

PIRUS2における機関リポジトリの統計データ収集方法について

PIRUS2は、電子ジャーナル等のオンライン利用統計の標準化を推進してきたCOUNTER (Counting Online Usage of Networked Electronic Resources) と密接に関連して、論文レベルの利用統計の標準化をめざすプロジェクトである。論文単位の利用統計の方式が標準化され、出版社、機関リポジトリ、分野別リポジトリ等に広く実装されることによって、利用データに基づく研究成果の新たな評価指標としての活用が想定されている。

PIRUS2では、DOI、ORCID、Institutional Identifierの識別子の利用が想定されている。こうした識別子の導入については、千葉大学を中心に進められてきたROATプロジェクトにおいても必要性が強く認識されているところであるが、導入の方式によっては、機関リポジトリその他の運用管理に大きな影響が予想される。ただし、PIRUS2の提案は、今後の運営基盤、特に経済的側面については、実施可能性、持続可能性に不透明なところがあると考えられる。例えば、リポジトリ側でのDOI入力負荷の問題、出版者側でDOIが振られていない場合の対応、わが国で多く使用されているDSpace, Eprints, Fedora以外のサーバソフトウェアへの対応、費用および参加のメリット（インセンティブ）あるいはデメリットといった問題があろう。このため、今後の推移を注視しつつ、必要に応じて適確な対応を採っていく必要があると考えられる。

PIRUS2の目標は、ある論文に関して発生する、出版者サーバ、機関リポジトリ、分野別リポジトリ等のすべての利用データを集約し、そこからクローラやボット等によるアクセス、一定時間内の重複リクエスト等を標準的な方式によって除去した後、アクセス論文単位の利用統計（回数）を生成することにある。想定されている方法は以下のとおりである。

(1) データの収集と集約

- ・ 出版者と仲介者（アグリゲータ等）のデータ
 - * 利用（ダウンロード）イベントをそのまま記録し、COUNTER標準に従いロボットアクセス、ダブルクリックによる重複を排除したうえで、雑誌単位のレポート（JR）を作成する。JRの統計は、手動方式の場合はMS-Excel/CSVで、自動方式の場合はSUSHIを用いて、公認団体が集める。
 - * 論文単位の利用統計についてはPLoS等が行い、未着手の出版者／仲介者もその需要

は承知しているが、比較可能な統計のための業界標準を適用する必要がある。

- * このため、PIRUS2 では一連の論文レベル利用統計 (AR) を提案する。
- * 出版者／仲介者からの論文単位のデータ収集については、SUSHI を推奨する。

・ 機関リポジトリのデータの収集

- * 「トラッカー・コード」をリポジトリ用のサーバソフトウェアに実装する。これは、論文本体のダウンロードについて、OpenURL 形式でログ出力を行うコードをリポジトリ用サーバソフトウェアに組み込むものであり、ダウンロードが行われるとメッセージがセントラルサーバに転送される。すべてのサーバソフトウェアに対応するのは難しいため、DSpace, GNU Eprints, Fedora を対象に開発を行った (PIRUS2 プロジェクトの website でソフトウェアを公開している)。
- * 上記の別方式として、参加組織はOpenURL形式のログをローカルサーバに送り、CCHがハーベスト (OAI-PMHを用いて) する方式、および処理までローカルサーバで統計結果のまとめをCCHで実施する (収集にはSUSHIを用いる) 方式を想定している。ただし、現時点ではSUSHIの利用は現実的でないと考えられているようである。

・ 論文単位の利用統計の構成要素

- * 論文単位の利用統計の形式としては、参加出版者から CCH に送付する AR (Article Report) 1 フォーマットがあり、AR1: 月毎、DOI 毎の論文ダウンロード完了数、AR1j: 雑誌単位の、月毎、DOI 毎の論文ダウンロード完了数を意味している。ただし、AR1 は開発中であり、COUNTER 標準になっているわけではない。
- * また、出版者とリポジトリ等のデータを集約した統計の形式の例として、AR2: 著者毎、月毎、DOI 毎の論文のダウンロード完了数 (一つの論文についてのレポート)、AR2a: 著者毎、月毎、DOI 毎の論文のダウンロード完了数 (ある著者のシステム内で利用可能なすべての論文を対象とするレポート) があげられる。
- * この結果、AR2 の形式としては、次のような構成が想定されている。

Article Report 2の形式 (提案)

<Journal>

<Publisher platform>

<Author name>

<ORICD>

<Institutional Identifier>

<Article title>

<DOI> ; = CrossRef DOI

※ DOIの活用にあたっては、CrossRefとの密接な連携（CrossMark - 剽窃検出ツール、CrossCheck等）が前提とされている

* なお、著者名の記述法、論文の異なるバージョン、複数著者への対応等の問題は認識されているものの、現時点では具体的な対応策は考えられていない。

• 出版者、リポジトリのデータの結合

* DOI を用いた出版者とリポジトリのデータの突合の可能性を探るために、The PIRUS2 Demonstrator と名付けた実験を行った。

* 結果として、出版者とリポジトリの両方にDOIがある場合はほぼ問題はないが、片方のみに DOI がある場合の突き合わせはかなり難しい。また、論文タイトルと第一著者の姓による CrossRef DB の検索も行ってみたが、結果は良くなかったとのことである。

* 重要な問題として、こうして得られた利用統計の何をどこまで誰に向けて公開するかがある（きわめて「政治的」な問題）。