リポジトリシステムの JAIRO Cloud 移行について

山本　都寛*

はじめに

明治大学では2007年度より機関リポジトリシステムとして「DSpace」[1]を利用してきたが、長らくソフトウェアやOSアップデートを行わずに運用を続けており、セキュリティや機能不足などの課題があった。古い構成が原因と思われる接続障害も頻発し始めたこと、今後機関リポジトリを幅広く普及させていくためにもシステム更新の検討を開始した。

システムの選定

システム更新にあたり、まずは現行システム（DSpace）をアップデートして継続利用するか、別のシステムへ移行するかを検討した。後者の場合の移行先システムとして、国立情報学研究所（以下、NII）とオープンアクセスリポジトリ推進協会（以下、JPCOAR）が提供する共用リポジト

*やまもと・くにひろ / 明治大学　学術・社会連携部　図書館総務事務室

1　オープンソースのデジタルアーカイブシステムを構築するためのソフトウェア。一般的に、学術機関リポジトリを構築するために使用される。DSpace.「About DSpace」. https://dspace.lyrasis.org/about/（参照 2022-12-12）

リサービス「JAIRO Cloud[2]」を候補とした。2022 年 4 月時点で日本の大学におけるシェアも大きく、スタンダートなリポジトリシステムの位置づけであると考え、機能面、費用面からも妥当だと判断したためである。

現行の機能・管理面の課題	現行システムのバージョンアップ	JAIRO Cloud
画面が分かりづらい	ある程度は改善される可能性有り	操作性を重視した分かりやすい画面構成
DOIの付与が不可	カスタマイズすれば可能	標準機能で可能
英語版ページがない	あり	あり
利用統計が実態と相違する	現行と同様。クローラーのアクセス数も含むため、実態と合わない。	クローラーのアクセスは排除され、実態に合った数値が取得できる。
ソフトウェア管理	本来は5～6年周期で再度アップデートが必要。その都度費用が発生	セキュリティアップデートや必要な機能追加はNIIで適宜実施。無償。
サーバ管理	学内サーバのため機器の更新や管理が必要	クラウド型システムのため機器管理が不要
運用費用（年額）		
年間利用料(保守費用)	非公開	非公開
サーバ関連費用		
計		

図 1　現行システムのバージョンアップと JAIRO Cloud 移行の比較

それぞれのパターンについて調査検討を重ね、図 1 の比較を行った。

画面インターフェースのユーザビリティが良い点や、DOI[3] の付与機能が標準で付属する点など、多くの点で JAIRO Cloud へシステム移行する形にメリットが多い結論となった。特に費用について、ここでは非公開としているが、現行システムの保守費用よりも JAIRO Cloud システム利用料のほうが安価となり、価格的に優位であった。なお、JAIRO Cloud の利用サポートは NII と大学間同士の相互サポートとなり、別途保守費用は発生しない。また、JAIRO Cloud は SaaS 型サービスであるため、課題であったサーバセキュリティもこちらで意識する必要がなくなり、インフラ管理から解放される点も運用上大きなメリットである。

以上の比較から、本学では次期リポジトリシステムに JAIRO Cloud を選定した。

2　JPCOAR.「JAIRO Cloud」. https://jpcoar.repo.nii.ac.jp/page/42（参照 2022-12-12）

3　国立国会図書館 .「国立国会図書館における DOI 付与」. https://www.ndl.go.jp/jp/dlib/cooperation/doi.html（参照 2022-12-12）

学内での検討

事務室内の関係者で JAIRO Cloud への移行可否の検討、および学内のリポジトリ運営方針を決定する学術成果リポジトリ運営部会への審議を図った。多くの大学での導入実績があることや、機能も充実し価格的にもメリットがあることから問題なく承認された。

システム移行作業方法の選定

システム移行作業については、業務委託するか、自身で行うかの選択肢がある。DSpace から JAIRO Cloud へのシステム移行については、NII によりドキュメントやツールが提供されていることと、業務委託の場合は 200 万円以上の作業費用が発生する見込みであることから、自身でデータ移行する形を選択した。

スケジュール

システム移行のスケジュールとしては、2022 年 6 月に利用申請後、2022 年 8 月に本番環境の利用開始、2022 年 12 月を本稼働ターゲットに設定した。

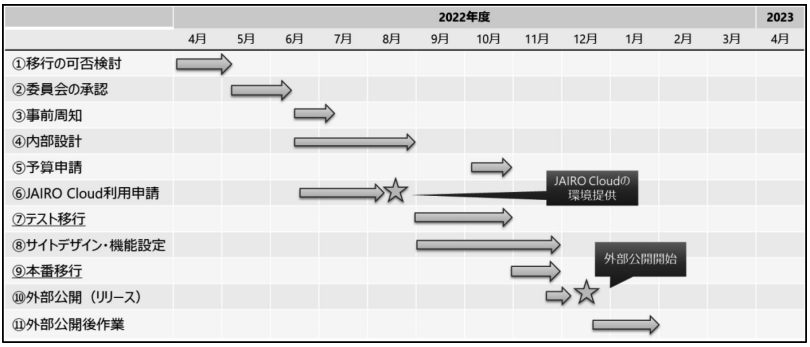

図 2　システム移行スケジュール

テスト環境の構築

データ移行をリハーサルするにあたり、自由に検証を行える環境があることが望ましいため、学内サーバに JAIRO Cloud テスト環境の構築を検討した。

JAIRO Cloud は NII が開発した「WEKO[4]」と呼ばれるリポジトリモジュールをベースに構築されている。WEKO には複数バージョンがあるが、2022 年 6 月時点で利用申し込み可能な JAIRO Cloud 環境[5]は、NetCommons2[6] 上で動作する WEKO2 バージョンで構築されていた。最新版かは不明だが、NetCommons2 と WEKO2 モジュールについては公開されており誰でも取得できるため、学内サーバにインストールして構築することとした。構築にあたってはドキュメント[7]が用意されていたため、それに倣い CentOS7 + MariaDB 構成で NetCommons2 および WEKO2 環境を用意した。構築した環境で少し遊んでみたところ、バージョンの違いからか JAIRO Cloud 本番環境と比較して制限事項はあるように思われたものの、データ移行の検証には問題ないと判断した。

4　WEKO.「WEKO」. https://weko.at.nii.ac.jp/index.php?page_id=0&pcviewer_flag=1&nc_session=nt60j2cropmcropmbnfsv5md21（参照 2022-12-12）

5　一部機関については新バージョンである WEKO3 環境に先行移行済み。他機関についても 2022 年度中に WEKO3 への移行がアナウンスされている。

6　NetCommons 公式サイト.「NetCommons」. https://www.netcommons.org/helpdesk/nc2_migrations（参照 2022-12-12）

7　meatwiki.「CentOS 7 with MariaDB」. https://meatwiki.nii.ac.jp/confluence/display/WEKO/CentOS+7+with+MariaDB (参照 2022-12-12)

テスト環境でのデータ移行リハーサル

大まかに図3のようなフローでデータ移行を行う。

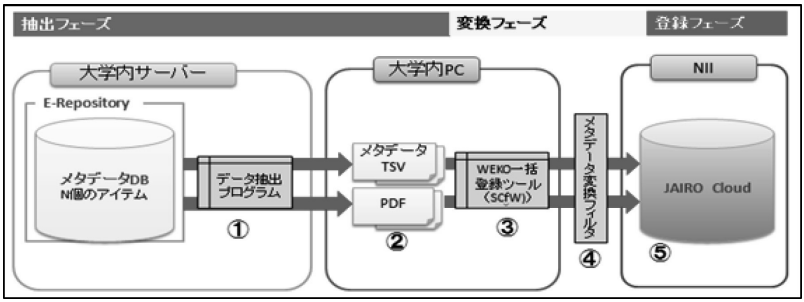

図3　JAIRO Cloudへのデータ移行フロー

DSpaceからのデータ抽出

上図①のように現行DSpaceから移行用データの抽出を行う。NIIから抽出用のツールが提供されており、基本的にはツールを実行するだけで移行対象のコンテンツファイル（本学の場合はほぼPDF）とメタデータがTSV形式で出力される。

しかしながら、本学のDSpaceは導入時にカスタマイズを実施しており、一部メタデータ項目がうまく取得できなかった。そのため、本学DSpaceの仕様に合わせてデータ抽出ツールをカスタマイズする必要があった。ツール内にソースコードが格納されているため、読み解いて本学仕様に必要なロジックを加筆修正した。なお、データ抽出ツールはPerlで作成されていた。

一括登録ツールでのコンテンツアップロード

JAIRO Cloudではコンテンツの一括登録ツールとして、「SWORD Client for WEKO」（以下、SCfW）が提供されている。当ツールは、「変換フィルタ」と呼ばれる入力データと出力データのマッピング定義を使用して、DSpaceメタデータ項目からJAIRO Cloud項目へデータ変換を行っ

た後、WEKO2 側で用意されている SWORD v2.0[8] に準拠したコンテンツ登録 API をキックしてアップロードを行う仕組みを提供するものと理解している。

まずは変換フィルタを作成する必要があるが、標準的な DSpace 用の変換フィルタは NII 側で用意されていた。ただし、前述したとおり本学 DSpace はカスタマイズされているため一部修正する必要はあったが、大きな負担とはならなかった。

2022 年 6 月時点で本学 DSpace 上のコンテンツ数は 17240 件あった。NII からの提供ドキュメント[9] によると、SCfW での一括登録時の入力データ件数は一回あたり 1 万件程度のロット毎にデータ登録を行うことが推奨されていたが、作業効率を考慮して本学では 5000 件ずつロット分割することとした。抽出したデータを基にテスト環境へ SCfW での一括登録リハーサルを行ったところ、データ変換や登録時にエラーとなるデータが散見された。

抽出データのクレンジング

抽出したデータを確認すると、データ型や入力内容が JAIRO Cloud 側で想定しているルールに合致しないデータが多く見受けられた。現在は精査された入力ルールに沿ったメタデータ登録を行っておりある程度データ精度は担保されているはずである。しかしながら、機関リポジトリシステム稼働当初は運用も安定しておらず、コンテンツ登録内容も整備されていなかったと思われること、メタデータの入力項目も比較的自由度が高いことから、内容にばらつきや不整合のあるメタデータが多く存在するものと推察できる。

DSpace 側のメタデータを修正する方法もあるが、件数が膨大なことと

8　Simple Web-service Offering Repository Deposit。リポジトリのコンテンツ操作（登録・更新・削除）を目的としたプロトコル。NII.「WEKO データ登録 API – SWORD API」. https://weko.at.nii.ac.jp/demo/weko/help/ja/html/api/WK24-03.html（参照 2022-12-12）

9　JPCOAR.「JAIRO Cloud への移行の手引き」. https://jpcoar.repo.nii.ac.jp/records/561（参照 2022-12-12）の「3.2.1 データの一括登録」を参照。

手作業で 1 件ずつ修正する必要があることから、抽出したデータをクレンジングする手法で対応することとした。繰り返しリハーサルを行うことや、本番環境での実施に備えて再現性を担保する必要もあることから、EXCEL ベースで修正するなどの手作業は極力避けるべきである。解決のため、NII から提供されているツールとは別にデータクレンジング用のスクリプトを独自で作成することとした。スクリプトは Python ベースで作成し、メタデータ内容の統一、登録時に必要なカラムの追加、5000 件ずつロット分割して TSV 出力する機能などを実装した。Python でのデータクレンジングには、pandas[10] というライブラリが非常に使い勝手が良く、表形式のイメージで柔軟なデータ加工が行えた。

クレンジング後のデータ一括登録

クレンジング後のデータで再度一括登録を行った。5000 件ずつの分割実行で、全データ登録完了までにおよそ 12 時間程度かかった。

データ一括登録後の作業

一括登録エラーとなるコンテンツの個別対応

PDF ファイルの容量が大きいデータなど、SWORD API 側の制限に引っ掛かり SCfW での登録が不可能なコンテンツが 22 件あった。これらについては、前述のデータクレンジングスクリプトで一括登録対象から除外し、全件データ登録後に個別に対応することとした。件数が少ないことから、一旦メタデータのみを一括登録し、PDF ファイルを GUI から手動でアップロードする形とした。

学内限定公開コンテンツの設定

現在はオープンアクセス方針が策定されたため、コンテンツの学内限定公開は受け付けていないが、過去の登録分については未だ学内限定公開で

10　NumFOCUS Inc. .「pandas documentation」. https://pandas.pydata.org/docs/（2022-12-12）

許諾を得ているコンテンツが残っている。JAIRO Cloud での学内限定公開の設定は、「サイトライセンス機能[11]」と該当コンテンツの本文ファイルの閲覧権限を「ログインユーザのみ」と設定することを組み合わせて実現できる。

ただし、本文ファイルの閲覧権限を「ログインユーザのみ」と変更するのは GUI 経由でしか行えない。学内限定公開としているコンテンツは 2022 年 6 月時点で 987 件あり、手作業での実施は現実的ではないため、画面を自動操作して対象コンテンツの設定を変更していく RPA スクリプトを作成して対応した。スクリプトは Python の Selenium[12] モジュールを中心に作成し、予期しない通信断などで中断された場合に備え、再実行しても同じ結果が得られるよう冪等性を担保する設計とした。リハーサルでは夜間に実行して 3 時間程度で完了した。

JAIRO Cloud 本番環境への移行

2022 年 8 月に JAIRO Cloud 本番環境が提供され、各機能をチェックした。テスト環境と比較して若干 UI が変更されている箇所や挙動の違う点は見られたが、データ移行についてはほぼリハーサルしたとおりの手順で問題ないように見受けられた。なお、本番環境へのアクセスは環境提供時点では本学 IP アドレスレンジからのみの制限をかけており、インターネット公開にはデータ移行後に一般公開申請を行う必要がある。

本番環境へのデータ移行については、2022 年 9 月末までの全件データ移行と、全件データ移行後から本稼働前までに登録された差分データ移行の 2 段階に分けて実施した。

全件データ移行

現行システムへの 2022 年 9 月末までの登録分を対象に、JAIRO Cloud

11　JPCOAR.「NC2_WEKO_Manual_ ユーザー利用手引書」. https://jpcoar.repo.nii.ac.jp/records/552（参照 2022-12-12）の「3.3.8.2.7. サイトライセンス認可」。

12　Software Freedom Conservancy.「Selenium」. https://www.selenium.dev/documentation/（参照 2022-12-12）

への全件データ移行を行った。データ抽出、データクレンジング、一括登録、個別対応コンテンツ登録、学内限定公開コンテンツの設定まで、テスト環境でのリハーサル時に実施した手順通りに実施し、エラーなく登録が完了した。データ件数は17338件あり、全件データ登録完了までにおよそ15時間程度、学内限定公開コンテンツの登録は約4時間で完了し、リハーサル時よりも時間がかかった。これは通信経路の違いが原因で、作業端末、テスト環境のサーバはともに学内にあることから学内通信のみで完結するのに対し、JAIRO CloudへはSINET（またはインターネット）経由での接続となるため、ネットワーク環境による通信速度の差であると考えられる。

差分データ移行

全件データ移行時点から2022年11月下旬までに現行システムへ新規登録されたコンテンツについて、JAIRO Cloudへ差分移行を行った。NIIから提供されているデータ抽出ツールは、任意の日付以降にデータ登録された分のみを抽出する機能も実装されているため、同機能を使用して差分データの抽出を行った。対象件数は49件で、全件移行時と同様にデータクレンジングを行い、SCfWでの一括登録を実施した。データ登録は7分程度で完了した。

JAIRO Cloudでの統計情報通知メールの運用

現行システム、JAIRO Cloudともに任意の対象者へアクセス数、ダウンロード数などの統計情報をメール通知する機能がある。現行システムは著者単位で設定できるが、JAIRO Cloudではコンテンツ単位での設定となり、現行と同様に著者単位へメール送信するためには作業工数が増える。

現行システムでは、ユーザ情報で保持する教職員番号と、コンテンツの教職員番号を紐づけて、著者を同定してメール通知を行うシンプルな仕組みとなっている（図4参照）。

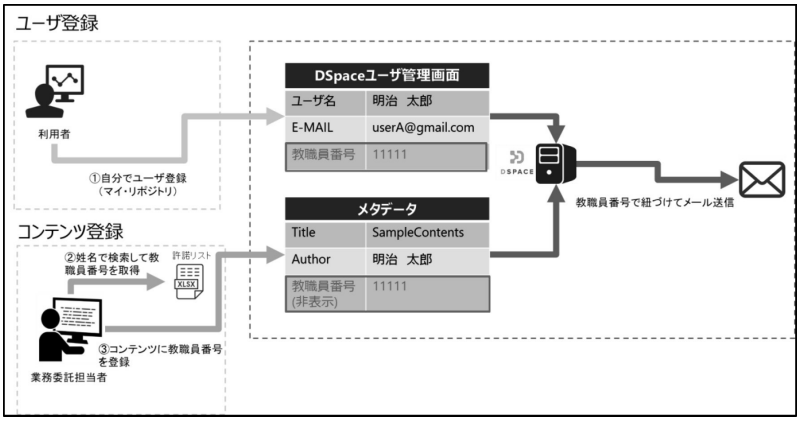

図 4　DSpace の統計情報通知メール送信フロー

一方、JAIRO Cloud では、以下のシステム上の制約事項がある。

①「著者名典拠[13]」の機能で著者の管理を行うが、著者自身では登録ができず管理者のみが登録処理が可能。

②著者単位ではなく、コンテンツ単位で統計情報通知メールの宛先を設定する必要がある。（= コンテンツ登録時に通知先メールアドレスを指定する必要がある）

①により、メールアドレス登録用画面を別途用意する必要があり、管理者にて登録作業を行う工程が増えることになる。また、②の制約から、現行システムのようにユーザ情報とコンテンツ情報で保持する教職員番号で紐づけてメール送信する仕組みは実現できず、コンテンツ一括登録時に著者名典拠から該当著者のメールアドレスを検索して指定する作業が必要となる。以上をまとめると図 5 のような運用フローとなる。

13　JPCOAR.「NC2_WEKO_Manual_ ユーザー利用手引書」. https://jpcoar.repo.nii.ac.jp/records/552（参照 2022-12-12）の「3.3.9. 著者名典拠」。

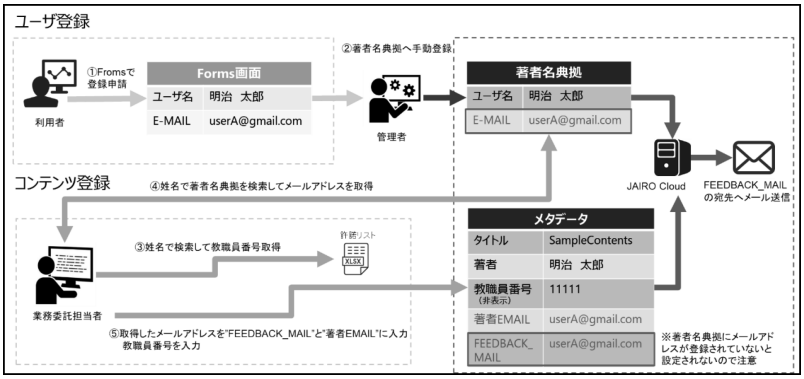

図 5　JAIRO Cloud の統計情報通知メール送信フロー

管理者の作業フロー

著者名典拠ではメールアドレスで登録コンテンツとの著者同定が行われる[14]。メールアドレス未登録の著者については、コンテンツ登録時に設定した著者と著者名典拠データの同定ができず、図 6 のように著者データが複数作成されている可能性がある。

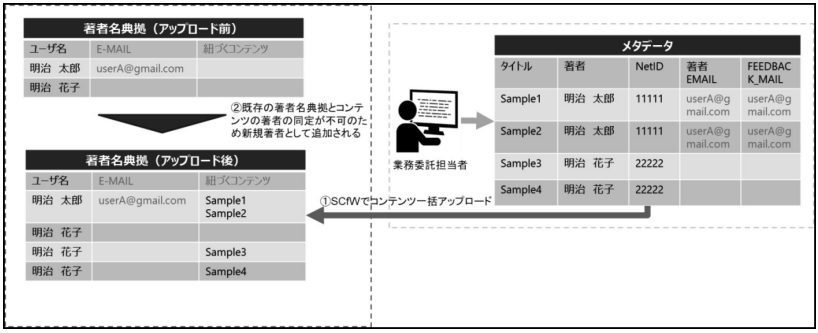

図 6　著者名典拠の著者登録イメージ

14　著者名典拠のユーザ情報に「外部著者 ID」として教職員番号を設定して著者同定に利用することもできるが、外部著者 ID はコンテンツ詳細画面で非表示にできず、教職員番号が公開されてしまうため望ましくない

そのような著者からメールアドレス登録申請があった場合には、著者名典拠の付替え機能[15]を利用して著者データを一つにまとめた後に、各コンテンツに対し統計通知メール送信先メールアドレスの登録作業が必要となる。以上を踏まえると管理者による著者名典拠登録作業は図 7 のフローとなる。

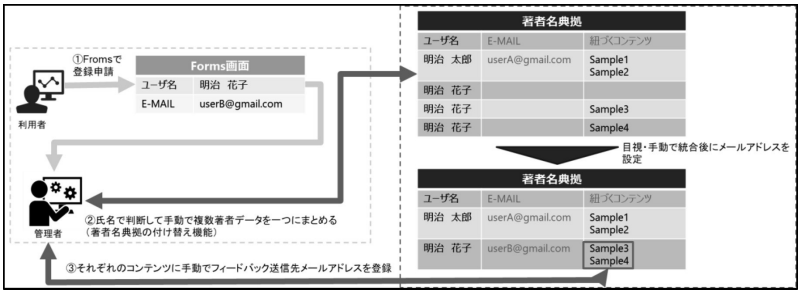

図 7　管理者による著者名典拠登録フロー

リポジトリ業務委託先への説明

本学ではコンテンツ登録作業を含むリポジトリ業務の大部分を業務委託しているため、委託先担当者へ JAIRO Cloud でのデータ登録方法の説明が必要となる。JAIRO Cloud については豊富にマニュアルやドキュメントが整備されてはいるが、本学独自の運用もあるため個別にマニュアルを作成して提供し、データ登録のデモを実施した。

データ登録手順が変更になることで業務委託費用が変動する可能性があるが、執筆時点では未定となっている。現行システムと比較してデータ登録作業が簡便化されるため、費用削減につながることを願う。

15　JPCOAR.「NC2_WEKO_Manual_ ユーザー利用手引書」。https://jpcoar.repo.nii.ac.jp/records/552（参照 2022-12-12）の「3.3.9.2. 付替え」。

データ移行完了後の作業

一般公開申請

データ移行が完了したため、JAIRO Cloud事務局へアクセス制限の解除を依頼した。本学ではCNRIハンドル[16]を継続利用するため、一般公開申請と併せてハンドル最大値も連絡した。公開予定日の14:40ごろに、アクセス制限の解除が完了し一般公開が開始された旨の連絡があった。

CNRIハンドルのリダイレクト先切り替え

コンテンツの公開URLが現行システムからJAIRO Cloudに変更となるため、CNRI事務局へハンドルのリダイレクト先の変更を依頼した。JAIRO Cloud利用申請時に、事務局へCNRIハンドルを継続利用する旨を通知すれば、切り替えに必要な情報が提供される。申請から2日程度でCNRI事務局から切り替え完了の連絡があった。

ハーベスト用ベースURL切り替え

現行システムからJAIRO Cloudへ移行となるため、IRDBによるハーベスト[17]のベースURLも変更となる。この変更手続きに関しては、一般公開が完了したタイミングでJAIRO Cloud事務局にて実施されていたため、本学としての作業は不要だった。

本番運用変換フィルタの作成

移行時に使用した変換フィルタを継続利用した運用も可能だが、入力データの項目名がDSpaceのものとなり将来的に混乱を招く可能性がある。JAIRO Cloudの項目名に合わせた方が運用管理もシンプルになるため、本学では移行用とは別に本番運用向けの変換フィルタを作成した。た

16　インターネット上に存在するデジタルオブジェクト等の資源に対して、永続的識別子を付与・管理・解決するための技術仕様。論文引用の際などハンドルURIを指定すれば永続的なアクセスが保証される。

17　学術機関リポジトリデータベースサポート.「IRDB ハーベスト仕様」. https://support.irdb.nii.ac.jp/ja/harvest（参照 2022-12-12）

だし、変換フィルタで定義したデータタイプ等が既存のアイテムタイプのものと少しでも異なると、一括登録時に同定されず新規アイテムタイプとして作成されてしまうため、作成には十分注意が必要となる。

今後の課題

DOI 付与について

JAIRO Cloud では DOI 付与の機能が標準で付属している。本学はこれまで CNRI ハンドルを付与してきたが、国際的に DOI を利用する場合が多いとの情報もある。付与対象とするコンテンツや条件の精査を行い、CNRI ハンドルと DOI の併用運用を検討することが望ましいと考えている。

統計情報通知メールの運用について

DSpace と比較して、管理者の手作業での対応が多くなる。許諾を得る際に併せて統計情報通知メールの必要有無、およびメールアドレスを取得できればシンプルな運用に落ち着かせることもできそうだが、包括許諾を進めている関係で難しい現状がある。運用を続けていく中で効率的に管理運用していく方法を模索したい。

システム移行時に苦労した点

現行システムと JAIRO Cloud の機能比較

現行システムがカスタマイズされていたため、JAIRO Cloud との機能比較にあたりカスタマイズ機能の洗い出しおよび仕様の理解が必要となった。JAIRO Cloud へ移行することで抜け落ちる機能がないかの判断は慎重に行った。幸い、現行システムのみにある機能については未使用であるものがほとんどだったため、大きな問題はなかった。

テスト環境の不具合

2022 年 6 月時点での NetCommons2 および WEKO2 公開モジュールで

は、ユーザ登録画面からSWORD用ユーザを登録しても、登録時に設定したパスワードでSWORD認証を突破できない。WEKOの認証は、DB上のユーザ管理テーブル格納されたパスワードハッシュ値と、入力したパスワードのハッシュ値を比較し、一致すればパスさせるシンプルな仕組みのように見受けられた。通常ログイン時とSWORDログイン時に入力したパスワードハッシュ値の値について、それぞれの機能を実装しているソースコードにデバッグ用のログ出力のロジックを追加して確認したところ、前者はユーザ管理テーブルの値と一致するが、後者は一致しないことがわかった。対応として、SWORDログイン認証時に取得したパスワードハッシュ値で、ユーザ管理テーブルのSWORDユーザのパスワード値を直接更新することで、強引ではあるがログイン可能とした。通常ログイン時とSWORDログイン時で同じ暗号化方式を使っているように見えたが、根本原因までは読み解いていない。

また、SCfWの登録時にエラーデータが存在すると、稀に画面表示が壊れる、次以降のコンテンツ登録が不可能となる障害が発生した。NetCommons2またはWEKO2のバージョンが最新でないことに関連したものと思われる。ソースコードやDBを調査したところ、登録処理がエラーで中断された場合に、内部的なコンテンツ番号などのシーケンスを管理するテーブルと、実際のシーケンスが不整合となることで、キー重複を起こしてクラッシュしていることがわかった。リハーサル時に何度か発生したことから、WEKOのソースに一部手を加えて原因を検知しやすくすること、および整合性を修復するためのSQL文を予め用意することで素早く環境復元できるよう準備した。

WEKO2利用申請受付締め切りの可能性

2022年度にはJAIRO Cloud側のシステムアップデート作業（WEKO2からWEKO3への移行）が予告されており、WEKO2の利用申請の締め切り時期が不確定な状況だった。DSpaceからWEKO3へのデータ移行ツールは提供されないことから、WEKO2の利用申請が早めに締め切られると自力移行を断念せざるを得ない状況であり、タイトなスケジュールでの学内調整が必要となった。なお、先方のシステム調整の準備が整わず、結果

として 2022 年 12 月時点でも WEKO2 利用申請締め切りはされていない。

終わりに

ドキュメントは豊富にあるものの、現行システムとの比較検討、および疑似環境構築のための公開モジュールが最新でないことに伴う想定外のエラーに大変苦慮したため、当報告書が他大学での JAIRO Cloud への移行検討、または自力移行する際の一助になれば幸いである。