文書構造化言語XMLを利用した文書管理手法の提案

A Method for Managing Documents Using XML

谷口 真也† Shinya TANIGUCHI
松下 誠† Makoto MATSUSHITA
井上 克郎†‡ Katsuro INOUE

†大阪大学大学院 基礎工学研究科
Graduate School of Engineering Science, Osaka University
‡奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

概 要

従来，紙媒体にて扱われていた文書を電子化する動きが急速に広がっている．文書を電子化することにより，文書を容易に参照することができ，検索，再利用などの二次的な利用を推進することが期待されている．しかしながら，現存する文書の多くは自然言語で記述されているために，文書の持つ情報を効率良く管理することが困難である．

そこで本研究では，構造化文書を用いて文書管理を効率化する手法を提案する．また，本手法に基づいて紙媒体で保管されている研究室内文書を対象とした文書管理システムの試作を行った．

1 まえがき

情報の保存と配布をどのように行うか，ということは永年の課題である．実際に，情報を保存するための媒体と情報を配布するための手段は共に過去の歴史においても数多く提案され，実現されてきた．現在では，この種の手段としては文書を電子化して保存，配布する方法が一般的である [1].

そのような電子化された文書の管理を効率よく行うためには，まず文書をどのように管理し，その為にはどのような情報が必要かということを知る必要がある．しかしながら，現存する文書の多くは自然言語で記述されており，その文書の持つ情報の意味を知ることやその情報を抽出することは非常に困難である．

そこで本研究では，文書管理を容易にするために構造化文書を用いる手法を提案する．また，本手法に基づいて，現在紙媒体として管理されている研究室内文書を管理するための文書管理システムの試作を行う．

まず最初に，研究室内文書を管理するために必要な構造について考察を行った．

次に，その構造をXMLを用いて定義し，研究室内文書の持つ情報を記述した．XML のタグに必要な情報を保持させることにより，文書中の情報を機械的に抽出することができるようにした．

最後に，構造化された研究室内文書の情報を利用して，研究室内文書を管理するためのシステムの試作を行った．本システムは Web を用いて実装しており，研究室内文書をデータベースに登録する機能と，キーワード等を用いた文書の検索や内容の表示を行う機能を持つ．

本システムを利用することで，研究室内文書の管理を効率良く行うことが可能となり，二次利用を行う際の手間を大幅に軽減することが期待される．

2 構造化文書を用いた文書管理

文書の管理を効率よく行うためには，その文書をどのように管理し，その為にはどのような情報が必要かということを知る必要がある．これにより,「一定期間にある人物が書いた文書の一覧が欲しい」,あるいは,「ある分類に属する文書の一覧が欲しい」といった要求に対し適切な情報を提示することが可能となる．

しかしながら，現存する文書の多くは自然言語で記述されているので，このような要求を満たすための情報は文書中の各所に散在している．このため要求を満たす際に必要な情報を得るために，文書のあちこちを参照しなければならず，場合によっては文書を人間が読んだ上でその情報を判断する必要が生じることもあり，大変な手間がかかる．

この種の手間を軽減するために，本研究では文書を構造化することによって，管理を行う．つまり，文書の情報を構造化して記述することにより，文書中に含まれる情報の属性などを明示的に示す．また，文書中からその情報を容易に抽出することが可能となる．

2.1 文書構造化言語

文書構造化言語とは文書構造を記述するための言語のことである．これらの言語では，原稿のファイルの中に，文字列のキャラクタコード以外のさまざまな属性情報（文字のタイプや組版情報，ハイパーリンク情報）などを，あらかじめ定義されたコマンドとして記述する．これらの言語の処理系は，そのファイルの中に記入されたコマンドを読みとり，そのコマンドで指定された通りに組版または表示を行う．これらの言語を利用することにより構造が明確で整合性を持つ文書を記述することが可能となる．

2.2 XML

本研究ではその文書の構造化を行うための手段として文書構造化言語XMLを利用した．XML(Extensible Markup Language)はインターネット上で文書やデータを交換したり，配布したりする際に用いられている，文書構造化言語である．XMLを用いて記述することにより人とアプリケーションがその目的にそって理解し処理できる形式でデータを表現できる．つまり，XMLはWebページ等を含めた文書データを記述する汎用的なデータ記述言語と言える [1] [2] [3] [4] [5].

XMLを用いる利点として，以下の2つが挙げられる．

まず，第一にデータ構造にあわせたタグの定義をすることが可能であるため，データ構造を自由に定義できまたその説明をすることが可能である．つまり，構造化したいデータに合わせて意味づけすることが可能となる．

第二に，XMLには将来的な拡張可能性があり，また標準化作業が進められているため，今後数多くのXML処理系が登場することが予想できる．XMLを利用してデータを記述することによりそれらの処理系を用いたデータ処理や文書データの相互交換も可能となる．また，プログラミング言語機能とXMLの記述内容間のインターフェースであるDOM(Document Object Model)の企画が制定されると文書処理に関する分散システムを利用することが可能となる [1] [2].

3 XMLを利用した文書構造の記述

この節では構造化文書を用いた管理手法について述べる．本研究では，文書はこれまでに学会や論文誌等で公表された論文や，研究室内で行われた輪講の際に取り上げた資料などを管理対象とした．これらの文書は，研究活動を行う上での重要な資産となっているが，現在，これらの情報は紙媒体で保存されているため，検索や再利用といった二次的利用を行うことが困難となっていた．

3.1 文書管理情報の定義

本研究では研究室文書を管理するために，論文，輪講という二つの区分を定義した．

前者は，論文という一つの文書に関連する文書の集合であり，論文という主体となるべき文書が実際に存在している．この区分に属する文書として，論文，論文発表時の資料，論文の参考文献等が挙げられる．

後者は，輪講に関連する文書の集合であり，主体となるべき文書そのものは存在していないが，いつ，どこで行われたかという情報を持っている．この区分に属する文書として，輪講発表時の資料，輪講に使用した文献等が挙げられる．

論文	タイトル，筆者，アブストラクト，日時， 公開場所 (論文が公開された場所)，保管場所， チーム (論文の所属する研究チーム)，ソース， 発表資料，参考文献， 版管理 (論文を作成する際に生じる版)
輪講	輪講対象 (輪講を行った対象)，輪講日時， 輪講文献，輪講資料

表 1　管理文書に必要な情報の定義

ただし，両方の区分に属する個々の文書は形態は異なるが，持っている情報はほぼ同様のものである．

表 1に各区分に必要な情報を定義したものを示す．

3.2　文書管理情報の構造

研究室内文書を管理するための最も簡単な方法は研究室内文書を一つ一つ個別に管理することであるが，これは管理する側と利用する側の両方に相当な負担がかかることが容易に想像できる．本研究では文書を管理するためのメタデータを定義する．メタデータは研究室内文書をある特定の意味を持つ集合として統括して管理を行う．このメタデータには統括する文書の種別と実際に統括する文書を示す情報が必要とされる．

従って，研究室内の文書管理情報は以下の 2 つに分類することができる．3.1 節で述べた 2 つの区分に含まれる文書がここでいう文書データにあたる．

- 文書データ - 研究室内文書に必要な情報を格納
- メタ文書データ - データを特定の意味を持つ集合に統括し管理

メタ文書データを利用することにより，研究室内文書を階層的に管理することができる．メタ文書データ，文書データを利用した研究室内文書の管理形態を図 1 に示す．図 1にあるように，最下層の文書データはプロジェクトデータによって管理され，そのプロジェクトデータはまたチームデータというメタデータによって管理される，という形式で階層的な管理をする．

本研究では，文書データを表 2に示す構造を持つデータとして定義した．表 2の構造の各要素に，論文を管理する際に必要である情報が保持される．ただし，チームに関する情報についてのみ，メタ文書データを利用することで情報として付加することが可能であるので文書データの構造には含まれない．

次に，メタ文書データを表 3に示す構造を持つ

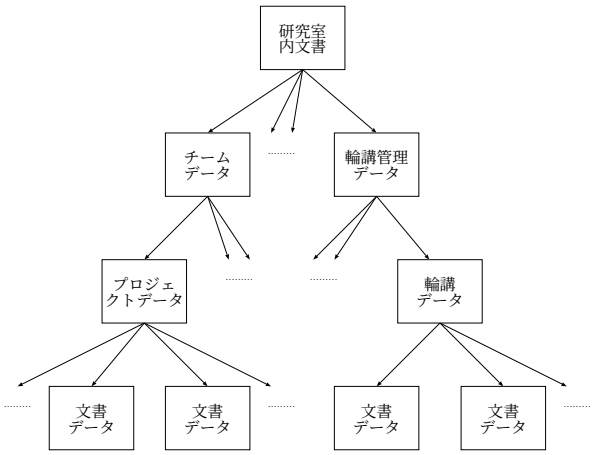

図 1　研究室内文書の管理形態

データとして定義した．表 3の構造の各要素に，メタ文書データに必要とされる，メタ文書データが統括する文書データの種別，メタ文書データが統括する文書データという情報が保持される．

3.3　XML を用いた文書管理情報の記述

XML のタグを利用することにより文書管理情報の構造の記述を行う．また，XML のタグに必要な情報を保持させることにより，文書中にある必要な情報の抽出が容易となるようにした．

項目名	説明
タイトル	文書のタイトル
筆者	文書を作成した人の名前
アブスト	文書のアブストラクト
日時	文書が公開，作成された日時
公開場所	文書が公開された場所
保管場所	実際の文書との対応
リンク	他の文書との関連 リンク先の文書の種類毎に属性 (ソース， 発表資料，参考文献，版管理) を持つ

表 2　文書データに必要な情報の定義

項目名	説明
名前	統括する研究室内文書の種別を識別
リンク	他文書へのリンク リンク先の文書の種類毎に属性 (メタデータ，文書データ，輪講文献， 輪講資料) を持つ

表 3　メタ文書データに必要な情報の定義

まず，文書データの構造を記述するために，タイトル，筆者，アブストラクト，日時，公開場所，保管場所，他文書との関連に関係する情報に対してそれぞれ，title，author，abstract，date，confer，keep，link-docというタグ付けを行った．さらにlink-docには属性情報として，href(リンク先のファイル名)，kind(リンク先の文書の種別)を定義した．

次に，メタ文書データの構造を記述するために，名前，リンクそれぞれにname，link-metaというタグ付けを行った．さらにlink-metaには属性情報として，href(リンク先のファイル名)，kind(リンク先の文書の種別)を定義した．また，メタ文書データのルートタグであるmetadataにも，type(メタ文書データの種別)という属性情報を定義した．

4 研究室内文書管理システム

XMLを用いて構造化した文書データを使って研究室内の文書の管理を行うためのシステムの試作を行った．本システムは研究室内文書の登録及び検索機能を有している．

登録機能では，研究室内で作成された論文，論文の所属するチーム，及び，プロジェクト，輪講に関連する情報の登録を行うことができる．

検索機能では，論文と輪講の情報を検索することができる．論文の検索ではタイトル，筆者等のデータをキーにして検索を行うことができ，その論文をダウンロードして利用することも可能である．

4.1 システム構成

本システムは図2に示すように，研究室内文書に関連するデータを保持しているデータベースとそのデータベースへ研究室内文書の登録をおこなうプログラム，データベースから検索を行うプログラムの3つの部分で構成されている．

このシステムはWeb上に実装した．研究室内文書のデータベースはWebサーバ上に構築されており，ユーザはそのデータベースに端末上のWebブラウザからCGIとして実装された登録・検索プログラムを起動してアクセスする．

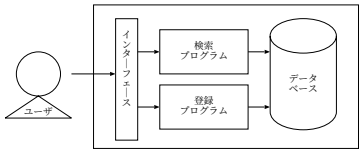

図2　システムの構成図

5 あとがき

研究室内文書から文書管理に必要な情報を抽出し，その構造をXMLを用いて記述した．その結果，文書を管理するために必要な情報を得ることができた．また，これに基づいて研究室内文書を管理するためのシステムの試作を行った．本システムは研究室内文書の登録と検索を行うことができる．本システムを使用することで，研究室内文書に関連する情報を容易に得ることができるので研究室内文書の二次利用にかかる手間がある程度軽減できると考えられる．

今後の課題として，利用者がより簡便に研究室内文書の管理をおこなうことができる使いやすいインターフェースにすることが挙げられる．

今後は本手法をソフトウェアプロセスの評価を行う際に利用される品質評価規格文書へ応用することを考えている [6].

参考文献

[1] XML/SGMLサロン: “標準XML完全解説”, 技術評論社, (1998).
[2] 池田 実: “特集『電子出版・電子新聞』 3.XMLの概要と応用”, 情報処理学会誌 Vol39 No.6, (1998).
[3] 村田 真: “XML入門〜HTMLの限界を打ち破るインターネットの新技術〜”, 日本経済新聞社, (1998).
[4] “Extensible Markup Language(XML)”, http://www.w3c.org/XML/
[5] “SGML/XML Cafe”, http://www.fxis.co.jp/DMS/sgml/index.html
[6] 松下 誠, 飯田 元, 井上克郎: “品質評価規格文書のモデル化とそれに基づく評価支援システム” 電子情報通信学会論文誌 D-I, Vol.J81-D-I, No.8, pp.986-993(1998).