平成 26 年　10 月 1　日
核 融 合 科 学 研 究 所

平成 26 年度　　核融合科学研究所　JAIRO Cloud 移行実験レポート

1.　はじめに

JAIRO Cloud とは、機関リポジトリを新たに構築する機関を当面の対象として、機関リポジトリのシステム環境を提供するサービスである。国立情報学研究所は、この JAIRO Cloud を平成 24 年度から運用している。JAIRO Cloud の今後の展開として、機関リポジトリの維持が困難な機関、JAIRO Cloud の先進機能の利用を望む機関に対しても、サービスの提供を検討している。その際に問題となるのが、既存のリポジトリシステムからの JAIRO Cloud へのデータ移行である。

本研究所は国立情報学研究所に協力し、データ移行の検証を行うべく、平成 26 年度より本実験に参加した。本レポートは、この実験における検証結果である。

2.　実験計画

2-1　実験概要

本移行実験は、核融合科学研究所（NIFS）の機関リポジトリシステム（現在公開中のシステムは外部委託の SaaS 環境で稼動しているが、SaaS 移行前の核融合研保有システムを使用する）から JAIRO Cloud に支障なくデータ移行ができるかどうかを確認するものである。国立情報学研究所（NII）が提供するデータ移行プログラム及びマッピング設定ファイル（フィルタ）、手順書、ワークシートをもとにデータ移行作業を行い、その評価及び問題点の指摘を行うことを主な目的としている。

データ移行においては、移行元システムからのデータ抽出プログラム、データ変換およびアップロード機能を有するツールにより、既存システムのデータを実際にロードすることによるテストを行う。試験に使用したシステムの構成を図 1 に示す。

この実験にあたっては、以下の 4 項目に着目して、実験を行う。

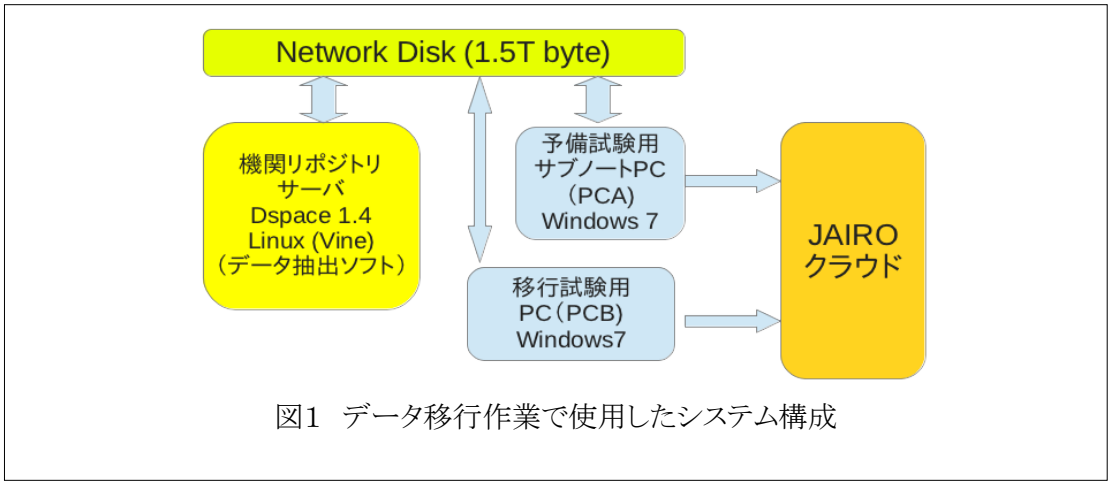

図1　データ移行作業で使用したシステム構成

1. 機関リポジトリサーバは文字コード EUC の環境下で稼動している。この環境と UTF-8 に基づいたデータ抽出プログラムの整合性を確認する。
2. データ移行では、既存サーバと移行で異なる計算機を使用し、かつ一時的ではあ

るがリポジトリシステムが保有する大量のデータを処理する必要がある。そこで、サーバと作業用 PC でネットワークドライブによる共有を行い、データはすべてネットワークドライブ上で取り扱うことを試みる。

3. 従来の移行手順では、データを 2000 件程度ごとに分割してのアップロードが推奨されていたが、2000 件を超えるデータの一括アップロードが可能か否かを確認する。
4. 大量のデータを移行する際には作業用 PC を長時間使用する必要がある。手元に長時間連続使用可能な Windows7 パソコンは、CPU に Atom を使用した小型のものしかなく、この非力な CPU での処理が可能かどうかを確認する。ただし、データ移行用フィルタの構築、修正には作業時の負担を減らす必要があることから、別途ノート PC を用意してこれを使用する。

2-2 データロード実験作業

データロード実験計画における作業は次のとおりである。

表 1 データロード実験作業一覧

作業項目	作業内容	作業主体
移行元リポジトリからのデータ抽出	移行元リポジトリからのデータ抽出プログラムをインストールし、データ抽出	核融合科学研究所
JAIRO Cloud 実験環境構築	データロード実験用の JAIRO Cloud 環境を構築	国立情報学研究所
フィルタ作成・修正	移行元システムのデータ項目と JAIRO Cloud のデータ項目とのマッピング設定	核融合科学研究所
サンプルデータロード	サンプルデータ（100 件程度）を JAIRO Cloud 実験環境にロードし、問題がなくデータがロードされたかどうかを検証	核融合科学研究所
大量データロード	大量データ（8482 件）を JAIRO Cloud 実験環境にロードする	核融合科学研究所
登録結果確認	大量データロードについて、問題がなくデータがロードされたかどうかを検証	国立情報学研究所・核融合科学研究所

2-3 スケジュール

実験のスケジュールは次のとおりである。

表 2 データロード実験スケジュール

		4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月
準備	移行元からのデータ抽出		⇒						
	JAIRO Cloud 実験環境構築			⇒					
	フィルタ作成・修正				⇒	⇒			
サンプルデータロード						⇒	⇒		
大量データロード								⇒	
登録結果確認									⇒

3. 実験の実施

3-1　実験の準備（データ抽出、フィルタ作成・修正）

3-1-1　機関リポジトリサーバからのデータの抽出

機関リポジトリサーバにデータ抽出用プログラムを導入して、データ抽出を行った。プログラムの設定ファイルに関しては、特に問題なく設定を記述できた。

3-1-1-1　サーバーの文字コード

核融合研のサーバは、EUC 環境で稼動し、リポジトリのデータベース内部では UTF-8 を使用している。そこで、データ抽出に関して、A. EUC 環境下での作業、B.言語設定を UTF-8 に切り替えての作業、の二通りについて試みた。

A. 言語設定を EUC にしての抽出。情報研から提供された抽出プログラムの文字コードを EUC に変換して、実行した。データの抽出自体は正常に行われ、UTF-8 でのメタデータファイルが生成された。ただし、メタデータファイル第一行の項目名で漢字を使用している項目では EUC が使われたため、文字化けを起こした。
B. サーバの言語設定を UTF-8 にしての抽出。こちらでは、特に問題なく抽出作業が行われた。

以降の作業では、UTF-8 環境下で抽出したメタデータを使用して作業することとする。

抽出プログラムが付与するメタデータの項目において漢字を使用した名称を使用していなければ、EUC 環境下でも問題なく抽出が可能と思われる。

3-1-1-2　データの出力先の選定

当該システムでは、ネットワークディスクによりサーバと作業用パソコンとでデータの共有を行っている。大量のデータコピーを避けるために、データはネットワークディスク上に置いて作業することとした。

まず、データ抽出プログラムで出力先をネットワークディスク上に設定して抽出作業を行った。データの抽出自体は問題なく行われたが、最後に全データをまとめて zip 圧縮するプロセスでエラーが発生し、プログラムが終了している。

次に、一旦サーバのローカルディスク上にデータを抽出し、その後、ネットワークディスクへコピーする手順を試みた。この手順では、zip 圧縮時の異常終了は起きず、全データを処理できている。

ただし、後の作業で全データを圧縮した zip ファイルは使用しておらず、このプロセスは不要と考える。

3-1-2　メタデータファイル内の項目前処理

核融合研のリポジトリでは、巻、号、ページをひとつのフィールドにまとめて記載している。これをデータ転送前に分割する必要があった。そこで、C 言語で簡易な分割プログラムを作成し、転送作業前に巻、号、ページを別個のフィールドに分割した。

3-1-3　サンプルデータロード用のメタデータファイルの作成

出力されたメタデータファイルをアイテムタイプでソートし、各アイテム毎に 20 件弱ずつを抽出して、サンプルデータロード用のメタデータを作成した。

また、抽出した項目を除外して、最終データロード用のメタデータファイルを生成した。

3-1-4　フィルタの作成

当初、先行して実験を行った筑波大学のフィルタを元にして、NIFS 用のフィルタを構築する予定であったが、NIFS リポジトリのメタデータ項目数が少ないことから、NII 提供の標準設定を使用してフィルタを構成した。フィルタ作成にあたっては、各アイテムタイプごとに使用している項目名を抽出する必要がある。ここでは、先に作成したサンプルロード用メタデータを元にして、アイテムタイプ毎に使用している項目をピックアップした。各アイテムタイプ毎に使用されている項目名をピックアップする支援ツールがあれば、作業がはかどる。

フィルタの編集にあたっては、SCfW を使用するためには JAIRO Cloud サーバに接続されている必要がある。オフラインでのフィルタ編集ができることが望ましい。

フィルタの項目では当初手入力によりフィルタを記述したが、後述のとおりタイプミスによる登録不良が多発した。登録前に、フィルタとメタデータファイルの整合性（一方のみに現れる項目名、未使用の項目名の抽出）を検証する支援ツールによる事前検証ができれば助かる。

3-2　データロード（サンプルデータロード、大量データロード）

3-2-1　サンプルデータロード

まず、70 件程度をピックアップしたサンプルデータのロードを実施した。ここでは先に述べたフィルタ内項目の記述ミスによる登録不良が多発した。これを修正後に、2 回目の検証アップロードを行う。本アップロードでは、後述の 2 件のデータを除いてアップロードが終了した。

アップロードの全データをチェックし、一部設定の見直し（複数登録可にすべき項目がそうなっていなかったのでフィルタを修正）の後にアップデート操作を実行したが、全データでエラーとなり処理できなかった。これに関しては、NII に問い合わせ、JAIRO Cloud 側のアイテム設定を同時に変更する必要があるとのことで、JAIRO Cloud 側の設定を修正し、再度アップロードを実施し、完了した。

この修正作業において、JAIRO Cloud での設定と SCfW のフィルタ設定での整合性をチェックするツールがあれば、有用であると考える。

ここまでの登録操作は、図１内、サブノートパソコン（PCA）にて実施した。

3-2-2　全件一括ロード

PCA から PCB にネットワークディスクを介してフィルタを転送し、PCB による全件一括登録操作に移った。

3-2-2-1　全件一括ロード初回

サンプルデータロードの後に、残りの 8300 件の一括アップロードを試みた。約 90 分経過時点で、2102 件目でエラーが発生した。ポップアップウインド上で処理継続の操作を行ったが、エラーが再度出たため、登録操作を中断した。クラウド上のリポジトリを見ると当該コンテンツが登録されていることを確認した。また、登録用 PC のログでは、当該コンテンツの登録が一旦完了した後に 600 秒経過してエラーが出ている模様であった。

なお、登録処理を中断したため、更新用のメタデータファイルは生成されてい

ない。

3-2-2-2　全件一括ロード第二回

第一回で登録された 2102 番目までのデータを除いたメタデータを作成し、二回目のコンテンツ登録（約 6300 件）を実施した。前回エラーが起きた時点を過ぎてもエラーにならないことを確認し、夜間に処理を継続することとした。約 6 時間後に遠隔操作で登録処理 PCB の状況を確認したところ、エラーなく処理を終了していたので、update メタデータを生成した。

3-2-3　不具合データチェック

大量データロード結果を精査したところ、ページ番号中に-（ハイフン）を含むアイテムがあった場合、SCfW の仕様ではそのハイフンを開始ページと終了ページの区切りとして処理してしまうことがわかった。レアケースであるため、ページ番号中のハイフンを別の文字に置き換えることで対処が可能である。ただし、近年はハイフンを含むページ番号が使われることも多くなってきており、今後は WEKO 側での対応が求められるかもしれない。

4. まとめ

DSpace　ver 1.4 からの JAIRO Cloud へのデータ移行を国立情報学研究所のサポートの元に実施した。今後、数十件程度の機関リポジトリから JAIROCloud へのシステム移行を受け入れるならば、以下に掲げるデータ移行に関する手順書の整備、および支援ツールの整備が望まれる。

4-1　手順書の整備

今回の実験において提示された手順書では、移行用プログラム上での操作はきめ細かに説明されていた。それと比較して、全体の作業の流れ、移行プログラムがどのように動作するかという箇所の説明が弱いと感じた。

移行作業において、中心となるのはメタデータの抽出、および、フィルタの作成であるが、作成したフィルタがデータ移行プログラムの中でどのように使われるか、抽出されたメタデータのうちどの項目は自動的に処理される（フィルタ作成時に組み込む必要がない）か、といった情報がほしい。これがあれば、メタデータの項目からフィルタに反映させる手順や方針を立てやすい。

また、今回の実験では、SCfW によるフィルタ作成の手順書が先に提示されたが、作業全体の手順書が先のほうがよいと思う。

一旦作成したファイルによりサンプルデータを移行した後、フィルタの修正を行う場合があるが、その際には JAIRO Cloud 本体と SCfW フィルタでの整合が取れている必要がある、との注意書きも必要である。同時に後述の支援ツールを要望する。

4-2　支援ツールの整備

今回の実験において、いくつかの支援ツールがあれば作業を確実に短時間で行えると感じた。実施状況の項目でも触れたが、以下に再掲する。

・抽出されたメタデータから、各アイテムタイプ毎に使用されている項目名を抽出する。
・オフラインでの変換用フィルタの編集。

・SCfW のフィルタとメタデータ項目の整合性検査（一方にしか存在しないメタデータ項目名の表示）。これによりミスタイプによる登録失敗を事前に防ぐことができる。
・JAIRO Cloud サーバのアイテムタイプ毎の項目と SCfW フィルタ設定の整合性確認。フィルタを修正した場合などでの不整合をあらかじめ確認することができるのと同時に、どちらのどの箇所を修正すれば不整合を解消できるか、の判断に有用。

また、リポジトリサーバからのデータ抽出プログラムにおいては、
・全データを一括して zip 圧縮する機能は不要（もしくはオプションで不使用にできる）。
・付加する項目名は非漢字が望ましい。（環境の文字コードの影響を受けにくくなる）。