資料１

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】

2006.3.24

総務省

Ⅰ　実証実験全体の説明

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。

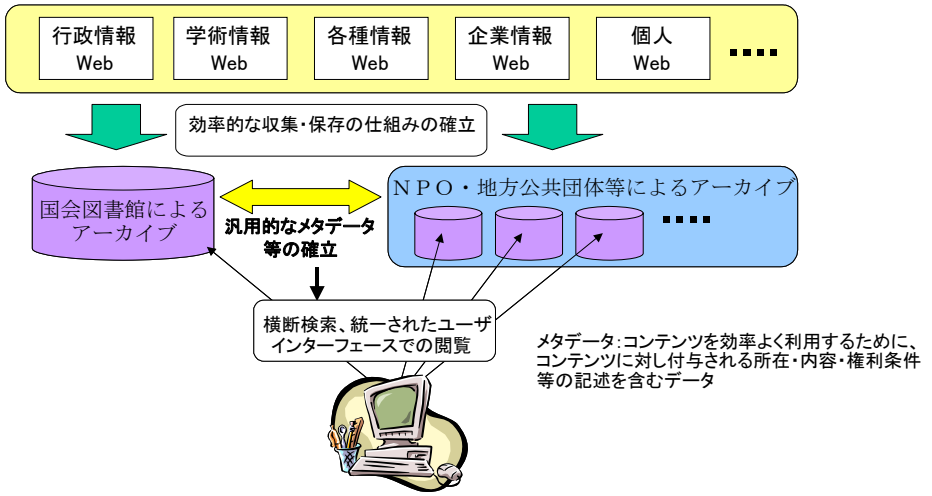
2　施策の概要

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年度予算額　約0．9億円、平成16年度予算額　1．4億円】

実験・調査概要

	16年度　実施項目	17年度　実施項目
コンテンツ収集	● 自動収集困難なコンテンツの収集に係る技術的対応の可否に関する調査研究	● 収集条件（起点URL・収集範囲・深さ・除外URL）の自動検出技術の開発
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	● メタデータの自動抽出技術の開発 ● ウェブ情報アーカイブに係るメタデータ体系のさらなる検証・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	● 異なるアーカイブ間のリンクナビゲーション機能の開発 ● 更新頻度と更新箇所の効果的な閲覧インターフェイスの開発

実験フィールド

コンテンツ数　約1220件

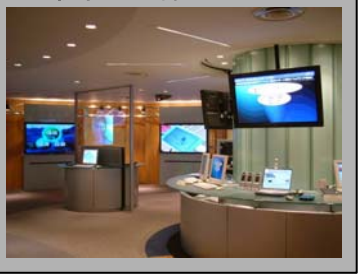
- 府省等サイトアーカイブ　約1000タイトル(90GB)
  (1.5TBの仮想ウェブから収集)
- 企業サイトアーカイブ　約220タイトル(10GB)

コンテンツ提供協力者(順不同)

- 国立国会図書館　・岡山県
- ヤフー株式会社　・日本電気株式会社　・富士通株式会社

JGNⅡ (20GB)
で2拠点を接続
[運用:情報通信研究機構(NICT)]

netCommunity
公開施設(関東)
●検証用端末



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

インターネット

実証実験関係者

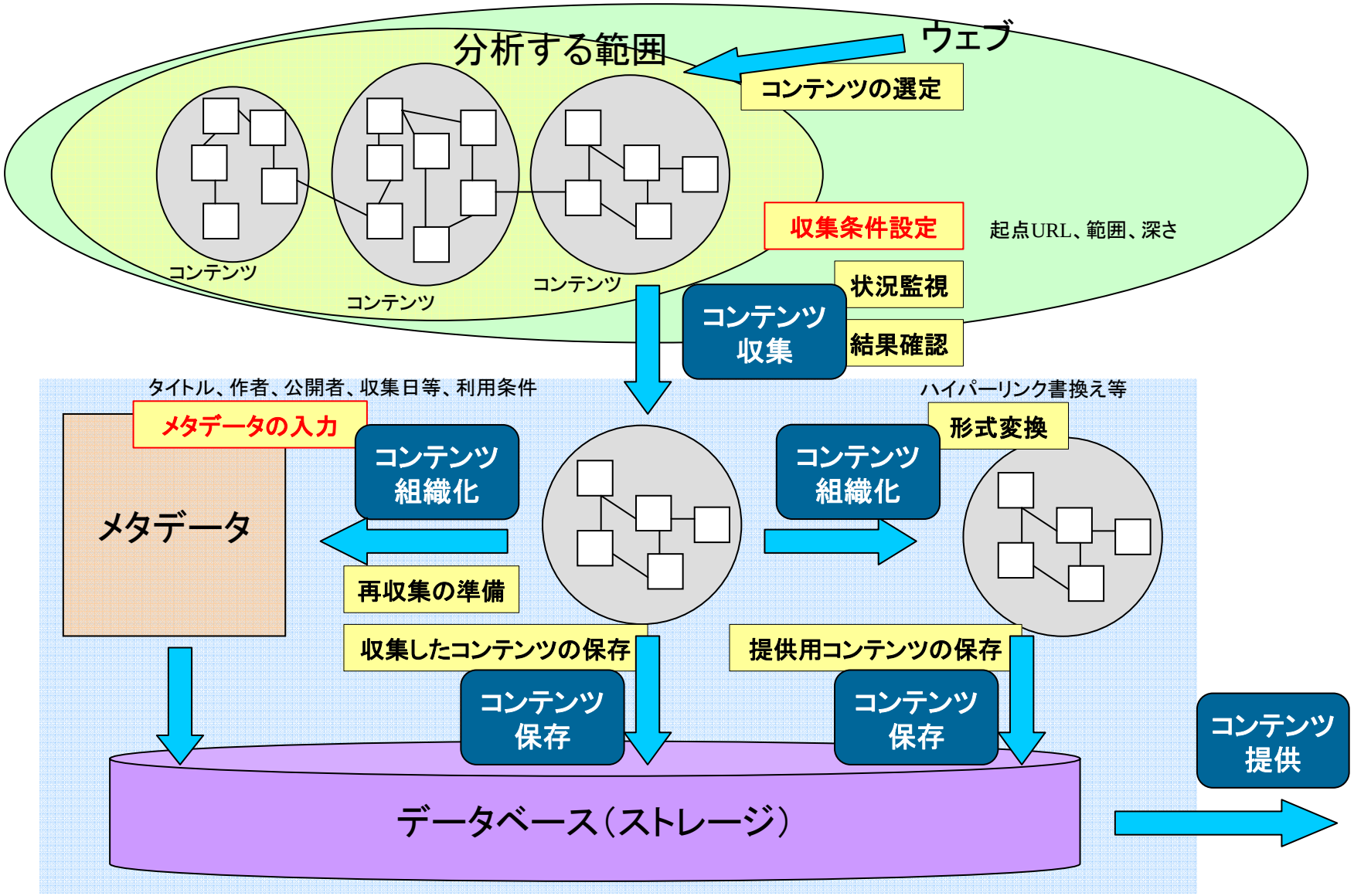
実証実験関係者は、インターネット経由でポータルサイトへアクセス可能

Ⅱ　システム説明

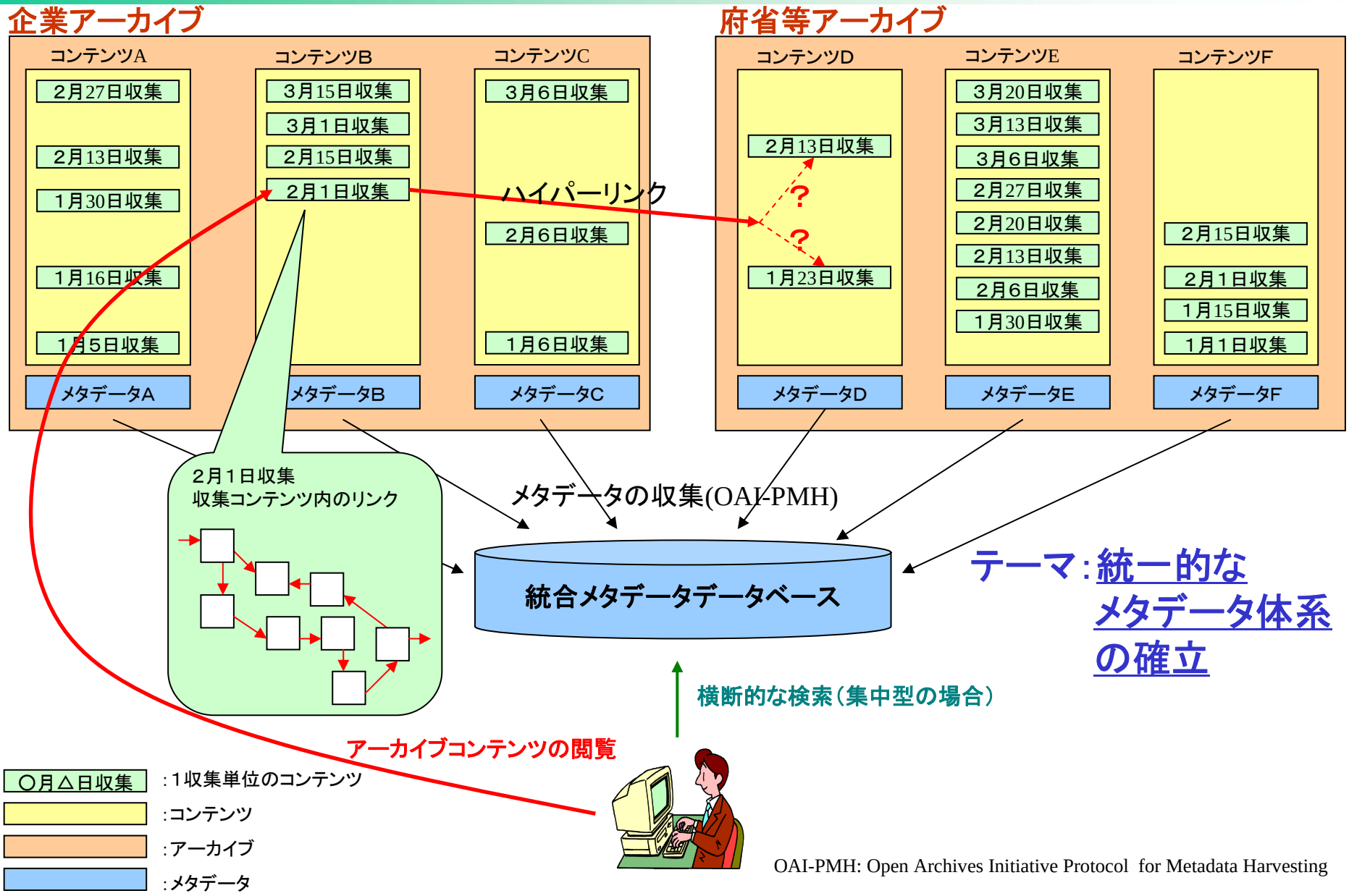
ウェブ情報アーカイブ業務（1）

機能	機能概要	各機能の流れ
コンテンツ収集	ウェブサイトにアクセスし、コンテンツの収集を行う機能	【収集①】収集実行 コンテンツの選定、収集条件の設定、ロボット収集 【収集②】収集状況の監視 収集の状況（ステータス等）を監視 【収集③】収集結果の確認 ウィルスチェック、収集したコンテンツの確認
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	【組織化①】メタデータの入力 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 【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（アーカイブ内リンクへの変換）等 【組織化③】再収集のための準備 ハッシュ値等を計算（次回収集時に更新の有無の判定基準とするため）
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	【保存①】収集したコンテンツの保存 収集したコンテンツを原本としてストレージに保存 【保存②】提供用コンテンツの保存 組織化したコンテンツを提供用としてストレージに保存 【保存③】再収集（更新版）コンテンツの保存 再収集し、更新検知されたコンテツは再度ストレージに保存
コンテンツ提供	利用者提供（検索・閲覧）を行う機能	【提供】収集・保存したコンテンツの提供 利用者に検索・閲覧サービスを行う

ウェブ情報アーカイブ業務（2）

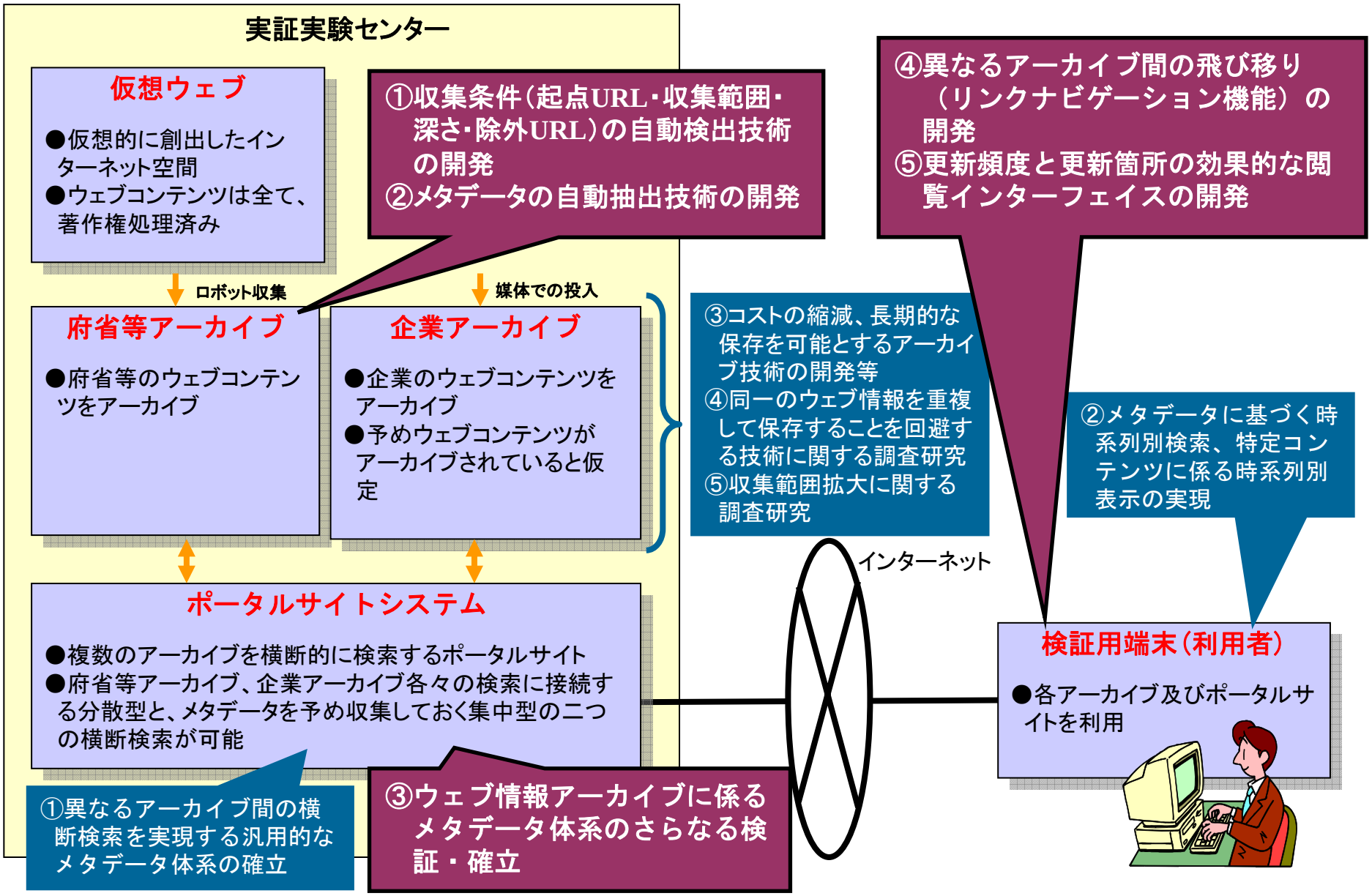


アーカイブ連携

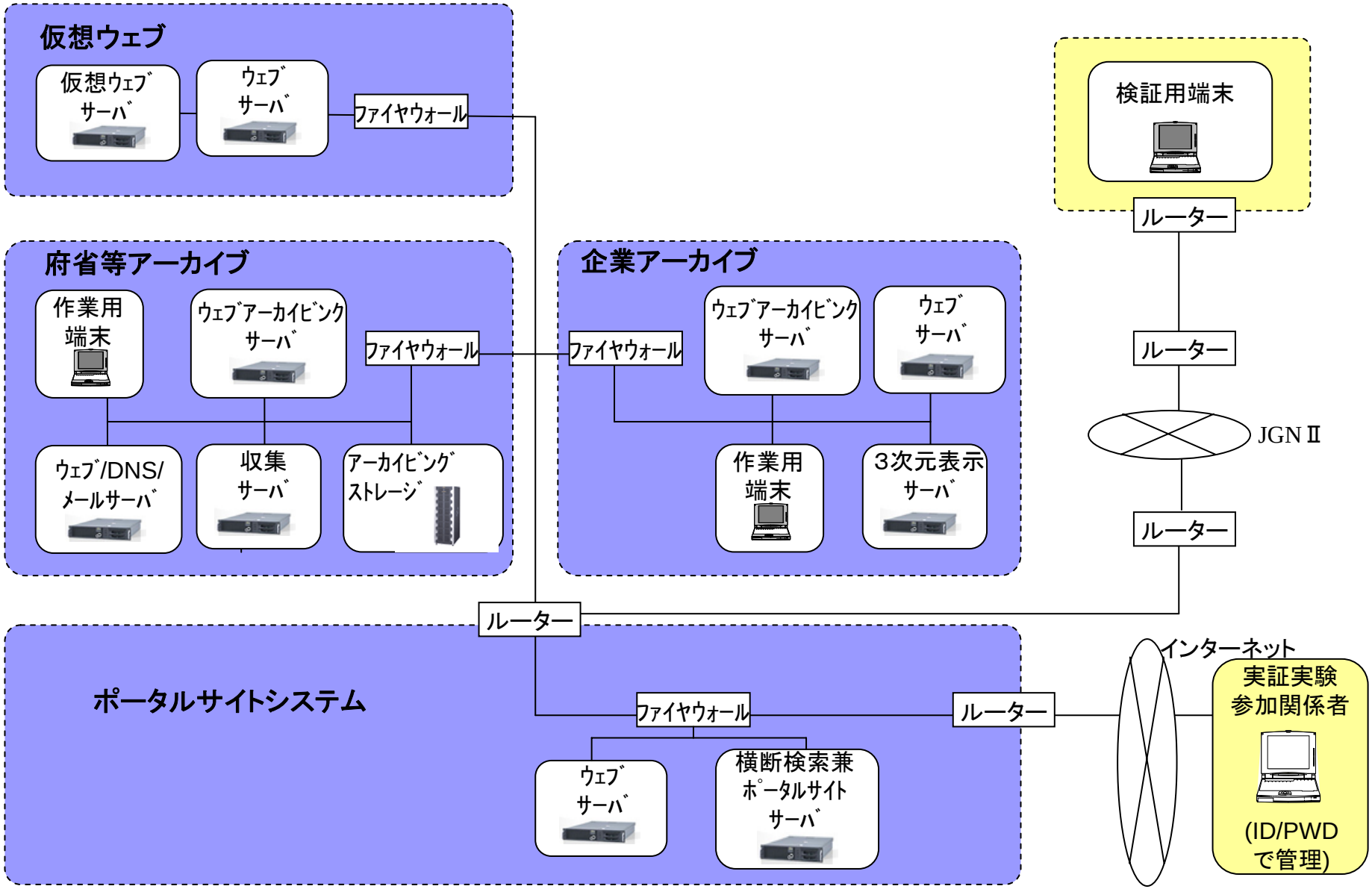


OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting

実験システムの全体構成



システム構成図


仮想ウェブ
仮想ｳｪﾌﾞ サーﾊﾞ
ｳｪﾌﾞ サーﾊﾞ
ﾌｧｲﾔｳｫｰﾙ
検証用端末
ルーター
府省等アーカイブ
作業用 端末
ｳｪﾌﾞｱｰｶｲﾋﾞﾝｸ サーﾊﾞ
ﾌｧｲﾔｳｫｰﾙ
ｳｪﾌﾞ/DNS/ メールサーﾊﾞ
収集 サーﾊﾞ
ｱｰｶｲﾋﾞﾝｸﾞ ｽﾄﾚｰｼﾞ
企業アーカイブ
ﾌｧｲﾔｳｫｰﾙ
ｳｪﾌﾞｱｰｶｲﾋﾞﾝｸ サーﾊﾞ
ｳｪﾌﾞ サーﾊﾞ
作業用 端末
3次元表示 サーﾊﾞ
ルーター
JGN Ⅱ
ルーター
ルーター
インターネット
ポータルサイトシステム
ﾌｧｲﾔｳｫｰﾙ
ルーター
ｳｪﾌﾞ サーﾊﾞ
横断検索兼 ﾎﾟｰﾀﾙｻｲﾄ サーﾊﾞ
実証実験 参加関係者
(ID/PWD で管理)

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

府省等アーカイブ

（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
（21）岡山県

※府省コンテンツに関しては国立国会図書館が収集したものを利用。
実証実験期間中に限り、コンテンツ利用許諾済。

約1000タイトル (90GB)

企業アーカイブ

（1） ヤフー株式会社
Yahooカテゴリ
（2） 日本電気株式会社
宣伝・広告
（3） 富士通株式会社
ホームページ
世界の車窓から
環境活動
社会貢献活動

※Yahoo NEC、富士通ホームページの情報を利用。
実証実験期間中に限り、コンテンツ利用許諾済。

約220タイトル (10GB)

メタデータ体系（1）

（1）ウェブ資源としてのメタデータ

ダブリン・コア(注)をベースに、国立国会図書館の動向等を加味して体系を定義

	概要	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	
作者	コンテンツの作者、作成機関	Creator
公開者	コンテンツの公開者、公開機関	Publisher
寄与者	コンテンツ作成、公開に係わる寄与者（実証実験では利用せず）	Contributor
説明	コンテンツの内容に関する説明	Description
分類	NDC（日本十進分類)。第三区分までを記述 。	Subject
公開日	コンテンツがウェブ上で公開された日付	Date
資源タイプ	コンテンツのタイプ（独自に規定）。「白書・報告書」「統計資料」等	Type
フォーマット	コンテンツのフォーマット(DCMI) text-html、application-pdfなど。	Format
識別子	URL	Identifier
情報源	コンテンツの情報源（実証実験では利用せず）	Source
言語	コンテンツの主要な記述言語	Language
時間的・空間的範囲	コンテンツの場所（緯度、経度）	Coverage
関係	参照先コンテンツとの関係（実証実験では利用せず）。「Is Part of」など。	Relation
権利	コンテンツの権利	Rights

（注）ダブリン・コア(Dublin Core)
ウェブ上のリソースを記述する共通のメタデータ標準などを開発、促進する組織であるDublin Core Metadate Initiative で提唱している、メタデータ記述に使う語彙の通称。略称DC。その語彙が共通の認識となるように、慎重な設計がされた基本語彙セットおよびそれらをサポートするメタデータ語彙が公開されている。

メタデータ体系（2）

（2）アーカイブに関するメタデータ

OAIS(注)、国立国会図書館等の動向等に基づき、体系を定義

メタデータ	概要	種別
収集範囲（ドメイン）	収集対象とするドメインの指定	収集に関する項目
収集範囲（ディレクトリ）	収集起点配下の、対象とするディレクトリ	
収集の深さ	リンクをたどる深さ（階層）	
再収集頻度	再収集を行う（更新有無をチェックする）頻度	
識別子	収集したコンテンツに付与するユニークなID	保存に関する項目
収集日	収集を行なった日付。再収集を行った場合には収集毎に付与。	
ハッシュ値（MD5）	収集体の特徴量としてのハッシュ値	
利用条件 (注)	利用者への提供・公開に関する条件(注)	提供に関する項目

（注）OAIS（Open Archival Information System ）

電子情報の保存システムモデル

（注）利用条件

誰が、いつ、どこで、何を、何の目的で、どのようにアーカイブを利用できるか等の条件 （実証実験では利用していない）

Ⅲ　成果公開デモ

成果公開デモ手順

本実証実験では、下記の５項目の実証と調査研究を実施中

実証項目

（１） 異なるアーカイブ間のリンクナビゲーション機能の開発
- ・【デモ】異なるアーカイブ間のリンク
- ・【デモ】異なるコンテンツ間のリンク

（２） 更新頻度と更新箇所の効果的な閲覧インターフェイスの開発
- ・【デモ】ページ更新頻度の可視化
- ・【デモ】ページ更新箇所の強調

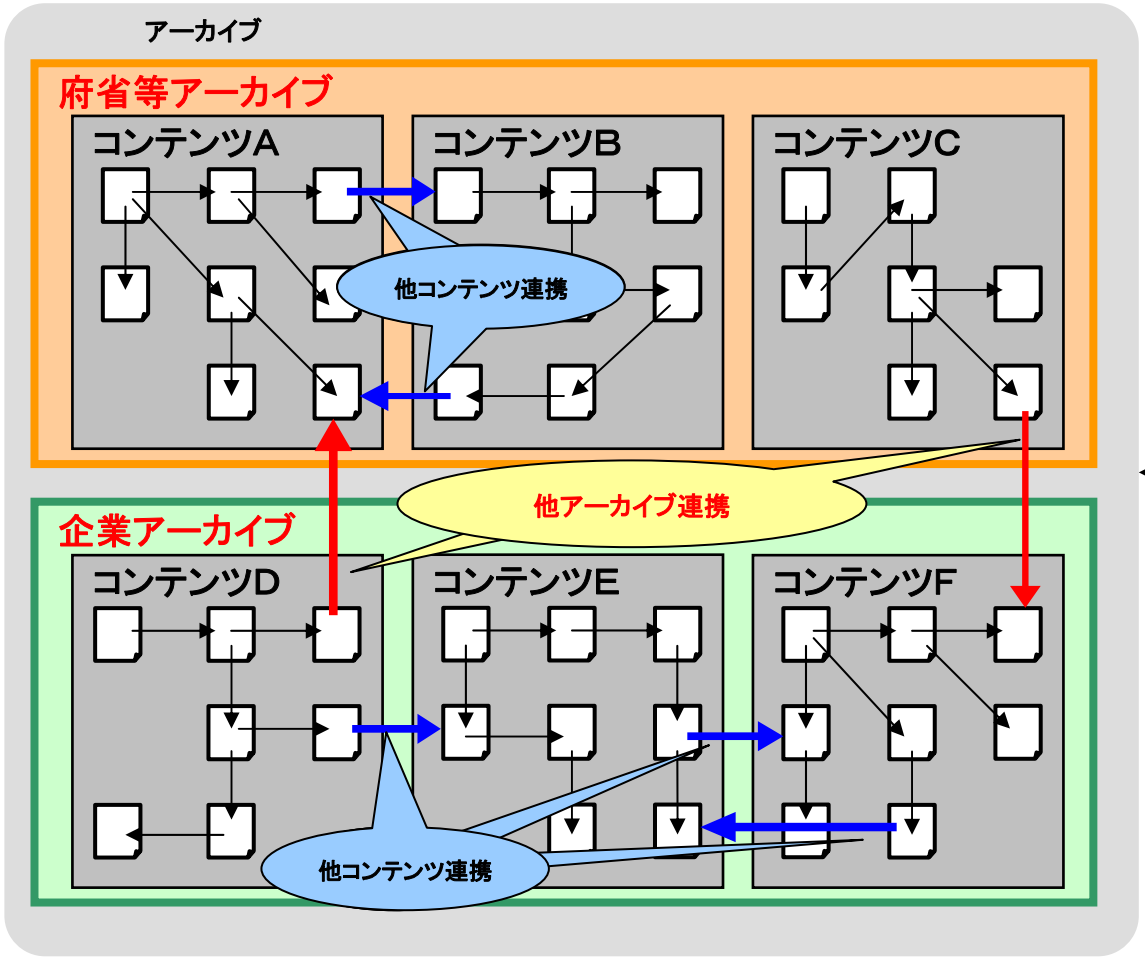
（３） 収集条件の自動検出技術の開発
- ・【デモ】ウェブからの収集条件（起点URL・収集範囲・深さ・除外URL）検出

（４） メタデータの自動抽出技術の開発
- ・【デモ】収集したコンテンツからのメタデータ抽出

（５） ウェブ情報アーカイブに係るメタデータ体系のさらなる検証・確立
- ・【デモ】メタデータ記述ツール

実証項目(1) 異なるアーカイブ間のリンクナビゲーション機能の開発

**実ウェブ上で実現されていたハイパーリンクを、アーカイブ内の「コンテンツ」間でも再現。
また、異なるアーカイブ間のハイパーリンクも可能。
各アーカイブのメタデータを収集して構築する「統合メタデータデータベース」の検索により実現。**



→:コンテンツ内のリンク、 →:他コンテンツにリンク、 →:他アーカイブのコンテンツにリンク

例. 他コンテンツ連携メッセージ

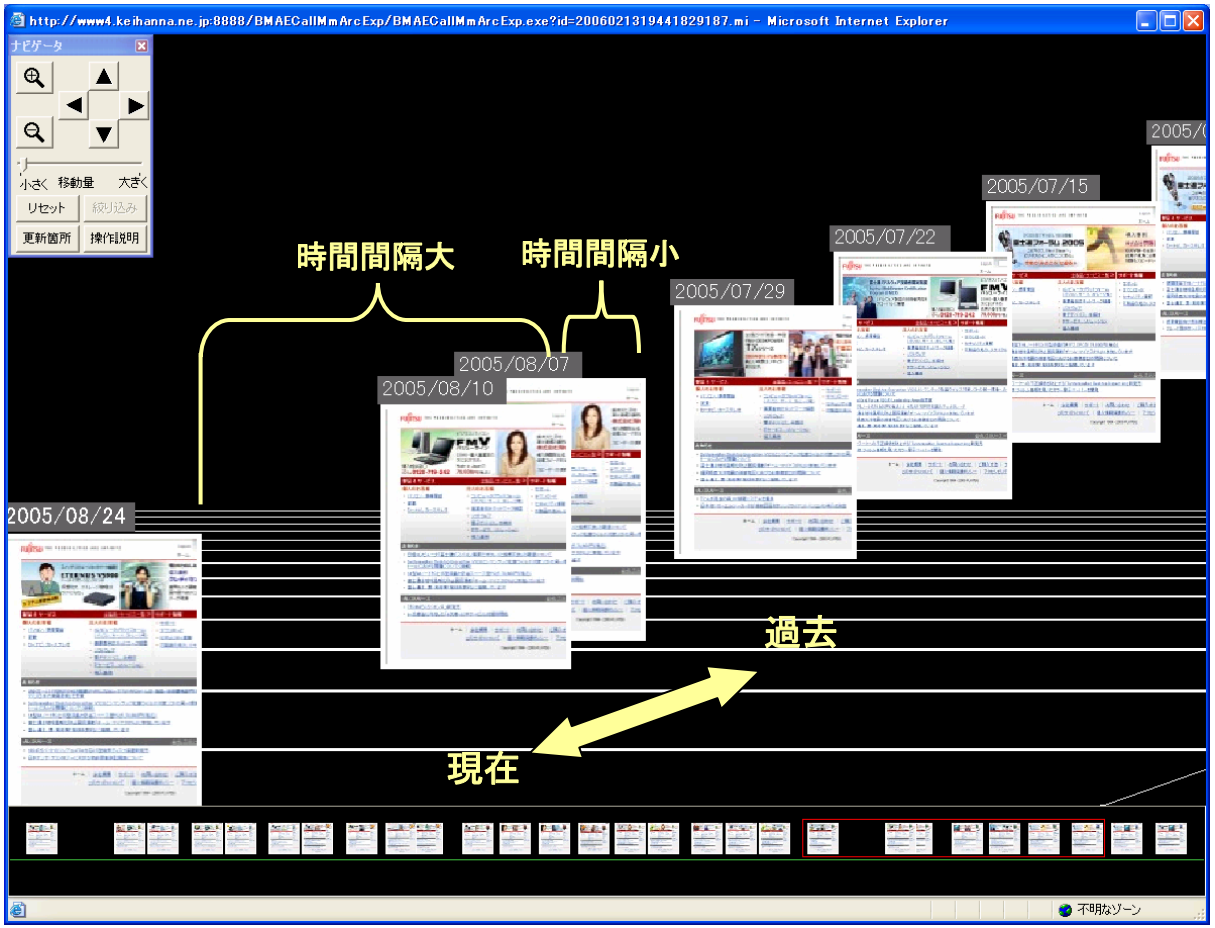
利用者:アーカイブ本文の閲覧

例. 他アーカイブ連携メッセージ

実証項目(2)更新頻度と更新箇所の効果的な閲覧インターフェイスの開発(1)

・ページ更新頻度の可視化
サムネイル間の距離で時間の長短を表現することにより直感的にページ更新頻度を把握。

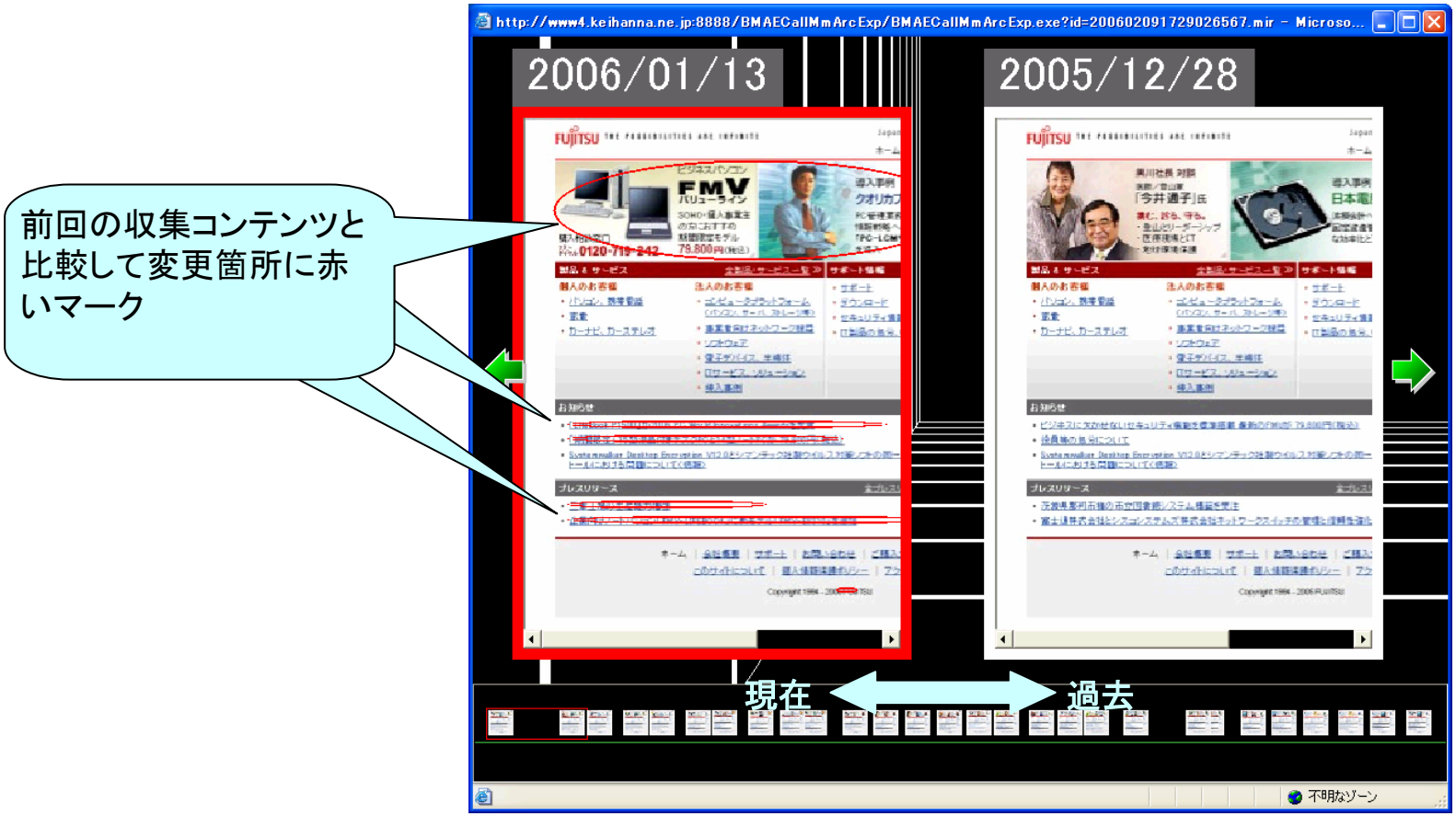
サムネイル表示(更新頻度表示)



実証項目(2)更新頻度と更新箇所の効果的な閲覧インターフェイスの開発(2)

・ページ更新箇所の強調表示
時間の前後でウェブページ内容を比べ、更新箇所を抽出することで、更新箇所を確認。

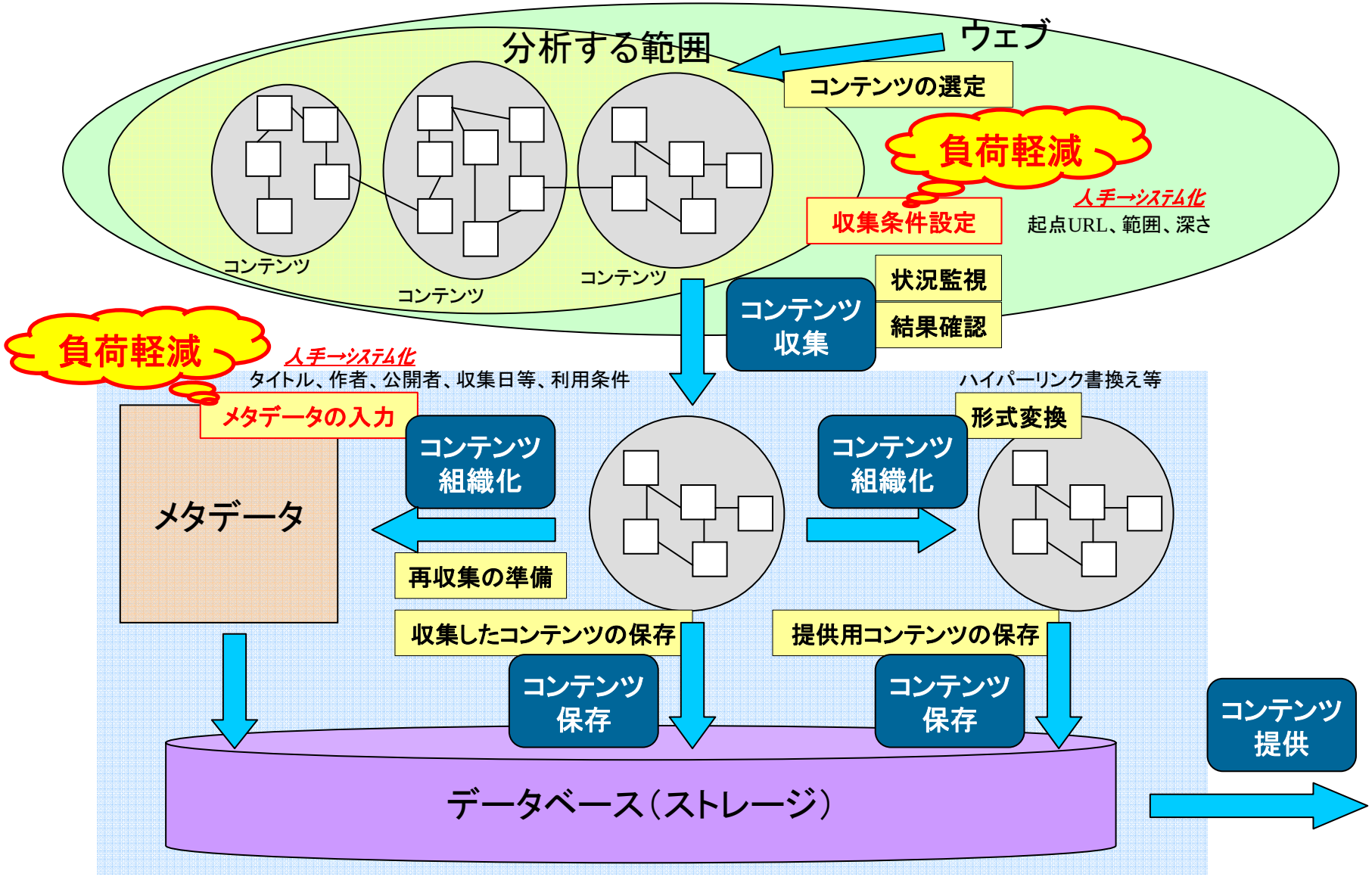
サムネイル表示(更新箇所表示)



平成16年度開発の3次元表示ソフトのUIの改良も実施

業務システム

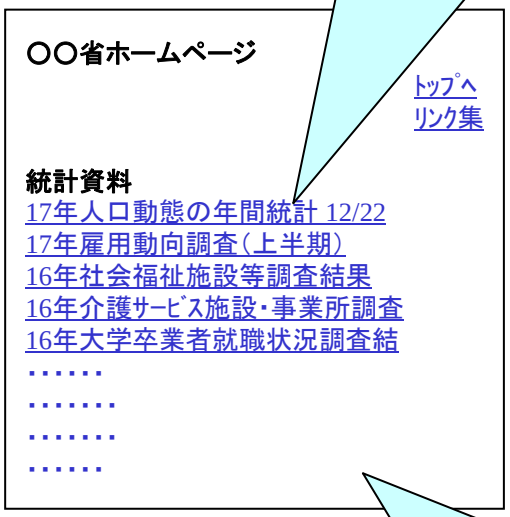
コンテンツ収集とコンテンツ組織化の一部をシステム化することにより、作業負荷を軽減。



実証項目(3)収集条件の自動検出技術の開発

収集条件として個々の情報の起点URL、深さ、除外URLを検出。

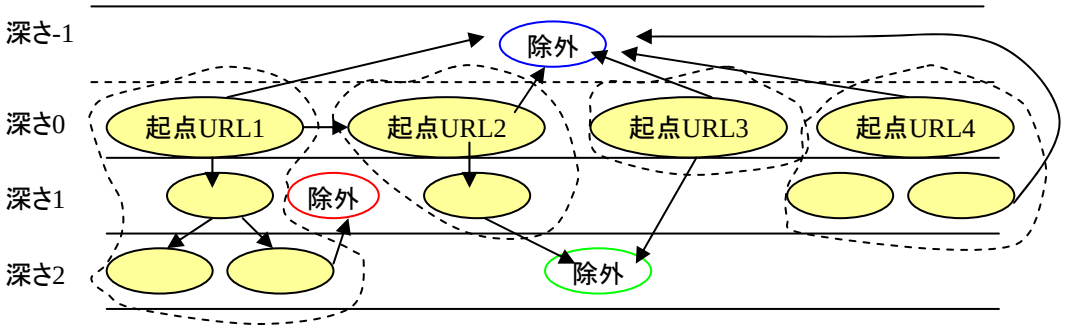
個々の情報を検出し、その起点URL、深さ、除外すべき範囲を自動判定（候補提示）



HTMLページの解析結果、下位のファイル群のディレクトリ構造等をもとに判定

コンテンツリストの例

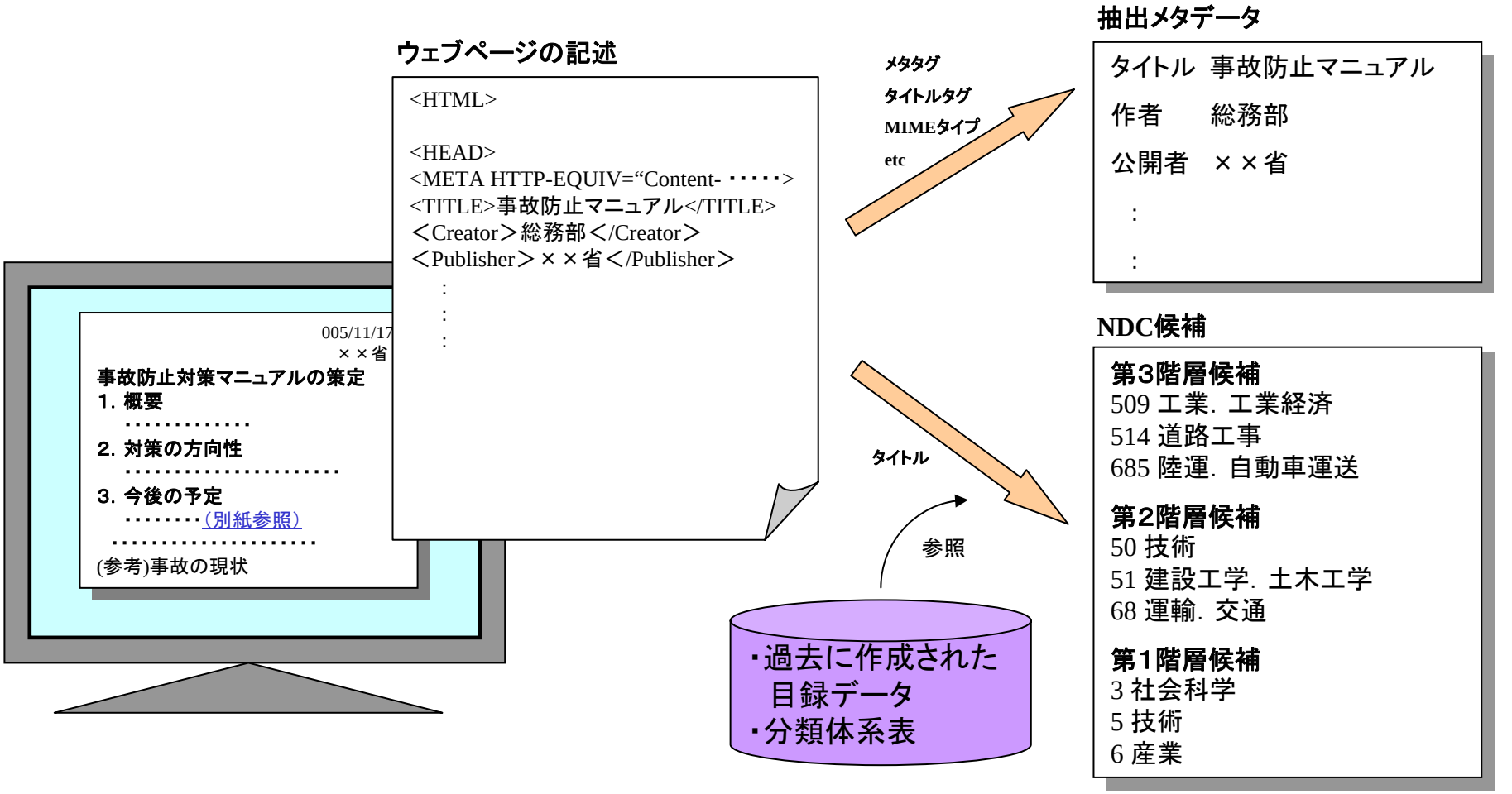
	コンテンツタイトル	起点候補URL	深さ	更新日時	除外URL
1	17年人口動態の年間統計 12/22	http://www.ooo/toukei/.....	1	2005/12/22	3件
2	17年雇用動向調査（上半期）	http://www.ooo/koyou/.....	3	2006/01/06	12件
3	16年社会福祉施設等調査結果	http://www.ooo/fukushi/...../	2	2006/01/13	0件
4	16年介護ｻｰﾋﾞｽ施設・事業所調査	http://www.ooo/kaigo/.....	2	2006/01/27	0件
5	16年大学卒業者就職状況調査結	http://www.ooo/syuusyoku	1	2006/02/03	1件



※点線の範囲が収集対象

実証項目(4)メタデータの自動抽出技術の開発

・メタタグ、TITLEタグ、またMIMEタイプ等コンテンツ内容・属性からメタデータを抽出
・人的に作成された過去の目録情報を経験値として、NDCの候補を提示

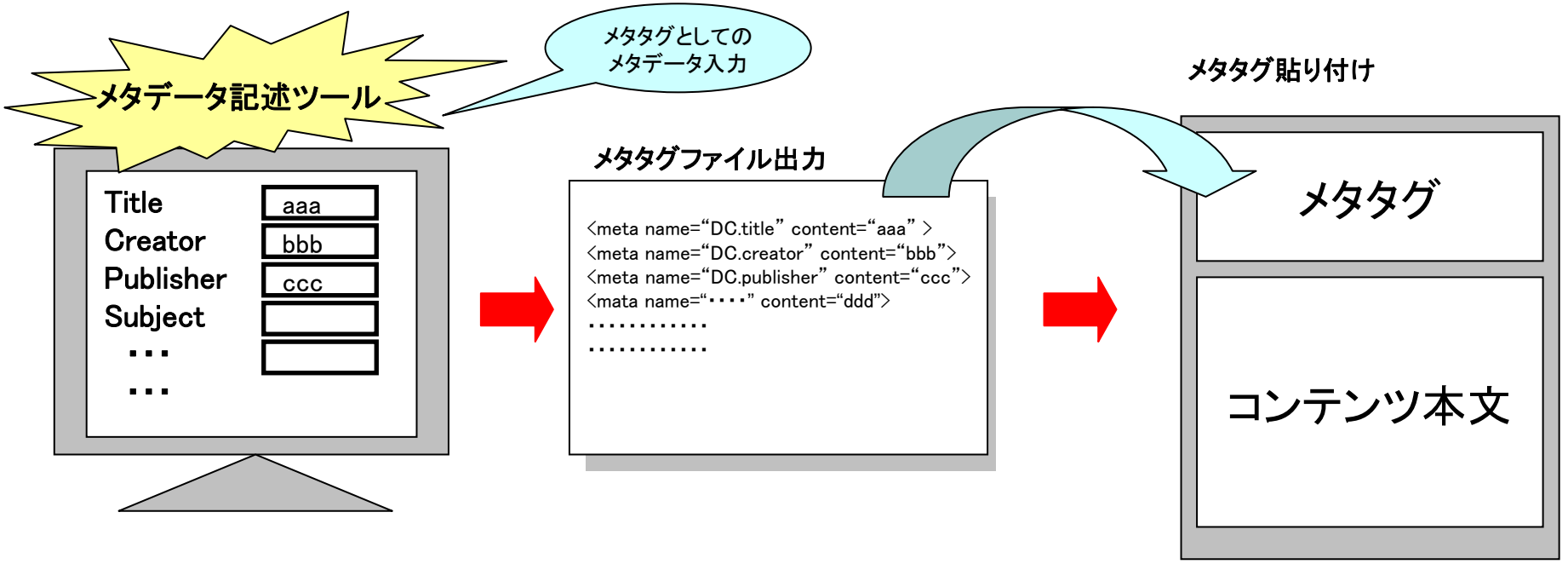


※NDCとは日本十進分類法のことで本の内容を3桁の数字で表現することを基本とし、“000”から“999”までの1,000項目に分類 します。

※MIME（マイム）とは、RFC（インターネットに関する技術の標準を定める団体であるIETFが正式に発行する文書 ）で定義されたメッセージ形式を拡張する標準仕様です。

実証項目(5)ウェブ情報アーカイブに係わるメタデータ体系のさらなる確立

・コンテンツ制作者自身が、コンテンツ内にメタタグとして埋め込むべきメタデータの記述ツールを開発
・ダブリン・コアを基本とするメタデータ体系
・利用条件としてクリエイティブ・コモンズに対応
・ロボット収集に関する許可条件にも対応



画面遷移

ログイン画面

ポータルトップ

利用案内

FAQ

実施概要

マルチメディア アーカイブ エクスプローラ

府省等アーカイブ検索

検索結果一覧

集中型横断検索

企業アーカイブ検索

検索結果一覧

全文検索

府省一覧

分散型横断検索

検索結果（サムネイル表示）

全文検索

企業一覧

書誌情報（サムネイル表示）