機関リポジトリログの標準処理・解析結果表示システムの構築

2015/06/12 NII学術情報基盤オープンフォーラム2015
機関リポジトリ推進委員会/技術WG　ログ・サブグループ
慶應義塾大学メディアセンター本部五十嵐健一

メンバー：
Takurou Kawamura*1, Kenichi Igarashi*2, Hiroshi Kato*3, Akira Maeda*3,
Toshihiro Aoyama*4, Kazutsuna Yamaji*3, Sho Sato*5
*1Hiroshima University *2Keio University *3National Institute of Informatics
*4National Institute of Technology, Suzuka College *5Doshisha University

ログ・サブグループとしての作業内容

- 機関リポジトリ・アクセスログの整理ために
- 機関リポジトリ・アクセスログから知るために

おっと、その前に

- ROATプロジェクト
  http://www.ll.chiba-u.jp/roat/

【業務の目的及び内容】

ROAT (Repositry Output Assessmennt Tool) を構築・運用してきた成果を前提として，従来の統計処理の妥当性の検証と改良方策の提案，多様な分析を可能とするためのレコード処理機能の検証およびシステム更新，カウント方法の標準化に関する国際連携の推進，ROATの利用促進ならびに利用機関に対する技術支援などを通じて，機関リポジトリへのアクセスの統計処理に関する標準的な方法の確立を目的とする。

（「機関リポジトリアウトプット評価の標準化と高度化」プロジェクトより）

機関リポジトリ
アクセスログの整理ために

機関リポジトリ
アクセスログの整理ために

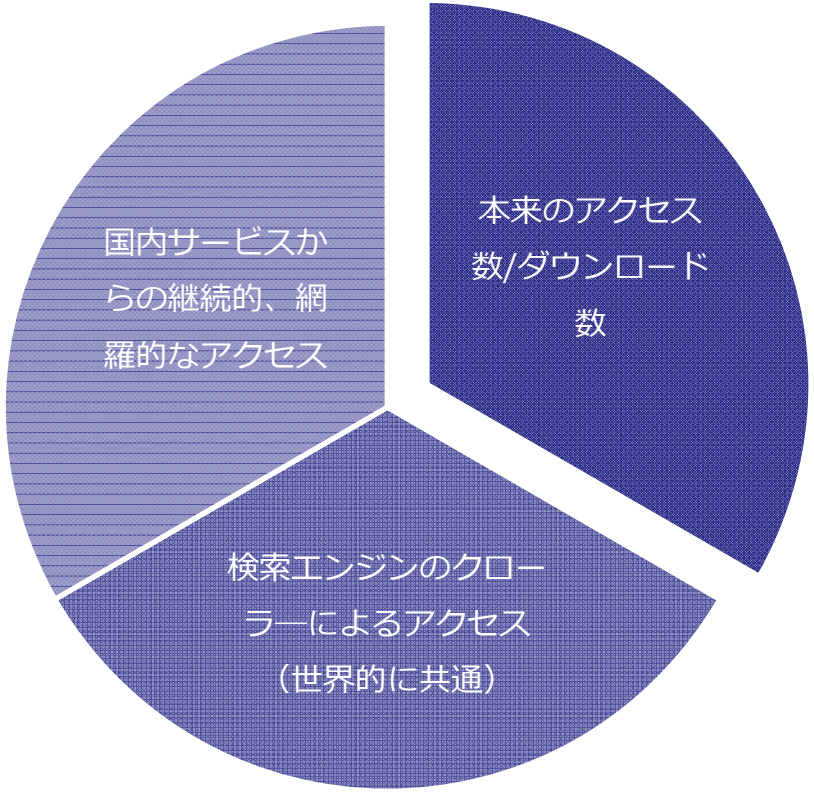
- “アクセス数”、“ダウンロード数”

今、各種統計に出す際に“アクセス数”や“ダウンロード数”って、どうやって出していますか？？

機関リポジトリとしての基準って？？

機関リポジトリ
アクセスログの整理ために

Webサーバへのアクセスログの内訳



機関リポジトリ
アクセスログの整理ために

- ロボットのリストとは？
  - 検索エンジンとしてアクセスしてくる際に、名乗るUserAgentや、IPアドレスを取りまとめたリスト
  - 定期的にメンテナンスしていく必要がある
  - みんなが同じリストを適用することに意味がある

基準の一つにできる！

機関リポジトリ
アクセスログの整理ために

- ロボットのリストをどう作ろう？
  - 使えるものを使おう！
    - COUNTERプロジェクトでUserAgentリストを公開している！
      http://www.projectcounter.org/r4/COUNTER_Robots_list_Jan2014.txt
    - JAIROのアクセス統計から維持しているリストもある！

採用してしまおう！

機関リポジトリ
アクセスログの整理ために

公開済！

https://bitbucket.org/niijp/jairo-crawler-list

- どういうもの？
  - UserAgentのリストと、IPアドレスのリストの２ファイルで公開
  - 更新は年に1回で、バージョン管理がなされており、利用して統計を出した際には、「JAIRO Crawler-ListのVersion xxxを使用した」と記述できます
  - **Jairo Cloudでも利用（されます、されるはず）**
  - 利用方法は、

JAIRO Crawler-Listは、「IPアドレスリスト」と「ユーザーエージェントリスト」の2ファイルに分かれています。 それぞれをダウンロードし、Webサーバーのアクセスログとマッチングさせ、マッチした場合は利用統計の対象から外すことで、クローラーからのアクセスを除外したアクセス集計を行えるようになります。 データ形式は後述のとおりです。IPアドレスは完全一致で、ユーザーエージェントはの部分一致でマッチさせる必要があります。マッチングにおいて大文字と小文字は区別します。（JAIRO Crawler-List公開サイトより）

詳細は是非一度サイトをご覧ください！

機関リポジトリ
アクセスログから知るために

機関リポジトリ
アクセスログから知るために

- 現状として
  - 各機関リポジトリごとに“アクセス数”や“ダウンロード数”を確認している。これはコレでもちろん有用。
  - ただ、基準は特になく、他の機関リポジトリと比較するのは難しい状況。

機関リポジトリ
アクセスログから知るために

- 他の機関リポジトリと比較すると嬉しいことがある？
  - 機関リポジトリの枠を超えて、アクセス数や、ダウンロード数の比較が見ることができる
  - 機関リポジトリの枠を超えて、コンテンツ間のアクセス分析ができるかもしれない？
  - アクセス分析により、コンテンツのレコメンドを提供できるようになるかもしれない？

・・・評価は必要だが、見てないと何とも言えない

機関リポジトリ
アクセスログから知るために

- 他の機関リポジトリと比較するには？
  - 同じ基準で数値を取ることが必要。
    → 公開されたJAIRO Crawler-Listが
    使えるんじゃないかな？
  - いろいろなシステムを採用しているけれど？
    → ロボットリストだけではない仕様の
    検討が必要かな？
  - 各機関リポジトリにかかる負担が大きいと、
    長続きしない・・
    → なるべく自動化していく必要があるな？

機関リポジトリ
アクセスログから知るために

- これまで
  - 打合せ　3回　internet会議　2回　メール審議随時
  - Open Repository 2015 へ参加
- 今後の進め方
  - 評価版システムの開発
    - まずはJAIRO Cloudのデータを検証で利用して評価を実施
    - 実際にデータを見ながら、実装する切り口を考える
    - JAIRO Crawler-List以外の、基準についても検討していく
  - 他のシステムへの適用を考える
  - 運用方法を考える

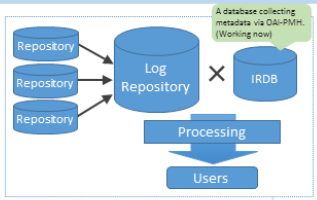
おまけ： Open Repository 2015報告資料

System for Cross-Organization Big Data Analysis of Japanese Institutional Repositories

Takurou Kawamura*1, Kenichi Igarashi*2, Hiroshi Kato*3, Akira Maeda*3, Toshihiro Aoyama*4, Kazutsuna Yamaji*3, Sho Sato*5
*1Hiroshima University, tkawa@hiroshima-u.ac.jp *2Keio University *3National Institute of Informatics
*4National Institute of Technology, Suzuka College *5Doshisha University

Introduction

We are developing a system that use access log of institutional repositories(IRs) and metadata of contents. Major work of the system is data analysis. This function enables cross-organization comparison and standardization of statistics. The system is also attached other functions to promote utilization of IRs.



Development

The project has just started. We are discussing the detail of the service, creating a prototype with the data of JAIRO Cloud on Microsoft Azure. Our current plan is as follows.

What's JAIRO Cloud?
It is a SaaS-type cloud service for IRs. The service have a repository software module, WEKO. More than 250 institutions in japan use it.

Design framework and functions → Develop and test → Launch for JAIRO Cloud → Extend to other repositories