平成27年度学術情報システム総合ワークショップ（2015/6/26）

CATデータの分析手法

一橋大学 学術・図書部 学術情報課
藤井　眞樹

目次

- １．自己紹介
- ２．CATのデータ
- ３．平成25年度のデータ分析
- ４．データ分析の方法
- ５．何を知りたいか，どう分析するか

１．自己紹介

業務経歴

- 最初（５年）

  大学図書館で目録担当（和洋図書雑誌何でも）

- NACSIS～NII（16年!）

  研究者ディレクトリ，広報，国際事業，研修事業

  最後の５年間　NACSIS-CAT/ILL

- 今（２年目）

  再び，大学図書館で目録担当

スキル

文学部卒。大学時代はワープロ。就職当初は汎用機時代…

- UNIX ？
- Perl ？
- スクリプト？
- Excel ？
- FileMaker Pro

- 最後？の手段・・・

2. CATのデータ

2. CATのデータ

WebUIPで見るデータ

個別版のデータ（CATPフォーマット）

hipai （pai形式）のデータ

別紙1参照…

hipaiのデータ

- hipai　日立製のCAT/ILLのデータをテキストファイルで取り出すツール
- BOOK（図書書誌）, SERIAL（雑誌書誌）, BHOLD（図書所蔵）, SHOLD（雑誌所蔵）, NAME（著者名典拠）, TITLE（統一書名典拠）などから出力可能。
- 参照ファイルのデータも出力可能。

  JPMARC, TRCMARC, USMARC, UKMARC, DNMARC ほか。
- パラメタファイルで条件設定をして出力。

3. 平成25年度のデータ分析

3．平成25年度のデータ分析

目録システムの将来への提案のために，CATの現状を知る。

- 結果は，「これからの学術情報システム構築検討委員会」の資料などに使われている。

3. 平成25年度のデータ分析

1. SOUCE=ORGの書誌データについての調査
2. BB書誌データについての調査
   i. 書誌数，VOL数，ISBN数，タイトル言語，
   ii. SOUCE=ORGの年代別件数，親書誌有の割合
   iii. SOUCE=ORGのUTL有の割合
   iv. SOUCE別件数
   v. SOUCE=ORGの作成館別件数
3. 著者名典拠についての調査

別紙2＆表1～4参照

4. データ分析の方法

大まかな手順 （paiデータを使う場合）

- CATのデータベースから，必要な項目だけを抜き出す。
  - hipaiでデータを抽出する時に必要項目だけを指定する。
  - paigrepで条件指定して必要なデータのみを抜き出す。
  - 全件出力したpaiデータから，grepで必要な項目だけ抜き出す。
  - Perlで必要な項目だけ抜き出す。

大まかな手順２　（paiデータを使う場合）

- １レコード１行にデータを成形して，エクセルにデータを突っ込む。

  ⇒フィルタで件数カウント

- UNIX側で必要な項目だけにする

  ⇒件数カウントしたり，diffで差分を出したり。

- FileMaker Pro にデータを突っ込んで，他のデータを掛け合わせる。

などなど，方法はいろいろ。

手を動かしてみよう（実習）

- 材料

  BOOK_BB.pai

  ＝書誌IDが<BB********>の書誌データ

  BHOLD_BB.pai

  ＝書誌IDが<BB********>の所蔵データ

- ツール

  Windowsのコマンド

  Perl

  Excel，Access

  その他，何でも（自分の使いたいものでOK）

何をする？

- とりあえずお手軽に。
  - BB書誌のSOUCE別作成館一覧
  - BB書誌の年代別件数一覧
  - BB書誌の言語別SOUCEの割合
  - BB書誌の所蔵館数の割合　　ほか。

5. 何を知りたいか，どう分析するか

たとえば，外部データの活用 に向けて

- 参照ファイル由来の書誌の割合（言語別，年代別，etc.）
- 参照ファイルの書誌と，BOOKへ取り込まれた書誌の違い（TRの取り方，書誌構造，etc.）
- 参照ファイルとして使用できるようになってから，どのぐらいの日数でBOOKへ取り込まれているのか。
- 参照ファイル由来の書誌には，どのタイミングで所蔵がついているのか。

たとえば，目録センター館制度 に向けて

- オリジナル書誌の作成館，言語別に，年代別に，etc.
- オリジナル書誌にどのぐらいの所蔵がついているのか
- オリジナル書誌への所蔵登録数が多いが作成はしていない所蔵館は？

他に何が使えるか？

- コード類ではなく，データの中身を比較検討するなら，個別版データを使う方がよいのでは？
- CATのログから何か見えてこないか？

  昨年のWSでの講義「学術情報サービスのログ分析」も参考にしては？