A4-2

XMLを用いた再構築可能な漢字文献データモデル

石川正敏★, 波多野賢治†, 天笠俊之†, 吉川正俊†‡, 植村俊亮†, 勝村哲也★

★島根県立大学 総合政策学部
†奈良先端科学技術大学院大学 情報科学研究科
‡国立情報学研究所 ソフトウェア研究系

DEWS2002, 岡山県倉敷市, 2002年3月5日

研究の背景

図書館などでは，保管している古典的な文献を電子化し，電子図書館などの形態で，インターネット上で公開することが増えてきた.

［電子図書館のサービス］

- 文献検索
- 文献閲覧

[利用者による文献の利用]

- 文献検索
- 文献閲覧
- 文献の一部を二次利用（引用など）

目的

利用者による電子図書館などで公開される文献の二次利用を可能にする.

- 文献の検索
  - 文字列検索，文献の見た目を利用した検索
- 文献の閲覧
  - 元の漢字文献の見た目を計算機上で再現
- 文献の一部の二次利用
  - 切抜き，貼り付け

漢字文献

日本・中国・韓国などの東アジア圏の文献

［画像情報］

山部赤人
田子の浦に
うち出でて見れば
白妙の
富士の高嶺に
雪はふりつつ

［テキスト情報］

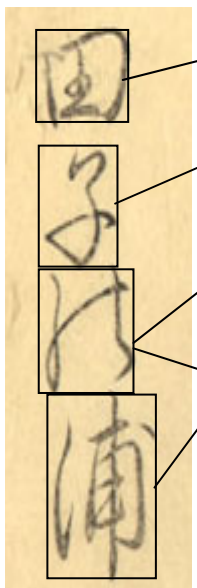
画像情報：符号化文字集合にない文字の表現が可能.
テキスト情報： 文字列検索などによる文献検索が可能.

→ 画像情報と関連するテキスト情報を一つXML文書として記述

漢字文献データモデルの構成

書誌情報（Title: 初日の富士，Subject: 百人一首，ID: XXX ）

表示情報



対応表 a

(位置情報)

内容情報 a

HEADER（書き下し文）

田子<外字>1</外字>浦・・・

BODY

対応表 b

(位置情報)

内容情報 b

HEADER（漢字属性情報リスト）

＜漢字情報＞ID:1，発音: の 関連情報: “の”の異体字＜漢字情報＞

・・・

BODY

漢字文献の検索

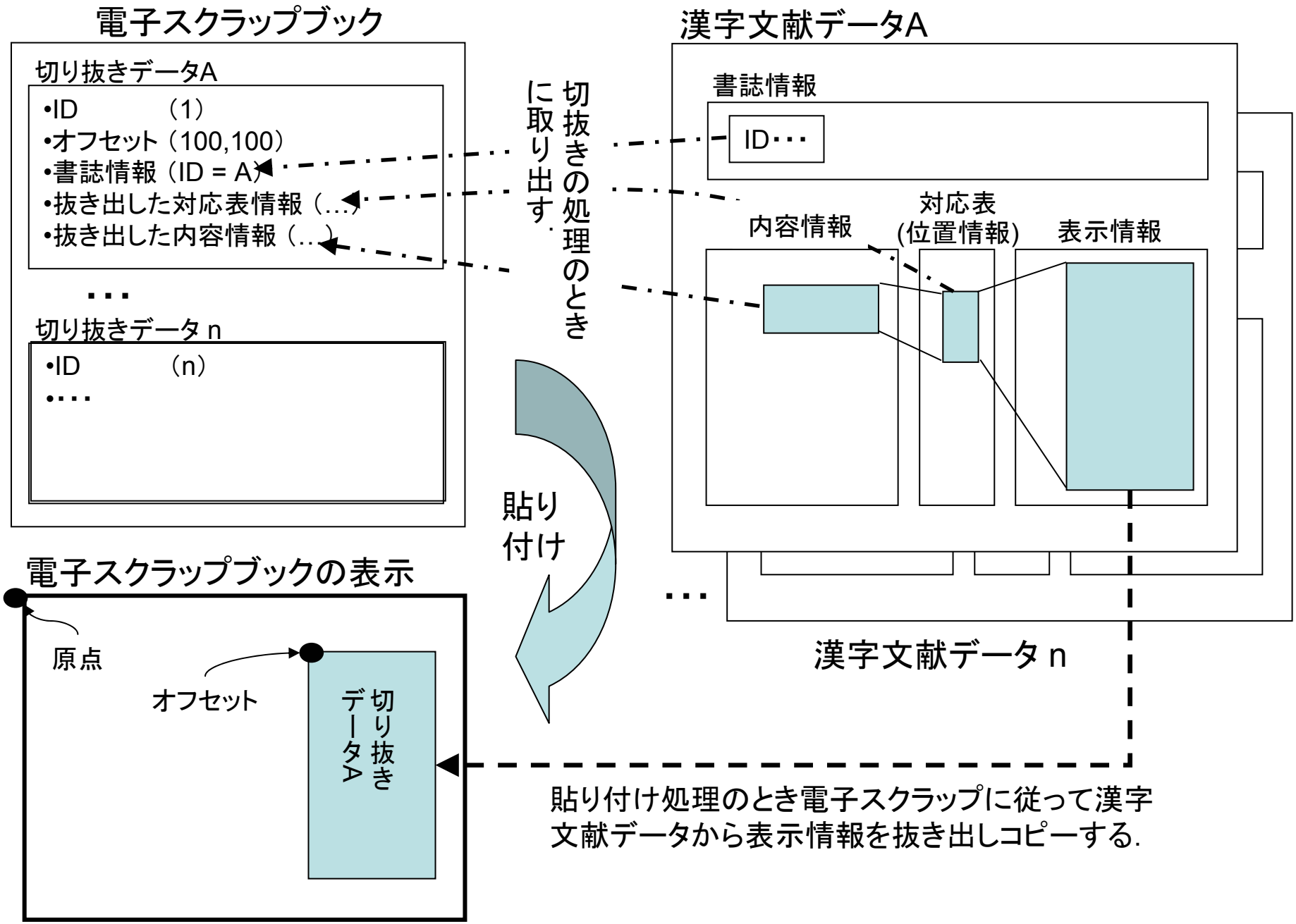
- 内容情報に対する検索
  - 符号化文字集合だけで構成された検索キー
  - 外字情報を含む検索キー
- 表示情報に対する検索
  - 表示情報の領域に着目した検索
- 内容情報，表示情報への検索の組み合わせ

電子スクラップブック

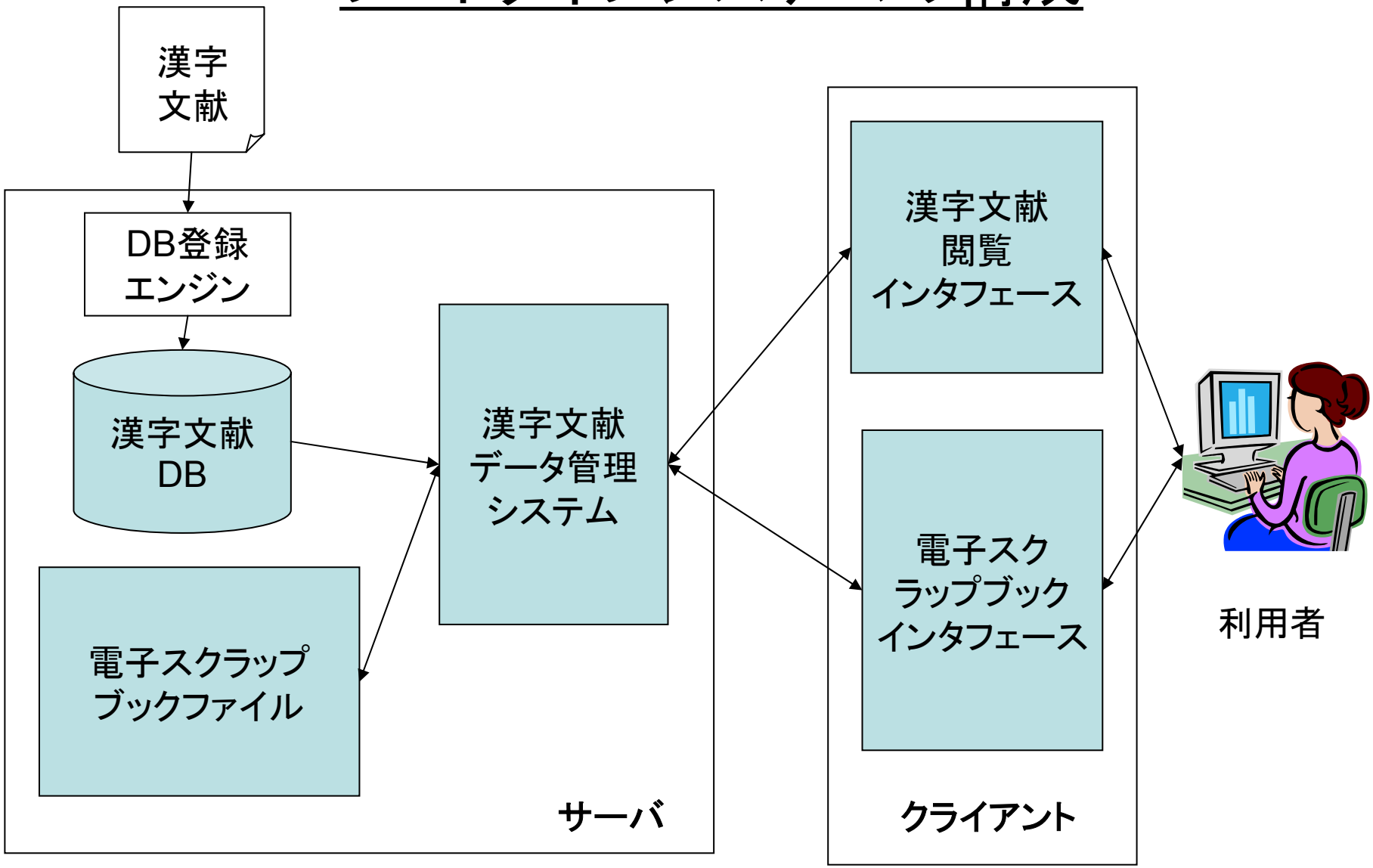
- 電子スクラップブック
  - 電子スクラップの集まり.
- 電子スクラップ
  - 漢字文献データから，利用者の必要とする情報を抜き出したもの.
- 操作
  - 切抜き

    漢字文献データから情報を取り出し電子スクラップブックに登録
  - 貼り付け

    電子スクラップブックに登録された電子スクラップに表示情報を表示


電子スクラップブック
切り抜きデータA
•ID （1）
•オフセット（100,100）
•書誌情報（ID = A）
•抜き出した対応表情報（…）
•抜き出した内容情報（…）
・・・
切り抜きデータ n
•ID （n）
•・・・
切抜きの処理のとき
に取り出す.
漢字文献データA
書誌情報
ID・・・
内容情報
対応表
(位置情報)
表示情報
・・・
漢字文献データ n
貼り
付け
電子スクラップブックの表示
原点
オフセット
切り抜き
データA
貼り付け処理のとき電子スクラップに従って漢字
文献データから表示情報を抜き出しコピーする.

プロトタイプシステムの構成



実行例(漢字文献データの閲覧)

実行例(電子スクラップブック)

まとめ

- 漢字文献データモデルの提案
  - 元の漢字文献の見た目を計算機上で再現
  - 文字列による内容情報への検索
- 電子スクラップブックの提案
  - 漢字文献からの任意の部分の再利用が可能.
- プロトタイプシステムの実装

今後の課題

- 外字を含む検索キーによる検索などの検索機能の充実
- 電子スクラップブックのメモ機能の追加
- 評価法