Dự thảo:

Chuẩn trao đổi tài liệu số hoá dựa trên Dublin Core Metadata

(Phiên bản 1.0)
dùng cho trao đổi dữ liệu trong các dự án CNTT

Cơ quan biên soạn:
Ban quản lý các dự án CNTT Thành phố HCM
Sở Khoa học và Công nghệ Thành phố HCM

Chủ trì:

TS. Hoàng Lê Minh
ThS. Nguyễn Khắc Thanh, ThS. Đào Quốc Hùng
Lê Phạm Hoàng Giàu, Võ Đức Cẩm Hải
Phạm Quốc Phương, Ngô Quang Tuấn Huy, Nguyễn Đức Tuấn

Phối hợp:

TS. **Nguyễn Chí Công**
Tổ trưởng Tổ chuyên môn, Ban Điều hành đề án 112 CP
TS. **Đỗ Văn Lộc**
Chánh Văn phòng CNTT, Bộ Khoa học và Công nghệ
ThS. **Nguyễn Long**
Tổng thư ký Hội Tin học Việt Nam
ThS. **Nguyễn Minh Hiệp**
Chủ tịch Liên hiệp thư viện các trường ĐH khu vực phía Nam

THÀNH PHỐ HỒ CHÍ MINH
2004

I. Sự cần thiết phải có chuẩn trong lưu trữ và trao đổi tài liệu số hoá

Bắt đầu từ năm 2004, thành phố Hồ Chí Minh sẽ triển khai mạnh mẽ các dự án CNTT của Chương trình mục tiêu ứng dụng và phát triển CNTT thành phố, thuộc bốn lĩnh vực lớn sau đây:

1. **Các dự án Tin học hoá quản lý hành chính nhà nước (Đề án 112)**
2. **Các dự án ứng dụng Hệ thống thông tin địa lý Tp. HCM (SagoGIS)**
3. **Các dự án ứng dụng CNTT trong các lĩnh vực khác**
4. **Các dự án đào tạo nhân lực, phát triển ngành Công nghiệp CNTT.**

Ban Quản lý các dự án CNTT (Ban QLDA CNTT) thành phố Hồ Chí Minh với nhiệm vụ tham mưu cho Sở Khoa học và Công nghệ giúp Ủy ban nhân dân thành phố Hồ Chí Minh tổ chức triển khai và quản lý toàn bộ các dự án CNTT nhìn nhận một thực tế: **để thực hiện có hiệu quả Chương trình CNTT, trách đầu tư dàn trải và thiếu hiệu quả, nhất thiết phải nhanh chóng xem xét và áp dụng chuẩn lưu trữ và trao đổi các tài liệu điện tử số hoá, tiến tới thống nhất các chuẩn trong trao đổi thông tin, dữ liệu giữa các hệ thống tin học**. Đây là một nhiệm vụ tương đối mới mẻ và khó khăn, do hiện nay có khá nhiều cách lưu trữ, trao đổi dữ liệu và thông tin đang được các công ty tin học trong nước sử dụng cho các doanh nghiệp và cơ quan chính phủ. Việc chấp nhận **hệ thống các chuẩn theo hướng mở, không phụ thuộc vào việc sử dụng các phần mềm lưu trữ và trao đổi thông tin** sẽ là nguyên tắc chủ đạo khi xem xét vấn đề định chuẩn để tránh vấn đề **phụ thuộc vào công nghệ và sản phẩm** do các nhà cung cấp đưa ra.. Xuất phát từ thực tiễn triển khai các ứng dụng CNTT và tin học hoá tại Tp. HCM, đặc biệt trong quá trình chuẩn bị đầu tư dự án "**Hệ thống thông tin – thư viện điện tử liên kết các trường đại học**", sau khi trao đổi với một số chuyên gia CNTT và thông tin – thư viện tại Hà nội và thành phố Hồ Chí Minh, Ban QLDA CNTT đề xuất xây dựng bản Dự thảo "**Chuẩn trao đổi tài liệu số hoá dựa trên Dublin Core Metadata**" để áp dụng trong các dự án CNTT của thành phố Hồ Chí minh, phục vụ việc trao đổi dữ liệu, thông tin, các tài liệu số hoá và là cơ sở nền tảng công nghệ để phục vụ tích hợp dữ liệu cho các Trung tâm tích hợp dữ liệu đang được xây dựng tại Thành phố Hồ Chí Minh: Trung tâm tích hợp dữ liệu cho các dự án 112, CityWEB, SagoGIS.

Tài liệu **Dự thảo Chuẩn lưu trữ và trao đổi** này sẽ được gửi cho một số chuyên gia CNTT, chuyên gia các ngành thông tin – thư viện, thương mại điện tử, GIS, một số cơ quan chuyên môn của trung ương và các địa phương xem xét, đóng góp ý kiến. Chúng tôi tin tưởng các kết quả triển khai trên thực tế của các chuẩn lưu trữ và trao đổi thông tin do Ban quản lý các dự án CNTT thành phố Hồ Chí Minh đề xuất trong Dự thảo sẽ là đóng góp thiết thực để các cơ quan chuyên môn và quản lý cấp trung ương: Ban chỉ đạo quốc gia về CNTT, Bộ Khoa học và Công nghệ, Bộ Bưu chính Viễn thông, Bộ Thương mại, Ủy ban Khoa học, Công nghệ và Môi trường của Quốc hội xem xét trước khi ban hành các tiêu chuẩn quốc gia.

Mọi ý kiến trao đổi xin gửi về địa chỉ info@itpmo.hochiminhcity.gov.vn

II. Chuẩn lưu trữ tài liệu số hoá (tài liệu điện tử toàn văn)

Xuất phát từ thực tiễn là hiện nay, chúng ta đang sử dụng các công cụ soạn thảo văn bản dựa trên phần mềm Microsoft Word, có khá nhiều tài liệu điện tử được tạo lập và lưu

trữ dưới khuôn dạng tài liệu **doc** của Microsoft. Tuy nhiên khuôn dạng **doc** không thích hợp cho trao đổi **văn bản hành chính** giữa các các cơ quan chính phủ, doanh nghiệp vì các lý do sau:

1. Tài liệu lưu trữ và trao đổi dưới dạng **doc** dễ dàng bị thay đổi nội dung, không có khả năng xác thực người tạo lập, người ký, con dấu đóng trên tài liệu và các thông tin khác kèm theo (bút phê của lãnh đạo, các bút tích khác)
2. Hầu hết các tài liệu - văn bản hiện hành đều không có phiên bản điện tử số hoá dạng **doc**. Việc sử dụng khuôn dạng **doc** như chuẩn trao đổi tài liệu điện tử đòi hỏi các cơ quan, doanh nghiệp phải tuân thủ quy trình soạn thảo, số hoá và lưu trữ tài liệu điện tử, hoặc bằng phương pháp nhập liệu, nhận dạng từ những tài liệu – văn bản bằng giấy. Đây là một **quy trình tin học hoá rất khó khăn và tốn kém, có thể gây nên những sự lãng phí rất lớn** cho chính các cơ quan, doanh nghiệp khi áp dụng tin học hoá.
3. Các tài liệu dạng **doc** thường chứa các thông tin ẩn, các macro, và có khả năng lây nhiễm virus rất lớn, do đó **không nên dùng để lưu trữ, trao đổi với các hệ thống khác**, trừ khi tài liệu đó đang được luân chuyển trong nội bộ một đơn vị, cơ quan để chờ xử lý, hoàn thiện và ban hành.

Với các lý do trên đây, việc chọn dạng tài liệu **doc** để lưu trữ và trao đổi là không phù hợp. Chúng tôi đề xuất chỉ sử dụng chuẩn tài liệu **PDF (Portable Document Format)** để lưu trữ và trao đổi tài liệu điện tử toàn văn giữa các hệ thống tin học với các ưu điểm như sau:

1. Tài liệu PDF có thể được hình thành từ các tài liệu **doc** một cách khá dễ dàng, giữ nguyên định dạng như tài liệu gốc. Ngoài ra các tài liệu do quét các văn bản như các hình ảnh số hoá cũng có thể lưu trữ dưới dạng PDF.
2. Tài liệu PDF **không thể thay đổi**, nhất là những văn bản, tài liệu do số hoá văn bản bằng giấy có chứa các bút tích, chữ ký, con dấu,
3. Sử dụng các tài liệu số hoá PDF, chúng ta **không cần có ngay chuẩn mã hoá tiếng Việt,** do các tài liệu có thể được số hoá từ các văn bản in trên giấy.
4. Tài liệu PDF có thể dễ dàng **đọc và in ra từ nhiều loại thiết bị**: PDA, máy tính IBM, MacIntosh, hệ điều hành Windows, Linux, UNIX, vv...

Với tiến bộ của công nghệ số hoá và lưu trữ tài liệu hiện nay, dung lượng của các tài liệu được quét vào máy và số hoá dạng PDF là khá nhỏ. Trên thế giới đã phát minh ra công nghệ tìm kiếm theo mẫu hình ảnh (image search engine) cho phép người ta có thể tìm kiếm toàn văn trong những văn bản số hoá quét vào máy tính và lưu trữ dạng PDF mà không phải dùng đến nhận dạng (xem thí dụ search inside the books tại Amazon website).

Tóm lại, chúng tôi đề nghị chọn

III. Phương thức trao đổi tài liệu số hoá

Để cho sự trao đổi các tài liệu số hoá dạng PDF được thuận tiện và dễ dàng, nên kèm theo các thông tin cơ bản về tài liệu như: tên tài liệu, tác giả, ngày ban hành, số hiệu, nguồn gốc, nơi lưu trữ, các thông tin vắn tắt về tài liệu, chú thích, v.v... Các thông

tin kèm theo này được gọi là các thông tin **metadata** về tài liệu.

Trong bộ tiêu chuẩn quốc gia của Mỹ, để mô tả các tài liệu điện tử, từ năm 2001 Chính phủ Mỹ đã chấp nhận sử dụng chuẩn mô tả thông tin metadata dự trên ngôn ngữ XML, ký hiệu chuẩn là **ANSI/NISO Z.39.85-2001.** Chuẩn này có tên gọi là **Dublin Core Metadata Element Set**.

Dublin Core Metadata Element Set gồm có **15** trường chính mô tả những thông tin quan trọng nhất, thường gặp và chung nhất trong phân loại, lưu trữ và trao đổi tài liệu điện tử. Từ các trường mô tả này, người ta có thể thêm vào các trường dẫn xuất để mở rộng tùy ý khả năng mô tả tài liệu của Dublin Core metadata.

Bản thân dữ liệu metadata có thể là một tập tin XML, có thể được lưu trữ trong một hệ quản trị CSDL, tuy nhiên để sử dụng đúng mục đích, người ta yêu cầu **tập tin chứa các thông tin metadata về tài liệu phải được kèm theo tài liệu** ngay khi bắt đầu đưa tài liệu vào lưu trữ, quản lý và trao đổi.

Sau đây là mô tả một quá trình trao đổi tài liệu điện tử toàn văn kèm theo thông tin metadata mà các hệ thống xử lý thông tin cần phải nhận biết và xử lý

- **Nhập liệu bằng tay:** hệ thống phải cho phép người dùng tạo lập và lưu trữ các thông tin metadata mô tả tài liệu bằng tay khi bắt đầu đưa tài liệu vào quản lý và lưu trữ trong hệ thống (chi tiết về các trường metadata nói ở ở phần sau)
- **Nhập liệu tự động:** hệ thống phải có khả năng tự động đọc các thông tin metadata được gửi từ bên ngoài tới hệ thống và xử lý theo cách thức giống như các thông tin này được người dùng nhập bằng tay vào hệ thống. (chi tiết về chuẩn mực trình bày thông tin metadata nói ở phần sau)
- **Xuất dữ liệu metadata:** hệ thống phải có khả năng xuất ra các dữ liệu metadata theo chuẩn mực thống nhất dùng để trao đổi với các hệ thống khác, kèm theo tài liệu điện tử toàn văn.
- **Phương thức trao đổi:** tài liệu điện tử toàn văn và các thông tin metadata kèm theo được khuyến cáo chỉ sử dụng **web service**. Tuy nhiên hệ thống phải có khả năng tiếp nhận các **tài liệu và thông tin metadata theo những cách truyền thống, trực tuyến và ngoại tuyến** khác, như trao đổi tập tin qua CD-ROM, E- mail, FTP, download từ Net, v.v....
- **Không khuyến cáo sử dụng** các mô hình **client/server,** các chuẩn trao đổi dữ liệu trên mạng phải sử dụng các phần mềm được viết riêng, các phương thức trao đổi dữ liệu trực tiếp từ CSDL như nhân bản dữ liệu (**database replication**), đồng bộ dữ liệu (**database synchronization**), các chuẩn đặc thù khác như Z.39.50, OAI harvest protocol, vv....

IV. Sử dụng Dublin Core Metadata cho mô tả văn bản hành chính

Sau đây là thí dụ sử dụng chuẩn **Dublin Core Metadata** mô tả các văn bản đã và đang được số hoá trên **hệ thống quản lý văn bản** của Ban Quản lý các dự án CNTT tại địa chỉ http://itpmo.hochiminhcity.gov.vn

Dublin Core Metadata	Trường con	Ý nghĩa sử dụng	Ví dụ
DC.Title		Tên của văn bản số hoá	Báo cáo kết qủa công tác Quý IV/2003.
DC.Creator		Tác giả (Người ký)	Hoàng Lê Minh
DC.Subject		Phân loại tiêu đề - đề mục	Công tác Ban QLDA
DC.Description		Trích yếu nội dung	
DC.Format		Định dạng tài liệu	digital
	Size	Kích thước toàn văn	245 KB
	Mime	Định dạng (doc, pdf,...)	text/pdf
DC.Publisher		Cơ quan ban hành	Ban QLDA CNTT
DC.Contributor		Người nhập văn bản	Ngô Quỳnh Linh
	Reviewer	Người sửa văn bản	Đào quốc Hùng
DC.Date		Ngày nhập văn bản	10/03/04
	Published	Ngày ban hành văn bản	03/01/04
	Updated	Ngày cập nhật văn bản	
DC.Type		Kiểu văn bản	Báo cáo
DC.Identifier		Định danh văn bản (nơi lưu trữ trên hệ thống)	1078478087975
DC.Language		Ngôn ngữ văn bản	
DC.Relation		Nơi nhận, Tài liệu kèm theo	Thường trực UBND Sở KHCN
DC.Source		Nguồn/số hiệu văn bản	24/BQLDACNTT-BC
DC.Coverage		Liên kết tài liệu toàn văn	**toanvan.pdf**
	Collection	Bộ sưu tập văn bản (theo người dùng)	Công văn đi
DC.Rights		Quyền tác giả	Văn bản đã được ký nháy, có thể ban hành
	Read	(Nhóm) có quyền đọc	everyone
	Write	(Nhóm) có quyền sửa	staff
	Delete	(Nhóm) có quyền xoá	manager
BODY		Ghi chú – bút tích -	Nhắc nhở các bộ phận viết báo cáo