, Salzburg Research mbH, South Bank University);
- due dei suddetti partner - Salzburg Research mbH e South Bank University - hanno ricoperto simultaneamente il ruolo di proprietari di contenuti e di sviluppatori;
- la Residencia de Estudiantes ha svolto anche il ruolo di coordinamento.

Obiettivo principale e approccio al progetto

A prescindere dalla loro collocazione, è un dato di fatto che musei, biblioteche e archivi ospitino un tale patrimonio documentario che attraverso la digitalizzazione non solo potrebbe essere conservato in futuro ma addirittura diventare accessibile a chiunque in qualsiasi parte del mondo attraverso una connessione a Internet. È altrettanto vero che la maggior parte di coloro che lavorano con questo materiale non dispongono delle risorse economiche o del *know-how* necessario a disseminare le proprie risorse attraverso la rete in maniera soddisfacente sia dal punto di vista della conservazione che da quello della presentazione.

Scopo del progetto COVAX [2, 4, 5, 6] era quello di analizzare e tracciare le soluzioni tecniche necessarie a fornire un accesso tramite Internet a descrizioni di documenti contenuti in raccolte di archivi, biblioteche e musei codificate in modo omogeneo e basate sull'applicazione del formato XML. XML (Extensible Markup Language) è il formato universale per documenti e dati strutturati sul Web (http:// www.w3.org/XML/). Con il formato XML è possibile riprodurre le informazioni in modo da consentire a qualsiasi altro sistema basato su XML di eseguire un gran numero di operazioni, quali ad esempio: estrarre dati; memorizzare e indicizzare le informazioni; eseguire ricerche strutturate su specifiche unità di informazione, sia all'interno di un singolo documento che in un gruppo di documenti; modificare la struttura di un documento; presentare le informazioni agli utenti.

Il formato XML è scalabile ed estendibile e può essere adattato a qualsiasi esigenza. Inoltre, esso presenta alcuni vantaggi che nessun altro tipo di approccio ha finora dimostrato di possedere: è un sistema *aperto* ed è supportato praticamente da *ogni* organizzazione di rilievo, è *semplice* e *potente*, ma facile da implementare, può essere utilizzato sul WWW.

Poiché i sistemi delle biblioteche devono consentire l'accesso e lo scambio di informazioni con altri sistemi, XML è la soluzione ideale per tali necessità. I riferimenti a banche dati bibliografiche (titoli, riassunti, riproduzioni di immagini, strumenti di ricerca ecc.) possono essere facilmente codificati in XML, veicolati attraverso il Web e visualizzati da chiunque mediante i *browser* Web di prossima generazione. Inoltre, XML non è progettato solo per il trattamento di dati di tipo testuale, ma integra anche altri tipi di informazioni ed è quindi in grado di descrivere e fare riferimento a dati multimediali, risultando di grande utilità per le

biblioteche e consentendo di eseguire numerose operazioni sulle informazioni, tra cui: immagini (ad esempio, riproduzioni); modelli 3D; presentazioni di musei in realtà virtuale; video; archivi sonori.

Utilizzando XML per offrire le proprie informazioni all'esterno, le biblioteche potranno aggiungere tutti questi tipi di informazioni alla propria offerta.

Il progetto ha dimostrato la propria fattibilità attraverso la realizzazione di un prototipo contenente un campione significativo di tutti i diversi tipi di documenti, finalizzato alla implementazione di un sistema globale per la ricerca e il recupero di informazioni. Tale prototipo si basa sull'assunto che nelle biblioteche, negli archivi e nei musei è custodita un'enorme mole di informazione che può essere resa disponibile al pubblico attraverso Internet grazie alla conversione dei record esistenti o alla creazione di nuovi record.

XML offre un ambiente omogeneo e normalizzato che consente di formalizzare e codificare documenti eterogenei in informazioni strutturate e, di conseguenza, facilita la costruzione di metodi comuni di ricerca e recupero di documenti e *database* di varie tipologie. L'utilizzo di XML consente anche l'applicazione di standard per il trattamento e la disseminazione di informazioni e documenti provenienti da archivi e musei, istituzioni che hanno un livello di sviluppo inferiore rispetto alle biblioteche che si riflette in una minore presenza in Internet di questo tipo di documenti. Altro fattore importante è la possibilità, grazie a XML, di convertire risorse elettroniche isolate già esistenti in archivi, biblioteche e musei in una rete di risorse informative distribuite che possono essere estese oltre il proprio *framework*. L'accesso alle informazioni da parte degli utenti viene così garantito a prescindere dalla loro ubicazione o dalle caratteristiche strutturali.

In generale, l'implementazione di XML nei sistemi informatici esistenti consente la standardizzazione, l'interoperabilità e l'interconnessione tra biblioteche, archivi e musei nei processi di trattamento, ricerca e recupero di qualsiasi tipo di descrizione o documento. D'altro canto, essendo un linguaggio *markup* standard con un rilevante tasso di implementazione in Internet, XML sta promuovendo la trasformazione e l'integrazione di standard specifici tipici, ad esempio, delle transazioni aziendali tramite sistemi elettronici e della gestione di servizi operativi come l'apprendimento *on-line*.

Per raggiungere gli obiettivi inizialmente fissati, il progetto COVAX si è articolato nelle seguenti fasi principali.

La prima fase del progetto è stata dedicata all'analisi dei *database* inizialmente disponibili e ha comportato:

1. lo studio dei diversi tipi di informazioni disponibili e delle strutture informative utilizzate da ciascun partner nella creazione di informazioni bibliografiche, archivistiche o museali;
2. la definizione di un gruppo di documenti (comprese copie digitalizzate e testi elettronici) rilevante ai fini del progetto e degli obiettivi prefissati;
3. l'analisi delle DTD (Document Type Definition) esistenti, il cui scopo è definire i blocchi strutturali ammessi in un documento XML, ovvero la struttura del documento attraverso un elenco di elementi consentiti, al fine di proporre una struttura comune per le informazioni di tutti i tipi di documenti, con particolare attenzione alla codificazione omogenea degli elementi e degli attributi comuni (nomi di persone, entità, nomi geografici, argomenti, date ecc.).

La seconda fase prevedeva la definizione e la progettazione del sistema, dei requisiti degli utenti e delle specifiche funzionali, con particolare attenzione alle

interfacce per la ricerca e il recupero delle informazioni al fine di garantire l'usabilità del sistema stesso a tutti i potenziali utenti COVAX (studenti, ricercatori e comuni cittadini). Le interfacce consentono di effettuare ricerche globali, ovvero in tutti i documenti e i database distribuiti, e specifiche, ossia limitate a un determinato tipo di documenti (ad esempio, solo quelli relativi ai musei) o a una determinata istituzione.

Nella terza fase del progetto si è proceduto allo sviluppo, all'implementazione e al collaudo degli strumenti destinati a convertire e/o interfacciare i sistemi esistenti dei vari proprietari di contenuti con il sistema COVAX: il numero dei componenti del sistema COVAX dipende infatti dal numero dei sistemi da integrare in esso. Ciò ha comportato la creazione di un motore di metaricerca e lo sviluppo di moduli, componenti di sistema, interfacce utenti, strumenti amministrativi e interfacce comuni in base ai requisiti degli utenti e alle specifiche funzionali individuate.

L'ultima fase è stata l'implementazione e la convalida del prototipo COVAX e la relativa verifica, convalida e valutazione di usabilità (da parte di un gruppo di utenti composto da studiosi, educatori, bibliotecari, archivisti ecc.) del software installato.

Risultati conseguiti

Obiettivo del progetto era studiare i sistemi e i *database* da integrare in COVAX. Per ciascun *database* da integrare sono state raccolte le seguenti informazioni: natura dei contenuti memorizzati, ambito geografico e istituzionale, supporto (testo, video ecc.), sistemi e ambienti informatici (hardware, software, standard supportati, espandibilità ecc.). Lo scopo di questa fase era l'individuazione di un metodo di integrazione dei dati, la proposta di uno schema di rappresentazione comune e la preparazione delle procedure di conversione o adattamento dei dati [3]. Ciò che più ha influito nel corso del progetto è stata l'esistenza di cinque diversi tipi di record MARC).

L'obiettivo del COVAX non era la realizzazione di una raccolta di documenti totalmente omogenea, quanto piuttosto la selezione di quei documenti che meglio si prestavano allo scopo del progetto, per provare la fattibilità della creazione di un sistema aperto per la ricerca e il recupero delle informazioni da qualsiasi tipo di documento.

Dalle raccolte dei partner COVAX abbiamo selezionato i record che meglio di altri mostravano una specificità e caratteristiche proprie ma, al contempo, si prestavano alla realizzazione di un gruppo di documenti che potesse dirsi significativo per l'utente che accede al sistema COVAX. Tale selezione ha incluso documenti che coprono un ampio spettro di ambiti disciplinari, tra cui letteratura, filosofia, storia, scienze tecniche (matematica, ingegneria, elettrotecnica), cibercultura, economia, giurisprudenza, pedagogia, psicologia, scienza dell'informazione, scienze navali e manifesti contemporanei di eventi culturali internazionali.

Circa l'ampiezza dei contenuti, è interessante sottolineare la sinergia creatasi tra le raccolte dei vari partner del progetto e i relativi contesti e ambiti scientifici: tra essi figurano, tra gli altri, produttori di supporti multimediali, biblioteche universitarie, istituti di ricerca, consorzi di biblioteche. Ciò ha permesso di ottenere una gamma molto ampia di documenti di diverse tipologie e livelli che ha contribuito a formare un quadro il più possibile completo dei contenuti.

L'architettura del sistema

La figura 1 rappresenta il diagramma delle funzionalità dell'applicativo, integrando due viste simultanee: hardware e software, in tre livelli principali.

COVAX è stato concepito per l'utilizzo con un *browser* Internet: l'accesso all'applicazione è possibile attraverso un *client* http, che consente al *client* di navigare nel sistema COVAX.

Il livello successivo, quello dei *server*, include almeno due tipi di *host* in grado di supportare operazioni COVAX: il *server* Web COVAX, che supporta le principali funzioni di ricerca e di report; il *server* Web dei partner, che supporta la memorizzazione e l'indicizzazione dei contenuti.

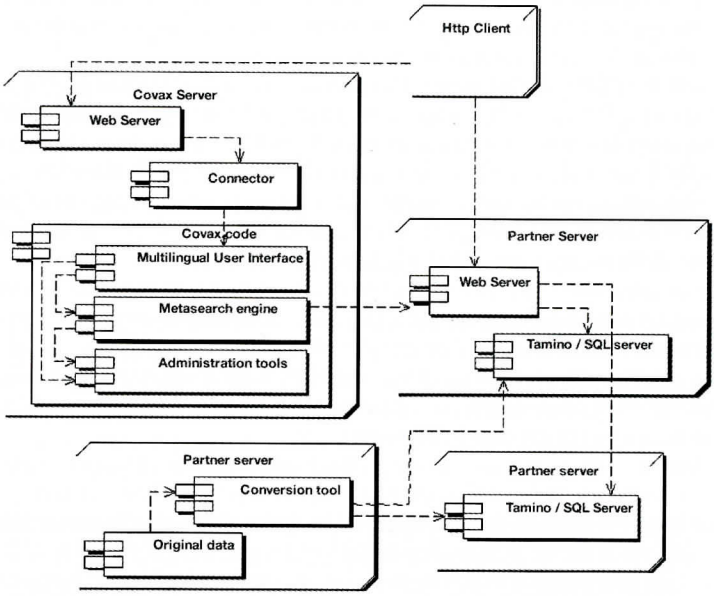

Fig. 1: Diagramma di sviluppo del sistema COVAX

Nella figura non sono riportati i dettagli dell'architettura fisica del sistema COVAX (ad esempio, i *firewall* o i *server* esterni) poiché essa non dipende (o almeno non dovrebbe dipendere) da installazioni particolari, ma deve essere adattabile agli ambienti più comunemente utilizzati. Questa filosofia ha guidato la progettazione dell'intero sistema. Ciascun sistema rilevante distribuisce i componenti del sistema COVAX e necessita soltanto della possibilità di accesso a Internet.

Il server COVAX consiste di tre componenti: un'interfaccia utente multilingue (Multilingual User Interface), un motore di metaricerca (Metasearch Engine) e strumenti amministrativi (Administration Tools). Questi tre elementi sono in grado di garantire tutte le funzioni necessarie del prodotto COVAX.

- *Interfaccia utente multilingue.* Supporta le interfacce multilingue per l'accesso alle funzionalità principali del sistema COVAX.
- *Motore di metaricerca.* Supporta le funzioni di metaricerca del sistema COVAX, le

quali consentono di gestire, in generale, l'interazione con i singoli siti di biblioteche inclusi in COVAX e, nello specifico, di effettuare ricerche globali all'interno del sistema COVAX.
– *Strumenti amministrativi*. Gli strumenti amministrativi del cuore del sistema COVAX (ovvero, il *server* di accesso) sono costituiti da una soluzione *middleware* e da componenti basati sui contenuti e orientati all'utente.

Processi di conversione

Data l'ampia diversità dei formati (alcuni dei quali proprietari) e delle strutture dei dati, il processo di conversione è stato in realtà il risultato di diversi processi di conversione condotti da ciascun partner. Questa fase prevedeva tre elementi importanti: il software e gli strumenti utilizzati per le conversioni, l'individuazione delle metodologie da applicare ai processi di conversione e le conclusioni che se ne sono potute trarre per archivi, biblioteche e musei.

Sono state utilizzate due categorie di strumenti e di software: quelli già esistenti – UseMARCON e MARCONV <http://lcweb.loc.gov/marc/marcdtd/usermanual.html> – e quelli appositamente creati dai partner COVAX per conversioni specifiche (tramite fogli di stile XSL, UltraEdit-32 v. 8.20, Borland Delphi 5, XMLSpy 3.5 ecc.). Per quanto concerne il metodo di conversione, i partner COVAX hanno definito dei processi per l'estrazione dei dati, per la conversione dei caratteri inclusi nei dati e per le specifiche delle mappature in tutte le raccolte e i formati.

In generale, date le numerose forme dei contenuti, non è stato possibile utilizzare un unico metodo di conversione dei dati. La soluzione migliore per i progetti di digitalizzazione che mirano alla condivisione di contenuti sembra al momento dipendere dalla natura del contenuto, dagli standard utilizzati, dalla disponibilità degli strumenti di conversione e dalla complessità delle mappature, delle codifiche e delle conversioni di caratteri necessarie.

Nel tempo a disposizione, è stato effettuato un numero di conversioni sufficiente per verificare la fattibilità della conversione dei dati esistenti in progetti che ne prevedono l'integrazione e l'accesso, ma non è stato possibile riutilizzare molte delle soluzioni al di fuori di un contesto locale, sebbene il progetto COVAX abbia fornito ai singoli partner esperienza e strumenti la cui utilità si estende ben oltre questo progetto.

Le conclusioni del progetto sottolineano la necessità di ulteriori approfondimenti – al di fuori dell'ambito del progetto COVAX – per convogliare le esperienze maturate, tanto diverse e preziose, in una forma che possa risultare utile per altri progetti.

Il prototipo consente l'accesso globale ai documenti memorizzati in database *distribuiti*, a prescindere dalla tipologia di documento e dalla sua ubicazione fisica; inoltre è consentito effettuare ricerche in tutte le lingue dei partner coinvolti. È in grado di funzionare con server XML e database esistenti diversi (Tamino e TextML già implementati, Dbxml già collaudato), è espandibile a nuovi server e database basati sul protocollo Z39.50 (questo protocollo è utilizzato per fornire informazioni a sistemi esterni in un standard di ampio utilizzo), consente l'implementazione di moduli esterni (ad esempio, per la connessione ai server Z39.50), di nuovi linguaggi ed è integrabile in un'ampia gamma di soluzioni per la gestione di istituzioni in possesso di materiale culturale.

Implementazione, convalida e valutazione di usabilità

Il prototipo (Fig. 2) è stato implementato in otto sedi diverse.

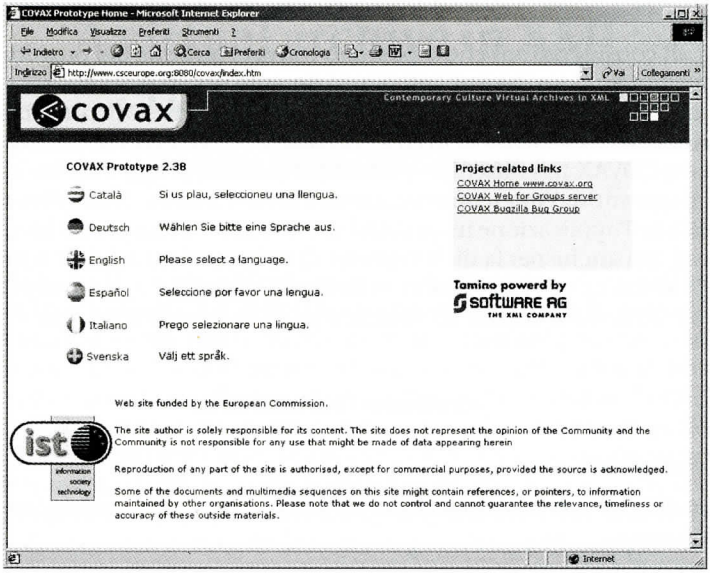

Fig. 2: Prototipo COVAX

Il Consorzio ha applicato processi diversi per valutare l'usabilità del sistema. Per quanto concerne i contenuti, la possibilità di effettuare ricerche in più ambiti disciplinari – combinando i documenti di biblioteche, archivi e musei (Fig. 3) – è stata largamente apprezzata da tutti gli utenti. Le prime impressioni sul prototipo COVAX sono state generalmente positive grazie anche all'interfaccia semplice e intuitiva che consente all'utente di connettersi facilmente al sistema.

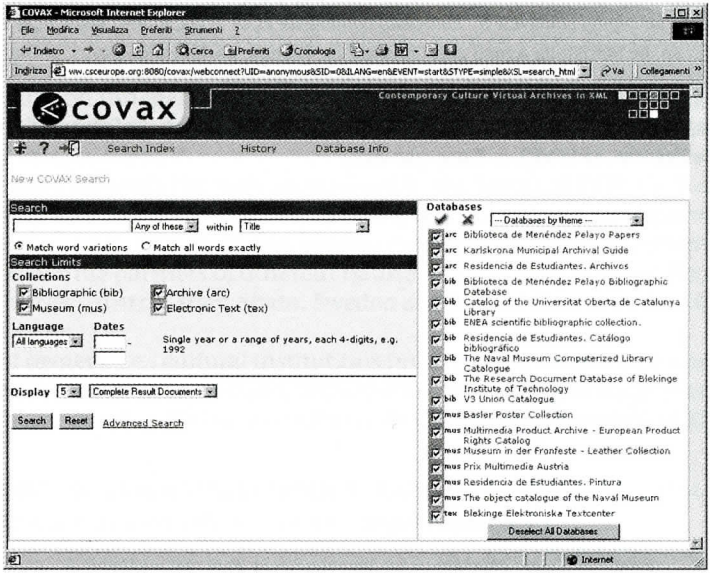

Fig. 3: COVAX: la pagina di ricerca

Conclusioni

Le principali considerazioni che si possono trarre al termine del progetto confermano il raggiungimento degli obiettivi inizialmente fissati.

Presumibilmente, l'evoluzione del software che supporta XML contribuirà a facilitare la creazione di strumenti di ricerca e di *database* contenenti documenti di testo.

Il sistema COVAX ha consentito pertanto una concreta applicazione del formato XML e ha garantito l'interconnessione tra i diversi sistemi. Ha inoltre dimostrato la validità dell'applicazione in qualsiasi ambiente, non solo per archivi, biblioteche e musei, ma anche per la distribuzione di informazioni relative a prodotti di *e-learning* e al settore turistico; inoltre ottimizza e consente l'accesso al patrimonio culturale a tutti i cittadini, uno dei maggiori punti di forza dell'Europa.

RIFERIMENTI BIBLIOGRAFICI

[1] ‹http://www.covax.org/›.

[2] Carlos Wert – Francisca Hernández. *COVAX Project*. «Cultivate interactive», 2001, n. 3, ‹http://www.cultivate-int.org/issue3/covax/›.

[3] Francisca Hernández – Peter Linde – Bob Mulrenin – Robin Yeates. *Converting heterogenous cultural catalogues and documents to XML: strategies and solutions of the Covax project*. In: *2001 in the Digital Publishing Odyssey: Proceedings of the 5th International Conference on Electronic Publishing (ELPUB 2001), Canterbury, UK, July 5-7 2001*, edited by Arved C.Hübler, Peter Linde and John W.T. Smith. Amsterdam: IOS Press, 2001, p. 65-82.

[4] Luciana Bordoni. *The COVAX Project*. In: *Atti del Workshop "Intelligenza Artificiale per i beni culturali e le biblioteche digitali", Bari, 25 settembre 2001*, dattiloscritto prodotto dal Dipartimento di Informatica dell'Università di Bari.

[5] Luciana Bordoni. *COVAX: a Contemporary Culture Virtual Archive in XML*. In: *Research and Advanced Technology for Digital Libraries: 6th European Conference ECDL 2002, Rome, September 16-18 2002: proceedings*, ed. by Maristella Agosti and Constantino Thanos. Berlin-Heidelberg: Springer, 2002, p. 661-662.

[6] Francisca Hernández [et.al.]. *XML for Libraries, Archives, and Museums: the Project COVAX*. «Applied artificial intelligence», 17 (2003), n. 8-9, p. 797-817.

A virtual archive in XML: the project COVAX

by Luciana Bordoni and Fabrizio Poggi

Arts and Humanities are sciences that are mainly based on the interpretation of cultural objects such as texts, paintings and works of arts, or historical/ethnological remains and monuments. Such objects are often unique, very valuable, fragile, irreplaceable and locally preserved in scientific collections at museums, in archives, or in urban and historic areas. Archives, museums and other cultural institutions do not simply conserve these objects. They also manage large documentations on them in the form of photo collections, expertise's, records, scientific studies and analyses. Both the objects themselves as well as the supplementary documentation are often accessible only through physical contact with users. Duplicates such as text documents (e.g., critical editions), or image documents (facsimiles, photographs) on paper are extremely expensive in terms of manpower, know-how and printing costs, and often cannot be justified for a small scientific audience. Electronic formats for object documentation in digital libraries might alleviate this access problem. The new challenge is now to provide these people with tools that are able to facilitate the fruition and investigation of the cultural heritage, so that even non-experts or communities of researchers may use up-to-date tools for both their personal work and for collaborative purposes. Technologically, the World Wide Web can serve both as a standard communication platform for such communities and as a gateway for document-centered digital library applications.

Therefore, improvement of the information systems for archives, libraries and museums should be considered from a global point of view. The necessary innovations should be introduced to guarantee the combined and common access to the information independently of its location, treatment type (bibliographical, archive or museum) or document typology, thus facilitating access to all information stored in them.

This paper presents the work carried out in the frame of COVAX (Contemporary Culture Virtual Archives in XML) financed by the European Commission in its IST (Information Society Technologies) Programme. The consortium kernel is composed of nine partners of different nature. They belong to five different European countries (Austria, Italy, Spain, Sweden and United Kingdom) and their roles in COVAX are:

- Content owners, i.e., cultural institutions holding their own information system that along the project have defined, implemented and validated the prototype (Residencia de Estudiantes, Universitat Oberta de Catalunya, University of Karlskrona

LUCIANA BORDONI – FABRIZIO POGGI, ENEA/UDA-Advisor, via Anguillarese 301, 00060 S. Maria di Galeria (Roma), e-mail bordoni@casaccia.enea.it; poggif@casaccia.enea.it.

Bollettino **AIB**, ISSN 1121-1490, vol. 44 n. 1 (March 2004), p. 19-28.

/Ronneby, Biblioteca de Menéndez Pelayo, London and South Eastern Library Region, replaced by South Bank University, ENEA).
- Developers, i.e., companies that have provided to COVAX the design and development of the prototype (Software AG España, S.A, Angewandte Informationstechnik mbH, Salzburg Research mbH, South Bank University).
- Two of the partners, Salzburg Research mbH and South Bank University, have played simultaneously the content owner and developer roles.
- Residencia de Estudiantes has also played the co-ordination role.