IIIFの導入による木簡画像データベースの連携強化

1　はじめに

奈良文化財研究所と東京大学史料編纂所（以下編纂所と略称）は、2009年に木簡字典データベース（開発当時）と電子くずし字字典データベースの連携検索システムを開発・公開し、高い評価を得た。その後も、文字画像入力によって類似文字画像を検索するシステム「MOJIZO」の開発・導入など、発展を続けている。しかしながら、この連携検索システムは、奈文研と編纂所の２機関の間でのみ運用が可能であり、他の機関に開かれたシステムではない。一方今日では、多くの機関がデータを蓄積・提供するとともに、共同研究を展開するようになっている。

そこで、編纂所との連携検索システム開発・公開の実績を踏まえつつ、近年の研究動向に対応して、連携の輪を国内外にさらに広げることを目指して、新たな連携検索システムの開発を計画した。

2　連携検索システムのコンセプト

奈文研と、従来から連携関係にある編纂所、共同研究を行っている国立国語研究所（以下国語研と略称）、また文字画像データの公開実績がある国文学研究資料館（以下国文研と略称）で、連携検索の方向性・コンセプトを協議した。この中で、次の３つの考え方が、重要な目標となると認識された。

A　開かれている　　データ提供側＝参加機関にも、利用者にも「開かれている」ことが重要であると考えた。参加機関に開かれているとは、具体的には当初の開発・参加機関以外の機関が、後から参加する場合にも、容易に連携できることである。 利用者に開かれているとは、具体的には利用しやすく、かつデータがオープンデータであることが該当する。

B　対等である　　参加機関の中で、中核となる機関が他機関のデータを吸い上げる体制ではなく、相互に「対等である」ことが重要だと考えた。対等でない場合、相互の「力関係」などの軋轢が生じやすい。参加機関が対等であることは極めて重要である。

C　継続的である　　一過性の開発ではなく、長期にわたって運用することが、研究インフラとして重要であると考えた。

以上の点をクリアするためには、連携専用のツール等を開発・提供するよりも、極力汎用性の高い方法を組み合わせることが妥当だと判断した。専用ツール・システムの開発は、短期的には利点も多いが、他機関の新規参入や継続的な維持には障害となる可能性が高い。

そこで、「連携検索システム開発」ではなく、「共通検索を可能にするフレームワークの構築」を目指すこととした。このコンセプト・方向性に基づき、従来より奈文研と木簡・簡牘に関する共同研究を推進している台湾中央研究院歴史語言研究所（以下史語所と略称）に連携検索および開発・研究への参加を呼びかけ、快諾を得た。

3　フレームワークの構築

史語所は、人文学の優れたデータベースを開発・公開しており、データベース開発の技術は高く、ノウハウも蓄積している。上記の国内４機関の他、史語所および情報系の研究者も交えた研究会を２回開催した。

連携検索には、近年人文情報学分野で国際的潮流となりつつあるIIIF（International Image Interoperability Framework：デジタルアーカイブの公開・共有のための国際的な枠組み）に準拠したデータを用いることが妥当だと考えた。ただし、IIIFにはデータ作成時点を含めて多くの機能が用意されているが、連携検索にはpresentation機能のみを利用することとした。すなわち、各機関がデータを構築する方法や個別のDBでのデータの持ち方まで規制する訳ではなく、あくまで連携検索用にIIIF準拠のデータを用意するというものである。

また、IIIF化したデータやサーバはそれぞれデータを保有する機関が用意することとした。参加機関の独自性・自主性を担保し、対等な連携を継続的に維持する上で重要だと思われる。また、各機関でのデータベース更新がシームレスに反映されることも期待される。

各機関の連携検索用データには、それぞれマニュフェストを公開する。これらをAPIを用いて横断検索して、連携検索の実現を図るという計画を立てた（図35）。

さて、多様な検索が可能になることは、ユーザにとって有意義と思われる。一方、ユーザごとのニーズは極めて多様であり、それらの要望に広く応えようとすると、

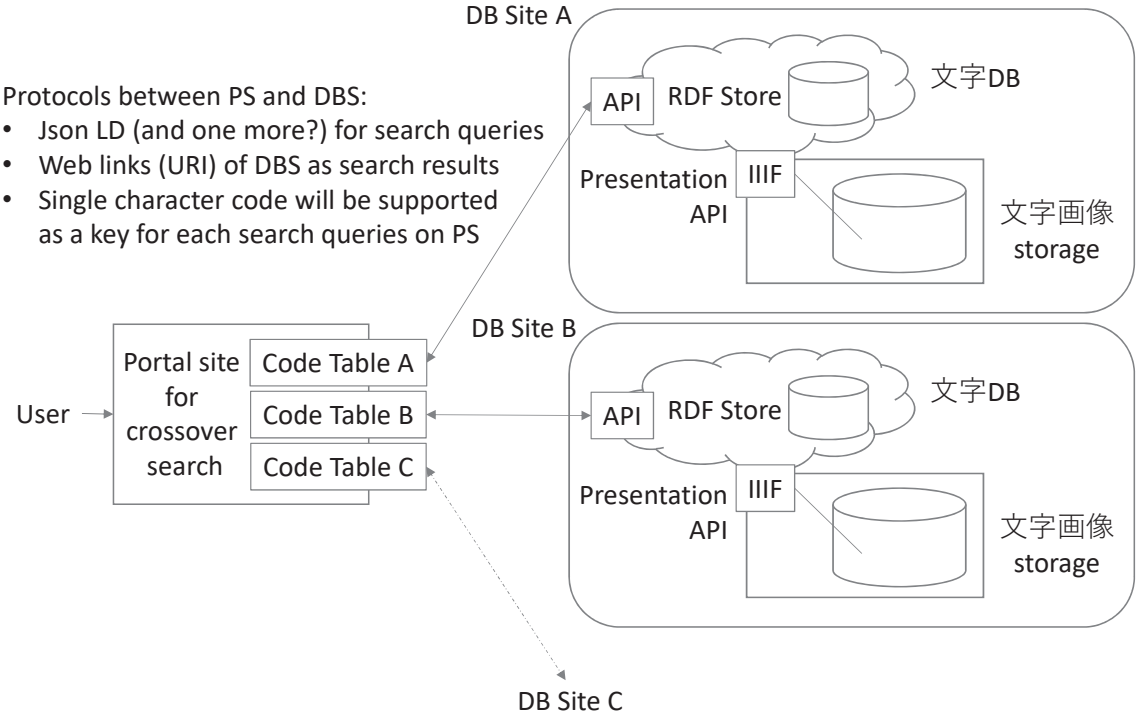

図35　本フレームワークによる連携検索の実現イメージ

データ項目も膨大になり、検索方法も複雑になる。結果的に、データの拡大や容易な検索を阻害する要因にもなりかねない。開かれたデータというコンセプトから、こうした状況は望ましくないと考えた。

そこで、連携検索の焦点を文字コードによる単文字画像の検索に絞り込むこととした。詳細で多様な情報は、リンクによってデータ提供機関のデータベースと接続して提供する。ユーザの多様なニーズには、連携検索データのオープン化で対応することを目指すことにした。

また、連携検索用データの質・内容に大きな差異が存在することは望ましくない。そこで、データに関する最低限の共通認識・ルールを以下のように整理した。

・各文字画像はIIIFに準拠したデータとする
・各文字画像に関するマニュフェストを公開する
・各文字画像には以下の情報を付与する
　画像提供機関
　各画像の固有ID（機関ごとのルールによる）
　文字コード（unicode）
　出典
・各文字画像には可能であれば以下の情報を付与する
　時空間情報
・各文字画像は以下の条件を満たすものとする
　300dpiまたは１文字あたり300p程度
　creative commmons BYSA　相当　以上
・有効な連携検索のために補足的な情報を整備する
　文字コードマッピングデータ
　連携検索用API

4　特徴と意義

本フレームワークの特徴は次のように評価できる。

①　各機関の独自性を尊重した対等な連携検索

参加各機関がデータを保持・公開し、それを共通に検索する、対等で開かれた枠組み。機関が持つ固有の課題・方向性をそれぞれの機関のデータベース等では維持したまま、連携検索が可能になる。

②　国際的な標準規格に準拠した枠組み

近年、国際標準となっているIIIFの規格に準拠しつつ、漢字画像に合わせてカスタマイズすることで、参加のハードルを下げた。海外の機関の参加も可能な、国際的な枠組みとなった。

③　オープンデータ化も含めた利便性の向上

連携検索の画像等はオープンデータとし、積極的な利活用を促す。連携検索を各機関のデータベースにリンクさせて、詳細な情報を容易に得られるようにする。なお、各機関のデータベースでの知的財産所有権は、各機関の方針により、機関ごとの独自性を尊重する。

この連携検索のフレームワークを提言し、現在参加を表明している機関（奈文研・編纂所・国文研・京都大学人文学研究所・史語所）以外の機関にも参加を広く呼びかけていきたいと考えている。

なお、本稿は科研費基盤（S）「木簡等の研究資源オープンデータ化を通じた参加誘発型研究スキーム確立による知の展開」等の成果を含む。

（馬場　基・高田祐一・桑田訓也）