資料３

ウェブ情報のアーカイブ化促進に資する技術の開発･実証

【説明資料】

2005.3.17

総務省

Ⅰ　実証実験全体の説明

実験目的　～ウェブ情報のアーカイブ化の促進～

デジタル時代の知識･文化が結集する貴重な資産であるものの、日々、消去が発生するウェブ情報について、アーカイブ化や利活用を促進するための技術･仕組みの構築･実証を実施する。

1　経緯

ウェブ情報にはデジタル時代の知識･文化が結集されており、それ自体がデジタル時代の貴重な文化遺産といえるが、日々の更新による消去･散逸が発生しやすい。こうしたことから、海外においてはウェブ情報のアーカイブ化が開始され、我が国においても平成14年度から国立国会図書館が実験プロジェクトを開始したところであり、e-Japan重点計画2004等においては、ウェブ情報のアーカイブ化の一層の推進に向けた取組を講ずることとされたところ。本施策は、こうした観点から、様々な主体によるウェブ情報のアーカイブ化とその横断的な利活用を可能とすることを目的とする。

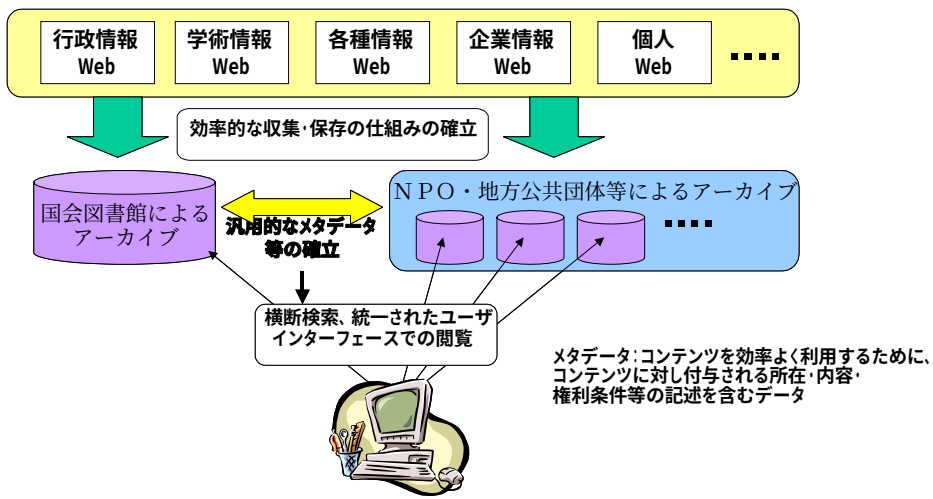
2　施策の概要

ウェブ情報を定期的に収集し、時系列に提供するためには、ＵＲＬや収集日等の情報をメタデータ化するなどウェブ情報を構造化して蓄積するとともに、それらの情報に基づく検索･閲覧等を可能とする必要がある。さらに、今後、地方公共団体やＮＰＯ等の様々な主体によるウェブ情報のアーカイブの構築が期待されることから、これらの標準的な技術･仕組みの構築が不可欠となる。このため、本施策においては、国立国会図書館と連携しつつ、

(1) ウェブ情報アーカイブの組織化及び大規模アーカイブの保存機能の開発･実証
(2) ウェブ情報アーカイブ間の連携･横断検索のための汎用的技術の開発･実証
(3) ウェブ情報の収集･保存･検索するための汎用的なメタデータ等の確立

を行う。

3　イメージ図



【平成１７年度予算予定額　２．１億円、平成１６年度予算額　２．５億円】

実験概要

１６年度　実施項目

コンテンツ収集

- 収集範囲拡大に関する調査研究

コンテンツ組織化

- 異なるアーカイブ間の横断検索を実現する汎用的なメタデータ体系の確立

コンテンツ保存

- コストの縮減、長期的な保存を可能とするアーカイブ技術の開発等
- 同一のウェブ情報を重複して保存することを回避する技術に関する調査研究

コンテンツ提供

- メタデータに基づく時系列別検索、特定コンテンツに係る時系列別表示の実現

実験フィールド

コンテンツ数　約８００件

府省サイトアーカイブ　約６５０コンテンツ(10GB)
(1.5TBの仮想ウェブから収集)

企業サイトアーアイブ　約１５０コンテンツ(10MB)

コンテンツ提供協力者

・国立国会図書館様

・富士通株式会社

JGN IIで
2拠点を接続

netCommunity
公開施設（関東）
●検証用端末

netCommunity

けいはんなプラザ
実証実験センター
●ポータルサイトシステム
●アーカイブ主体A、アーカイブストレージ
●アーカイブ主体B、3次元表示サーバ
●仮想ウェブシステム　など
公開施設（関西）
●検証用端末

Internet

実証実験関係者

実証実験関係者は、インターネット経由でポータルサイトへアクセス可能

II　システム説明

実験システムの全体構成

実証実験センター

仮想ウェブ

●実証実験のために、仮想的に創出したインターネット空間
●実証実験では、この仮想ウェブからウェブコンテンツを収集
●ウェブコンテンツは全て、著作権処理済み

ロボット収集

アーカイブ主体A

●ウェブコンテンツを収集･アーカイブする機関
●アーカイブ主体Aは、仮想ウェブに存在するウェブコンテンツを収集･アーカイブ
●本実証実験では、府省のウェブコンテンツをアーカイブ

④同一のウェブ情報を重複して保存することを回避する技術に関する調査研究
⑤収集範囲拡大に関する調査研究

アーカイブストレージ

●複数の保存装置を階層的に用いることによりコスト低減、長期的にコンテンツを保存する技術を開発･実証
●アーカイブされたコンテンツの改竄･消失を防止する技術の開発･実証する

③コストの縮減、長期的な保存を可能とするアーカイブ技術の開発等

媒体での投入

アーカイブ主体B

●アーカイブ主体Aとは別の機関とし、本実証実験では、企業のウェブコンテンツをアーカイブ
●アーカイブ主体Bを設置する趣旨は、共通のメタデータ体系に基づく他のアーカイブ主体（本実証実験ではアーカイブ主体A)との連携を実証する為
●アーカイブ主体Bは予めウェブコンテンツがアーカイブされているものとすし、仮想ウェブから収集はしない

ポータルサイトシステム

●複数アーカイブ主体を横断的に検索するポータルサイト
●アーカイブ主体A,B各々の検索に接続する分散型と、メタデータをあらかじめ収集しておく集中型の二つの形態の横断検索が可能

①異なるアーカイブ間の横断検索を実現する汎用的なメタデータ体系の確立

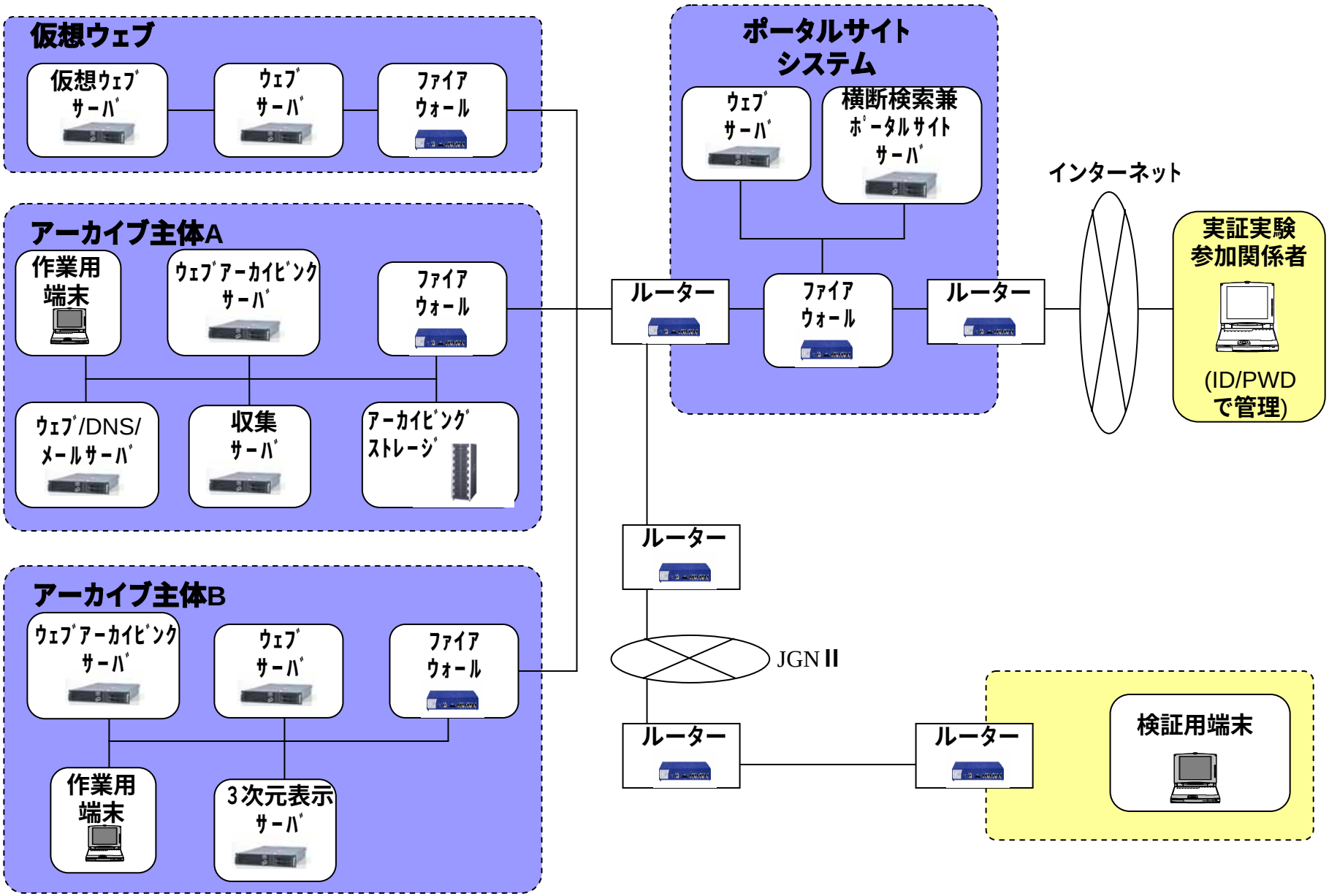
Internet

検証用端末（利用者）

●ポータルサイトシステムを通じてアーカイブ主体を利用する一般利用者用端末

②メタデータに基づく時系列別検索、特定コンテンツに係る時系列別表示の実現

システム構成図



業務システム　～コンテンツの収集から提供まで～

機能	機能概要
コンテンツ収集	仮想ウェブ上の各機関サイトにアクセスし、ウェブコンテンツの収集を行う機能
コンテンツ組織化	メタデータの付与／ハイパーリンクの書換え等を行う機能
コンテンツ保存	収集したコンテンツ、提供用のコンテンツをストレージに保存する機能
コンテンツ提供	利用者提供（検索・閲覧）を行う機能

収集

機能	収集業務の流れ
コンテンツ収集	**①収集実行** 収集の起点URL、各種収集条件（収集範囲、深さ、再収集頻度等）を設定し、ロボットにより収集
コンテンツ組織化	**②収集状況の監視** 個々の収集対象に関する、収集の状況（ステータス等）を監視
コンテンツ保存	**③収集結果の確認** ウィルスチェック等を行った上で、意図した収集が行われたか否かを確認
コンテンツ提供	

組織化

機能	組織化業務の流れ
コンテンツ収集	**①メタデータの入力** コンテンツ毎に、タイトル、作者、公開者、収集日等のメタデータを付与、また利用条件を設定
コンテンツ組織化	**②形式変換** 収集したコンテンツのリンクの書換え（アーカイブとしてのリンクへの変換）処理等
コンテンツ保存	
コンテンツ提供	**③更新検知のための情報の計算** 再収集を行った場合に更新の有無を判定するために、現世代のコンテンツに対してハッシュ値等を計算

保存

機能	保存業務の流れ
コンテンツ収集	**①収集したコンテンツの保存** 収集したそのままのコンテンツを原本として階層型ストレージに保存
コンテンツ組織化	**②提供用コンテンツの格納** 形式変換等を行ったコンテンツをウェブアーカイビングサーバのストレージに格納
コンテンツ保存	**③再収集（更新版）コンテンツの保存・格納** 再収集し、更新検知されたコンテツは再度階層型ストレージに保存、形式変換後ウェブアーカイビングサーバのストレージに格納
コンテンツ提供	

コンテンツ概要

本実証実験システムでは、以下のコンテンツを利用

アーカイブ主体A
（府省コンテンツのアーカイブ）

① 内閣官房
② 首相官邸
③ 警察庁
④ 総務省
⑤ 消防庁
⑥ 法務省
⑦ 財務省
⑧ 文部科学省
⑨ 文化庁
⑩ 厚生労働省
⑪ 社会保険庁
⑫ 農林水産省
⑬ 経済産業省
⑭ 資源エネルギー庁
⑮ 特許庁
⑯ 中小企業庁
⑰ 国土交通省
⑱ 海上保安庁
⑲ 高等海難審判庁
⑳ 環境省

※国立国会図書館が収集したものを利用投入
著作権は処理済

約650タイトル (10GB)

アーカイブ主体B
（企業コンテンツのアーカイブ）

① 世界の車窓から
② 環境活動
③ 社会貢献活動

※富士通ホームページの情報を利用
著作権は処理済

約150タイトル (10MB)

メタデータ体系

(1) ウェブ資源としてのメタデータ

Dublin Core(*1)等をベースに、国会図書館や海外の動向を加味して体系を定義

実証実験システムにおける名称	実証実験システムにおける内容	記述内容	Dublin Core 対応項目
タイトル	コンテンツの名称		title
その他のタイトル	サブタイトル、関連タイトル		
作者	コンテンツの作成者、機関	府省名、一部部署名まで記述	creator
分類	NDC分類(*2)	NDC第3区分までを記述	subject
説明	コンテンツの内容に関する説明		description
公開者	ウェブ上での公開者、公開機関	府省名を主に記述	publisher
寄与者	協力、貢献している人・組織等	(16年度は記述していない)	contributor
公開日	コンテンツがウェブ上で公開された日付	明記されていないものは推測で記述	date
資源タイプ	資料のジャンル	「白書」、「審議会資料」、「統計資料」等を記述	type
フォーマット	データの形式	主には「html」,「pdf」を記述	format
資源識別子	ウェブ上での起点(コンテンツトップページ) URL		type
情報源	元となる情報への参照	(16年度は記述していない)	source
言語	コンテンツの主たる記述言語	日本語、英語等	language
関係	関連するリソースへの参照	(16年度は記述していない)	relation
時間的・空間的範囲	コンテンツの範囲、対象(場所、時代等)	一部コンテンツに関し位置(緯度、経度)を記述	coverage
権利管理	コンテンツの権利に関する情報。	コンテンツトップページにコピーライトが明記されている場合に記述	rights

*1 Dublin Core　:ウェブ上のリソースを記述するメタデータの国際標準

*2 NDC: Nippon Decimal Classification (日本十進分類)

メタデータ体系

(2) アーカイブに関するメタデータ

以下のメタデータおよび記述内容を規定し、付与

実証実験システムにおける名称	実証実験システムにおける内容	記述内容
収集ドメイン	収集する範囲（ドメイン）	各サイトのドメイン（サーバ）名を指定
収集ディレクトリ	収集する範囲（ディレクトリ）	範囲を限定できる場合に指定
収集の深さ	収集の際、リンクをたどる深さ	コンテンツごとに必要な深さを設定
再収集頻度	再収集を行う頻度	16年度は再収集を行わない設定
利用条件	利用者への提供の条件	16年度は、未設定
収集日	収集を行った日	仮想ウェブからの収集日

III　成果公開デモ

成果公開デモ手順

本実証実験では、3項目の実証と2項目の調査研究を実施中

実証項目

(1) 異なるアーカイブ間の横断検索を実現する汎用的なメタデータ体系の確立
- ・実証内容の説明（分散型・集中型横断検索方式）
- ・【デモ】分散型・集中型横断検索と結果閲覧

(2) メタデータに基づく時系列別検索、特定コンテンツに係る時系列別表示の実現
- ・実証内容の説明（時系列表示実現方式）
- ・【デモ】アーカイブ主体Bの検索とサムネイル表示

(3) コストの縮減、長期的な保存を可能とするアーカイブ技術の開発等
- ・実証内容の説明（アーカイブストレージの適用、改竄・消去防止）
- ・【解説】アーカイブストレージの動作解説

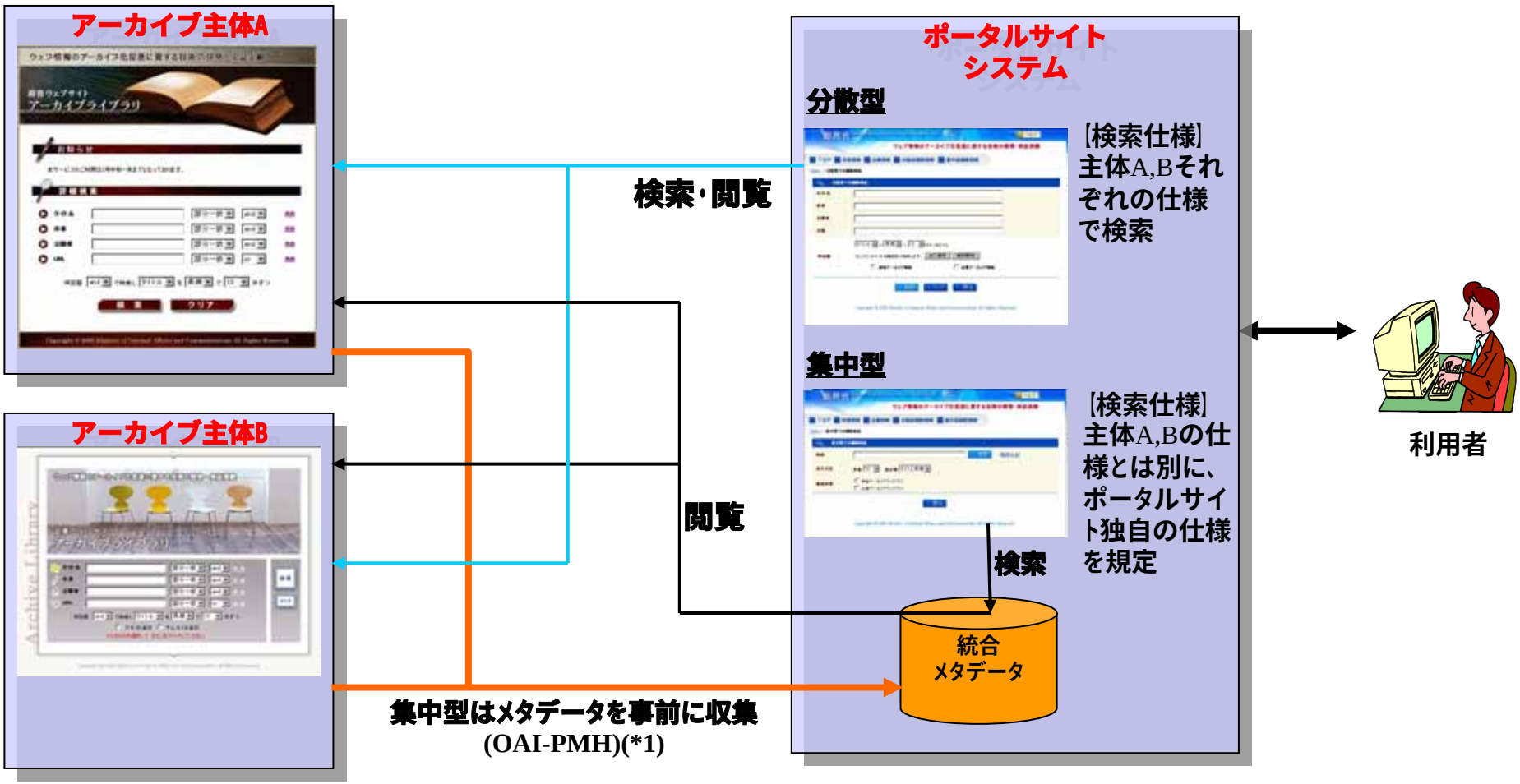
研究項目　（報告書にて別途報告）

(4) 同一のウェブ情報を重複して保存することを回避する技術に関する調査研究

(5) 収集範囲拡大に関する調査研究

横断検索　システム概要

今後多数のアーカイブの出現を予想
⟶ 複数のアーカイブに対し、ワンストップでの検索を可能とすることが必要
横断検索の実現方式
⟶ 分散型と集中型の2つの方式の検討



(*1)OAI-PMH : Open Archives Initiative Protocol for Metadata Harvesting（メタデータ収集のためのプロトコル）

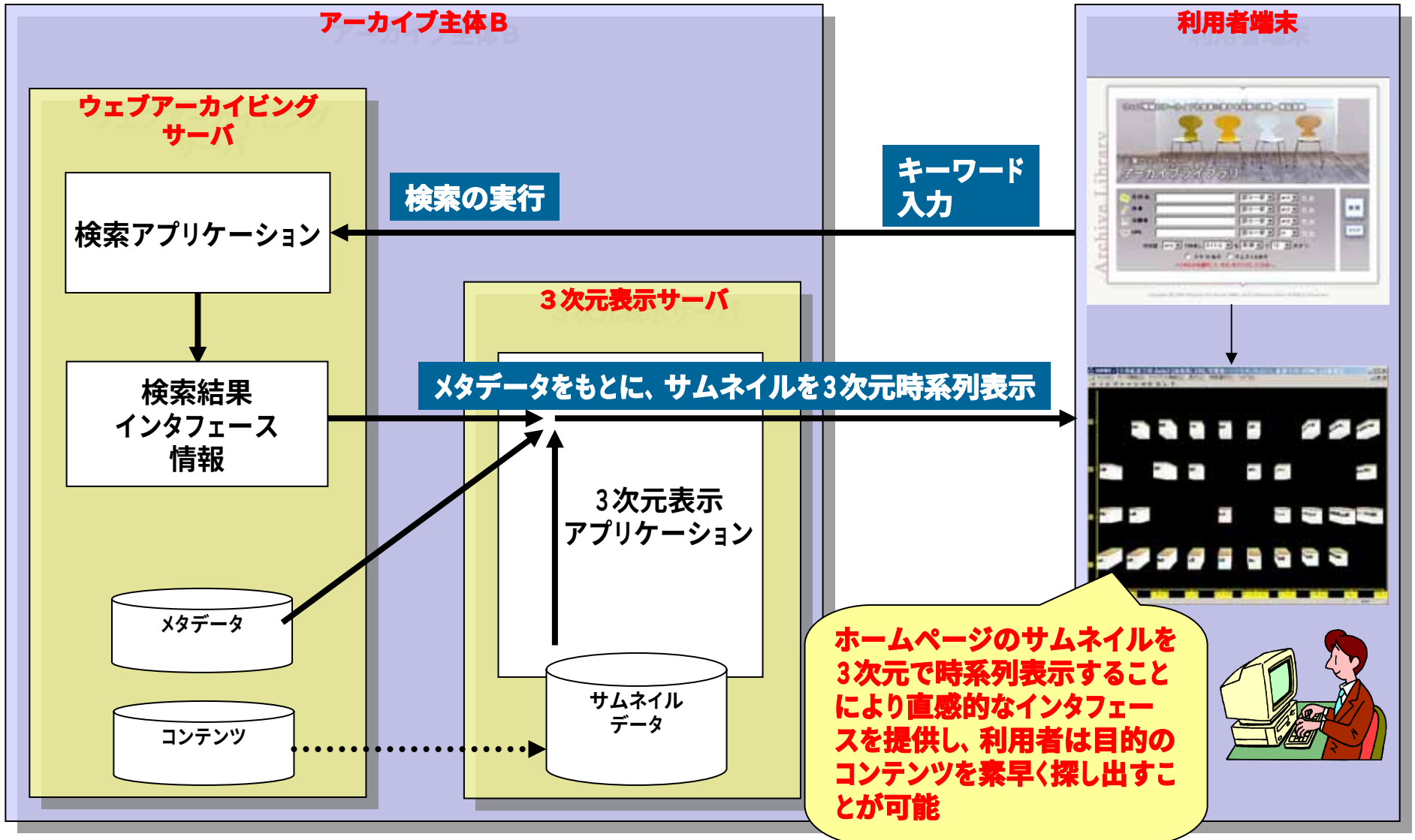
集中型と分散型の横断検索比較

・集中型、分散型それぞれに特徴があり、一長一短
・両方式が補完的に組合されていく傾向

	集中型	分散型
検索精度	○データを一元化するので、検索の詳細仕様統一可能	△表記揺れ対応などの仕様は相手システムに依存
検索機能の拡張性	○ﾒﾀﾃﾞｰﾀ自体を一元化するので自由な検索機能が実現可	△検索機能自体の拡張は困難
検索先との連携	△検索先の閲覧サー ビスとの連携には課題が存在	○検索先の閲覧サービと連携が容易
サーバ負荷、ネットワーク負荷	○検索時の相手サーバ、ネットワークへの負荷小	△検索ごとに負荷
コスト	△OAIでの提供、収集に関する実装が必要	○簡易的なものであれば低コスト
運用性	△メタデータ収集に関する運用の追加が必要	○相手サービスをそのまま利用するため運用が容易

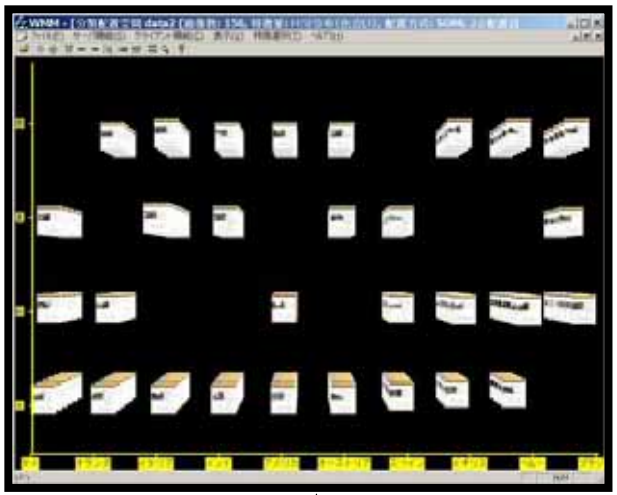
時系列別検索　システム概要

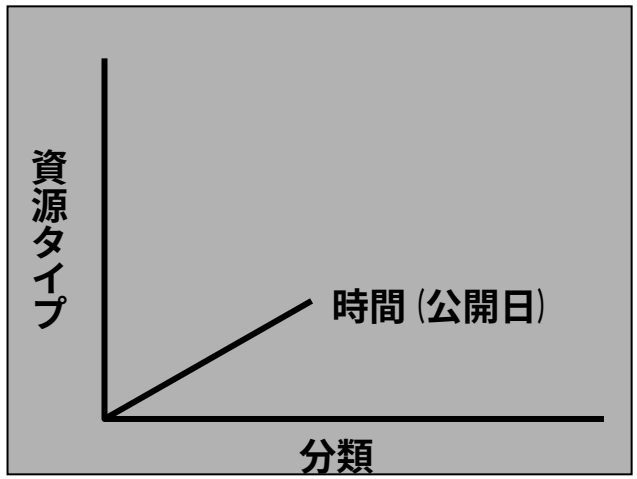
メタデータに基づく時系列検索･閲覧機能の実現

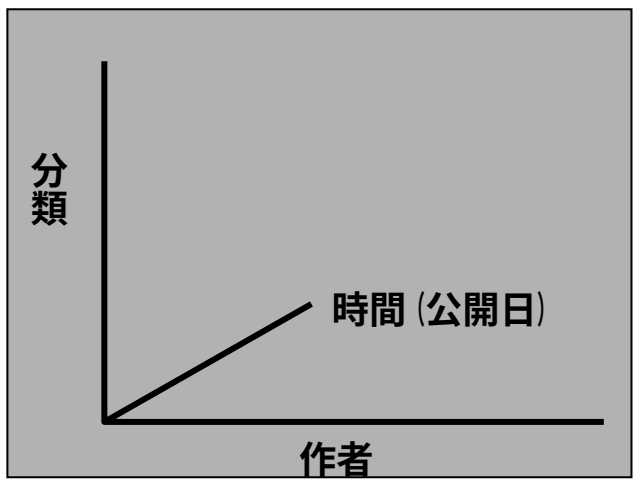


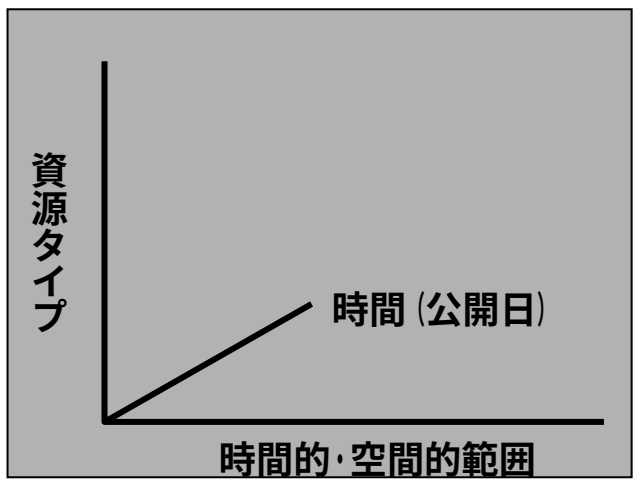
時系列別検索　インタフェース概要

利用者の検索目的にあわせて、3次元の軸をメタデータ項目から選択することが可能









※1）表示された3次元空間内をウォークスルーのイメージでコンテンツを検索。単純な縦移動、横移動、奥手前への移動も可能

※2）コンテンツ表示の拡大縮小も利用者の自由に行うことが可能

※3）2次元表示も可能。その場合、レイアウト、色合いなどが類似したサムネイルを抽出することも可能

※4）本実証実験では、システム側で事前に用意した軸の組合せの中から選択可能

アーカイブストレージの実証概要

1. 低コストの保存を可能とする技術の実証

⟶ 階層管理機能によるコスト低減

①一次階層ストレージ：高速アクセスが可能なディスクアレイ(一次階層ストレージ)情報を保存
②二次階層ストレージ：一次階層ストレージの情報を速やかに長期保存性に優れ、低価格な大容量LTOテープライブラリ(二次階層ストレージ)に自動コピーし、原本情報として確実に保存

※新しい情報はディスクアレイにキャッシュしコンテンツの確認に利用。テープには、収集した全てのコンテンツを保存

2. 保存したウェブ情報の改竄･消失防止技術の実証

⟶ 情報の改竄防止

①追記型構造(WORM *1)：操作ミスや故意の改竄／削除から原本データを守る為、一度書き込まれたデータを変更できない追記型構造を採用
②専用API(*2)：権限を持たないサーバ・クライアントからのアクセスを確実に防止する為、専用APIを持つ機器からのみアクセス可能とする仕組の採用

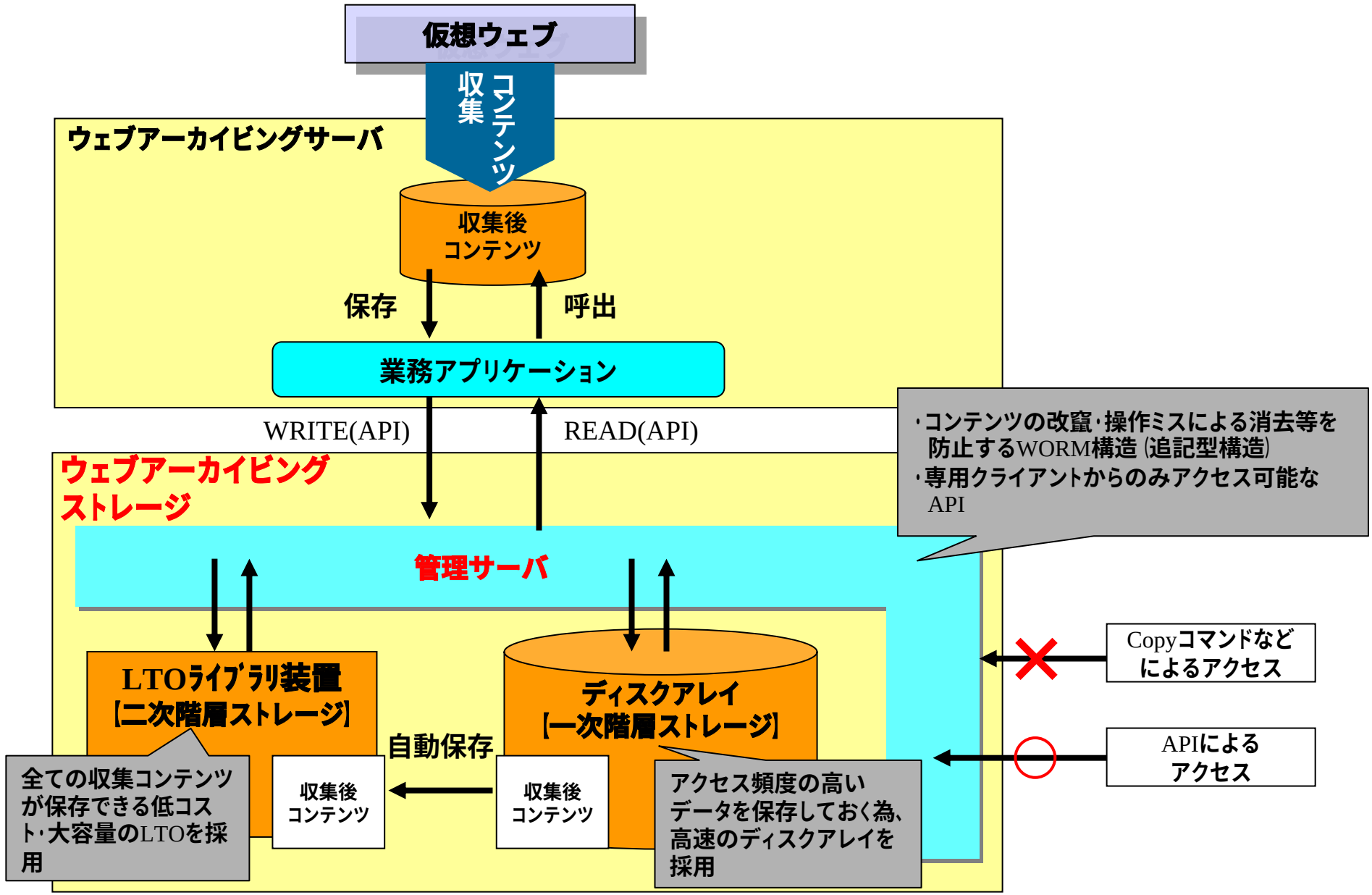
*1 WORM：Write Once Read Many(一度だけ書き込むことができ、消去／変更のできない記憶メディア)
*2 API：Application Program Interface（アプリケーションからアクセスするためのインタフェース）

3. 長期的な保存を可能とする技術の実証

⟶ XML形式による保存情報の標準化

○保存するコンテンツには、保存に関する管理情報をXML形式で自動的に付与。そのため、二次階層ストレージであるテープ媒体に保存した情報を他のシステムで読出し・再利用することが可能

アーカイブストレージの適用



画面遷移

ログイン画面

ポータルトップ

利用案内

実施概要

FAQ

階層ストレージ

集中型横断検索

府省アーカイブ検索

検索結果（府省アーカイブ）

その他テーマ

分散型横断検索

企業アーカイブ

検索結果（企業アーカイブ）

検索結果（時系列表示）