IM-Workflow
Ver.7.2

クローラ 仕様書

2010/07/30 初版

＜＜　変更履歴　＞＞

変更年月日	変更内容
2010/07/30	初版

＜＜　目次　＞＞

1　はじめに 1
1.1　目的 1
1.2　前提条件 1
2　システム概要 2
3　動作仕様 3
3.1.1　インデックス作成対象 3
3.1.2　閲覧可能権限 6
3.1.3　添付ファイル情報のインデックス化 7
3.1.4　過去案件参照可能ユーザを変更した場合のインデックス再作成 9
3.1.5　最終クローラ起動日時の保存 10
3.1.6　エラー発生時の動作仕様 11
3.1.7　その他の注意事項 12
4　IM-Workflowクローラの拡張 13
4.1　リスナーの呼び出し 13
4.2　リスナーの設定 14
4.3　リスナーの作成 14
4.3.1　実処理の記述 14

1 はじめに

1.1 目的

IM-Workflow クローラは、intra-mart と全文検索エンジンサーバ Solr(以下 Solr サーバと略します)との連携が行われている環境で、intra-mart の IM-Workflow の案件情報を収集し Solr サーバに文書として登録する機能を持つバッチプログラムです。

本資料では、IM-Workflow クローラの動作仕様と拡張方法について解説します。

1.2 前提条件

IM-Workflow クローラを動作させるためには、IM-Workflow が動作する

・ intra-mart WebPlatform/AppFramework (アドバンスト)版以上

の製品がインストールされている必要があります。

また、intra-mart とは別に Solr サーバを構築し、設定を行う必要があります。

構築手順については、「IM-ContentsSearch セットアップガイド」 をご参照ください。

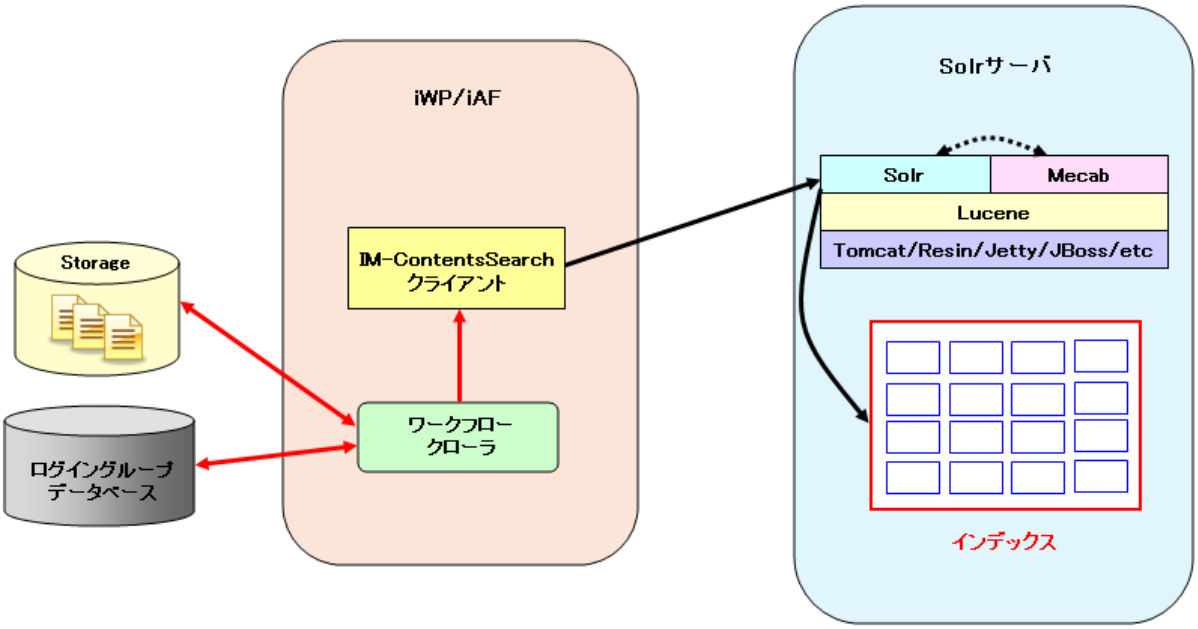
本資料では全文検索機能固有の用語、IM-Workflow 固有の用語が使用されています。

その為、前提知識として全文検索機能と IM-Workflow の基本的な仕様を理解している必要があります。

全文検索についての十分な知識を有していない場合は、「IM-ContentsSearch セットアップガイド」 を、IM-Workflow についての十分な知識を有していない場合は、「IM-Workflow 仕様書」 を事前にお読みください。

2 システム概要

IM-Workflow クローラは、intra-mart 上で intra-mart Batch Server により実行されるバッチプログラムです。
バッチプログラムを起動すると、インデックス作成対象の IM-Workflow の案件情報を取得し、IM-ContentsSearch クライアントAPIを利用して、Solr サーバに IM-Workflow の案件情報を文書として登録します。
文書を登録し、インデックスが作成された IM-Workflow の案件情報は、intra-mart の全文検索画面からの検索が可能となります。
以下に IM-Workflow クローラのシステム構成とバッチ登録情報を示します。



IM-Workflow クローラ システム概要

バッチ名	IM-Workflow 全文検索クローラ
バッチ ID	IMWSolrCrawler
実行プログラム言語	JAVA
実行プログラムパス	jp.co.intra_mart.system.workflow.solr.MatterCrawler

バッチ登録情報

3 動作仕様

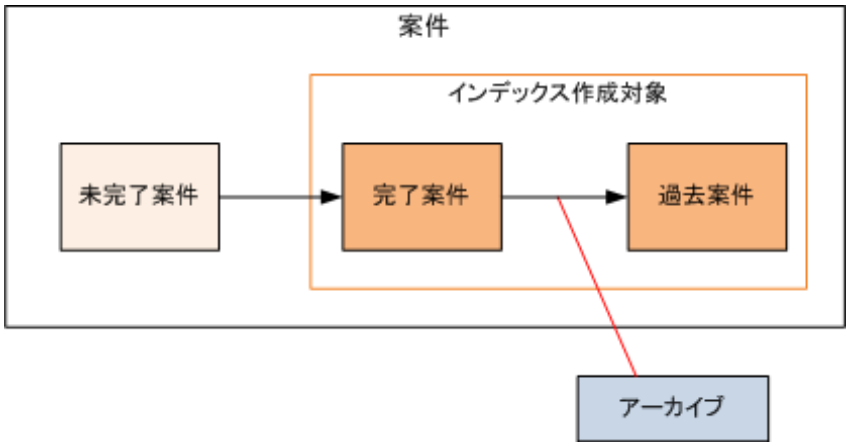
この章では、IM-Workflow クローラの動作仕様について詳しく解説します。

3.1.1 インデックス作成対象

IM-Workflow クローラがインデックスの作成対象とする IM-Workflow 案件は、最後に IM-Workflow クローラが正常終了してから今回起動するまでの間に完了した案件となります。
（初回起動時の対象期間は 3.1.5 最終クローラ起動日時の保存 をご参照ください。）

以下の状態の案件が作成対象になります。

- 完了案件
- 過去案件



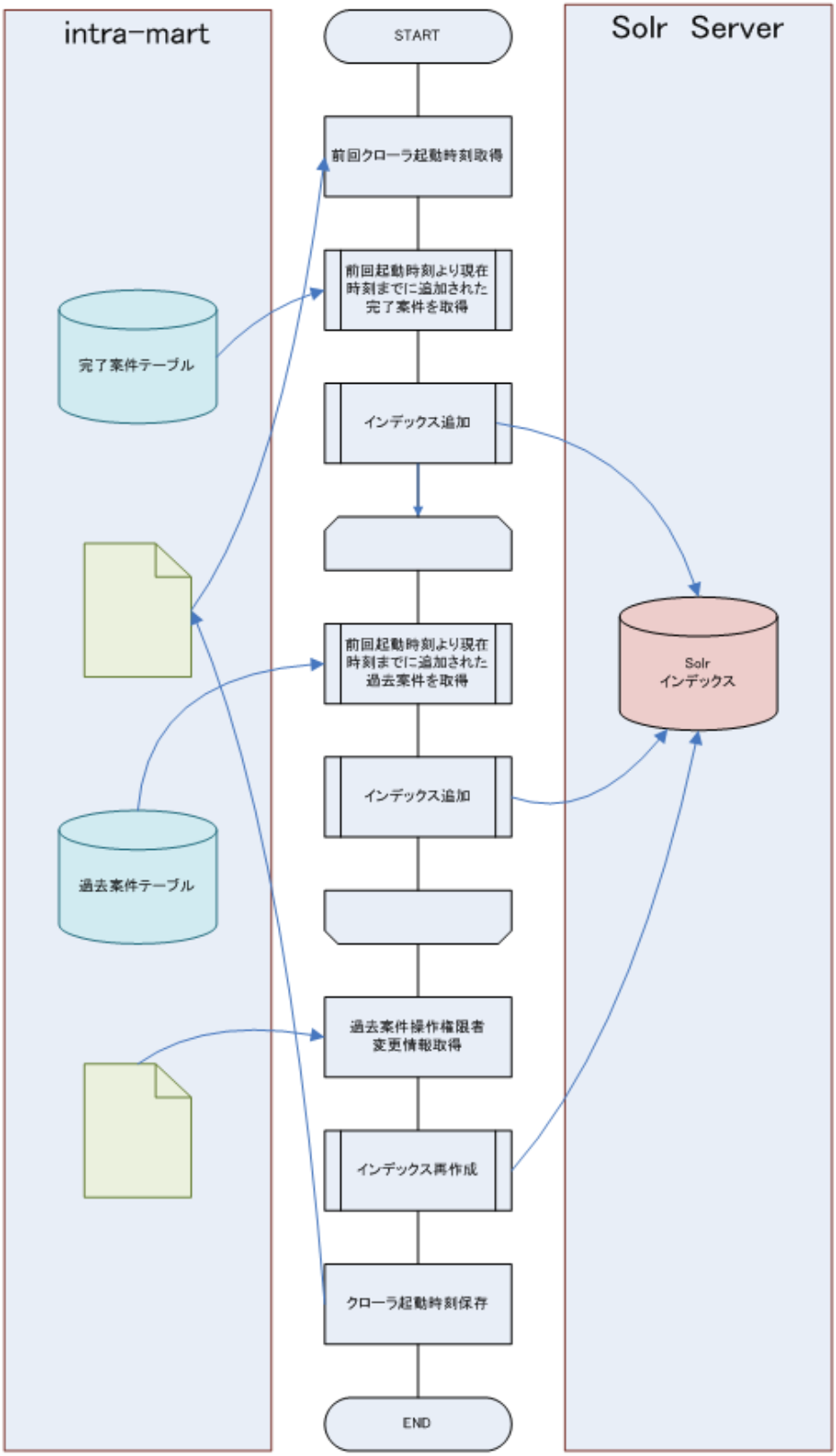
インデックス作成対象の案件状態

状態	説明
未完了案件	申請により案件が開始した後、まだ案件が完了していない状態
完了案件	案件が完了した状態
過去案件	完了案件が「アーカイブ機能」により アーカイブ領域に退避された状態

IM-Workflow の案件状態

3.1.1.1 処理フロー

IM-Workflow クローラの処理フローは以下になります。



IM-Workflow クローラの処理フロー

3.1.1.2 登録情報

IM-Workflow クローラでは、案件情報を以下の内容で登録します。

※テキスト (text_ngram または text_morph) に登録したデータが検索時に検索対象となるフィールドです。
Solr スキーマ定義に関する詳細は、「IM-ContentsSearch プログラミングガイド」 3.1 イントラマート標準 Solr スキーマ定義 をご参照ください。

項番	フィールド名	格納されるデータ	設定値	備考
1	id	文書を一意に識別する ID	imw_システム案件 ID	
2	type	文書のタイプ	imw 固定	
3	url	詳細画面 URL	workflow/common/switch/switch_content_detail.jssp	IM-Workflow 詳細画面を呼び出す為の URL を設定
4	id_original	詳細画面を表示するために 必要なパラメータ	・ログイングループ ID ・システム案件 ID ・ユーザデータ ID ・画面種別	IM-Workflow 詳細画面を呼び出す為のパラメータを設定
5	title	タイトル	案件名	
6	text_ngram	N-gram 用テキストデータ	案件名 フロー名 【添付ファイルがある場合】 ・ファイル名 ・ドキュメント内テキスト**※**	N-gram を使用する場合のみ設定 **※ファイルフォーマットがサポート対象の場合は登録**
7	text_morph	形態素解析用テキストデータ	上記に同じ	形態素解析を使用する場合のみ設定
8	sids_allowed	閲覧可能権限	【完了案件】 ・処理権限者、参照者、確認者 【過去案件】 ・処理権限者	**※詳細は次章を参照**
9	sids_denied	閲覧不可権限	設定なし	
10	record_date	登録日	文書の登録日時	
11	*_string	文書タイプ固有の文字列データ	IM-Workflow 固有のデータ	
12		matter_number_string	案件番号	
13		apply_base_date_string	申請基準日	
		apply_auth_user_name_string	申請者名	
14		flow_name_string	フロー名	

3.1.2 閲覧可能権限

全文検索で IM-Workflow 案件情報を閲覧可能なユーザは、以下になります。
完了案件、過去案件で閲覧可能ユーザは異なります。

【案件状態】

■ 完了案件

1. 処理者（処理権限者）
2. 参照者
3. 確認者

ユーザの種類	説明
処理者（処理権限者）	案件上のノードに対して、本人として処理権限を持つユーザ
参照者	案件に対して、参照権限を持つユーザ （案件操作権限者）
確認者	案件に対して、確認権限を持つユーザ

■ 過去案件

1. 処理者（処理権限者）

ユーザの種類	説明
処理者（処理権限者）	案件上のノードに対して、本人として処理権限を持つユーザ

※ アーカイブバッチの標準では、過去案件の参照可能ユーザを処理権限者としています。
そのため、過去案件の参照可能ユーザを追加するには、案件退避処理リスナーをカスタマイズする必要があります。

※ 過去案件の参照可能ユーザを変更した場合は、インデックスの再作成が必要になります。
⇒ 詳細は 3.1.4 過去案件参照可能ユーザを変更した場合のインデックス再作成をご参照ください。

3.1.3 添付ファイル情報のインデックス化

IM-Workflow 案件にドキュメントが添付されている場合、IM-Workflow クローラでは案件情報に加え、添付ファイル情報もテキストフィールドに登録します。
登録される内容は以下の通りです。

- ◆ ドキュメントのファイルフォーマットがサポート対象外の場合
  - ・ ファイル名

- ◆ ドキュメントのファイルフォーマットがサポート対象の場合
  - ・ ファイル名
  - ・ ドキュメント内テキスト

3.1.3.1 ドキュメント内テキストのインデックス化

ドキュメント内テキストがテキストフィールドに登録された場合は、ドキュメント内テキストのインデックスが作成され、全文検索の検索キーワードにドキュメントに含まれる単語を指定することで、該当の案件を検索結果として取得することが可能となります。

但し一つのフィールドに索引付けする単語数は Solr サーバ側の設定に依存するため、案件情報にサイズの大きいドキュメントや複数のドキュメントが添付ファイルとして登録されている場合、ドキュメントの最後までインデックス化されない可能性があります。
ドキュメント内のインデックス化されていない単語を検索キーワードに指定した場合には対象案件を取得することはできませんので、ご注意ください。

テキストを抽出する際の制限事項に関しては、「IM-ContentsSearch プログラミングガイド」 6.2.2　テキストを抽出する際の制限事項　をご参照ください。

3.1.3.2 テキストを抽出することができるファイルフォーマット一覧

テキストを抽出できるファイルフォーマットの一覧を下表に示します。
ファイルフォーマットと**ファイルの拡張子が下記表に一致しない場合は、サポート対象外とみなされます**ので、ご注意下さい。

ファイルフォーマット	ファイルの拡張子
テキスト	txt
PDF	pdf
HTML	html,htm
XML	xml
ZIP	zip
Microsoft　Word	doc, docx※
Microsoft　Excel	xls, xlsx※
Microsoft　PowerPoint	ppt, pptx※

表 3.1.3.2-1 サポート対象ファイルフォーマット一覧

※これらの拡張子のファイルは、テキストの抽出方法にオプションツールを使用する場合のみ対応しています。
詳細は「IM-ContentsSearch プログラミングガイド」 6.2.2 テキストを抽出する際の制限事項 をご参照ください。

3.1.3.3 添付ファイルがサポート対象外の場合の動作仕様

サポート対象外のドキュメントが添付されていた場合は、警告ログを出力しファイル名のみインデックスを作成して処理を継続します。

#出力されるログの例

[WARN] j.c.n.i.s.SolrManager - 次のファイルは、テキスト抽出対象外です:。 [ファイル名]

3.1.3.4 添付ファイルからのドキュメント取り出しに失敗した場合の動作仕様

ドキュメントからのテキスト抽出に失敗した場合は、警告ログを出力しファイル名のみインデックスを作成して処理を継続します。

#出力されるログの例

[WARN] j.c.n.i.s.SolrManager - ファイルからテキストを抽出できません: [ファイルのパス]

3.1.3.5 パスワード付添付ファイルの場合の動作仕様

添付ファイルにパスワードが設定されている場合、IM-Workflow クローラはドキュメントからテキストの抽出は行いません。警告ログを出力し、ファイル名のみインデックスを作成して処理を継続します。

またパスワード付zip ファイルの場合は、暗号化方式により動作が若干異なります。

- ZIP の暗号化方式が ZIP 2.0 互換の場合
  zip ファイル名、ZIP 内のファイル名を登録し、警告ログを出力して処理を継続

#出力されるログの例

[WARN] j.c.n.i.s.u.ZipExtractor - ZIP 内の次のファイルは、暗号化されているため、処理されませんでした [Zip ファイル名]

- ZIP の暗号化方式が ZIP 2.0 互換以外（AES-128bit・AES-256bit など）の場合
  zip ファイル名を登録し、警告ログを出力して処理を継続

#出力されるログの例

[WARN] j.c.n.i.s.SolrManager - ファイルからテキストを抽出できません: [ファイルパス]

3.1.4 過去案件参照可能ユーザを変更した場合のインデックス再作成

過去案件参照可能ユーザ(過去案件操作権限者)を変更した場合は、以下の CSV ファイルを作成して変更した日付とシステム案件 ID を記述してください。次回のクローラ起動時にこの CSV ファイルを読み込み、該当するシステム案件 ID に対してインデックスの再作成を行います。

CSV ファイルの名称と場所は以下になります。

%StorageService%/storage/workflow/data/ログイングループ名/solr/arc_matter_reindex.csv

フォーマットは「変更日付(yyyy/MM/dd), システム案件 ID」です。

```
2010/06/30,ma_5hx2rpiicqe1q3o
2010/07/08,ma_5hx2qcql78bqm3o
2010/07/20,ma_5hx2r98fct61v3o
```

変更日付が 最終クローラ起動日時から今回のクローラ起動日時までの期間に該当するデータがインデックス再作成の対象になります。この変更日付を編集することで、インデックス再作成の対象にすることが可能です。

3.1.5 最終クローラ起動日時の保存

IM-Workflow クローラでは処理が正常終了した場合、クローラの最終起動日時をファイルに保存します。
次回起動時にはこのファイルを参照し、前回の起動日時より後に完了した IM-Workflow 案件を登録対象として抽出します。

最終起動日時保存ファイルは、クローラの初回正常終了時に以下のディレクトリ内に作成され、次回からは日時の更新のみが行われます。
日付のフォーマットは、**yyyy/MM/dd hh:mm:ss** 形式です。

%StorageService%/storage/workflow/data/ログイングループ名/solr/**last_crawling_date**

このファイルの日時を編集することで、指定した日時より後に完了した IM-Workflow 案件をインデックス作成対象にすることが可能です。

3.1.5.1 初回起動時の動作仕様

初期状態の場合、上記ディレクトリ及びファイルは存在していません。
この場合、既定値として”**2000 年1月1日 0 時 0 分 0 秒**”より後に完了した IM-Workflow 案件をインデックス作成対象とする動作になっています。
初回起動時の対象日時を明示的に指定したい場合は、上記ファイルを作成して対象日時を上記フォーマットにて設定してください。

3.1.6 エラー発生時の動作仕様

IM-Workflow クローラ実行中にエラーが発生して異常終了した場合の動作について説明します。
クローラ実行中に予期しないエラーが発生してクローラが異常終了した場合、データベースとは異なり **Solr サーバ**に対してロールバック処理を行い、インデックス情報をクローラ実行前の状態に戻すことはできません。
従って、今回のクローリングでエラーが発生する前に作成したインデックス情報はそのまま残ることになります。
但し、異常終了時には以下の処理は行いません。

- 登録済インデックスの即時反映
- インデックスの最適化
- 最終クローラ起動日の更新

3.1.6.1 エラー発生時の作成済インデックス

クローラ実行中にエラーが発生して異常終了した場合、同一クローリングにてエラー発生前に作成したインデックスの即時反映処理は行われていないため、すぐに検索結果としてエラー発生前に登録した案件情報が取得されることはありません。しかしSolrサーバで何れかのタイミングで登録データのフラッシュが行われた場合、検索結果に反映される動作となります。

3.1.6.2 エラー発生後の次回起動時の動作仕様

異常終了後、次に IM-Workflow クローラを起動した時は、最後に正常終了した日時より後に完了となった IM-Workflow 案件情報が登録対象となります。

前回のクローリングで登録済みの案件情報があった場合、再び同一案件情報を登録することになりますが、Solr サーバでは登録時に同一IDの文書が既に存在した場合、文書を上書きする動作となっていますので再度登録処理を行っても問題はありません。
また最終クローラ起動日を編集することで、任意の日時以降に完了した案件情報を登録対象にすることも可能です。

3.1.7 その他の注意事項

案件完了後に IM-Workflow クローラを実行すると、この案件が全文検索の検索結果に表示され、案件名リンクを押下すると処理詳細画面が開きます。その後にアーカイブバッチを実行すると、この案件が完了案件から過去案件になります。この状態で再度、全文検索を行い、検索結果に表示される案件名リンクを押下すると画面に以下のメッセージが表示されます。

対象の案件が既に処理されたか、削除された可能性があります。

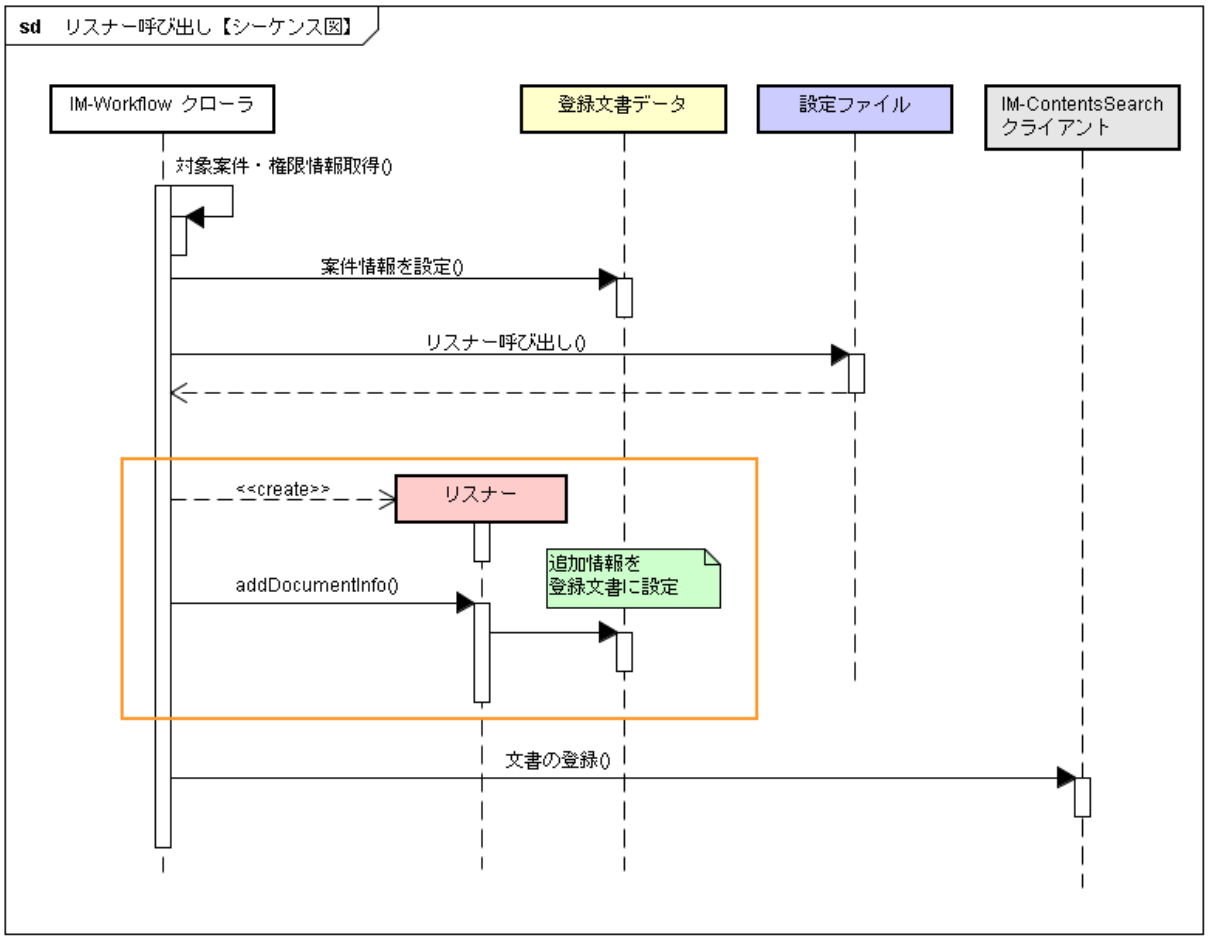
この場合は、IM-Workflow クローラを再度実行してください。実行後はメッセージが表示されず、過去案件詳細画面が開きます。

4 IM-Workflow クローラの拡張

IM-Workflow クローラでは、IM-Workflow の案件情報を「3.1.1.2 登録情報」の内容で Solr サーバに登録しますが、IM-Workflow クローラ用のリスナーを作成して追加することで、登録情報として独自の項目を追加することが可能です。
この章では、独自の項目を登録情報に追加するためのリスナーの作成方法について説明します。

4.1 リスナーの呼び出し

IM-Workflow クローラは案件単位で案件情報を文書として登録していきますが、登録処理の直前に設定ファイルに定義されたリスナーを呼び出し、リスナーが存在した場合はリスナーの項目追加メソッドを実行します。
リスナーが IM-Workflow クローラに呼び出されるシーケンスを以下に示します。



リスナー呼び出しシーケンス

4.2 リスナーの設定

作成したリスナークラスは、以下のファイルの<**listener**>タグ内に<**imwcrawler-add-listener**>タグを追加し listener-class に指定します。

Server Manager/conf/ system-install.xml

（設定例）

```
<listener>
  ・・・・・・・・・・
  <imwcrawler-add-listener>
      <listener-class>jp.co.intra_mart.sample.workflow.purchase.listener.WorkflowCrawlingAddListener</listener-class>
  </imwcrawler-add-listener>
</listener>
```

4.3 リスナーの作成

作成するリスナーは以下のインタフェースを実装している必要があります。

- **jp.co.intra_mart.foundation.workflow.listener.IWorkflowMatterCrawlingAddListener**

4.3.1 実処理の記述

登録情報に独自の項目を追加するための実処理は **addDocumentInfo** メソッド内に記述します。
addDocumentInfo メソッドの引数には、文書登録用オブジェクト、ログイングループ ID、ロケール ID、システム案件 ID、ユーザデータ ID が渡されます。

リスナーでは以下の処理を行います。

① 引数のパラメータ情報を元に、アプリケーションデータを取得する。
② この案件情報が追加対象となるかの判定を行う。
③ 引数の文書登録用オブジェクトにアプリケーションデータを追加する。
- 検索対象となるテキストフィールドにデータを登録したい場合
  IntramartSolrInputDocument #addText メソッドを使用
- 表示目的等の格納用データを登録したい場合
  IntramartSolrInputDocument # addField メソッドを使用してダイナミックフィールドにデータを追加します。

※いずれのフィールド値にも null 値を設定することはできませんので、ご注意ください。

表示用に登録したデータを専用のレイアウトで検索結果画面に表示したい場合には、上記以外に以下の処理を行う必要があります。

- 文書種別の追加
  IntramartSolrInputDocument # addType メソッドを使用
- 業務テンプレートの作成
- 追加した文書種別情報を設定ファイルに設定

- 業務テンプレートの作成・設定ファイルへの設定方法についての詳細は、「IM-ContentsSearch プログラミングガイド」 3.2.2 業務テンプレートの作成， 3.2.3 文書種別情報の定義 をご参照ください。
- 業務テンプレートの作成・設定ファイルへの設定方法のサンプルは、「IM-Workflow プログラミングガイド」 7.3 IM-Workflow クローラリスナーの作成 をご参照ください。

IM-Workflow Ver.7.2
クローラ 仕様書

2010/07/30 初版

TEL: 03-5549-2821
FAX: 03-5549-2816
E-MAIL: info@intra-mart.jp
URL: http://www.intra-mart.jp/