機関リポジトリのアウトプット評価　プロジェクト最終報告書

Final Report of the Projects for Output Evaluation of Institutional Repository

千葉大学

Chiba University

2013.2.28

執筆者：
佐藤義則（東北学院大学文学部教授，東北大学附属図書館研究開発室客員研究員）
竹内比呂也（千葉大学文学部教授，千葉大学附属図書館長，千葉大学アカデミック・リンク・センター長）
武内八重子（千葉大学附属図書館利用支援企画課）
竹内茉莉子（千葉大学附属図書館利用支援企画課）

はじめに

本報告書は，国立情報学研究所次世代学術コンテンツ基盤構築事業(CSI)の枠内で継続的に行われてきた，機関リポジトリのアウトプット評価のための基盤構築にかかる活動を総括するものである。時系列に沿ってプロジェクト名を列挙すると，

・「機関リポジトリの評価システム」（2006-2007 年度）
　代表機関：千葉大学，分担機関：三重大学
・「機関リポジトリ評価のための基盤構築」（2008-2009 年度）
　代表機関：千葉大学，連携機関：東北大学，金沢大学，北海道大学，大阪大学
・「機関リポジトリアウトプット評価の標準化と高度化」（2010-2012 年度）
　代表機関：千葉大学，連携機関：東北大学，筑波大学

となる。このように，都合 7 カ年にわたり，千葉大学を中心に複数の大学の参画の下，この活動は実施された。

この間，日本の機関リポジトリに見られた大きな変化は，機関リポジトリの整備・運用が大学図書館にとって特別の業務ではなくなったということである。その証拠として，文部科学省が毎年実施する『学術情報基盤実態調査』において，2008 年度から機関リポジトリに関する調査が行われるようになったこと，また，朝日新聞出版が刊行する『大学ランキング』においても，2010 年度版から機関リポジトリに関する指標が含まれるようになったことを挙げることができる。機関リポジトリ発展の初期においては，機関リポジトリを設置する大学数，機関リポジトリによって提供されるコンテンツ数といったインプットに係る指標がその発展状況を示す指標として意味を持っていた。しかし，今日のように，設置機関数が 200 を超え（共同リポジトリ参加機関を含めるともっと多くなる），コンテンツ数も 100 万件を超えるようになると，それがどのくらい使われているのかといったアウトプットに係る指標が，機関リポジトリによるコンテンツの収集，蓄積の価値を示す指標の一つとして関心をひくようになるのは当然のことである。また，それと同時に各大学が公表している機関リポジトリへの「アクセス数」に関し，その数値の妥当性についての議論を引き起こすことにもなろう。また，英国で展開された Research Assessment Exercise (RAE)のように，アクセス数が業績の評価と結びついてくればなおさらである。

アクセス数の算出に利用できるアクセスログ分析ツールとしては, Google Analytics, AWstats，WebAlyzer 等の複数の分析ソフトウェアが存在しているが，同じログファイルをこれらの分析ソフトウェアを使って処理し，結果を比較すると，元のデータが同じでも結果は全く異なるものとなるのである。共通のツール，共通の方法論で分析

を行わずして利用統計の比較を行うことは全く無意味であるからである。

このような状況を見越して，機関リポジトリのアウトプット指標としてのアクセス統計を，アクセスログをもとに一定のルールに基づいて算出し，リポジトリ間で相互比較できるようにするための基盤整備をめざしたのがここで対象とする一連のプロジェクトであり，その具体的な成果が ROAT（Repository Output Assessment Tool）である。ROAT は，電子ジャーナルやデータベースのアクセス統計のための標準である COUNTER Code of Practice の考え方を採用しているが，このことは，図書館が提供する電子情報資源へのアクセス数の評価という枠組みの中に，電子ジャーナルなどと同様に機関リポジトリも位置づけることも意味している。

ROAT のこのような取組みは，国際的にみても先駆的と言えるものであった。ROAT をアウトプット評価のためのツールとして試験的に提供しはじめたのは 2008 年度であるが，その当時，標準化された機関リポジトリ利用統計を出力できるシステムとして世界で唯一の存在であったと言っても過言ではないと思われる。それゆえ，国際展開の可能性についても議論，検討され，実際，ドイツ，フランスの専門家との意見交換も行ってきた。しかしながら，日本の機関リポジトリに対して行っているのと同じサービスの提供を多言語で行うことは現実的には極めて困難であり，これは断念せざるを得なかった。

ROAT のもう一つの大きな特徴は，これを機関リポジトリに関するコミュニティで共有すべきツールと位置づけ，特に維持管理に手間と労力がかかるロボットリストに維持管理を共同で行うという提案をしてきたことにあった。日本では，デジタルリポジトリ連合（DRF）を中心として，機関リポジトリの諸問題を解決するためにコミュニティ・ベースでさまざまな活動を行ってきた実績があり，プロジェクトチームとしては，その活動の延長として ROAT の活動についてもコミュニティからのサポートを期待したわけである。しかしながら，機関リポジトリに係る活動が図書館にとって日常的になるにつれ，図書館全体の業務とのバランス，あるいは，プライオリティの観点からリポジトリの活動も位置づけられるようになり，各図書館が直接的な利益とはならない活動にマンパワーを割くことが困難になったのか，期待したような共同管理は実現できなかった。

しかしながら，このような問題があるにせよ，このことはこれまでの ROAT の成果を否定するものではなく，まとめられたガイドラインは，現実的かつバランスのとれた優れた指針となったと自負している。

ガイドラインでも取り上げたように，単にファイルのアクセス数をカウントするのではなく「コンテンツ」のアクセス数をカウントするためには，同一著作が異なるファイル形式で並行して存在する場合や，異なる複数の機関に所属する研究者の共同研究がそれぞれの機関リポジトリに登録された場合等，合算が必要な場合がある。この

ような合算を自動的に行うためには，DOI や URN などの識別子を用いることが不可欠と思われる。日本においても，ジャパンリンクセンターが発足し，機関リポジトリに登録されたコンテンツに対しても識別子を付与する環境が整いつつある。そのような環境の変化を活かしつつ，このプロジェクトが示した基本的な考え方が実現され，より精緻な利用統計，アウトプット評価が実現することを願っている。

このプロジェクトの実施期間においては，千葉大学が全体の取りまとめ，サーバの設置, 管理などを行ってきたが, ROAT の開発における中心的な役割を担われたのは，佐藤義則教授（プロジェクト発足時は三重大学教授, 平成 19 年より東北学院大学教授，東北大学附属図書館研究開発室客員研究員）である。佐藤教授は，当初の開発のみならず, 継続的に利用者からの問い合わせにも対応するなど献身的に関与された。また，このプロジェクトの必要性をいち早く認識してプロジェクトの形にまとめられ，さらには，国際的な動向に関して最新の情報収集に務められたのは土屋俊教授（プロジェクト発足時は千葉大学教授・附属図書館長，平成 23 年より大学評価・学位授与機構教授）である。また，多忙なスケジュールを調整して，訪問調査に応じてくださり，また国際ワークショップの際には来日された，お二人の専門家，すなわち，Joachim Shoepfel 氏（リール第 3 大学），Ulrich Herb 氏（ザールラント大学）にもお礼申し上げる。ネット時代ではあるが，時間をかけて対面で意見交換をし，相互理解を深めることができたことは極めて重要であった。

また，千葉大学附属図書館においては，加藤晃一（現京都大学附属図書館），森一郎（現信州大学附属図書館），武内八重子がこの活動に携わってきた。また，報告書のとりまとめにあたっては，武内八重子，竹内茉莉子が担当した。記してその貢献に感謝したい。

2013 年 1 月

千葉大学附属図書館長
竹内　比呂也

1. 機関リポジトリのアウトプット評価に関連する取り組み

機関リポジトリのアウトプット評価を行う際に必要となるのは，各機関が同じ枠組みで処理を実行し，比較可能な利用統計を作成することである。そのためには，アクセスログに対して同一基準によるフィルタリングを行うこと，その手段が様々な機関リポジトリのプラットフォームに対応していること，加えて，機関横断的な利用統計をどのように整備するかということが課題である。こうした課題に対して，これまで各国で様々な取り組みが行われてきた。本章では，機関リポジトリのアウトプット評価のために必要な要件を明確にするために，こうした取り組みについて概観する。

1.1 Interoperable Repository Statistics（IRS プロジェクト）[1) 2) 3)]

2005 年 6 月～2007 年 5 月まで, JISC の助成によりサウサンプトン大学(イギリス)，タスマニア大学（オーストラリア)，ロングアイランド大学（アメリカ)，キー・パースペクティブ社（イギリス）が行ったプロジェクトである。

IRS プロジェクトは，柔軟で使いやすく，相互利用可能な機関リポジトリの利用統計のために，機関リポジトリ管理者や研究者への調査を行い，二つのシステムを構築した。

その一つは，各リポジトリで個別に使用するためのオープンソースのソフトウェアであり“IRStats”と名付けられ EPrints 向けパッケージが公開されたものである（現在はリンク先不明)。システムのベースとした AWStats による基本的機能（ウェブページのアクセスログを読み込み，MySQL または PostgreSQL により利用イベントのデータベース（匿名形式で個々の利用の記録を作成）を構築したうえで，集計およびグラフ・図表の作成処理を行う）に加えて，IRStats では機関リポジトリで利用されたコンテンツをメタデータの形式で把握できるようにするためのモジュールが組み込まれた。AWStats の機能では利用されたコンテンツは URL でしか表せないが，このモジュールにより IRStats では論文別，著者別の集計が可能となった。作成された集計や図表は必要に応じて簡単にローカルサイトで公開することができ，単純な利用回数（例．どのコンテンツが何回ダウンロードされたか)，利用ランキング（例．最も多くダウンロードされた資料／著者のトップ 10)，アイテムごとの月別ダウンロードグラフ等が利用できる。また，ビットストリーム（本文ファイル）のダウンロードの分析をどの単位で行うか（個々のレコードかコレクション全体か）や，分析の対象機関，求めるグラフや図表の選択のためのインターフェースが用意された。

IRS が構築したもう一つのシステムは，OAI サービスである。これは，メタデータを含む IRStats の利用イベントのデータベースに相当する内容を”OpenURL Context Objects”[4) 5)]の形式に変換したうえで, それらを OAI-PMH プロトコルを用いて収集（ハーベスト）し，集中的に分析処理を行い，その結果を各リポジトリに返すという一連

のサービスを指している。サウサンプトン大学の Tim Brody による"Citebase Search"のウェブサイト [6]では，論文の引用頻度に加えてダウンロード回数の表示が実現されたが，OAI サービスの活用によって，それぞれのリポジトリは単にログファイルをハーベスト可能な状態に置くだけで，プラットフォームの違いを問わず，同じ枠組みで処理された利用統計を入手することができ，さらに"Citebase Search"のように文献データベース上でその結果を表示できるようになることが想定された。しかし，残念ながら OAI サービスの成果は明確には示されなかった。システムログの分析には予期せぬデータの出現などからどうしても単純な機械的作業だけで行うことができない側面があるため，システム資源や人的資源の面で非効率と判断されたのかもしれない。

利用統計の OAI サービスが構想された背景には，英国の大学研究評価（Research Assessment Exercise; RAE）方式の見直しの動きがある。英国における各大学への研究資金の配分は，個々の研究に対する競争的な研究補助金のほかに，研究分野ごとの個人の研究業績の審査と学科自体に対する評価による学科のランクづけに基づいて，大学全体への配分額が決定される方式になっている [7]。このうち個人の研究業績の審査については， 2007 年頃までピアレビュー方式で行われてきたが，時間や費用の面から，ビブリオメトリクスをもとにした数的指標による評価の導入が検討された [8]。Citebase においては，論文や著者ごとの引用回数とダウンロード回数が実験的に表示されていたが，この表示は RAE における数的指標の活用との連動が意図されている [7][9]。Citebase において引用回数とともにダウンロード回数が示されているのは，ダウンロード回数を学術的影響に関する新たな指標として捉えているからに他ならない。利用は「引用に先行するため，学術的影響のより初期の指標として役立つ」[10]ものであり，雑誌論文の範囲を超えた，雑誌論文の著者だけに限定されないコミュニティ全般の学術情報の流通を表すと考えられるからである。この点で，IRS プロジェクトは，出版のサービス，機関リポジトリ，リンク・リゾルバ等から大規模な引用データ，利用データ, 書誌データを収集し, 分析を行っている"MESUR プロジェクト"（1.2 参照）と共通する方向性を持つ。

2007 年 6 月に行われた JISC Digital Deluge Conference において, IRS プロジェクト（JISC）と EU Knowledge Exchange プロジェクト（JISC-UK, DINI-ドイツ, DEFF-デンマーク, SURF-オランダ）は，リポジトリ統計が公正である条件を担保し，論文・データセット・雑誌論文の利用に関して多目的利用が可能かつ適切な報告を可能とするために次のように相互運用性に関する提案を行った [11]。

(1)　ウェブログにおいて「利用イベントの痕跡」とする解釈の統一

クローラーおよび人手によらないダウンロード行動を除去する標準化された方法により，60 分以内の繰り返しがない e プリントのアブストラクト（コンテンツの詳細情報ページ）または要素のビットストリーム（本文ファイル）に対する要求。

(2)　「利用イベントフォーマット」の統一

OAI による交換と長期保存のために，匿名形式による個々の利用者の記録の実現。

(3)　メタデータのようにログデータも共有するための OAI の活用

メタデータにはダブリンコア，ログイベントには OpenURL ContextObjects を使用。

1.2 Metrics from Scholarly Usage of Resources（MESUR プロジェクト）[12)]

2006 年 10 月～2008 年 10 月まで，アンドリューW メロン財団の助成により，Johan Bollen を研究代表とするロスアラモス国立研究所(LANL)の電子図書館調査･試作チームが行ったプロジェクト。利用データ．引用データ，書誌データを大規模なセマンティック・ネットワークに統合し，考えられる広い範囲の指標を調査し，それらの学術的影響度，（指標間の）関係，妥当性を検証することを目標とした。

MESUR プロジェクトに先駆けて，Bollen らはカリフォルニア州立大学（CSU）のリンキング・サーバに記録された大規模な利用データの解析を行った [13)]。CSU に 23 あるリンキング・サーバのうち 9 サーバから，167,204 人による 2,133,556 件の資料の利用データを集めている。しかしこの利用データは，引用データ，書誌データとは正式には統合されておらず，学術コミュニティを代表する有効なサンプルとは言えなかった。これに対し，MESUR プロジェクトでは関連するメタデータと参照データを含む，少なくとも 5,000 万件の資料，約 7,000 万のユーザと著者の結合，世界の主要出版者（社）と大規模な図書館コンソーシアムを含む学術機関の有効なサンプルといったデータの集積が図られ，そのオントロジーに従って，学術的な実体に関連する約 30 億から 50 億の意味的関係の記述がもたらされた。

MESUR プロジェクトは四つのフェーズで進められた [14)]。

［フェーズ 1］学術コミュニケーション・プロセスのモデル化

MESUR プロジェクトは，しばしば分離して利用されることのあるさまざまな情報源（例．利用データ，引用データ，書誌データ）について関係を特定するオントロジーで表現することにより，学術コミュニケーション･プロセスのモデルを明確化する。

［フェーズ 2］参照データセットの構築

形成されたオントロジーを統合のための枠組みとして使用し，大規模な学術データを細かな粒度の意味的ネットワークに集積，組織化する。この意味的ネットワークは，その後の利用に基づく数的指標の研究において，参照データセットとして役に立つ。

［フェーズ 3］特徴付け

この規模と粒度で学術的な参照データセットを構築する初めての試みの一つであるので，学術コミュニティの重要な区分と構造的特性に関するわれわれの知識を深めるために，形成された意味的ネットワークの構造と特性を調査する。

［フェーズ 4］数的指標の定義と検証

広い範囲の利用に基づく数的指標を定義し，形成された参照データセットをもとにその妥当性，信頼性，学術的な相互関係性を明らかにし，それらの意味論および適切な応用に関する指標をまとめる。大規模な参照データセットを入手することは，利用データのばらつきではなく数的指標そのものに起因する様々な数的指標間の類似点と相違点をもたらすだろう。さらに，数的指標間の相関構造は，学術的影響度のさまざまな要素と，どのようにしたらそれらをもっとも正確に測れるかを明らかにするために，量的に調べることができる。

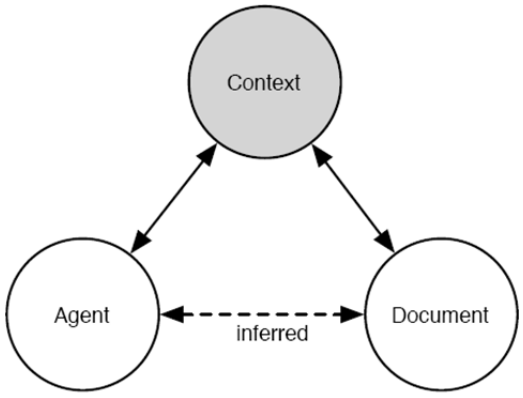

図 1　MESUR のオントロジーにおけるコンテクストの位置づけ

（Bollen,J. et al. “MESUR: usage-based metrics of scholarly impact.” Proceedings of the ACM International Conference on Digital Libraries, 2007, 474 より）

フェーズ 1 から 4 に至る具体的な手法は Bollen らの文献[13) 14) 15)]に詳しく紹介されているが，具体的には，フェーズ 1 で”エージェント”（著者，利用者，機関などの主体），”ドキュメント”（論文，雑誌，予稿集，図書などの文書）を”コンテクスト”（利用，引用，測定，共著といった文脈）が結びつけ（図 1），そのコンテクストは”事象（event）”と”状態（state）” を分ける。事象はドキュメントに対するエージェントの動作を表し，状態はドキュメントやエージェントの持続的，定常的な状態を表すといった関係を特

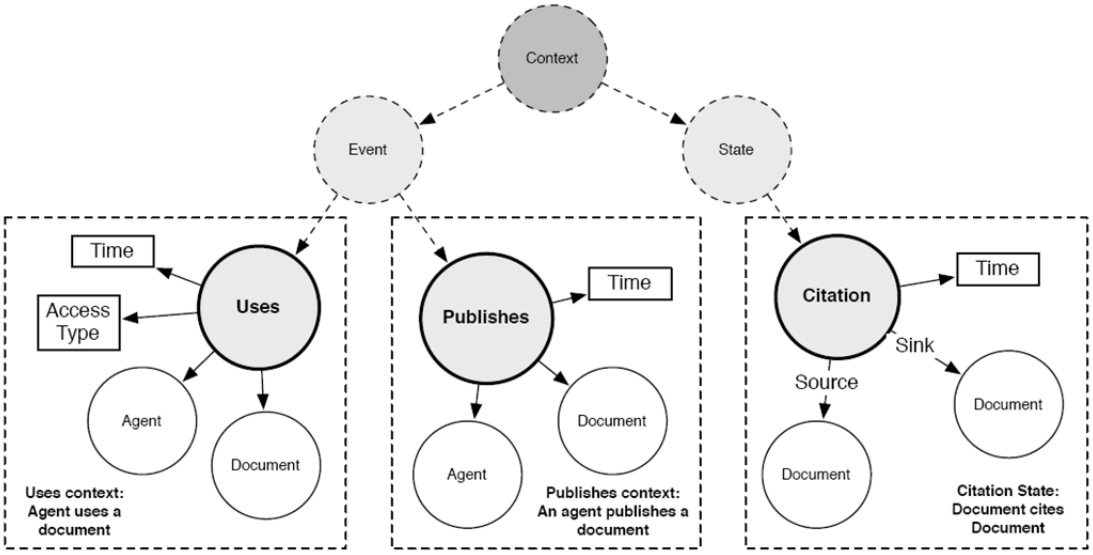

図 2　利用，引用，出版の関係表現のための，MESUR のオントロジーに位置づけられたコンテクスト・クラスの使用

（Bollen,J. et al. “MESUR: usage-based metrics of scholarly impact.” Proceedings of the ACM International Conference on Digital Libraries, 2007, 474 より）

定するオントロジーで表現することにより（図 2），学術コミュニケーション・プロセスのモデルを明確化する。フェーズ 2 でこのオントロジーを使用して大量のデータを組織化して参照データセットを作成，フェーズ 3 で作成した参照データセットの意味的ネットワークの構造と特性を調査し，フェーズ 4 では確立した参照データセットによる利用に基づいた多面的な学術的インパクトの数的評価が有効になるとしている。

MESUR プロジェクトでは，利用に基づく数的指標が相当な可能性を持っているにもかかわらず受け入れられない原因を，利用データや得られた数的指標への理解不足によるものとし，その解決のために大規模な参照データセットが構築された。また得られたデータと，従来から使用される引用に基づく数的指標（インパクトファクター等）を比較する予備的な調査が実施され，利用に基づく数的指標が十分な可能性を持つことが示された。

1.3 Publisher and Institutional Repository Usage Statistics（PIRUS，PIRUS2 プロジェクト）

第 1 期：2008 年 9 月～2009 年 1 月，第 2 期 2009 年 10 月～2011 年 5 月で行われた JISC の助成を受けたプロジェクト。

PIRUS[16]の目標は，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，オンラインの雑誌論文を提供するすべての事業体（出版者，アグリゲータ，機関リポジトリ，主題リポジトリその他）が導入可能な論文レベルの利用統計の方式を標準化し，グローバルなレベルで，記録，報告，統合できるようにすることにあった。PIRUS では，大多数のリポジトリと出版者での実現を視野に入れ，複数のシナリオが用意された。PIRUS の主要な成果として以下の 4 点があげられる [17]。

a. 個別文献利用報告「文献レポート 1：フルテキスト文献ダウンロード成功数」用の COUNTER に準拠した XML 概念の実証プロトタイプ。リポジトリと出版者が共通して利用できるもの。理念的には，このレポートは，著者個人と機関の両方に対して作成できる。しかし，実務的には著者個人についてのレポートは生成が容易で短期的に実現のある目標となる一方，機関など（例えば，助成機関）についてのレポートはより複雑であり，長期的な目標と見なすべきである。
b. リポジトリに実装させるトラッカーコード。利用統計の作成と集約の責任および集約を行う関係出版者への送付の責任を担う外部当事者へ，またはローカルのリポジトリサーバへ，メッセージを送る。
c. 個別文献利用統計の作成，記録，および集約のための様々なシナリオ。現行のリポジトリの大部分をカバーする。各リポジトリは自組織の技術と実装に合ったシナリオを選択することができる。

d. 必要に応じて（一部のカテゴリーのリポジトリについて）利用統計を作成し，他社のために利用統計を収集，集約する中心機構についての要件の決定。

PIRUS は JISC に対して，技術面，組織面，経済面，政治面での更なる研究と開発が必要であることを提言し，また COUNTER やリポジトリ，出版者やベンダーに対してもそれぞれの立場で果たすべき役割について提案した。

PIRUS2[18) 19)]では，PIRUS の成果を受けて，出版者や機関リポジトリの利用データを収集・解析して利用統計を作成・提供する中央処理機構（Central Clearing House; CCH）のための技術的，組織的，経済的モデルを含むプロトタイプサービスの構築により，出版者やリポジトリなど関連機関に信頼性のある個々の論文の利用統計を可能にすることが目標とされた。具体的には，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，そこからクローラー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)を生成することである。

PIRUS2 では，DOI，ORCID，Institutional Identifier といった識別子の利用が想定された。こうした識別子の導入については，ROAT プロジェクトや金沢大学が中心に進めてきた「オープンアクセス環境下における同定機能導入のための恒久識別子実証実験」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，導入の方法によっては，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。PIRUS2 の提案でも，今後の運営基盤，特に経済的側面については，実施可能性，持続可能性には不透明なところがあり，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 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)などが問題になると考えられた。なお，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 と名付けた実験が行なわれた。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く，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ことが報告されている。

こうした一連の PIRUS プロジェクトの成果を受け，2012 年 11 月には，COUNTER から PIRUS 実務指針（PIRUS Code of Practice）のドラフト[20)]が公表された。その概要は以下の通りである。

a. PIRUS 実務指針と COUNTER の関係

COUNTER は PIRUS 実務指針の開発を行うとともに, 継続的運営や導入についても責任を担う。PIRUS は COUNTER 実務指針に準拠している。ただし，PIRUS 実務指針の実装は，電子情報資源の COUNTER 指針リリース 4 準拠の要件ではない。PIRUS

実務指針は，むしろ COUNTER に基づいた標準を提供するものであり，より粒度の小さいアイテムのレベルでの利用の記録と報告を望む組織が実装できる。

b. 監査とPIRUS準拠組織

出版者やアグリゲータ等の場合は，各組織の報告に対する独立した年次監査をもとに，PIRUS準拠の認証を行う。機関リポジトリ，主題別リポジトリといった，これまでCOUNTERとは直接関係のない組織の場合は，生の利用データの収集，それに続く収集した生データのCOUNTER統計への変換処理，データ提供組織へのCOUNTER統計の提供について責任を負う第三者，すなわち中央処理機構あるいは他の（国または地域の）統計集約サービスへの参加を通じてCOUNTER準拠の資格が得られる。なお，この場合の監査の対象は第三者のサービスだけとなる。

c. 中央処理機構の役割

中央処理機構は，COUNTERによって監督される。この機構は二つの主な機能を有している。◎

① 論文単位の統合PIRUS利用統計を作成するために，出版者，アグリゲータ，リポジトリその他から個々の論文レベルの利用統計を収集し処理する

② 個々の論文（将来的には，他のアイテムも）に関する認証済みの統合PIRUS利用統計の中心情報源となる

d. 他の統計処理機構の役割

国または地域の統計の集約・統合サービスのような他の統計処理機構は，機関リポジトリから利用データを集約し，結果としてのPIRUS準拠の論文利用統計を中央処理機構に提供するという重要な役割を担うことが想定されている。PIRUSの二次処理機構に対する認可のモデルは，英国の機関リポジトリ統計のための全国処理機構としてのIRUS-UKである。なお, IRUS-UKはオープンアクセス研究論文のデポジット, 収集, 整理，公開のための社会的，技術的基盤の形成を目指すUK RepositoryNet+の一部を構成する組織でもある。

e. 利用統計の対象

主たる対象は，一般の研究論文（ショートコミュニケーションを含む）とレビュー論文であるが，論説，書評，博士論文についても受理可能としている。また，利用のカウントはHTMLおよびPDFについてのみ行うこととされている。

注目すべき点として，NISO/ALPSPのJAV（Journal Article Version）ワーキンググループが定義している論文の七つの版[21]のうち，以下の五つだけが対象となっており，著者のオリジナル原稿（Author's Original）やピアレビュー段階の投稿原稿（Submitted Manuscript Under Review）は除外されていることが挙げられる。

- 受理された原稿（Accepted Manuscript）
- プルーフ（校正段階の原稿; Proof）

- 公開版（Version of Record）
- 訂正後の公開版（Corrected Version of Record）
- 改版後の公開版（Enhanced Version of Record）

なお，このドラフトでは，検索エンジン等（ロボット）からのアクセス，運営者等による内部利用，LOCKSSのキャッシュのためのダウンロードについて，除外するように指示されているが，公開されているロボットリストには少なくとも日本に固有のロボット等が含まれていないため，今後の拡張が必要と考えられる。

f. 三つのシナリオ

利用データと統計の転送のために，参加機関のタイプに合わせて次の三つのシナリオが用意されている。

- シナリオ（A）: フルテキスト論文がダウンロードされた際に，メッセージ（生の利用データ）がリモートサーバに送られる。
  - 対象組織：リポジトリおよび小規模出版者
  - プロトコル： 「トラッカー」; リポジトリのソフトウェアに組み込んだプログラムにより，ロボットや内部からのアクセス等を除去済みの生の統計データを OpenURL key value pair strings（URLs）形式に変換し，リモートサーバ（IRUS-UK 等の二次処理機構）に転送する。プログラムとしては，IRUS-UK によって，DSpace（1.8.0, 1.8.1, 1.8.2）向けパッチと EPrints 向けアドオンが準備されている。
- シナリオ（B）: フルテキスト論文がダウンロードされると，生の利用データイベントのレコードがローカルに格納され，オンデマンドでリモートサーバからハーベストできるようにされる。
  - 対象組織： リポジトリ
  - プロトコル： OAI-PMH；リポジトリのソフトウェアに組み込んだプログラムにより，ロボットや内部からのアクセス等を除去済みの生の統計データを OpenURL コンテキスト・オブジェクト（XML）形式に変換し，リモートサーバ（IRUS-UK 等の二次処理機構）に転送する。プログラムとしては，IRUS-UK によって，DSpace（1.8.0, 1.8.1, 1.8.2）向けパッチが準備されている。
- シナリオ（C）： フルテキスト論文がダウンロードされると，生の利用データイベントのレコードがローカルに格納される。利用データは，COUNTER の規則に従って処理され，オンデマンドでリモートサーバからハーベストできるようにされる。
  - 対象組織： 出版者

- プロトコル： SUSHI（Standardized Usage Statistics Harvesting Initiative Protocol）；生データではなく，集計処理後の PIRUS 論文レポート 1（XML）形式のファイルを中央処理機構に転送する

g. 最終的な論文レポート

PIRUS 準拠の出版者やリポジトリが著者および所蔵機関に対して，個々の論文の利用統計を報告するための標準フォーマットとして，PIRUS 論文レポート 2 と PIRUS 論文レポート 3 が指定されている。これらのレポートについては，他のフォーマットでの提供の有無に関わりなく，XML で利用可能としなければならないとされている。

- PIRUS 論文レポート 2: 異なる情報源を統合した，著者，月，DOI ごとのフルテキスト論文に対する達成されたリクエスト数
- PIRUS 利用レポート 3: 著者向けの，月ごとの個々の論文に対する達成されたリクエストの概要

1.4 Open-Access-Statistik（OA-Statistics）

第 1 期 : 2008 年 5 月～2010 年 12 月，第 2 期 : 2011 年 4 月～2013 年 4 月でドイツ研究振興協会 (DFG) の支援により行われた DINI(German Initiative for Network Information)によるプロジェクト [22]。ザールラント大学，ゲッティンゲン大学，フンボルト大学，シュツットガルト大学が参加，第 2 期はさらに参加機関が増えている。

第 1 期は，異なるサービス間で利用データを交換するための基準を策定すること，利用データを収集，処理して異なるサービス間で交換するためのインフラ構築，COUNTER，LogEc や IFABC の基準に基づいた利用情報の処理や，リポジトリのための追加サービスや実施要項の作成を目標とした。OAS のインフラでは データプロバイダー（リポジトリなど）とサービスプロバイダーに分かれている [23]。OAS データプロバイダーは，蓄積したアクセスログについて IP アドレスなどの利用者情報がわからないよう匿名化し，固有のドキュメント ID を付与して OpenURL Context Objects に変換し，OAI-PMH でハーベストできる状態にする。利用の多いリポジトリソフトウェアである DSpace と WebDoc 用には，データプロバイダー用のソフトウェアパッケージが OA-Statistics のサイトで提供されている。一方，OAS サービスプロバイダーはデータをハーベストし，COUNTER，LogEc や IFABC の基準にそって処理を行う。具体的には，人以外の（ロボット）アクセスを除き，重複したドキュメント（異なるシステムに置かれた同じドキュメントへのアクセス）を特定し集計する。その後，統計データはリポジトリに戻され，リポジトリ側で表示することができるようになる。

第 2 期は，ドイツのリポジトリへの OAS インフラの更なる拡大と標準化された利用統計を提供すること，正確な数的指標と付加価値サービスにより科学出版の著者と利用者のオープンアクセスの受容を高めること，国際的に比較可能な利用統計のための

協調，永続的なサービスインフラを提供することが目標として掲げられている。

2. ROAT プロジェクトのアウトプット評価へのアプローチ

各機関リポジトリのサービス状況を適切に評価するため，および複数の機関間で相互比較をするために，統一された基準で集計することの重要性は明らかである。しかし，実際にどのように統一された基準を構築し，またそれらをいかに安定的に運用できるか，ということが課題であった。そのような課題をふまえたうえで，各機関リポジトリの運営者が自主的に実行でき，かつ比較可能なアウトプット評価環境の創出を目指したのが今回の一連のプロジェクト（以下，ROAT プロジェクト）である。

この章では，ROAT プロジェクトの経緯や成果，および，ROAT の具体的な機能を記述する。

2.1 ROAT プロジェクトの活動

ROAT プロジェクトでは，機関リポジトリのアウトプット評価はアクセスログの分析に基づくものとして，2006 年度に千葉大学のアクセスログを用いて予備的な分析を行った。また，その際に開発したプログラムをなるべく早い段階で公開し，それを用いて各大学が独自にログの分析を行い機関リポジトリの相互比較を実現することを計画した。

しかしながら，各大学のログの形式や使っているアプリケーションによってデータの処理方法を変える必要があること（そのために三重大学において DSpace のログの統計処理のためのプログラムを作成した），また，利用可能なアクセス解析ソフトウェア間で十分な統一化が行われておらず，カウント基準が必ずしも同一ではなく単純な相互比較が不可能なことなどが明らかとなったため，基準の統一化のための概念的な検討をまず行うこととした。

2.1.1 ROAT プロジェクトで明らかになったこと

ログデータをもとに当該ウェブの利用状況を把握する際には，生のアクセスログから不適当なデータを除外することが必要である。具体的には，利用に結びつかなかったリクエストの除外（HTTP のステータスコードによって判断する），直接的な利用とは関係しないクローラー，収集ロボット，SPAM 等によるアクセスの除去，一つのウェブページを構成する断片的なファイル群（例えば， GIF または JPEG によるマスコットアイコン等）のカウント対象からの除外，同一利用者からの連続する複数回の要求（ダブルクリック現象）の制御，さらに業務上で生じる内部利用分の除去等が必要であることが，分析の過程で明らかになった。

さらに，2007 年 12 月には，NII において，関係大学の参加を得て「機関リポジトリアウトプット評価ワークショップ」を開催し，アウトプット評価についての基本的な考え方ならびに解決すべき様々な問題点を説明し，各大学のアクセスログの提供を求

めた。14 大学にログファイルの提供を依頼したところ 11 大学の協力を得られたので，2008 年 1 月にデータの分析を行った。このアクセスログの分析結果として，アクセス数， PDF ファイルのダウンロード数，アクセス元別アクセス数，アクセスの経路，利用頻度の高いコンテンツについての定量的データを得た。それらの検討から，さまざまな国からのアクセスがあること，多様なコンテンツが使われていること，様々な種類の機関からのアクセスがあることなど，機関リポジトリの利用についての側面を明らかにすることができた。つまり，このような分析が，機関リポジトリの効果を示すためには有効であると思われるような結果を得ることができた。

同時に，いくつかの課題も浮き彫りになった。まず，ログデータ収集に関するコンセンサスが機関リポジトリを運営している機関間で成立していないという点である。2007 年度の事業においてデータ提供していただいた 11 大学においても全期間のデータを保持していたのはわずか 2 大学のみであった。したがって，多くの機関においてログデータの活用および重要性に対する認識とログローテーション等の運用技術に関連する知識の不足が懸念され，これらの点に関する啓発の必要性が認識された。また，アウトプットの比較分析の技術的側面では，コンテンツのアクセス回数について一つの機関リポジトリ内でのアクセスランキングをメタデータ付きで出力する機能は作成済みであるものの，このままでは複数の機関リポジトリを比較するためには適用できないという問題があることがわかった（2007 年度の分析ではメタデータとの照合は 1 件ずつ手作業で行った）。

2.2 機関リポジトリアウトプット評価システムの構築

上記に挙げた ROAT プロジェクトで検討した結果や明らかになった内容は，「機関リポジトリアウトプット評価システム」（通称 ROAT）という名称で，千葉大学に置かれたサーバにより 2008 年度より試行的に提供された。ROAT は機関リポジトリのアクセスログに対して一定の処理を行うことにより，統一された基準での機関リポジトリのアウトプット指標が出力可能なシステムである（図 3）。登録申請はログイン画面にリンクされたフォームから行うことができ，機関リポジトリ単位での利用とすること以外は条件を課していない。2013 年 1 月時点での登録機関数は 33 機関である。

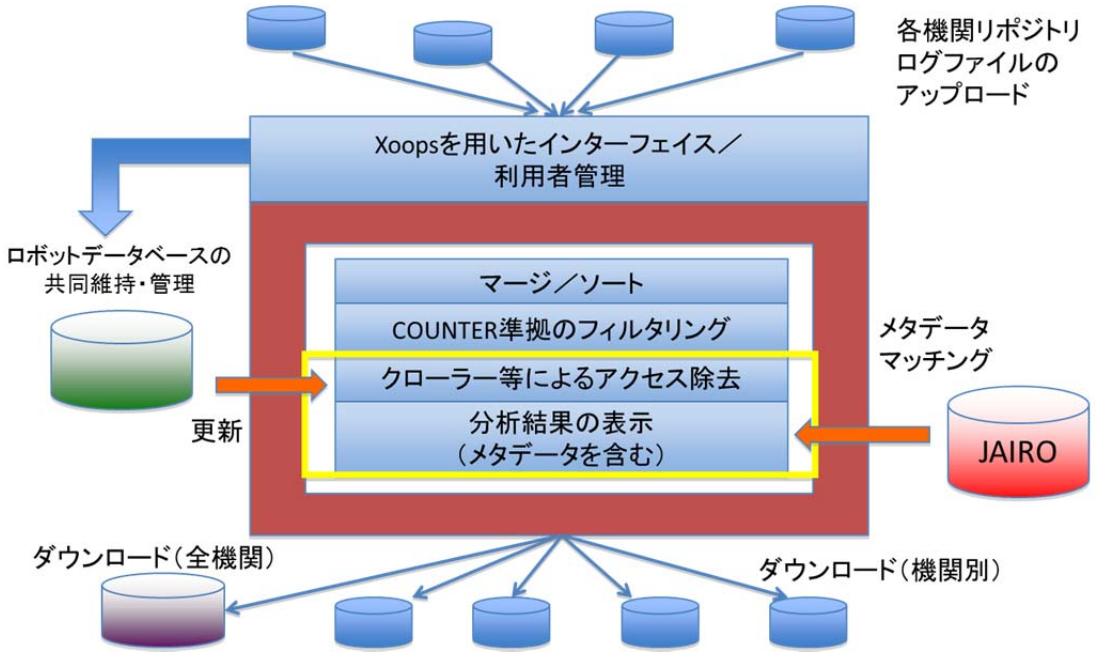

図 3　ROAT の概念図

以下では ROAT が具体的にどのような機能を有し，各機関リポジトリの運営者が自主的に実行でき，かつ比較可能なアウトプット評価環境をいかにして実現したかについて記述する。

2.2.1 使いやすいインターフェースを備えた，共同利用可能なシステム

ROAT は Xoops をベースとした使いやすいインターフェースを備えており，利用にあたって各機関がソフトウェアをインストールする等の作業が必要ないため，気軽に使えるシステムとなっている（図 4)。また複数機関が利用することを前提としたシステムであり, 管理者権限ではユーザ管理や, 全機関の統計結果の管理等が可能である。

ROAT では二度にわたって大規模なシステム改修を行ったが，その際に，アップロード当日中であれば, アップロードしたログファイルの情報を画面に表示する機能や, 各機関のログを自動的に ROAT へアップロードする機能，および統計結果をダウンロードする際に TSV ファイルに含まれる項目を適宜選択できる機能などを順次追加することで，よりユーザにとって使いやすいシステムの構築を目指した。

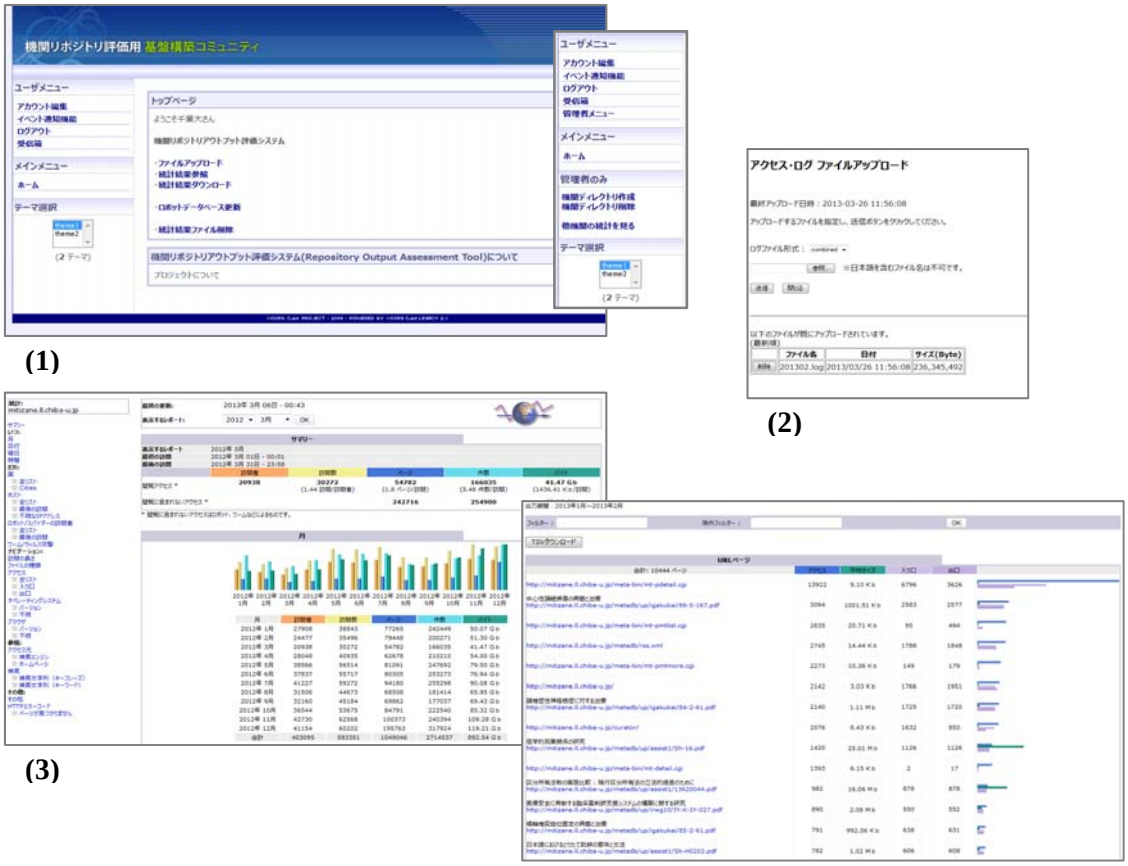

図 4　ROAT の画面例

(1) 利用者用のメニュー画面（左）と管理者用メニュー（右），(2) アクセスログファイルアップロード画面，(3) 統計結果参照画面，(4) 統計結果ダウンロード画面

2.2.2 標準的なカウント方法（COUNTER）に準拠した利用統計

2.1.1 でも述べたように，機関リポジトリの生のアクセスログには，実際にそのファイルを利用したとは言えない不適当なデータが含まれている。適切に機関リポジトリの利用状況を把握するためには，それらを除外し，「真のアクセスログ」ともいうべきデータを抽出することが必要である（図 5）。

ROAT では，機関リポジトリにおける主たる登録対象が雑誌等に掲載された論文であることから，電子ジャーナルの利用統計で一般的な COUNTER 実務指針（COUNTER Code of Practice）[24] に準拠したカウント方法を採用することとした。これにより，ある論文が電子ジャーナルと機関リポジトリの両方で公開されている場合に，両方のアクセス数の比較あるいは合算が想定できるからである。

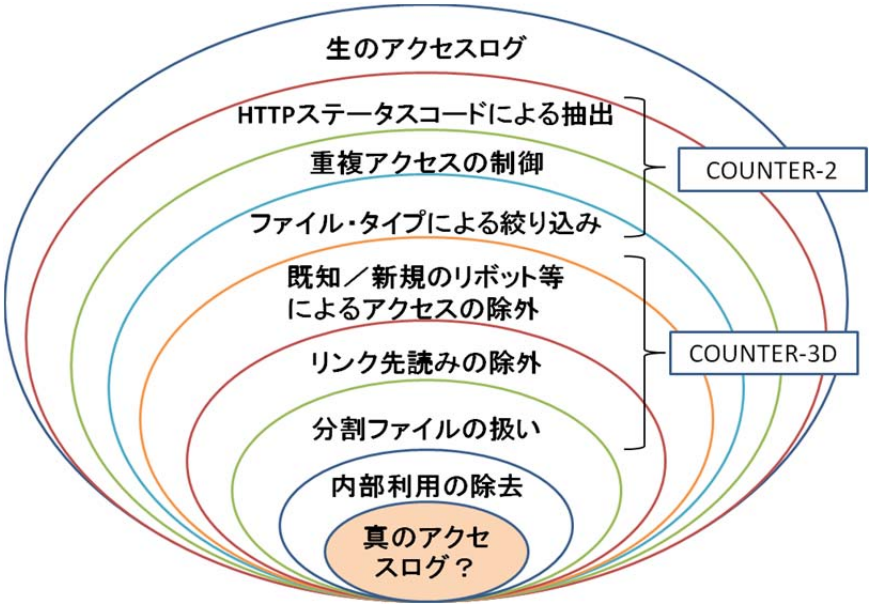

図 5　アクセスログからのデータ抽出処理

具体的には，ROAT では COUNTER 実務指針に準拠し，下記(1)～(4)のフィルタリング処理を実装した。

(1) 重複アクセスの制御

同一のアクセス対象に対する同一ユーザ (IP アドレス) からのアクセスについて，HTML ファイルは 10 秒以内，PDF ファイルは 30 秒以内である場合，ダブルクリックによって生じた重複アクセスとみなし，利用統計から最初のものを除外する。

(2) HTTP ステータスコードによる抽出

HTTP Status Codes が 200（OK）および 304 （Not Modified; サーバ側のコンテンツとブラウザのキャッシュが同一で，後者が使用された場合）であるもののみを集計する。

(3) ファイルタイプによる絞り込み

一つのウェブページを構成する断片的なファイル群（image，gifs，style sheets 等）へのアクセス回数も，利用カウントを不必要に増大させる可能性がある。ログをファイルタイプ別に分けたうえで，これらに対するアクセスは利用統計から除外する。

(4) 検索エンジン等によるアクセスの除去

機関リポジトリを含むウェブサーバへのアクセスには，検索サイト等が情報を収集するための，いわゆるロボットによるアクセスが含まれるが，ロボットによるアクセスは利用統計から除外する。

なお，ROAT ではカウント対象として，ユーザをアクセス元の IP アドレス単位で捕捉することとし，アイテムの種別についてはアクセスログ中の URL に対応する国立情報学研究所が運営する JAIRO (Japanese Institutional Repositories Online, http://jairo.nii.ac.jp/)のレコードから取得している。

2.2.3 統計から排除すべきクローラー等の情報の維持管理の共同化を実現するための環境の整備

上記 2.2.2 中の(4)で挙げた，検索サイトからのアクセスについては，既知のサイトを登録するブラックリスト方式で管理するのが一般的である。しかし，ブラックリスト方式の場合，次から次に生まれる新種のロボットや，既知の場合であっても IP アドレスの変更に対処しなければならないため，常に更新を行っていく必要がある。

ROAT では，独自に開発したアクセスフィルタによって HTTP ステータスコードによる絞り込みと重複アクセスの除去を行ったうえで，フリーウェアの AWStats を用いてロボットやクローラーの排除とファイルタイプ別の分析を行っているが，ロボット等の排除に関して AWStats 本来の機能では国内や新種のロボットに加えリポジトリ事業に固有のハーベスタ等を排除できないため，その作業を各機関が共同で行うことができるように，ロボット等によるアクセス元の追加指定を ROAT 上から申請できる機能を用意した。

また，各利用機関からアップロードされたアクセスログから抽出したユーザエージェントを，既存のロボットデータベースの該当項目と比較し，データベースに存在しない場合は新種のロボットの可能性があるものとみなし，ファイルに出力する機能も作成した。

2.2.4 さまざまなプラットフォームへの対応

現在，国内の大学では DSpace，EPrints，WEKO など様々なリポジトリ・システムを使用している。ROAT では，これらのシステムにおいてほぼ共通に利用可能な Apache の Combined 形式のアクセスログの使用を原則としているため，導入システムの種類に関わりなく利用することができる。また，この形式のログを使用することにより JISC の利用統計レビュー[25]で機関リポジトリのアクセス数を集計する際に必須とされた項目である「ユーザ」,「セッション」,「アイテムの識別」,「リクエスト種別」,「リクエスト日時」の取得が可能となっている。

2.2.5 メタデータとログデータのマッチング機能

ROAT では定期的に JAIRO の論文メタデータを OAI-PMH によるハーベスティングで取得している。この JAIRO から JuNii2 形式で出力されたメタデータをもとにデ

ータベースを構築し，フィルタリングおよび集計処理後のログデータからこのデータベース内のレコードを参照することにより，論題の表示を実現した。すなわち，分析結果に書誌情報が付与されたことで，どの文献が何回アクセスされたという情報の取得が可能となった。

また，2.2.2（3）で挙げたファイルタイプによる絞り込みについても，当初の機能のままではAWStatsで排除されないファイルタイプはカウント対象になってしまうため，アクセス数を増大させてしまう可能性や，反対に，必要なファイルであるが指定されたファイルタイプに該当するために排除されてしまう可能性があった。しかし，このマッチングの過程で「本文フルテキスト」と指定されているファイルを特定できるため，それをカウント対象とすることによって過剰なカウントやカウント漏れの回避が可能となった。

さらに，メタデータとのマッチング機能により，複数ファイルからなるフルテキストやデータベース構造による過剰カウントを発見し，対処することもかなり容易となった。機関リポジトリのシステムでは，一つのコンテンツを複数のフルテキストファイルで表現している場合がある。しかし，本来アクセス状況を知るにあたりカウントされることが期待されるのは「コンテンツあたりの」アクセス数であり，「ファイル毎の」アクセス数ではない。あるコンテンツが何らかの事情から上・下の二つに分割されている場合であっても，メタデータの表示によって把握および必要に応じた合算が可能となる。ただし，メタデータの表示だけでは解決できない場合も残っている。例えば，一つの論文の本文と図が一つのPDFとして提供されている場合には，本文だけを読んだ人，図だけを見た人，本文・図とも利用した人のアクセス数はファイルがまとまっていれば3回であるが，本文と図が二つのPDFファイルに分割されている場合に分割により4回とカウントされることになる。

2.2.6 統計収集機能

機関ごとにAWStatsを用いた結果の表示を実現するとともに，機関別・機関横断別の統計結果のダウンロードを実現した。

2.3 機関リポジトリアウトプット評価のためのガイドラインの作成

2.1.1にも挙げたが，ROATプロジェクトの過程で，ログデータ収集に関するコンセンサスが機関リポジトリを運営している機関においてそれほど成立していないことが明らかとなった。そのため，アクセスログの他機関への提供ともとれるROATの使用について，プライバシーの観点から差し控える機関が出ることも想定されたため，ROATと同様の機能（ロボットリストの同一性を除く）を実現するための機能要件を

「機関リポジトリアウトプット評価のためのガイドライン」としてまとめ，2012 年 2 月には第 2 版 [26]を公開した。なお，本報告書では，最新の改訂版である第 3 版を収録している。

3. アクセスログの分析と考察 [27)]

ROAT プロジェクトがアクセスログのフィルタリングやカウントの手法を検討する過程で，いくつかの課題が浮かび上がってきた。

3.1 ロボットアクセス

ROAT においてロボットアクセスとしてフィルタリングすべき対象を抽出することを目的として，15 の機関リポジトリの協力を得て，各機関リポジトリの PDF ファイルへのリクエストを行った IP アドレスとユーザエージェントの出現頻度の調査を行った。15 機関のアクセスログはいずれも 2008 年 12 月のログを使用したが，機関リポジトリごとに出現するロボットの種類や頻度，アクセスログ全体に占める割合が大きく異なることが確認された。（図 6）

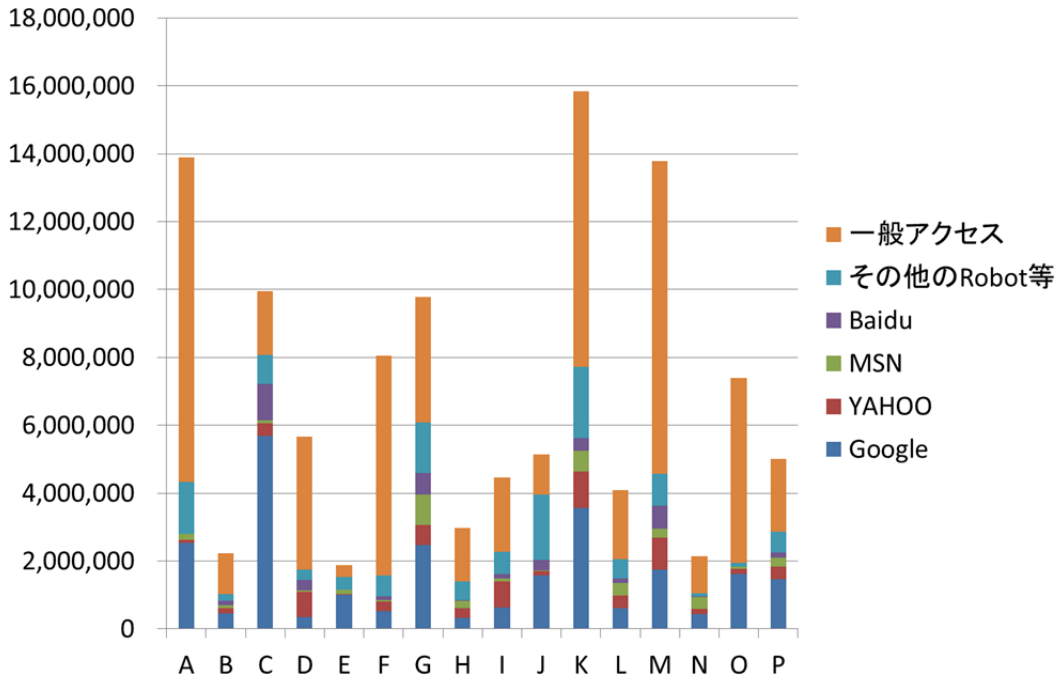

図 6 各機関リポジトリのアクセスログに含まれたロボットアクセスの実数
（2008/1/1-12-31，ただし G は 1/6-12/31，L は 4/1-12/31，O は 6/23-12/31）

さらに抽出されたロボットについて詳細を比較すると，15 の機関リポジトリのすべてに出現したロボットの IP アドレスは 3 件のみであったが，半数以上（8 以上）の機関リポジトリに出現した IP アドレスは 87 件あった。また各サーチエンジンのロボットは複数の IP アドレスを使用しており，今回の調査においては Yahoo で 532 件，Google で 466 件，MSN で 421 件の IP アドレスが確認された。

ロボットアクセスをフィルタリングするにはブラックリスト式に対象となる IP ア

ドレスやユーザエージェントを抽出する方法をとるが，一部の機関リポジトリのアクセスログから作成したブラックリストを全機関に用いるといった運用は適さないことが明らかとなった。

3.2 フィルタリングの効果

ROAT に設定したフィルタリングの効果を確認するため，千葉大学学術成果リポジトリ（CURATOR）の 2010 年 1-3 月のアクセスログを使用して調査を行った。

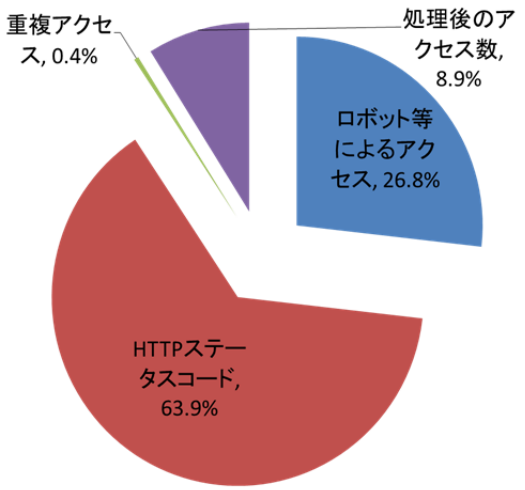

図 7 各処理段階におけるフィルタリングの効果（千葉大学学術成果リポジトリ 2010 年 1-3 月のアクセスログの処理結果）

ROAT のアクセスログ処理については第 3 章で述べたとおりであるが，それぞれの過程で除去されるレコード件数の割合を確認した（図 7）。測定の結果，元のレコード件数に対して，クローラー等のアクセス排除によって 26.8%，ファイルステイタスによる抽出によって 63.9%の除去が行われた。一方，多重アクセスの除去では，わずか 0.4%しか行われなかった。当然のことながら，これらの結果は処理の順番によって変化するが，0.4%という小さな数値は，このフィルタリングの実効性に疑問を投げかけるものであった。

続いて，前述した多重アクセス除去のための秒数設定の妥当性を確認するために，秒数を 1 秒から 50 秒まで 1 秒単位に変化させた場合の，フィルタリングの効果を測定した。結果は図 8 に示すとおり，10 秒前後の設定値以降は全般になだらかな曲線となっており，実際に 10 秒の設定で 50 秒の設定の場合の 83.7%（学外からのアクセス），67.4%（学内からのアクセス）の除去が行われていた。このことから，少なくとも COUNTER 実務指針の 30 秒（PDF ファイルの場合）という設定については，見直しが必要な可能性が高いと言える。こうした指針については，インターネット接続環境の進展や多様な利用形態に関して継続的な観察に基づく見直しと対応が必要になると考えられる。

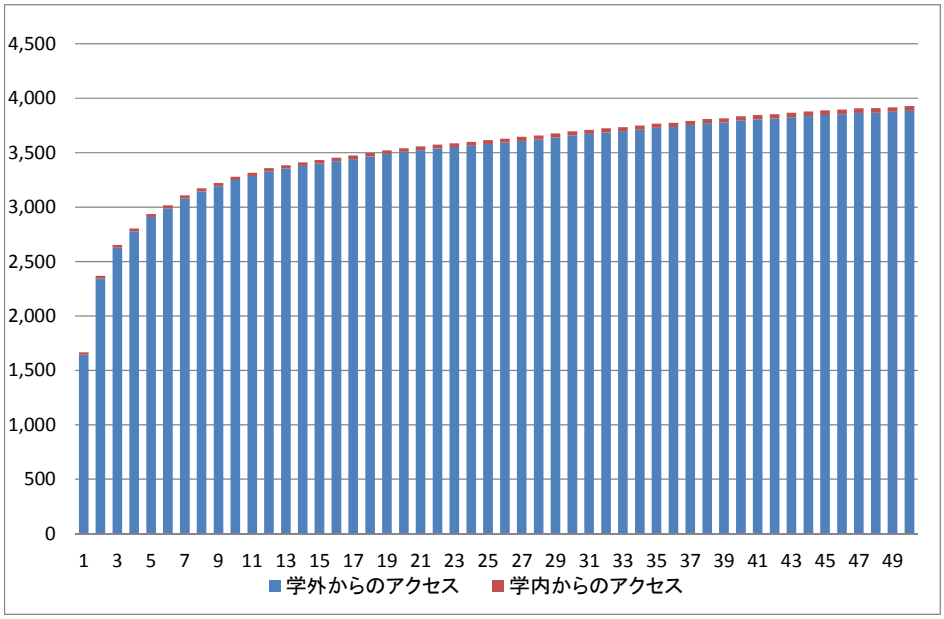

図 8　設定秒数の変化にともなう多重アクセス削除の効果

3.3 利用者単位の集計の可能性

利用データを元にした分析の中で，利用者数あるいは利用者がウェブサイト上で費やした時間量や閲覧したページ数も評価の指標として考えられる。こうした算定を行うためにはセッションの識別が不可欠であり，その実証的検討としてクッキーを利用した場合と IP アドレスだけによる場合のデータから推定される利用者数，セッション数の比較を行った。

検証は千葉大学学術成果リポジトリにクッキーの設定を追加し，2010 年 3 月 3 日から 4 月 14 日までの 42 日間のアクセスログを収集した。クッキーの有効期間は利用者の把握という目的から 1 ヶ月に設定した。

クッキーを拒否する設定にしている場合でもアクセスは有効となるため，どの程度クッキーが付与されていたかは問題であるが，実験期間中の 277,093 件のリクエストのうち 215,105 件（77.6%）にクッキーが付与されていた。またクッキー付きの本文，メタデータ，検索画面および検索結果へのアクセス件数に絞ってカウントしたところ，IP アドレスとクッキーの組み合わせ 7,264 件のうちクッキーと IP アドレスが 1 対 1 で対応しているケースは 5,087 件（約 70%）であった（表 1）。

その他の場合について検討すると，一つの IP アドレスに複数のクッキーが対応しているケース（1,288 件）は，例えば大学でプロキシアドレスを設定して異なる端末で一つの IP アドレスを共有している状況が想定される。この場合は IP アドレスだけで解析するとクッキーを利用した場合よりもセッション数を少なく見積もってしまう可能性があると考えられる。ユーザがクッキーを意図的に消去している場合が想定

表 1　クッキーと IP アドレスの対応状況

Cookieあり	7,264	100%	Cookie:IPアドレス = 1:1	5,087	70%
			Cookie:IPアドレス = n:1 (n>1)	1,288	18%
			Cookie:IPアドレス = 1:n (n>1)	456	6%
			Cookie:IPアドレス = n:m (n>1, m>1)	433	6%

されるが，アクセスログからその状況を確認することはできない。一方，一つのクッキーに複数の IP アドレスが対応しているケース（456 件）では，プロバイダーが端末に対して動的にグローバルアドレスを割り当てている例が多いことが想定された。このケースでは IP アドレスだけで解析するとセッション数を多く見積もってしまう可能性が考えられ，この点でもクッキーによるユーザ（ブラウザ）の識別が有効であると言える。

次にタイムアウト値推定のため，すべてのアクセスについて同一のクッキーが付与された次のリクエストとの時間差を集計（ケース 1）したうち，前回のリクエストが PDF ファイルであるものだけを抽出（ケース 2）して比較した（図 9）。すべてのアクセスに比べて，検索の最終目的物と考えられる論文本文（PDF ファイル）にアクセスしてから次の操作に移るまでには時間をかけていることがわかる。また，20〜25 分

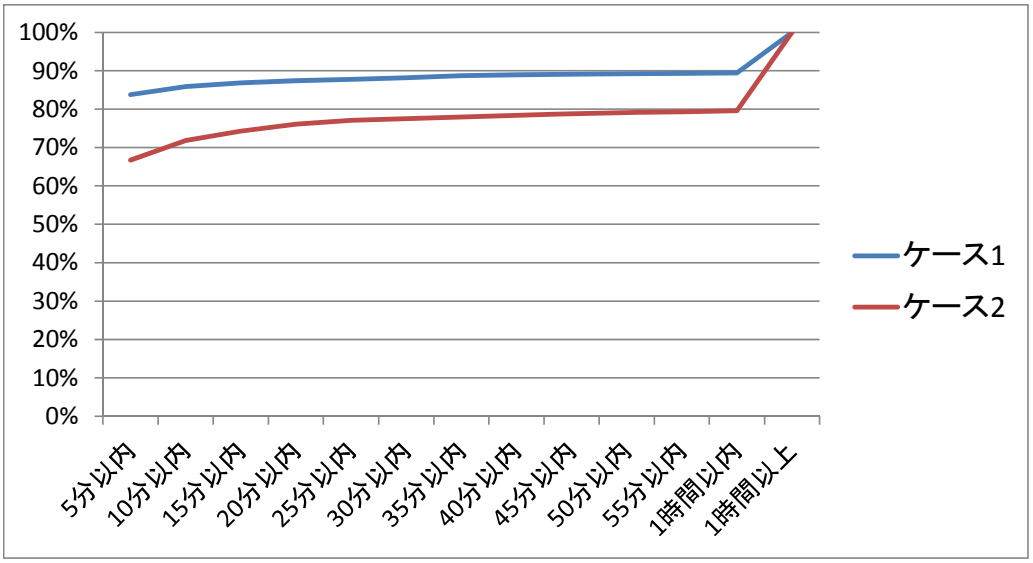

図 9　同一のクッキーが付与された次のリクエストまでの時間差の分布
ケース 1：すべてのアクセス，ケース 2：PDF ファイルから次のアクセス

以降はケース 1，ケース 2 とも上昇率が変わらないことから，機関リポジトリの利用にかかわるセッション時間は 20 分前後と推定できると考えられる。

最後に，タイムアウト値を 300 秒（5 分）から 2400 秒（40 分）まで変化させた時に得られるセッション数を推定したところ（図 10），5 分あるいは 10 分では明らかに短

すぎ，変化が収束するのはやはり 20 分程度と考えられた。興味深いのは，クッキーを使用した算定と IP アドレスを使用した算定にほとんど変化が見られないことであった。これは短い時間では IP アドレスが変化することがほぼないためと考えられ，セッション数については IP アドレスのみによるログでも算定可能と言えるだろう。

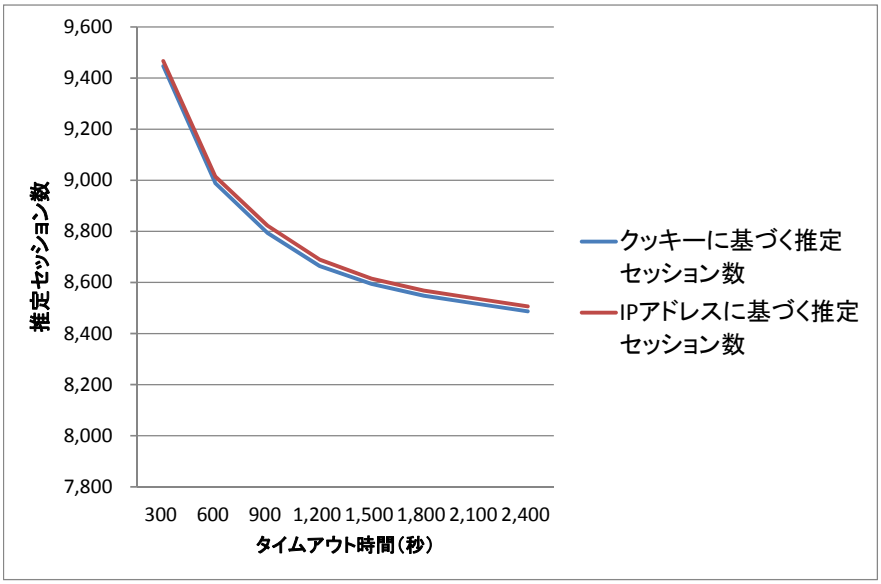

図 10　セッション時間（秒数）の推定値：クッキーによる算定 vs. IP アドレスによる算定

3.4 本文ファイルの URL が動的に生じるシステムへの対応

ROAT では，ログデータの集計結果に対応するコンテンツの情報を付与するために，JAIRO から取得したメタデータの本文 URL を集計結果とマッチングさせている。しかしリポジトリ・システムの中には，メタデータのページから本文ファイルをダウンロードする時に，URL が動的に発行されるものが存在する。こうしたシステムでは，アクセスログには動的に発行された URL が記載され，JAIRO などにハーベストされるメタデータ中の「フルテキスト URL」とは異なるという現象が起こり，ROAT の処理では正確な利用統計を得ることができなかった。

ログデータとメタデータのマッチングに使用する項目として ROAT では本文 URL を最適と判断したが，すべての機関リポジトリが DOI のような重複しない文献 ID を持ちそれがアクセスログにも記載される，さらに JAIRO でハーベストするメタデータにも記載されることになれば，文献 ID をマッチングに使用することができる。この点については，ジャパンリンクセンター（JaLC）の DOI が期待される。

3.5 著作単位でのカウント

同一著作が異なるフォーマットで蓄積されている（例えばhtml版とPDF版がある）場合，それぞれへのアクセスを1アクセスと捉えてカウント数を合算する必要がある。また，一つの著作で各章が別ファイルになっているような場合，著作単位でカウントしようとするとカウント数を合算する必要があるケースもあり，合算の要否は条件によって異なることが考えられた。ROATでは，JAIROのメタデータ中に「本文フルテキスト」として指定されているファイルが複数存在する場合にはそれぞれを別のものとしてカウントできるようにし，必要に応じて後から合算できるようにした（図11）。この処理は手作業に依存せざるを得ないものとしている。

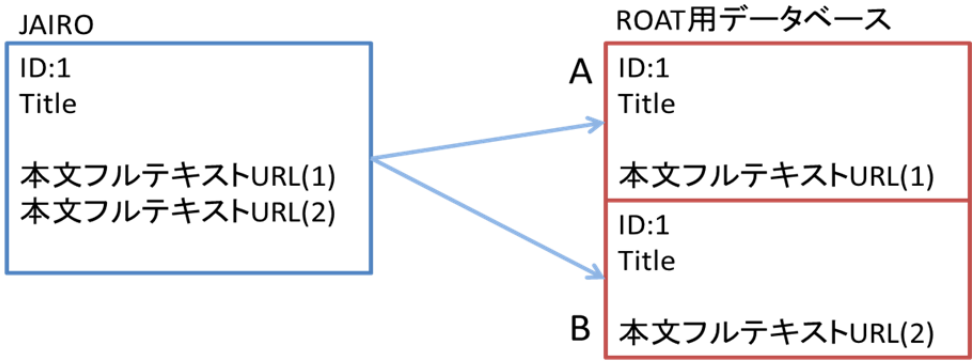

図11　本文フルテキストURLが複数あるデータのメタデータ処理

さらに，ある機関リポジトリに登録されている著作が，共著者所属機関の機関リポジトリにも登録されていることが考えられ，著作単位での真の利用数を測るためには横断的に集計するための方策が必要と考えられる。この点についても，ジャパンリンクセンター（JaLC）の活動により横断的に集計できる環境が整うことを期待したい。

参考文献等

1) IRS: Interoperable Repository Statistics. “Facilitating trust-worthy repository use statistics.” http://irs.eprints.org/ (accessed: 2013-02-27)

2) IRS: Interoperable Repository Statistics. “About the project.” http://irs.eprints.org/about.html (accessed: 2013-02-27)

3) IRS: Interoperable Repository Statistics. “Project Evaluation.” http://irs.eprints.org/evaluationreport.pdf (accessed: 2013-02-27)

4) Bollen, Johan; Van de Sompel, Herbert. “An architecture for the aggregation and analysis of scholarly usage data,” *Opening Information Horizons: 6th ACM/IEEE-CS Joint Conference on Digital Libraries 2006*, Chapel Hill, NC, USA., 2006-07-11/15. 2006.5, p. 298-307. http://arxiv.org/abs/cs.DL/0605113 (accessed: 2013-02-27)

5) *ANSI/NISO Z39.88-2004 (R2010): The OpenURL Framework for Context-Sensitive Services*. 2010.5, 104 p. http://www.niso.org/apps/group_public/download.php/6640/The%20OpenURL%20Framework%20for%20Context-Sensitive%20Services.pdf (accessed: 2013-02-27)

6) "Citebase Search" http://www.citebase.org/
ただし現在は，この URL にアクセスすると The SAO/NASA Astrophysics Data System http://adsabs.harvard.edu/ へリダイレクトされる。 (accessed: 2013-02-27)

7) Brody, Tim et al. “Incentivizing the open access research web,” *CTWatch Quarterly*. 3 (3), 2007.8, p. 42-50. http://www.ctwatch.org/quarterly/articles/2007/08/incentivizing-the-open-access-research-web (accessed: 2013-02-27)

8) Universities UK. *Research Repot: The Use of Bibliometrics to Measure Research Quality in UK Higher Education Institutions*. [2007], 40 p. http://www.universitiesuk.ac.uk/highereducation/Documents/2007/Bibliometrics.pdf (accessed: 2013-02-27)

9) Harnad, Steven. “Open access scientometrics and the UK Research Assessment Exercise,” *11th Annual Meeting of the International Society for Scientometrics and Informatics*. Madrid, Spain, 2007-06-25/27, http://arxiv.org/abs/cs/0703131 (accessed: 2013-02-27)

10) Brody, Tim; Harnad, Steven; Carr, Leslie. “Earlier web usage statistics as predictors of later citation impact,” *Journal of the American Society for Information Science and Technology*. 57 (8), 2006.6, p. 1060-1072. http://onlinelibrary.wiley.com/doi/10.1002/asi.20373/pdf (accessed: 2013-02-27)

11) University of Southampton & Key Perspectives [Inc.] “IRS: Interoperable Repository Statistics,” *JISC Digital Deluge Conference*. Manchester, 2007-07-05/06.

http://irs.eprints.org/papers/IRSpresentation.pdf (accessed: 2013-02-27)

12) MESUR. "Metrics from Scholarly Usage of Resources." http://www.mesur.org/MESUR.html (accessed: 2013-02-27)

13) Bollen, Johan; Van de Sompel, Herbert. "An architecture for the aggregation and analysis of scholarly usage data," *Proceedings of the Joint Conference on Digital Libraries 2006*, Chapel Hill, NC., USA, 2006-06-11/15. http://arxiv.org/abs/cs.DL/0605113 (accessed: 2013-02-27)

14) Bollen, Johan; Rodriguez, Marko A.; Van de Sompel, Herbert. "MESUR: usage-based metrics of scholarly impact," *The 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL07)*, Vancouver, Canada, 2007-06-17/22. http://www.mesur.org/Documentation_files/JCDL07_bollen.pdf (accessed: 2013-02-27)

15) Bollen, Johan; Van de Sompel, Herbert; Rodriguez, Marko A. "Towards usage-based impact metrics: first results from the MESUR project," *Proceedings of the Joint Conference on Digital Libraries 2008 (JCDL08)*, Pittsburgh, PA., USA, 2008-06-16. http://arxiv.org/abs/0804.3791 (accessed: 2013-02-27)

16) Merk, Christine; Windisch, Nils K. *Final Report: JISC Usage Statistics Review.* Joint Information Steering Committee, 2008.9, 24 p. http://repository.jisc.ac.uk/250/1/Usage_Statistics_Review_Final_report.pdf (accessed: 2013-02-27); ［日本語訳］「最終報告書：JISC 利用統計レビュー」 http://www.nii.ac.jp/irp/archive/translation/pdf/Usage_Statistics_Review_Final_report.doc (accessed: 2013-02-27)

17) PIRUS Project Team. *Final Report: Developing a Global Standard to Enable the Recording, Reporting and Consolidation of Online Usage Statistics for Individual Journal Articles Hosted by Institutional Repositories, Publishers and Other Entities (Publisher Metadata and Interoperability Projects 3)*, 2009.1, 20 p. http://www.jisc.ac.uk/media/documents/programmes/pals3/pirus_finalreport.pdf (accessed: 2013-02-27)

18) Shepherd, Peter & Needham, Paul. *Publisher and Institutional Repository usage Statistics: The PIRUS2 Project Final Report*, 2011.10, 74 p. http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-download_wiki_attachment.php?attId=170&download=y (accessed: 2013-02-27)

19) The PIRUS2 Project. "PIRUS2 end of project seminar programme and presentations." http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-index.php?page=PIRUS2+End+of+Project+Seminar+Programme+and+presentations (accessed: 2013-02-27)

20) PIRUS. *The PIRUS Code of Practice for Recording and Reporting Usage at the Individual Article Level: A COUNTER Standard. Draft 1, Release 1*. 2012.11, 15 p. http://www.projectcounter.org/documents/pirus_cop.pdf (accessed: 2013-02-27)

21) NISO/ALPSP Journal Article Versions (JAV) Technical Working Group. *Journal Article Versions (JAV): Recommendations of the NISO/ALPSP JAV Technical Working Group*. 2008.4, 27 p. http://www.niso.org/publications/rp/RP-8-2008.pdf (accessed: 2013-02-27)

22) DINI. "Why OA-Statistics?" http://www.dini.de/projekte/oa-statistik/english/ (accessed: 2013-02-27)

23) DINI. "The technology: how does OA-Statistics work?" http://www.dini.de/projekte/oa-statistik/english/the-technology/ (accessed: 2013-02-27)

24) COUNTER Online Metrics. *The COUNTER Code of Practice for E-Resources: Release 4*. 2012.4, 29 p. http://www.projectcounter.org/r4/COPR4.pdf (accessed: 2013-01-11)

25) Merk, Christine; Windisch, Nils K. *JISC Usage Statistics Review: Final Report*. JISC, 2008.9, 24 p. http://ie-repository.jisc.ac.uk/250/1/Usage_Statistics_Review_Final_report.pdf (accessed: 2013-01-11)

26) 千葉大学附属図書館『機関リポジトリアウトプット評価のためのガイドライン 第 2 版 』 2012.2. http://www.ll.chiba-u.jp/roat/document/GL_v2.pdf (accessed: 2013-02-27)

27) この節の内容は，次の発表を基にしている。
佐藤義則，西佳祐，森一郎，竹内比呂也，土屋俊「アクセスログ分析の方法論的課題」『第 58 回日本図書館情報学会研究大会発表要綱』2010.10, p. 129-132.