

EADによる情報検索システムの構築

京都大学研究資源アーカイブ

五島 敏芳

京都大学研究資源アーカイブの五島です。12月から京都大学へ移りまして総合博物館に横付けされている所に所属をしていて、研究資源アーカイブを担当する事になっております、8年間ほどお世話になるかと思います。そのおかげで実はこのプロジェクトとKEK史料室さんにも今日の話をする内容ではご迷惑をおかけしてしまって、大変申し訳ありませんでした。国文研のデータベースの不具合にすぐ私が国文研に居た時は管理者権限持っていたのですが、今、その管理者権限行使することが出来ずにいるので、後でこっそり行使してしまうんですけれども、始終不具合が起きてしまっているので大変申し訳ありません。ただそれでも尚且つ、こちらで報告させて頂けることに感謝申し上げます。

タイトルがプログラムの方で「EAD-情報検索システム構築」となっていますけれども、ちょっと改めさせていただいて「EADによる情報検索システムの構築」とさせていただきました。同じ物です。

最初のこういう類の話というのは、今日縷々実は他の先生方がお話をされておりておりますので、具体的な事を少し紹介して私の報告に変えたいと思います。

EADという呪文の説明の前にアーカイブズの資料というものは、どうやって見つけてどうやって使える様にするかという情報検索システムの事について少しだけお話しをしておきたいと思います。

松岡先生が、図書館の世界でのお話をしたり、PORTA（国立国会図書館デジタルアーカイブポータル）のお話など画面で紹介されていたと思うのですが、そこで出てくるのは上澄みなんですね、Googleよりは性能がいいかも知れないけれど、全般的にある限られた領域のものを探し出してくるというものだと五島は理解しております。有象無象を集めるというよりは見つけやすくする、その基盤というのは実はそれぞれのコミュニティで持っている情報がきちんと探し出せるようになっている事が大事です。

アーカイブズの世界の場合は見つけるためのツールというのが検索手段、finding aid と呼ばれるもので、日本では典型的には資料目録と呼ばれて来たものがそうだと思います。目録という言葉自体使われていますから、その説明はいいかと思います。

ダイブしなくとも物が探せるようになるのが一番理想的ではあります。

なんですが、実は揚足を取るかのように、図書の資料目録を例に出すのは良くはないのですけれど、図書の資料と比べると、アーカイブズの資料というのは非常に扱いづらい、要するに面倒なんですね。

図書の資料の場合は、図書 1 冊だったらその 1 冊が何か 1 つの独立した情報を持っていて、その中の書誌事項は決まりきったのですが、アーカイブズの資料とはそうではないと言うのは今日もうこの場でも説明があったかと思います。出所があるとか、群として存在しているとか、量が多いとか、後は 1 個 1 個取り上げてしまうとそれぞれ断片的であったり、また同じ様な配布資料が何回も、それは会議としてよくないのかも議事録の作り方が悪いという事なのでしょうけれど配布資料で同じ物が出てくる、重複する部分も現実には沢山あると思います。そうなると 1 個 1 個の中身だけを見ても仕方がない、前後の脈略であるとか構造も把握しないといけない、そうしないとアーカイブズの資料は理解出来ないというのが前提にあるとするとその為の探すツールというのは図書の目録では足りなくて、前後の脈略の説明が中に組み込まれていないといけないと言う事があると思います。

階層に関して、具体的にどの様な階層があるか説明を簡単にていきたいと思います。

アーカイブの資料は出所があって、その単位で捉える一番上の collection、fonds と呼ばれるまとまりなんですが、その中の例えば組織であれば分業の体系、人間であれば機能等で分けられていると思うんですね。

まとまりとしての理解の仕方は一番上の 1 個だけ理解しても勿論良いと思うのですね、まとまり全体、段ボールで何箱かあって、中を開けていないがどこから送られて来たというのは判る状態、それが 1 つ。

中を開けた場合、括りがあり、括りのまとまり、或いは意味で分類されていてここからここまでが何か判っている物がそれぞれあったりする。そういう物を一々記述しないといけないという面倒な事があります。今言った

パターンだと大雑把に分けられている、箱が何箱かあって箱の単位で分けられているでもいいですし、出所があって sub fonds、sub group と言うのは 1 つのパターンですし、それ以外にもう少し中が判ったというのがあったとすれば、そこまで区画として分けられた様な何か捉え方も出来るでしょう。それ以外に片端から 1 点ずつリストにしていった等の捉え方もあるかもしれません。全ての階層でという事があるかもしれません。そういうたった幾つかの捉え方もあるし、図書の場合は 1 冊に 1 レコードと対応するのですがアーカイブの資料はそいつたデータの捉え方が出来なくなってしまう、collection があるとすると、例えばこれは五島所蔵史料なのですが古本屋で買ってしまった物 1 個これが collection だとしたら、これに対する概要のデータ 1 個記述を作ると 1 個記述レコードが出来ますね、これで終わりにしてももちろん良いし、中を 1 枚 1 枚取り出してレコードを作ってもいいと思いますが概要データでレコードは確実にある訳です。その次に札が見えるでしょうか 1 とか 2 とか 3 とかありますが、纏りによって大よそ意味的にも区画的にも分けられていると考えて、ある series だと仮に捉えたとすると、series の概要データが 3 つ出来るわけですね、一応階層構造この様に捉えること出来るのですが、足すと 1 個の資料のデータに対して合計すると 4 つ出来ます、目録レコードという言い方が正確かわかりませんがレコードが 4 つ出来る、作業の進行によってデータが増えるという不思議な状態が起こるその集積体がアーカイブの資料の目録なのです。

やっと EAD の話なのですが、EAD とは何かというと図書の目録の場合は、図書の目録を作るにあたり、目録の規則が前提にあります、その規則に則りデータを書いていく、その書いていって図書だったら OPAC(Online Public Access Catalog) に載せます。その段階で加工しないといけない、加工する上に企画も揃えないといけない OPAC の場合は ISBD(International Standard Bibliographic Description) 記述があり国際標準があります、国際標準は考え方なので規則までは細かく定められてはいない、それについて書誌的事項としてどうとれば良いか目録規則として存在しています。日本目録規則でもいいですし NCR でも良いですしあるとすると、それに対して OPAC で具体的に実現するのは MARC(The Machine-Readable Cataloging format) というのがあります。

(資料 7) 表下側に ISBD に対して MARC とありますが、そのアーカイブ版みたいなものです EAD というのは。

アーカイブ版の機械可読目録、電子化するための構造の標準、デファクト標準と言うのが EAD(Encoded Archival Description) と言うことが出来ます。ほかに EAC や EAG など存在しますが細かく説明しだすと時間が足りなくなってしまいますが、図書の目録で著者であるとか編集者から本を探したいという場合もあると思います、その時に著者の情報が別途入っていて、ペンネームと実名とバラバラになっているとしてそれが同じ人という情報があると、ペンネームの方で引いても実名で書いた本や論文が出てくるようになると便利ですね、そういった人に対する情報をまとめているのが実は典拠レコードといいまして、そのアーカイブ版、図書館典拠レコードの規則は忘れてしましましたが、MARC については書誌の情報以外に典拠レコードの何か部分があったかと思います。それに対応する物が Encoded Archival Context (EAC) と言うことが出来ると思います。Encoded Archival Guide (EAG) というのは何処にどんな史料があるか、どういった所が持っているか、持っている所に関する大雑把な情報をまとめたもので、つい最近 α 版のDTDの中身を解析することが出来たのですが、今は説明がまだ出来ないので申し訳ありません、これは省略します。それは何で書かれているかと言うと XML、または元々は SGML が基になっている標準汎用マークアップ言語 (SGML) で書かれる物です。EAD というのはそのセット、部分集合サブセットという言い方も出来るかと思います。DTD 文書形定義またはスキーマ (schema) 文法や記号等に規程、制御している、XML / SGML を使わなくとも EXCEL 等を使う方が便利じゃないのと言われるのですが、先程の階層的な把握と言うのをやらないといけないので、その為にはデータを入れ子に扱わないといけないので、具体的に例を挙げないとピンと来ないかもしれませんのですが、どう階層制を填補するのか実は EAD の情報検索システム構築で一番大変だった所でもあります。それを実現できる枠組みと言うことだけ、大雑把に頭に入れておいて頂ければと思います。

この間にタグのセットがあるのですが、EAD というのは 3 つ位の大雑把な情報のまとまりがあって、その中にさらに細かい情報がセット出来る様になっています。ですから単純に CSV の様に、或いは EXCEL の縦横の表の様

にどこの列に何を入れ行が増えていくイメージとは違うのですね。

それもやろうと思えば出来なくはないのですが、どこかに階層か何か、何番目に何処の中に入っていると情報を別途入れてやらないと CSV の情報などに開く事が出来ない、と言う場合は先程階層的に 3 つか 4 つ位の把握の仕方がある事を申し上げましたけれど collection のレベルと series のレベルだけの情報を引っ張り出す記述をする事も出来ますし、collection のレベルと一覧のリストの組合せる事も出来ます、フルのセットというのもデータとして収める事は出来る、先程階層的な把握をすると言ったその階層的な幾つかのパターンを全てカバーできる枠組みが実は EAD の枠組みです。

と考えると何故 EAD なのかアーカイブズの資料の目録の難しさを克服出来たものだからと、特にコンピューターの資料目録、オンラインの資料目録で実現出来るから EAD を採用するのが良いのではないかと言う事です。

実はこれだけではアーカイブズの資料足りないんですね、メタデータとして使うには EAD だけでは不足しますが差当たって、何処にどんな情報があるか EAD に関して言えば、目録の情報に関して言えば EAD があれば足りるだろうと言う事です。

SGML、XML であることは重要な理由なりますし、何より CSP などだと、EXCEL の表 1 セルの中に本で言うと何十頁となる記述を 1 セルに収めることができるとか、収められないですよね、今は出来る様になってしまったのですが、前は 256 文字しか確かに入らなかったと思います。

実は図書館の資料目録を電子化した時 OPAC は、その制約にだいぶ縛られていました。それを壊して超えることが出来たのは実は EAD というより XML だった。尚且つ、入れ子の表現が出来るという事は、電子的検索手段としてマルチレベル記述というのが可能になる、つまり先程の階層的にデータを取扱うことが出来るようになりました。またそれだけではなく、これは EAD というより本文を電子化するための TEI (Text Encoding Initiative) という人文系で使われている物が別にあるのですけれども、そこの中で電子記録、電子書類として取扱えるような部分があり、先程の EAD ヘッダーと呼ばれる資料目録のバージョン管理のような事がそこで出来るのですね、というものを兼ね備えているので EAD を使うのがよろしいのではないかという事です。

では実際に OPAC みたいに出来るかというと相当困難が最初ありました。コンピューターがそこまで使えない、私が使えないというだけではなく、コンピューターにも駄目な部分もどうしてもあるので、コンピューターにさせる作業を限定的に考える事をやらないといけなかったと言うのは 1 つあります。

もう 1 つは、情報検索システムはアーカイブの資料で重要な脈略であるとか構造を、ズタズタにするのが大前提にあるのです。検索結果として現れてくるものは、どれか 1 個出てくるわけですね、それが 1 つの単位になっていて、脈略や構造の提示などとは完全に相反するものだと、そうなると困るというので、階層的な提示というのが出来るか確認をして次にキーワード検索とその表現へという形で話を進めて行き、ここの場でも使ってもらえるような所に、やっと至ったという話です。

その組合せが、参考文献で私が「記録と史料・第 18 号」に記事として書かしてもらった物の中にもありますけど 2 つのコンポーネントがありますて、表現の所、それと後は実際の検索システムという形で組んであります。実は EAD は資料目録のデータを作ってしまえば、一種の画一化・企画化なので、機関とか、機関の中の資料群の違いは、全部吸収して横断的に検索する事は、自動的に可能になるのですね。

(資料 13) 全体像をこの様に出しましたが、上側に外部サーバーとありますが、ある EAD の検索システムが 1 つ動いていたとして、それとは別の所に EAD のデータ上のほうに置いてあったとしても、集めてきて中に取り込んで別の資料だと、別の資料 collection ですよという形で自分の所で検索システムがあれば収蔵資料目録のデータを併せて検索する事でそのまま総合目録・ユニオンカタログと同じ事が実現出来てしまします、それをここで研究機関、研究資料のアーカイブズについても行うという事で、国文研に仕組みを入れてもらっていたので、そこを使う予定だったのですけれども急に居なくなって本当にご迷惑をおかけしました。

どの様な内容かを紹介します、全文の表紙は要するに概要があって、中身の一覧等があるのですが全部プリントアウトして行くと冊子形の資料目録と同じ物が出来上がります。それを Web のブラウザの画面上で表現をするのが全文の表示の所です。これはあった方がおそらく良いでしょう。前後スク

ロールして見る事が出来ますから。

もう 1 つ、(資料 21) そっけない画面が出てきてしましましたが、本当はもっとインデックスを充実させないといけないと思っております、アルファベット順であるとか、事項の主題等でそこをクリックすると検索結果がズラーッと出てくるなど、一覧が本当は必要であると思っております。以前そういったご指摘を総研大アーカイブズプロジェクトからも頂いた事があります。参加している基盤機関だけの collection 一覧を出してほしいとかそういうのがありました。出来るようにはなっているのですが今それをサーバーに登録する事が出来ないので止まっています。

詳細検索であるとか簡易検索の項目などはこういったものかなというのを挙げておきましたので今は省いてしまいます。

簡易検索から(資料 23)仮に「物理」というキーワードを入れて検索をクリックすると、こういった検索結果(資料 24-26)が出てきます。国文研の史料情報共有化データベース、全国アーカイブズ総合目録の検索結果の所では約 800 の機関が参加して頂いているのでこのように出てきますこれをスクロールしていくと、中にこの様な画面が出てきます(資料 26)。

赤の矢印が出ている所が実は検索結果として当たった所なのですが、例えば早川幸男先生寄贈史料で、箱 B303 b の中のある series と言うより具体的には file が出てきました、さらにその file の中に item をとっていて、その item の中でこの 1 点が該当しましたとこのように階層的に表示出来るようになりました。collection レベルのデータも series レベル或いは sub series レベル、file レベル、item レベルと順々にこの様に出せるようになっているのが、情報検索システムの中で階層的な表示を両立させる苦労した点でした。どれか 1 つ選ぶと詳細な表示が出てきます。(資料 27-28) ここも少し工夫をした所で、上にどんな物が入っているか分かるようになってきました。また画面左側にナビゲーションのパネルを作成しまして、例えば今概要の所へラインを引きましたが、概要を選択すればその上位の属している collection レベルの情報に移ることが出来ます。

それ以外別途に画像・音声・映像等がこの資料については有ると、予めまとめておく事も出来ますし、それを選ぶと、ここでは別のメタデータ、データの為のデータを使わないといけないのですが、EAD 以外の METS (Metadata

Encoding & Transmission Standard) というものを使っています。画像のファイルとなると、ある冊子に頁があるわけですが、その 1 頁 1 頁が 1 つの本だとまとめておくデータも別途必要になるので、それも XML でコントロールすることをやってきました。ここまででは現在総研大のアーカイブズプロジェクトあるいは KEK 史料室では実現出来ていませんが、やろうと思えば出来ます。

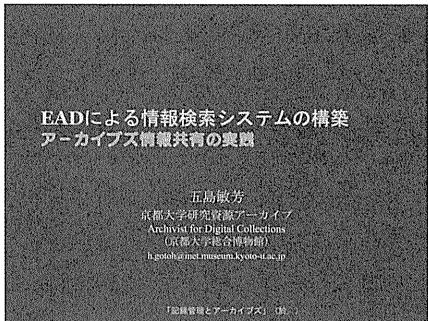
例えば古文書の場合、タイトルが表示されている所をクリックすると説明の所に戻るというように、行き来出来るようにしてあります。

機関を超えて実現出来れば自然と全国アーカイブズ総合目録、アーカイブズユニオンカタログが出来上がるだろうと考えております。今の所国文研でしか実現できていないのですが、仕組みとしては、私の所属している京都大学研究資源アーカイブで同じ物が出来るかと思います。差当り学内版ですが。

全国アーカイブズ総合目録の概要など言葉で示している物は今は省略したいと思います。

もしここにいらっしゃる方でまだ参加されていない所がございましたらご紹介いただければと思います。具体的に所蔵資料を、公文書でなくても研究資料のまとめなど色々なケースがあるかと思います 1 つの枠にまとめて載せてみると先程のようにひっかける事ができます。1 個 2 個であっても。しかし Google のように有象無象がでてくるノイズの多い検索結果ほど大きいことはないと思います。実務的に行っていくと実感していただけると思います。そこで面白い実例をみつけたら上の人に訴えかける事が出来るかと思います、そのためのツールとして使ってみる、その為にはまずここに入力するか、同じ規格の物を作らないといけないので是非ご参加いただければと思います。

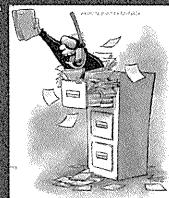
最後に宣伝ですが京都大学研究資源アーカイブにて学内で同様の事を行っています。ここでの横断検索を考えています、出来れば総研大のアーカイブズプロジェクトで出来上がっているような EAD のデータを横断検索対象として繋げさせてほしい、実証実験も予定されておりますので、国立公文書館さんにもお願いをしなければなりません、ご協力いただければと思います。学内版で 2009 年 4 月始動予定です。体を壊さない程度に両方とも頑張りたいと思います、以上です。



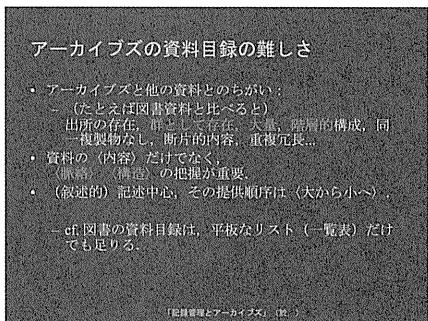
はじめに

- アーカイブスの資料は、見つかなければ、保存も利用もできない。
- その、資料を見つけるための工具（ツール）が、「検索手段」*finding aids*である。
- 日本では、その典型が（伝統的に）「資料目録」といえる。

イラストは、オーストラリア国立図書館ブックレット“Keep the knowledge – MAKE A RECORD”より

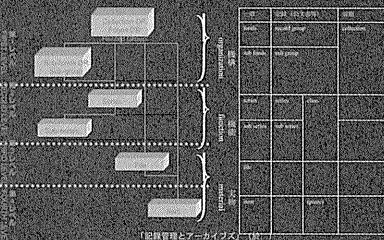


「記録管理とアーカイブス」(絵)

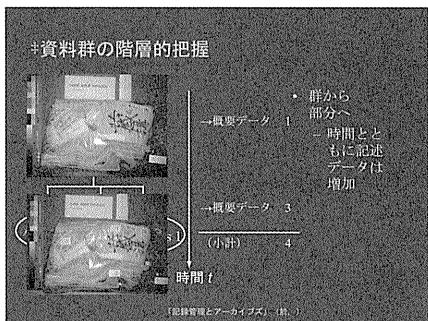


(図) つづきの文部の網に加筆: OAC Working Group, OAC Best Practice Guidelines for Encoded Archival Description Version 2.0, CDL, 2005

†資料群の階層構造モデル



「記録管理とアーカイブス」(図)



EADとはなにか？

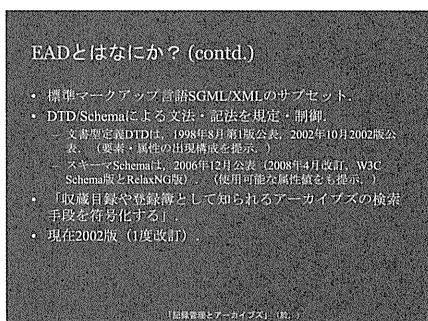
- アーカイブスの検索手段（資料目録）を電子化するためのデータマスク（非実上の）国際標準（規格）
- EAD, *Encoded Archival Description*（符号化永久保存記録記述）
- EAC, *Encoded Archival Context*（符号化永久保存記録コンテキスト）
- EAG, *Encoded Archival Guide*（符号化永久保存記録[収蔵者]便覧）

• 記述の国際標準との対応:

EADは、アーカイブス/MARC（機械可読データ）

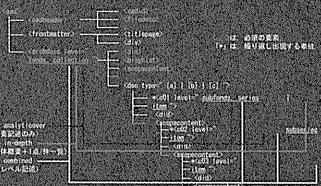
文書館	記述	EAD	MARC
図書館	記述	ISAD	EAD
図書館	画像レコード	ISAR	EAC
収蔵機関	ICA-ISDIAB	EAG	
機能	ICA-ISDF	?	
図書館	ISBD	MARC	

「記録管理とアーカイブス」(図)



EADデータの構成と類型

- 記述水準・深度と各記述部分の組合せ。



「記録管理とアーカイブス」(図)

なぜ EADなのか?

- アーカイブスの資料目録の難しさを、克服できた。
→ コンピュータでの資料目録（電子的検索手段）で。
 - (1) SGML/XMLであること。
 - (2) 大量の記述データを扱えること。
 - (3) 既存の検索手段としてマルチレベル記述を表現できること。
 - データの階層構造化が可能
 - 人間可読のデータ構造の実現
 - (4) 既存の検索手段としてだけでなく電子記録・電子書類として扱うための枠組みがあること。

「記録管理とアーカイブス」(終...)

アーカイブスの情報検索システムへの挑戦

- EADにいたるまでの困難。
 - コンピュータ利用を、あえて限定的にしかねることも、
「検索システム」の特性による矛盾。
 - もじもじと検索システムの結果表示は、順序を寸断してデータを
断片化する。（←階層構造の提示と相反する。）
- 戦術：まず基本的表现（階層的提示）を確保し、つぎに
キーワード検索とその表現へ。

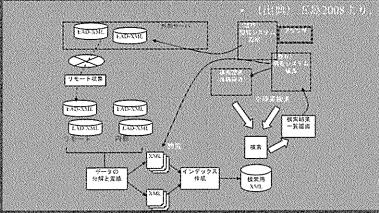
「記録管理とアーカイブス」(終...)

EAD-XML検索システムの紹介

- 資料目録EAD/XML化の実践例
- (A) 基本的表現・配信システム
 - 1) 資料目録全体の全文表示。
 - 2) 資料目録各部分の語譲的表示。
- (B) 情報検索システム
 - 1) 特定の資料群の検索。
 - 2) 多数の資料群の検索。
 - 3) 検索の取扱者の資料群の検索。

「記録管理とアーカイブス」(終...)

EAD-XML検索システムの全体像



「記録管理とアーカイブス」(終...)

EAD-XML検索システムの紹介

- 資料目録EAD/XML化の実践例
- (A) 基本的表現・配信システム
 - 1) 資料目録全体の全文表示。
 - 2) 資料目録各部分の語譲的表示。
- (B) 情報検索システム
 - 1) 特定の資料群の検索。
 - 2) 多数の資料群の検索。
 - 3) 検索の取扱者の資料群の検索。

「記録管理とアーカイブス」(終...)

(A) 1) 全文表示

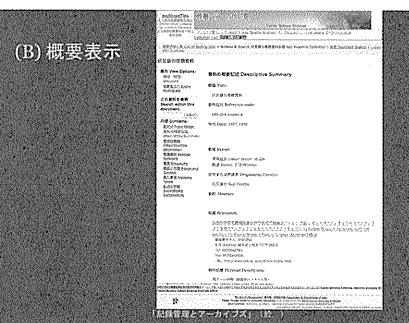
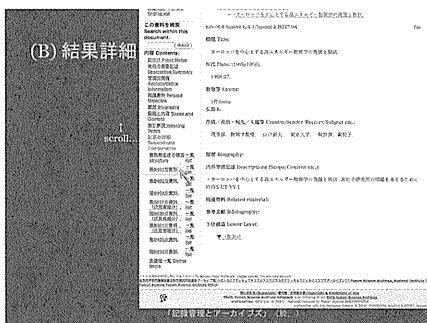
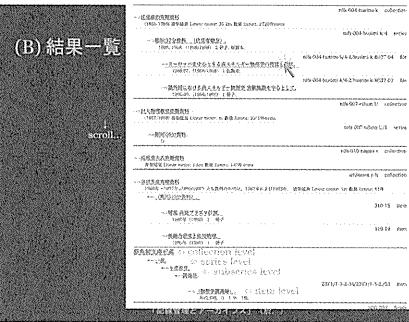
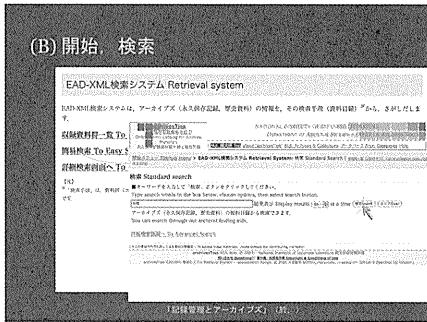
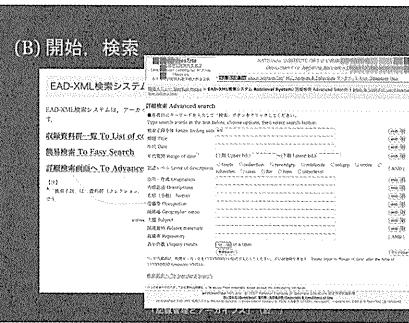
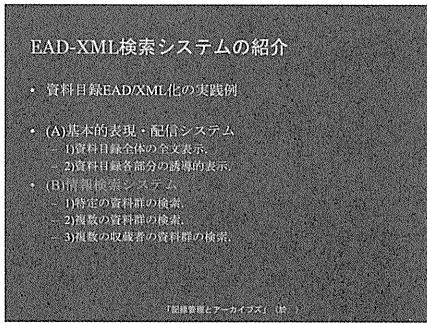
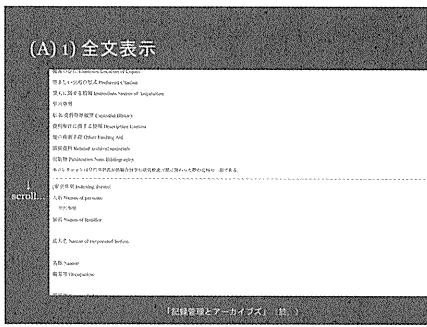
「記録管理とアーカイブス」(終...)

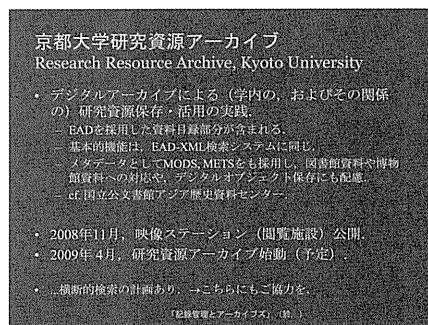
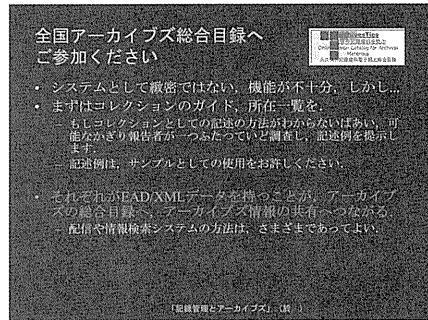
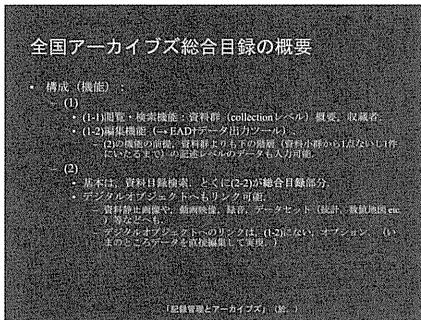
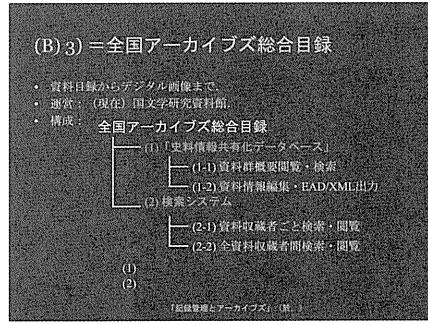
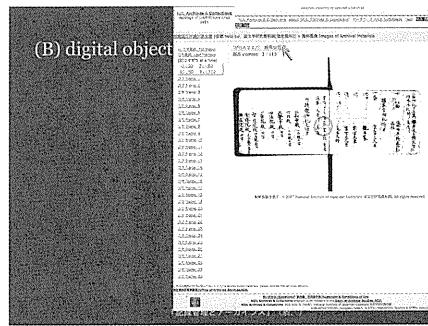
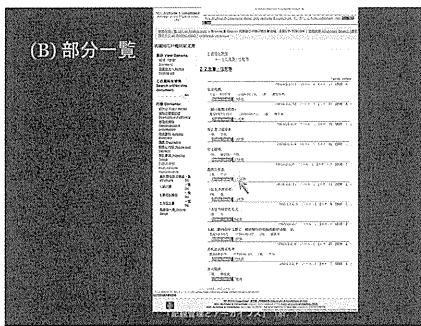
(A) 1) 全文表示

「記録管理とアーカイブス」(終...)

(A) 1) 全文表示

「記録管理とアーカイブス」(終...)





参考文献・謝辞

- 五島敏男、「日本におけるアーカイブズのオンライン総合目録構築における記録と史料」、18、2008年3月、pp.1-17。
- 本報告は、国文学研究資料館史料館におけるアーカイブズの情報検索システムに関する研究と、EADを日本に導入する個人研究者とともにになります。これまでたどりつくのに、実に多くの人たちにお世話になりました。ありがとうございました。
- 報告の機会を貰えてくださった総研大、大学共同利用機関アーカイブズプロジェクトおよびKEK史料室に感謝いたします。

「記録管理とアーカイブズ」（緒）