正倉院文書研究資料のXML/XSLTによる記述と統合

後藤 真，柴山 守
大阪市立大学 大学院文学研究科
大阪市立大学 学術情報総合センター

正倉院文書は、東大寺の正倉院に伝来した８世紀の文書群の総称である。本文書は、背面再利用と１９世紀初頭からの「整理」作業のため、その奈良時代の帳簿の形態が著しく損なわれ、論理構造と物理構造の差異という特徴をもつ。この研究は、主に (1)写真版マイクロフィルム、(2)『大日本古文書』、(3)『正倉院文書目録』を相互に参照しながら分析が進められる。本研究は、この帳簿形態の復原過程や関連史料の各々の実体を XML/XSLT (eXtensible Markup Language / eXtensible Stylesheet Language Transformations)で記述することを目指す。すなわち、各史料の実体、及び史料の統合化の過程において、知識構造化ルールを XML/XSLT を用いて記述することである。知識構造化ルールに基づく処理例外から新たな事象の発見が可能となり、新たな知見を得る機会になる。統合化された XML 文書は、論理構造を復原する XSL(eXtensible Stylesheet Language)に基づき、Web ブラウザ上に表示され、また Word マクロ機能により物理構造が反映された「短冊」として復原できる。

本論文では、すでに提案した正倉院文書復原過程への XML/XSLT 導入を拡張し、実用化する「正倉院文書研究支援システム」を新たに提案する。また、古代史料構造に即した形態で XML 記述を提案することによって、歴史学・史料学とのコラボレーションをいっそう深化させることが目的である。

Description and Integration of “Syoso-in Monjo” Research Materials using the XML/XSLT

Makoto GOTO†, Mamoru SHIBAYAMA‡
Graduate School of Literature, Osaka City University†
Media Center, Osaka City University‡

The "Syoso-in Monjyo" is a generic name of a set of documents that were handed down in the Syoso-in of Todaiji Temple from the 8th century. The "Syoso-in Monjyo" exhibits a unique dissimilarity between their logical structure and physical layout. In order to analyze the historical facts, the researchers consult and identify information scattered mainly in 3 materials; the microfilm images which show the physical layout, “Dai-nippon Komonjyo” which has the document body and structures the logical form, and the catalog which describes pasting and sticking between the fragmentary documents.

This study attempts to apply the XML/XSLT (eXtensible Markup Language /eXtensible Stylesheet Language Transformations) for integrating these materials and develops techniques that could contribute towards a better understanding and preservation of similar historical documents.

The authors have described the integration rules for the materials using the XSL/XSLT and also examined procedures for formulating a rule-based logic. Besides each material had written by XML expressions, the integration has been achieved using the XSLT programming which describes the rule-based logic. Based on the XML expression as a result of being integrated, the user can browse an arbitrary and original logical or physical form of the document. The rule-based logic using XML/XSLT provides a model for researching the “Syoso-in Monjyo” study.

1．はじめに

最近、デジタル文書の生成・流通・交換・提供においてXML(eXtensible Markup Language)技術が注目されている[1]。XMLは、文書の実体と共にその階層構造が記述できる。また、タグで囲まれた要素には属性（XML Schema）が定義できる。

歴史学研究におけるXMLの適用としては、安澤[2]がSGML-XMLの応用としてのEAD規格について、また五島[3]がXMLを利用した史料記述の可能性を論じている。筆者らは以前、正倉院文書の比較的単純な例を用いて、構造の異なる複数史料を統合化した中間表現としてのXML文書の記述とその利用について検討した[4]。また、史料の実体と共に研究プロセスそのものにXML/XSLTを導入した事例も同様に検討した。本報告では、その事例をもとに、すべての正倉院文書にかかわる情報のデータベースの構築、および、その情報をもととしたXML化を目指す。さらにそれをくみこんだ総合的な研究支援システムである「正倉院文書研究支援システム」を提示し、歴史学とのコラボレーションをはかる。また、東京大学史料編纂所が「奈良時代フルテキストデータベース」として、正倉院文書をデジタルデータ化しているので、編纂所との共同研究も視野に入れている。

2．正倉院文書

正倉院文書とは、東大寺の正倉院に伝来した文書群の総称である[5]。詳細は以前報告したとおりである。正倉院文書研究はよりいっそう深化し、さまざまな情報をもとに研究が進められている[6]。正倉院文書はその歴史的特性により、さまざまな構造と要素が複雑に絡み合い、散在するようになっている。その歴史過程を記したものが図1である。正倉院文書は論理構造と物理構造が錯綜し、また、その研究の材料となる史料が散在していることも特徴である。『大日本古文書』1巻～25巻、『正倉院古文書影印集成』、『正倉院文書目録』『正倉院古文書目録』、マイクロフィルム、『正倉院文書拾遺』、『正倉院宝物銘文集成』と多岐にわたるこれらの情報を、随時参照しながら、正倉院文書に関わる研究は行われる。この作業は非常に煩雑なものである。これらの情報を統合し、必要に応じて抽出するツールが完成する事は、正倉院文書研究の支援になるのみでなく、一般的な歴史資料の情報統合を行う一つの方法となる可能性をみる。

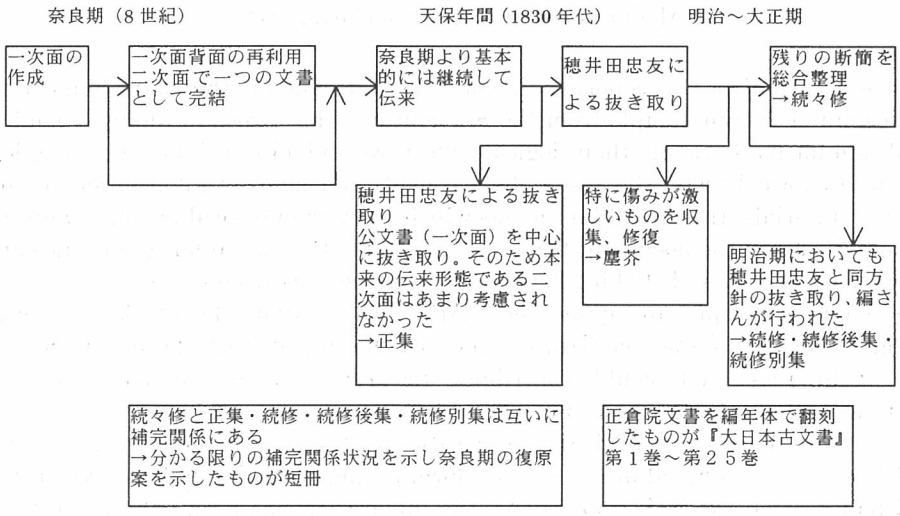

図1　正倉院文書の歴史過程

正倉院文書の構造を考える上で、日本古代史料の一般構造を考える必要がある。古代史料論に関しては石上[7][8]の研究がある。古代史料論には史料体論、歴史情報伝達行動論、歴史情報テキスト構造論、歴史的情報過程論、歴史情報群がある。史料体論では史料をメッセージ（文字・文字列や図像など）・付加メッセージ（史料体が形成されたのち一定時間を経過してから付加されたメッセージ）・搬送体（メッセージを積載・定着させている“素材”を積載・定着の仕方も包括して把握する概念）、様態（メッセージ・付加メッセージが搬送体に積載・定着され、定位されている状態）と４つの要素として定義する。歴史情報伝達行動論は史料体をそれ自体として完結したものとしてとらえず、時系列的に展開される歴史的時空間における情報伝達行動の多元的・多重的な動的過程の中に位置づけるものである。歴史情報テキスト構造論は、史料上のテキストを生成過程状態と終了状態、特殊構造と一般構造、表層構造と規定構造などに定義する。歴史的情報過程論はいかにして歴史情報が定着し、伝存されていくかの過程を問題とする。歴史情報群は歴史の時間順行過程において再構成される様態を論じる。たとえば正倉院文書は、「正倉院に収納された」という歴史的過程によって、写経所にあった文書が再構成されたものである。

この古代史料論にもとづいて、正倉院文書を定義し、XML のタグ付けを定義すると図２のようになる。まずは、テキストとメタデータ（搬送体・歴史情報群情報・歴史情報過程情報）に分類する。次にテキストを表層構造を中心に定義する。そこに環境要素と付加メッセージの情報を加える。そして、メタデータの部分に料紙体の情報として料紙の貼継や、切断などの情報を表紙や汚れ、軸などをタグ付けする。さらに歴史情報群としての情報を付加することによって、正倉院文書 XML は定義される。

文書の属性　〈正倉院文書〉
　所属　〈所属〉〈成巻〉
　　帙　〈所属〉
　　巻　〈所属〉
　　紙　〈紙〉
　　料紙情報
　　料紙自体の情報
　　紙質・大きさ・表裏
　　巻子・料紙の調製
　　転用料紙
　　色移り、墨移り
　　汚れ、しみ
　　料紙面構成の情報　〈切断〉〈接続〉
　　料紙の貼継
　　料紙調製当初か、文面改鼠による切断・貼継か
　　切断の時期
　　写経機関による貼継か、文書整理の際の貼継か
　　界線
　　墨界・押界・折界（折れ皺との区別）
　　界幅（左端・右端の幅、平均界幅）
　　横界線の位置・本数（端点の天辺からの位置、界の高さ）
　　表紙〈浜名郡輸租帳、御野国戸籍、写経目録に例〉〈表紙〉
　　軸　〈軸〉
　　軸木・題籤軸附のための紙端の加工（料紙切除、料紙畳み込み）、糊跡
　　右軸か左軸か
　　軸附の時期（文書作成当初、文書作成後）

継目裏書・継目裏印、継目封(「封」、人名文字、記号(○など)) (〈紙〉と〈紙〉との間に示されるテキスト情報として表記)
丁附け
印章
印影
捺印の方式
テキスト
一次史料(正税帳)情報
紙面の割り附け
書式
文字定位 〈line〉
裏の使用〈a〉
表層構造
首部・郡部・尾部-郡配列順序 〈line_s〉〈line_e〉
前期繰越・当期収入支出等・次期繰越一虫喰算による欠失部復原
文面の改変
擦消・重書、抹消 〈del〉〈ins〉
付加
未修古文書の巻帙を表示する附箋〈付箋〉
基底構造
国府・郡家における財政事務処理と民部省の勘会の財政システム
地方財政の諸要素・諸制度
二次史料情報(紙背を写経機関文書として使用)のテキスト構造
表層構造
首部・郡部・尾部-郡配列順序 〈line_s〉〈line_e〉
文書・帳簿の構造
基底構造
写経・造寺・造仏事業の事務処理システム
環境要素
廃棄の経緯と状況
二次利用面の順序値一二次利用用料紙(断簡)の産出過程
二次利用面の利用空間・利用行為属性
文書群・帳簿群の再構成

図2 古代史料論に基づく正倉院文書の構造とタグ付けの定義
(石上[8]pp11-12の図をもとにして作成
ゴシックが史料の属性を、明朝が該当するタグをしめすものとして表記した)

3. XML/XSLTによる記述

紙を単位とする構造を物理構造と定義する。この物理構造に関する情報は、写真(以下、D1という)、『正倉院文書目録』(D3)から得る。また、料紙の接続に関わらず、文書としての形態を整えた構造を論理構造と定義する。『大日本古文書』(D2)より論理構造を情報を得る。

この論理構造の表示と物理構造としての表記を同時に実現させ、支援ツールとして実現したシステムの概要は、図3のとおりである。

まず、正倉院文書研究者は、図2の史料構造のルールにおいて入力を行って、正倉院文書データベースを作成する。このデータベースであるD1、D2、D3は、各々RDB、XML、XML/RDB形式である。D2に関しては東京大学史料編纂所が現在入力を行っており、2002年3月から順次公開されており、そのデータを援用させていただくこととな

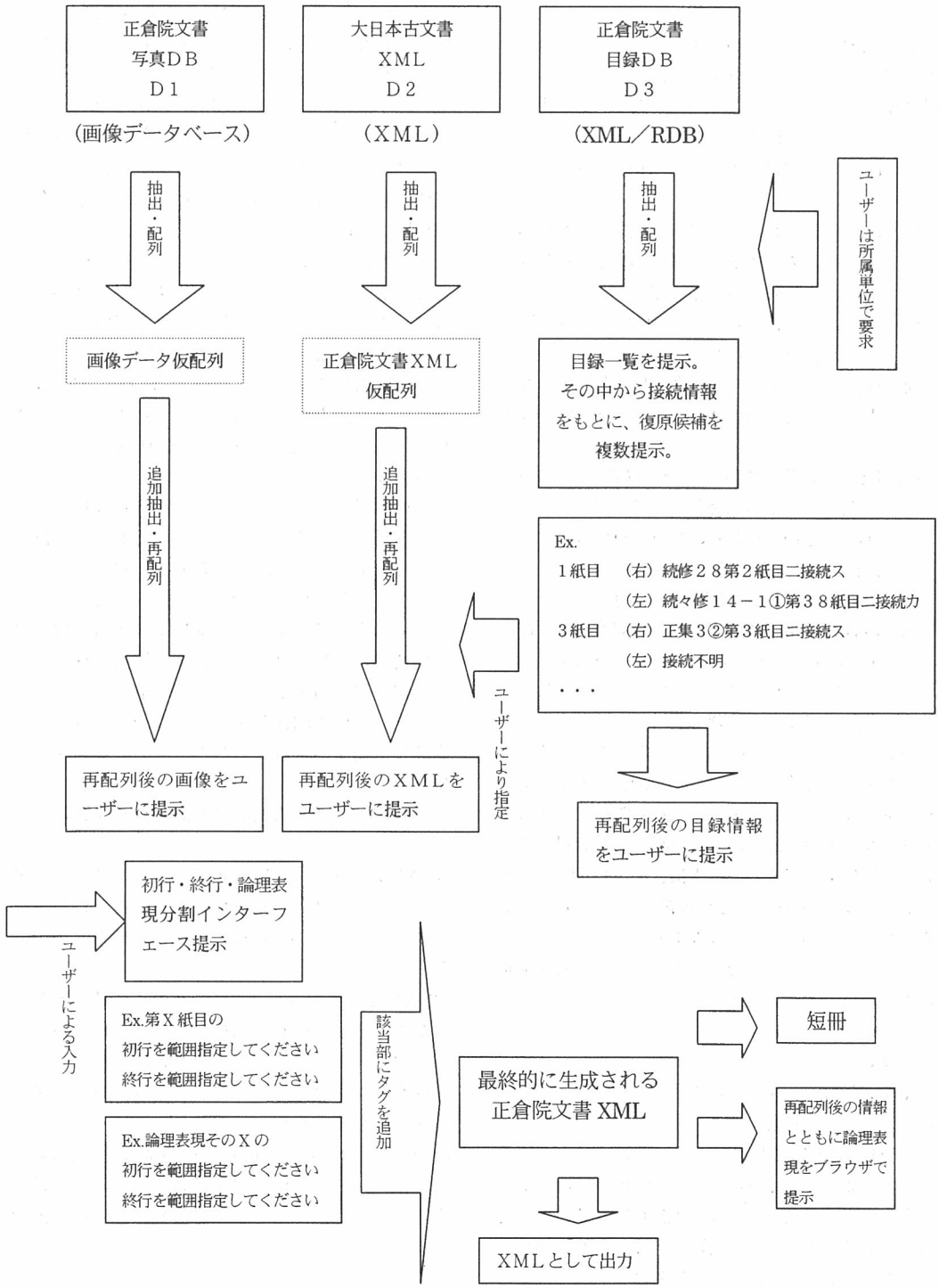

図３　正倉院文書研究支援システム概念図

った。

この変換の後、各 XML 形式から XML 化のための「構造」を持たせる変換が必要になる。これを構造化ルールと呼ぶ。この構造化ルールに基づくシステムの概要は図 X のとおりである。

(1) D1 から紙番号、所属、先頭行、最終行を読み込む。
(2) D2 から、D1 に対応する史料名、大日本古文書巻、大日本古文書頁、論理文書を読み込み、物理構造の<紙>を構造化する。
(3) D3 から D1 に対応する史料名、料紙情報、接続情報を読み込む。
(4) D1 の所属、D2 の史料名、大日本古文書巻、大日本古文書頁、D3 の史料名整合性を検査する。
(5) D2 の論理文書から、物理構造である<紙>の範囲の中での論理構造を定義する。
(6) D1 の先頭行と D2 の行 1、D1 の最終行と D2 の行 n の整合性を検査する。
(7) 料紙情報、及び接続情報を付加する。
(8) 検査の結果、整合性がなければ、XSLT の message 機能により、エラー表示するか、又は応答を要求する。D1 のすべてのレコードが終了するまで繰り返す。

正倉院文書データベースからは、この構造化ルールをもとに中間表現である XML 文書を生成する。構造化ルールは XSLT で記述する。図 4 は、本構造化ルールに基づいて生成された中間表現である XML 文書である。本 XML 文書において、基本的には紙を基準にした物理的なタグにおいて構成されるが、その一方で論理 ID を用いることによって、物理構造である紙単位を超えた、論理構造の表現を行うこととした。

4．考察

XML/XSLT を用いて記述した結果、以下に示すような有効性を確認した。

(1) XML/XSLT による記述は、他のプログラミング言語においても同様に記述できるが煩雑になるなどの問題を含む。しかし、XML/XSLT の利用では、物理と論理の異なる構造を同一の階層的な空間に写像し、物理/論理構造を必要に応じて抽出した事例に見られるように、階層構造を基本にした文書全体、或いは部分における要素の探索・判断・抽出・複写・排列などが可能であり、複数史料の統合化や部分抽出に容易に対処できる。

(2) 史料を記述した実体である XML 文書は、データの変換や表示のための制御を行う XSL/XSLT からは完全に独立性を保持しており、XSL/XSLT の記述に基づいて、実体とは関わりなく、適切な表現形態に変換できる。

(3) 実現した支援システムでは、構造化ルールに従って、従来とは異なる史料間の関係や構造が含まれている場合の新たな事象の発見が可能になり、これは新たな知見を得る機会にもなりうる。

(4) XML/XSLT 表現は、Web ブラウザ（IE5.5 以上）に組み込まれた XML パーサにより、実行される。したがって、一般的には XML/XSLT の実行に特別な処理系を用意する必要はない。

(5) RDB 形式データベースからの構造変換の手続きは容易に実現でき、史料入力の段階では、研究の内容や史料の形態によって XML や RDB 形式データベースとの使い分けが自由に行える。

以上のように、実現した支援システムでは、階層構造を持つ複数文書からの部分構造の抽出や併合、他のタグ構造を持つ文書への構造変換などにおいて、XML/XSLT は他のプログラミング言語に比較して効率的に行えるなど、その有効性を確認した。

一方、本システムにおける問題点や課題は、つぎのとおりである。

(1) XML 文書が well-formed 形式のために、文書構造やデータ型の定義に曖昧性がある。これを解決するためには、XML Schema の採用が必要である。

(2) XML 記述では、タグの整合性を保持することが基本である。それには相当の労力や注意力が要求される。従って、これらの煩雑性をなくし、利便性を向上させるようなユーザインターフェイスの開発が必要である。

(3)前述の(3)に示した message 機能などにおいて、会話型機能の向上やタグを意識することなく構造化ルールが記述できるツール開発は、今後の課題である。

```
〈正倉院文書〉
〈成巻〉〈成巻番号〉続々修 18-6〈/成巻番号〉
〈成巻本文〉
〈成巻題目〉御願奉写等雑文案〈/成巻題目〉
〈紙〉
〈大日本古文書本文〉〈大日本古文書巻頁〉14-365〈/大日本古文書巻頁〉
〈本文〉〈論理文書 id="0001-1"〉
〈line_s〉可奉写一切経〈del〉律并〈/del〉〈ins〉一部経律并〈/ins〉三千四百卅四〈ins〉三〈/ins〉〈a〉巻〈/a〉〈wari〉之中卅九
〈/wari〉〈ins〉八〈/ins〉〈wari〉巻今所〈/wari〉〈/line_s〉〈br /〉
〈line〉合可用紙六万七千〈del〉八百四〈/del〉〈ins〉九百三十五〈/ins〉張〈/line〉〈br /〉
〈line〉　六万〈del〉三千九百五十一張〈/del〉〈ins〉〈del〉四千八十二張〈/del〉〈ins〉四千六百六十六張〈/ins〉〈/ins〉経紙
〈/line〉
〈line〉　　五万九千〈del〉七百卅九〈/del〉〈ins〉九百〈gt set="mojikyo" name="002712" /〉九〈/ins〉張見写料〈/line〉
〈/論理文書〉〈/本文〉
（中略）
〈/紙〉
（中略）
〈/成巻本文〉
〈/成巻〉
〈/正倉院文書〉
```

図４　正倉院文書 XML の中間表現

図１　段階的整理

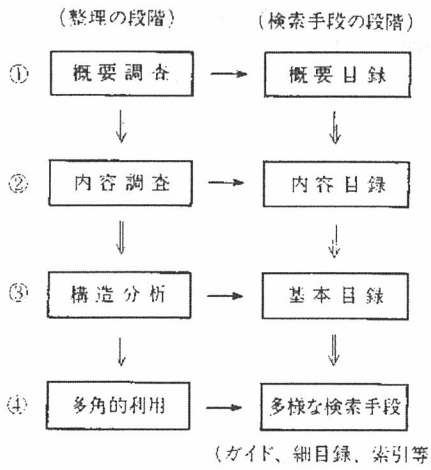

図２　記録史料調査の作業手順

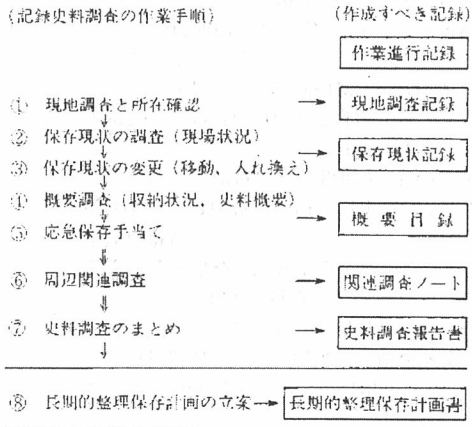

図５　史料調査の段階と記録

（安藤[9]pp112（左図）,114（右図）より引用）

５．正倉院文書 XML の汎用化

正倉院文書研究支援システムは、正倉院文書に即した形式でつくられている。しかし、その一方で同システムは、一般の歴史資料へと汎用化できる可能性をも持っている。それは、2-2 でも示したとおり、古代史料の一般的な理論に基づいているという点と、一般的な史料の記録形式をも踏まえた形式であることを、その根拠としている。安藤[9]は図５のように記録史料の「段階的整理」の手法を提案している。歴史資料に関わる情報は、目録が３種類、画像データが１種類と、本文情報

が出来上がり、さらにそれに「構造分析の結果」が加わるはずである。これらのデータはそれぞれが、正倉院文書研究支援システムの D1（画像データ）、D2（本文情報）、D3（目録情報）の情報と合致することとなる。これに図２の表をもとに XML を定義することによって、本システムは歴史資料一般の形式にも適用可能ではないかと考える。ただし、タグ付けがあまりに煩雑となり、作業が膨大なものになる可能性があることには留意しておきたい。

６．　結論

正倉院文書は、奈良時代の帳簿の形態が著しく損なわれ、非常に複雑な構造を持つ。また、関係情報が分散されている。このため、この諸情報を統合しながら研究をする必要がある。史料の XML 化による記述は勿論のこと、関連史料の統合化と構造化を行う構造化ルールの記述においても XML/XSLT を用いた。

この結果から、構造化ルールで生成された中間表現である XML 文書から、XSL/XSLT を用いて物理構造や論理構造が抽出でき、構成できる。この構造化ルールの記述は、正倉院文書研究者が、正倉院文書を研究する上で行っている関係する諸情報のなかから、必要なものを抽出し、統合するという過程を一般化、あるいはモデル化することになる。たとえば、帳簿作成者ごとに微妙に異なると言われる、帳簿作成の方法もより明らかになるであろうし、文書作成の時期による差も明白となるだろう。また、ルール自体の理論化は、歴史研究者が明確に意識せずに行っている作業を理論化することでもあり、「古代帳簿論」の一方法として寄与できるのではないか。また、XML 記述を正倉院文書の史料体構造にもとづいて行っているが、このことは逆に言えば XML による構造化により、現在の古代史料論に提言ができるようになる。

歴史学研究にかかわって資料統合を行った研究というのはこれまであまり多くなされてこなかった。この事例が歴史資料の統合的な把握のひとつの例となれば幸いである。

最後に本論文を作成するにあたって、大阪市立大学大学院文学研究科の栄原永遠男教授、日本史研究室、写経所文書研究会の方々に多くのご教示をいただいた。ここに末文ながら謝意をあらわす次第である。

〔参考文献〕

[1] World Wide Web Consortium http://www.w3c.org/　参照 2001.08.20
[2] 安澤秀一:エンコーデッドアーカイバルデスクリプション EAD：SGML-XML の応用形として、情報処理学会研究報告、2001-CH-51, Vol.2001, No.67, pp.17-24, 2001
[3] 五島敏芳: XML を利用した史料記述の可能性―「国際標準：記録史料記述の一般原則」ISAD(G) 第２版とデータベースをめぐって―, 情報知識学会人文・社会系部会、第１６回「歴史研究と電算機利用ワークショップ」資料、2001
[4] 後藤　真、高山 典史、柴山　守：正倉院文書の XML による構造化と復原の検討、情報処理学会研究報告、2001-CH-51, Vol.2001, No.67, pp.31-38, 2001
[5] 栄原永遠男:『奈良時代の写経と内裏』、2000、塙書房、など
[6] 栄原永遠男：正倉院文書関係文献目録、『正倉院文書研究』, １，２，３号, 1993
[7] 石上英一:『日本古代史料学』、東京大学出版会、1997
[8] 石上英一:「日本学データベースと史料編纂所データベース」(前近代日本の史料遺産プロジェクト　Japan Memory Project　第３回国際研究集会「日本学研究と史料学の国際化」2002 年 6 月）
[9] 安藤正人『記録史料学と現代』、1998、吉川弘文館