D-008

インターネット空間多地点観測ヘテロデータシェア向けカタログ生成システムの開発

Development of Catalogue Generation System for Heterogeneous Data from Multiple Observation Sites in Cyberspace

岩田翔汰 † Shota Iwata
中平 勝子 † Katsuko T. Nakahira
北島 宗雄 † Muneo Kitajima

1 はじめに

インターネット空間の観測は，情報格差の分析の精度向上や，情報格差に関する新たな指標の開発に繋がることが期待される．UNESCO は Lena Resolution の中で，言語多様性の保護を訴え，サイバー空間へのユニバーサルアクセスを推奨し，インターネットにおける言語多様性のモニタリングを提案している [1]．インターネット空間の観測は，様々な機関によって行われるため，各機関の観測方針によって指標用データが異なることが予想される．そのため，データ収集頻度や観測対象のデータは多様となりヘテロな様相を示す．

伊藤らは，情報格差分析に利用するクロールデータや統計情報などを 1 次処理したデータを研究者間で流通させるために必要なデータ形式 Information Trade Handling Format（以下，ITHF）および，情報格差を分析するシステムを提案している [2]．また，筆者らは，ITHF を流通させる基盤の設計と共有システムの設計を提案している [3, 4]．この共有システムにおいて，観測されたデータはカタログにより管理されるが，固定フィールドの場合，新規観測に伴うフィールド追加を整合的に行うために複雑なデータベース構造の変更が必要になる．また，観測データの利用目的は多様なため，カタログは単純である必要がある．様々な目的による利用や，新規観測に伴うフィールド追加に対応するため，本稿では，データベース構造の単純化，項目結合を活用した検索機能を付加したユーザインタフェースの実装を行う．

2 多地点観測ヘテロデータ

インターネット空間の観測は，様々な機関によって長期にわたって行われる．各機関が行った部分的な観測結果を統合することで，インターネット空間全体の実体を反映することができる．これまで，観測データは収集方法や管理方法が異なるため，機関ごとに独自に収集 · 管理されていた．それらのデータを，共通フォーマットである ITHF ファイルに変換し，所在情報カタログに集約することで，収集されたデータを他のユーザが再利用することが可能になる．しかし，同じ ITHF であっても，機関によってファイルに含まれるヘッダやデータが異なり，ヘテロな様相を示す．

† 長岡技術科学大学

3 カタログ生成システム

インターネット空間多地点観測ヘテロデータシェア向けカタログを用いて ITHF ファイルへアクセスするには，ITHF ファイルの検索を可能にするため，各項目がデータベース化されることが必要である．しかし，インターネット空間を観測したデータ（クローリングデータ等）は，データサイズが大きく，またデータの所有権の問題があり，1 つのサーバに集約して共有する方法は適さない．したがって，各データのヘッダ情報をカタログとして共有する．

インターネット空間を観測したデータはヘテロな様相を示し，また，新たな観測指標に対応するため，新たな記録方式でデータが収集される可能性がある．そのため，カタログを生成するためのデータベースのフィールド全てをあらかじめ定めることはできない．そこで，筆者らは ITHF 概要ファイルというカタログに必須となる項目を決定し，管理する方法の検討を行った [3]．本稿では，ITHF 概要ファイルに該当する必須項目を固定項目としてデータベースに格納し，それ以外の必須ではない項目を可変項目として，XML 形式でデータベースの 1 フィールドに格納する．

[固定項目] カタログが持つべき情報として，まず，観測日時，観測者，観測機関，観測総量が挙げられる．これらの項目が存在することで，インターネットを観測したデータの質が担保される．次に，持つべき情報として，データの位置情報が挙げられる．インターネット空間を実際の国や地域と関連付けて観測するために必要な項目である．

[可変項目] インターネット空間の利用のされ方は国や地域によって特徴があることが考えられるため，位置情報に加えて国や地域の特性を示す統計データも必要である．また，何を観測したデータなのかということがわからないと，データの分析を行うことができないため，観測する内容に関する項目もカタログに記される必要がある．また，新たな記録方式で観測が行われた場合に記録される基本情報も可変項目として格納する．

これら項目は，全て ITHF ファイルに含まれる HDU のヘッダから取得する．また，ITHF は複数の HDU が含まれる構造であるため，データベースを正規化して簡素にするために，ITHF ファイルの詳細情報や所在地情報を格納するテーブルと，各 HDU の詳細情報を格納するテーブルの 2 つのテーブルを分けて設置する．そして，検索時にはそれらのテーブルを結合して結果を表示するように実装を行う．

4　検索インタフェース

検索を行うことは単なる手段であり，ある目的のために検索を行っているに過ぎない．したがって，インタフェースがシンプルであることが好まれる [5]．本システムも ICT に詳しいユーザ，そうでないユーザの誰もがユーザになるため，キーワード入力フォームのみが存在するシンプルな検索インタフェースと，必須項目以外の様々な情報に対して検索を行うことができる詳細検索インタフェースの 2 つを用意する．検索インタフェースの作成は HTML，Perl，JavaScript（JQuery），CSS を用いて行う．

ICT に詳しくない人も含め幅広いユーザを対象とするため，シンプルな検索インタフェースのキーワード入力フォームは必須項目の内容に対応するキーワード（国名，取得機関名，データ種類等）に柔軟に対応できるようにキーワード修正や検索項目の判別を行う．そして，ユーザが自分の検索に間違いがないか判断するために，検索結果は，該当する HDU を含む ITHF ファイルの必須項目の一覧を即時に表示するようにする．この際，データベースへかかる負荷に留意し，単純なクエリで検索を行うようにシステムを構築する．

情報検索に慣れている ICT に詳しいユーザに対しては，詳細検索用インタフェースを提供する．ある特定の ITHF ファイルを必要とするユーザを想定するため，検索にある程度の時間がかかっても細かなクエリを指定して検索が行えるように実装を行う．詳細検索用インタフェースには，複数の国名を同時に選択して検索できる機能や，ユーザが行いたい分析方法を選択するとそれに対応するデータを持つ ITHF ファイルが検索される機能，検索項目名と要素のペアを入力して必須項目以外の情報を検索する機能を実装する．検索項目名と要素をユーザに入力させることで，ITHF ファイルによって内容が異なるヘテロな情報や，新規観測に伴い追加された項目についての検索を可能とする．

双方の検索においても，検索結果の各 ITHF ファイルの行をクリックすることで，ITHF ファイルのカタログ情報の詳細が表示されるように実装する．そして，カタログ詳細画面から，ITHF ファイルのダウンロードが行えるような機能を設ける．

ITHF ファイルはファイルサイズが大きいため，ダウンロードするという意思決定を慎重に検討することが考えられる．この意思決定の支援となるように，カタログ詳細画面における情報の表示位置を工夫する．まず，最も意思決定に関わると考えられる必須項目の情報を最も上部に表示する．次に，XML として格納されている基本情報を表示する．このとき，ダウンロードの意思決定に関わると考えられる，観測データ取得に関わるデータを上方に表示し，続いて，それ以外の各 HDU に関する統計情報等の項目を表示する（図 1）．ITHF ファイル内の HDU の数によっては，画面に表示する項目が多く見づらくなってしまうため，基本情報の項目は，HDU ごとにボタンで切り替えて表示する．基本情報の項目は HDU の種類ごとに異なり，また，新規項目が追加されることがあるため，カタログ詳細画面に表示する項目や順序についてはシステム管理者が事前に定める必要がある．

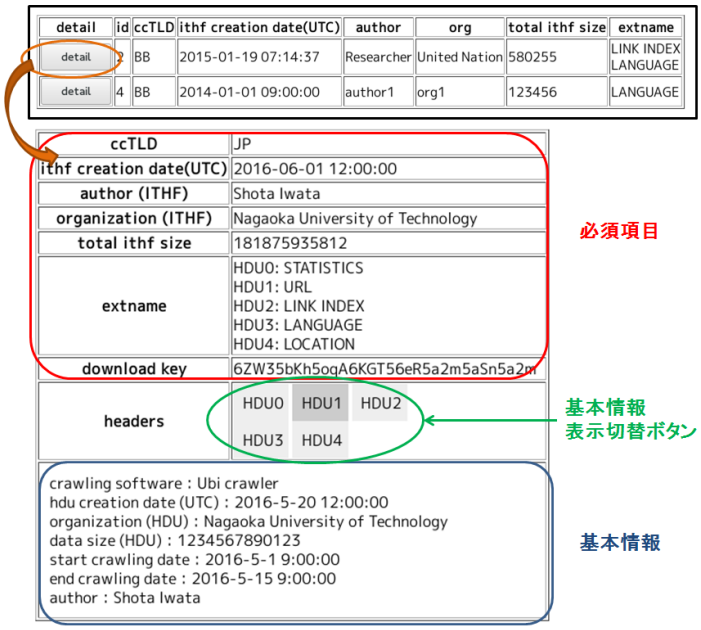

図 1　カタログ詳細画面実装例

5　まとめと今後の課題

機関によって設定状況が異なるヘッダ項目や，新規観測に伴う新たな項目へ対応が可能な単純化されたデータベースの設計を行った．そして，インターネット空間多地点観測ヘテロデータシェア向けカタログ生成システムの提案を行った．

今後は，実際の運用状況を想定したデータベースを作成し，必須項目やそれ以外の XML 部分の検索にかかる時間を測定し，システムの性能の検証を行う必要がある．そして，多くのユーザに対して使いやすいシステムになるようにデザインや操作性を工夫し，ユーザビリティの高いシステムを作っていく必要がある．また，インターネット空間の観測を情報格差の研究に役立てるために，本システムの多言語を行うことも今後の課題である．

参考文献

[1] UNESCO:INTERNATUONAL MEETINGS ON MULTILINGUALISM IN CYBERSPACE 2008-2014 FINAL DOCUMENTS

[2] 伊藤公，中平勝子，三上喜貴：大容量データ流通のためのファイルシステムの開発・評価，情報処理学会第 76 回全国大会，第 4 分冊，pp.805-806，2015

[3] 岩田翔汰，中平勝子，北島宗雄：サービス利用状況に着目した情報格差観測データ流通基盤の設計，第 14 回情報科学技術フォーラム，第 2 分冊，pp.171-174，2015

[4] 岩田翔汰，中平勝子，北島宗雄：情報格差多地点観測ヘテロデータ共有システムにおけるデータ取得手順，情報処理学会第 78 回全国大会，第 1 分冊，pp.647-648，2016

[5] Marti A. Hearst(著)，角谷和俊，田中克己（訳）：情報検索のためのユーザインタフェース，共立出版株式会社，2011