|  |
| --- |
| 統計等情報流通連携基盤システム  実装詳細仕様書  株式会社横須賀テレコムリサーチパーク |

目次

[1 はじめに 3](#_Toc387405388)

[1.1 統計等情報流通連携基盤システムの概要 3](#_Toc387405389)

[1.2 本書の構成 4](#_Toc387405390)

[2 統計等情報流通連携基盤システムの構成 5](#_Toc387405391)

[2.1 Frontend Server 6](#_Toc387405392)

[2.2 Developer Site Server 6](#_Toc387405393)

[2.3 User Information Server 7](#_Toc387405394)

[2.4 Graph Database Server 7](#_Toc387405395)

[3 データ規格 8](#_Toc387405396)

[3.1 統計情報を記述するボキャブラリ 8](#_Toc387405397)

[3.1.1 データセットとボキャブラリの規定方針 8](#_Toc387405398)

[3.1.2 ボキャブラリ定義 9](#_Toc387405399)

[3.1.3 統計情報のRDF化と登録方法 24](#_Toc387405400)

[3.2 「政府データカタログサイト」メタデータ記述ボキャブラリ 26](#_Toc387405401)

[3.2.1 ボキャブラリの設計方針 26](#_Toc387405402)

[3.2.2 ボキャブラリ定義 26](#_Toc387405403)

[3.2.3 「政府データカタログサイト」メタデータの登録方法 28](#_Toc387405404)

[4 実装の詳細 29](#_Toc387405405)

[4.1 API機能 29](#_Toc387405406)

[4.1.1 SPARQL-Based Command 29](#_Toc387405407)

[4.1.2 Security Management Comamnd 30](#_Toc387405408)

[4.2 開発者サイト機能 31](#_Toc387405409)

[4.2.1 スタティックページ表示部 32](#_Toc387405410)

[4.2.2 ニュースページ 35](#_Toc387405411)

[4.2.3 フォーラムページ 36](#_Toc387405412)

# はじめに

本書は、「情報流通連携基盤の統計情報・データカタログへの適用実証」（以下「本実証と呼ぶ」）において構築した統計等情報流通基盤システム（以下「本システム」と呼ぶ）の実装仕様を示す。

## 統計等情報流通連携基盤システムの概要

「情報流通連携基盤の統計情報・データカタログへの適用実証」（以下「本実証と呼ぶ」）において構築した統計等情報流通基盤システム（以下「本システム」と呼ぶ）は、情報流通連携基盤システム外部仕様書[[1]](#footnote-1)（以下「外部仕様書」と呼ぶ）が規定するSPARQL-Based Commandにより、アプリケーションからのRDFデータに対する検索・取得要求を受け付ける「API機能」と、開発者を支援するためのAPIドキュメント、サンプルプログラム、ライブラリ等を提供する「開発者サイト機能」からなるシステムである（図 1）。なおAPI機能には、次世代統計システム利用API[[2]](#footnote-2)やデータカタログAPI[[3]](#footnote-3)から取得した情報をRDF化した結果を保持する機能を含む。

本システムが提供するデータは以下の通りである。

* 統計センターの「次世代統計利用システム[[4]](#footnote-4)」が提供する統計情報のうち以下のもの
  + 次世代統計利用システムAPIから取得する統計表・関連情報
    - メッシュ統計、国勢調査小地域集計
    - 平成12年から現在までの以下のデータを収録
      * 国勢調査（500mメッシュ、1kmメッシュ、小地域）
    - 統計表は都道府県毎に分割されているため、およそ2,300件存在し、含まれるデータ数はおよそ1億1千万件である
  + HTMLやExcel等のデータから抽出する統計表・関連情報
    - 日本標準産業分類の一覧
      * 平成5年に改正されてから現在までの分類（計3回）
        + HTMLで公開されている範囲を対象としている
    - 日本標準職業分類の一覧
      * 平成9年に改正されてから現在までの分類（計2回）
        + HTMLで公開されている範囲を対象としている
* 内閣官房が提供する「政府オープンデータカタログサイト試行版[[5]](#footnote-5)」が提供するデータカタログのメタデータ

これらのデータの総サイズは46GB、 RDFトリプル総数は7.7億件であった。



図 1 統計等情報流通連携基盤システムの位置づけ

なお、「API機能」は、アプリケーションに対してデータの検索・取得機能を提供する機能であり、ユーザインタフェースを持たない。一方「開発者サイト機能」は、開発者に対してAPIドキュメント、サンプルプログラム、ライブラリ等を提供する機能であり、Webページとして構築した。

## 本書の構成

本書の構成は、以下の通りである。

* 第2章では、本システムの機能構成と各機能を提供するサーバのスペックを示す。
* 第3章では、本システムが利用するデータ規格と、それに基づくRDFデータの生成方法を示す。
* 第4章では、本システムの実装の詳細を示す。

# 統計等情報流通連携基盤システムの構成

本章では、統計等情報流通連携基盤システムの構成と、構成要素である各サーバのソフトウェア構成を示す。

本システムは、以下の4つのサーバから構成される。システムの構成図を図 2に示す。また、1分間に100件程度のリクエストを受け付け、それぞれのリクエストへの応答時間を1秒程度とするためには、それぞれのサーバのスペックを表 1のように設定することが望ましい。ただし、この値は本実証で扱ったデータと実証中に受け付けたリクエストから算出したものであり、応答時間は受付クエリの複雑度に依存することに注意が必要である。

* Frontend Server（api.odstat.jp）
  + 外部のアプリケーションに対して、情報流通連携基盤システム外部仕様書が規定するAPI（Application Programming Interface）を提供する。
* Developer Site Server（developer.odstat.jp）
  + 開発者サイトをホスティングする。
* User Information Server
  + 開発者のユーザ情報を保持する。
* Graph Database Server
  + RDFデータを格納し、SPARQLで検索する機能を提供する。



図 2 システム構成図

表 1 サーバスペック一覧

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Frontend Server | Developer Site Server | User Information Server | Graph Database Server |
| CPU | Intel Xeon 1.7GHz相当 1コア | Intel Xeon 1.7GHz相当 1コア | Intel Xeon 1.7GHz相当 2コア | Intel Xeon 1.7GHz相当 8コア |
| Memory | 1.7GiB | 1.7GiB | 6.5GiB | 68GiB |
| Storage | 8GiB | 8GiB | 60GiB | 120GiB |

以下、それぞれのサーバのソフトウェア環境を示す。

## Frontend Server

Frontend Serverは、情報流通連携基盤システム外部仕様書が規定する、SPARQL-Based CommandのAPIを提供する。本実証では、表 1の環境でFrontend serverを構築した。

表 2 Frontend Server ソフトウェア環境

|  |  |
| --- | --- |
|  | ソフトウェア名・バージョン |
| Operating System | Amazon Linux AMI release 2013.09 |
| リバースプロキシ | nginx 1.4.3 |
| Ruby実行環境 | Ruby 2.1.0p0 |

## Developer Site Server

Developer Site Serverは、情報流通連携基盤の開発者サイトをホスティングする。本実証では、表 3の環境でDeveloper Site Serverを構築した。

表 3 Developer Site Server ソフトウェア環境

|  |  |
| --- | --- |
|  | ソフトウェア名・バージョン |
| Operating System | Amazon Linux AMI release 2013.09 |
| リバースプロキシ | nginx 1.4.3 |
| Database | PostgreSQL 9.3.1 |
| Ruby実行環境 | Ruby 2.0.0p247 |

## User Information Server

User Information Serverは、情報流通連携基盤の開発者サイトに登録されたユーザ情報を管理する。本実証では、表 4の環境でUser Information Serverを構築した。

表 4 User Information Server ソフトウェア環境

|  |  |
| --- | --- |
|  | ソフトウェア名・バージョン |
| Operating System | Amazon Linux AMI release 2013.09 |
| Database | MongoDB 2.4.9 |

## Graph Database Server

Graph Database Serverは、RDFで記述されたデータを格納し、情報流通連携基盤のSPARQL 1.1準拠APIで検索するための機能を提供する。本実証では、表 5の環境でGraph Database Serverを構築した。

なお、この環境でGraph Database Serverを構築する場合，Graph Database Systemのインデクス用に1トリプルあたり39バイト必要である。従って、ディスク容量は、格納するデータ本体に加えて、トリプル数×39バイト以上必要である。

表 5 Graph Database Server ソフトウェア環境

|  |  |
| --- | --- |
|  | ソフトウェア名・バージョン |
| Operating System | Amazon Linux AMI release 2013.09 |
| Graph Database System | Virtuoso Open-Source Edition 6.1.8 |

# データ規格

本章では、本システムが扱うデータの規格を示す。主に、本システムが扱うRDFデータが利用するボキャブラリとデータの準備手段を、統計データとデータカタログのそれぞれについて述べる。

## 統計情報を記述するボキャブラリ

### データセットとボキャブラリの規定方針

統計情報のRDF化に際しては、世界的に広く利用されつつあるRDF Data Cube Vocabularyを採用した。RDF Data Cube Vocabularyでは、多次元の項目から特徴付けられるデータを、統計データを交換するために広く使われているSDMXのキューブモデルで表現することによって、データの比較を行いやすくすることを目指している。

RDF Data Cube Vocabularyでは、すべての統計データはデータセットに所属する。データセットの構成方法をRDF Data Cube Vocabularyは規定していないが、一般的には統計の種類毎に行う場合や、統計の種類と統計実施年の組ごとにデータセットを作成して利用されている。本システムでは、公開されている統計情報のデータの性質から、国勢調査の実施年ごとにデータセットを作成した。

データセットを統計の種類毎に構成するためには、データセットの構造定義で利用する軸を指定する必要があるため、実施年ごとに軸が共通である必要がある。しかし、現実の統計データでは、実施年によって軸に含まれる項目（codelist）が違うことがある。軸に含まれる項目が変わる場合には、同じ統計表でも実施年をまたいで軸を共通化できないため、データセット構造定義を共通化できない。このため、データセット構造定義を実施年ごとに作成する必要が生じ、結果的にデータセットも実施年ごとに作成した。

　統計データの統計軸の情報や調査実施年、エリアなどの情報は、統計データに付与する方法やデータセットに付与する方法などが存在する。本システムでは、公開されている統計表になるべく忠実にRDF化作業を行うことを基本方針とた。1つのデータセットに複数の異なる観測データが存在する場合は、データセットに統計軸の定義やエリアなどの情報を紐付けた。

統計軸の名前（例えば性別）や統計軸に含まれる項目の名前（例えば、男性、女性など）には、同じ概念を様々な別の言葉で表現する可能性があるため（例えば、男、女など）、これに対応できるようdc:titleやrdf:labelを採用せず、skos:prefLabelを採用した。このため、一つの概念に複数の名前の表記がある場合にはskos:altLabelを複数関連づけることによって、別名の記述を行うことが可能である。

### ボキャブラリ定義

統計データを表現するために必要な構造は、以下の10種類である。

1. データセット
2. データセット構造定義
3. 統計軸（次元）
4. 統計軸項目
5. コードリスト
6. 統計データ
7. 単位
8. 基本単位次元
9. 基準項目
10. 基準項目コードリスト

以下、これらを記述するためのボキャブラリを示す。なお、これらの記述に共通して利用する名前空間の接頭辞を図 3に示す。

図 3 統計情報の記述に利用するボキャブラリの接頭辞

PREFIX dct: <http://purl.org/dc/terms/>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

PREFIX owl: <http://www.w3.org/2002/07/owl#>

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

PREFIX sdmx: <http://purl.org/linked-data/sdmx#>

PREFIX sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#>

PREFIX sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#>

PREFIX sdmx-attribute: <http://purl.org/linked-data/sdmx/2009/attribute#>

PREFIX qb: <http://purl.org/linked-data/cube#>

PREFIX odstat: <http://odstat.jp/>

PREFIX odstat-dim: <http://odstat.jp/dim#>

PREFIX odstat-data: <http://odstat.jp/data#>

PREFIX odstat-code: <http://odstat.jp/code#>

PREFIX odstat-ref: <http://odstat.jp/ref#>

PREFIX odstat-vocab: <http://odstat.jp/vocab#>

PREFIX odstat-area: <http://odstat.jp/area#>

PREFIX odstat-unit: <http://odstat.jp/unit#>

#### データセット

##### 概要

2000年、2005年、2010年の以下の統計情報を、それぞれ実施回ごとにデータセットを作成している。

データセットを表す属性名は「dataset」から始まり、統計データ名の英語表記と、調査実施年を表す4桁の数字を付与する命名規則としている。

データセットの一覧は以下の通りである。

* 65歳以上親族のみの一般世帯数（2000年のみ存在）
  + 2000年  
    odstat-data:datasetNumberOfPrivateHouseholdsWithRelatedMembers65YearsOfAgeAndOver2000
* 世帯の家族類型別一般世帯数
  + 2000年  
    odstat-data:datasetNumberOfPrivateHouseholdsByFamilyTypeOfHousehold2000
  + 2005年  
    odstat-data:datasetNumberOfPrivateHouseholdsByFamilyTypeOfHousehold2005
  + 2010年  
    odstat-data:datasetNumberOfPrivateHouseholdsByFamilyTypeOfHousehold2010
* 世帯の経済構成別一般世帯数
  + 2000年  
    odstat-data:datasetNumberOfPrivateHouseholdsByEconomicTypeOfHousehold2000
  + 2005年  
    odstat-data:datasetNumberOfPrivateHouseholdsByEconomicTypeOfHousehold2005
  + 2010年  
    odstat-data:datasetNumberOfPrivateHouseholdsByEconomicTypeOfHousehold2010
* 世帯人員別一般世帯数
  + 2000年  
    odstat-data:datasetNumberOfPrivateHouseholdsByHouseholdMembers2000
  + 2005年  
    odstat-data:datasetNumberOfPrivateHouseholdsByHouseholdMembers2005
  + 2010年  
    odstat-data:datasetNumberOfPrivateHouseholdsByHouseholdMembers2010
* 住宅の建て方別世帯数
  + 2000年  
    odstat-data:datasetNumberOfHouseholdsByTypeOfBuilding2000
  + 2005年  
    odstat-data:datasetNumberOfHouseholdsByTypeOfBuilding2005
  + 2010年  
    odstat-data:datasetNumberOfHouseholdsByTypeOfBuilding2010
* 住宅の種類・所有の関係別一般世帯数、延べ面積（2010年には「延べ面積」なし）
  + 2000年  
    odstat-data:datasetNumberOfHouseholdsAndAreaOfFloorSpaceByKindOfResidenceAndTenureOfDwelling2000
  + 2005年  
    odstat-data:datasetNumberOfHouseholdsAndAreaOfFloorSpaceByKindOfResidenceAndTenureOfDwelling2005
  + 2010年  
    odstat-data:datasetNumberOfHouseholdsByKindOfResidenceAndTenureOfDwelling2010
* 在学学校・未就学の種類別在学者数・未就学者数（2000年のみ存在）
  + 2000年  
    odstat-data:datasetNumberOfPersonsAttendingSchoolAndPersonsNeverAttendedSchoolBySchoolLevelOfPersonsAttendingSchoolAndTypeOfPersonsNeverAttended2000
* 年齢別（５歳階級、４区分）、男女別人口
  + 2000年  
    odstat-data:datasetPopulationByAgeAndSex2000
  + 2005年  
    odstat-data:datasetPopulationByAgeAndSex2005
  + 2010年  
    odstat-data:datasetPopulationByAgeAndSex2010
* 産業別（大分類）・従業上の地位別就業者数
  + 2000年  
    odstat-data:datasetNumberOfEmployedPersonsByIndustryAndEmploymentStatus2000
  + 2005年  
    odstat-data:datasetNumberOfEmployedPersonsByIndustryAndEmploymentStatus2005
  + 2010年  
    odstat-data:datasetNumberOfEmployedPersonsByIndustryAndEmploymentStatus2010
* 男女別人口総数及び世帯総数
  + 2000年  
    odstat-data:datasetNumberOfPopulationBySexAndNumberOfHouseholdsInMunicipality2000
  + 2005年  
    odstat-data:datasetNumberOfPopulationBySexAndNumberOfHouseholdsInMunicipality2005
  + 2010年  
    odstat-data:datasetNumberOfPopulationBySexAndNumberOfHouseholdsInMunicipality2010
* 職業別（大分類）就業者数
  + 2000年  
    odstat-data:datasetNumberOfEmployedPerson2000
  + 2005年  
    odstat-data:datasetNumberOfEmployedPerson2005
  + 2010年  
    odstat-data:datasetNumberOfEmployedPerson2010

##### 名前空間の接頭辞

名前空間の接頭辞は以下の通りである。

* odstat-data: http://odstat.jp/data#

##### 属性

データセットの属性は表 6の通りである。

表 6 データセットの属性を記述するボキャブラリ

|  |  |  |
| --- | --- | --- |
| **属性** | **値域** | **説明** |
| rdf:type | rdfs:Class | qb:DataSetクラスを指定する |
| qb:structure | qb:DatasetStructureDefinition | データセットの構造定義のURLを示す |

#### データセット構造定義

##### 概要

データセットの構造を定義する。1つのデータセットに対して1つのデータセット構造定義が存在する。

データセット構造定義を表す属性名は「dsd」から始まり、統計データ名の英語表記と、調査実施年を表す4桁の数字を付与する命名規則としている。

##### 名前空間の接頭辞

名前空間の接頭辞は以下の通りである。

* odstat-data: [http://odstat.jp/data#](http://odstat.jp/data)

##### 属性

データセット構造定義を記述する属性は表 7の通りである。

表 7 データセット構造定義を記述するボキャブラリ

| **属性** | **値域** | **説明** |
| --- | --- | --- |
| rdf:type | rdfs:Class | qb:DatasetStructureDefinitionクラスを指定する |
| qb:component | 空白ノード | 子ノードにqb:attributeとqb:componentAttachmentの属性で結ばれた空白ノードか、qb:dimensionの属性で紐付けられた軸の項目を持つ空白ノードを複数持つ |
| (qb:componentに接続する空白ノードの属性) qb:dimension | qb:DimensionProperty | データセット内で利用する次元を指定する |
| (qb:componentに接続する空白ノードの属性) qb:attribute | qb:AttributeProperty | データセット内で利用する属性を指定する。 本実証では、qb:componentAttachmentと組み合わせてデータの単位の記述位置を表すために利用している |
| (qb:componentに接続する空白ノードの属性) qb:componentAttachment | qb:ComponentSpecification | データセット内で利用するコンポーネントがどのレベルで定義されているかを表す。本実証では、qb:attributeと組み合わせてデータの単位がqb:Observationで定義されていることを表すために利用している |

#### 統計軸（次元）

##### 概要

データの統計軸（次元）を表し、本実証では統計データの集計内容を表す。本実証のデータでは、「男女別人口総数及び世帯総数」、「世帯の家族類型別一般世帯数」などが存在する。

##### 名前空間の接頭辞

名前空間の接頭辞は以下の通りである。

* odstat-dim: [http://odstat.jp/dim#](http://odstat.jp/data)

##### 属性

データの属性は表 8の通りである。

表 8 統計軸（次元）を記述するボキャブラリ

|  |  |  |
| --- | --- | --- |
| **属性** | **値域** | **説明** |
| rdf:type | rdfs:Class | qb:DimensionPropertyクラス、qb:CodedPropertyクラスを指定する |
| skos:prefLabel | rdfs:Literal | 次元の名称を指定する |
| rdfs:comment | rdfs:Literal | データの統計軸（次元）に関するコメントを指定する |
| qb:codeList | qb:CodedProperty | データの統計軸（次元）に含まれる統計軸項目のリスト（コードリスト）を指定する |
| rdfs:range | rdfs:Class | 統計軸のクラスを指定する |

#### 統計軸項目

##### 概要

統計軸に含まれる項目を表す。本実証のデータでは、「人口総数」、「世帯総数」、「男」、「女」などが存在する。

##### 名前空間の接頭辞

名前空間の接頭辞は以下の通りである。

* odstat-vocab: http://odstat.jp/vocab#

##### 属性

データの属性は表 9の通りである。

表 9 統計軸項目の属性

|  |  |  |
| --- | --- | --- |
| **属性** | **値域** | **説明** |
| rdf:type | rdfs:Class | skos:Conceptクラスと、統計表毎に 定義された統計軸のクラスを指定する |
| skos:prefLabel | rdfs:Literal | 項目のタイトルを表す |
| odstat:consistOf | 空白ノード | 子ノードにはodstat:derivedVocab で 基準項目を指定する |
| skos:exactMatch | skos:Concept | 統計軸項目と完全に一致する基準項目が存在する場合に指定する |

#### コードリスト

##### 概要

データの次元に含まれている統計軸の項目の一覧を表す。

##### 名前空間の接頭辞

名前空間の接頭辞は以下の通りである。

* odstat-codes: [http://odstat.jp/codes#](http://odstat.jp/codes)

##### 属性

データの属性は表 10の通りである。

表 10 コードリストを記述するボキャブラリ

|  |  |  |
| --- | --- | --- |
| **属性** | **値域** | **説明** |
| rdf:type | rdfs:Class | skos:ConceptSchemeクラスを指定する |
| skos:prefLabel | rdfs:Literal | コードリストの名称を指定する |
| skos:hasTopConcept | skos:Concept | コードリストに含まれる統計軸項目を 列挙する |

#### 統計データ

##### 概要

数値データ、軸、単位を含む統計データを表す。

##### 名前空間の接頭辞

名前空間の接頭辞は以下の通りである。

* odstat-data: [http://odstat.jp/data#](http://odstat.jp/data)

##### 属性

データの属性は表 11の通りである。

表 11 統計データを記述するボキャブラリ

|  |  |  |
| --- | --- | --- |
| **属性** | **値域** | **説明** |
| rdf:type | rdfs:Class | qb:Observationクラスを指定する |
| qb:dataSet | qb:DataSet | 統計データの所属する データセットを指定する |
| odstat-dim:\* | skos:Concept | qb:DataSetDefinitionで指定した次元に対応する統計軸項目を格納する。属性名は、データセット構造定義で指定される |
| sdmx-dimension:refArea | skos:Concept | 統計データを集計した地域を示す |
| sdmx-attribute:unitMeasure | odstat:Unit | 測度の単位を表す （人、世帯など） |
| sdmx-measure:obsValue | xsd:decimal (値が非公開の場合は"-"などが格納される) | 数値データを格納する |

#### 単位

##### 概要

統計データの単位を表す。

##### 名前空間の接頭辞

名前空間の接頭辞は以下の通りである。

* odstat-unit: [http://odstat.jp/unit#](http://odstat.jp/vocab)

##### 属性

データの属性は表 12の通りである。

表 12 単位を記述するボキャブラリ

|  |  |  |
| --- | --- | --- |
| **属性** | **値域** | **説明** |
| rdf:type | rdfs:Class | odstat:Unitクラスを指定する |
| skos:prefLabel | rdfs:Literal | 単位の表記名を表す |
| rdfs:range | rdfs:Class | メジャーの値域を表す |

#### 基準項目次元

##### 概要

調査票などから抽出した、基準となる次元を表す。本実証のデータでは、「性別」、「職業分類」などが存在する

統計実施年に依らず、一意である。

##### 名前空間の接頭辞

名前空間の接頭辞は以下の通りである。

* odstat-ref: [http://odstat.jp/ref#](http://odstat.jp/data)

##### 属性

データの属性は表 13の通りである。

表 13 基準項目次元を記述するボキャブラリ

|  |  |  |
| --- | --- | --- |
| **属性** | **値域** | **説明** |
| rdf:type | rdfs:Class | qb:CodedPropertyクラスを指定する |
| skos:prefLabel | rdfs:Literal | 次元の名称を指定する （存在する場合としない場合がある） |
| rdfs:comment | rdfs:Literal | データの基準項目次元に関するコメントを指定する |
| qb:codeList | qb:CodedProperty | データの基準項目次元に含まれる コードリストを指定する |
| rdfs:range | rdfs:Class | 基準項目軸を表すクラスを指定する |
| qb:concept | skos:Concept | skos:prefLabelを格納したConcept記述が存在する場合に指定する |

#### 基準項目

##### 概要

統計軸に含まれる項目の基準となる項目を表す。本実証のデータでは、「男」、「女」、「不明」などが存在する。

統計実施年に依らず一意である。

##### 名前空間の接頭辞

名前空間の接頭辞は以下の通りである。

* odstat-ref: [http://odstat.jp/ref#](http://odstat.jp/vocab)

##### 属性

データの属性は表 14の通りである。

表 14 基準項目を記述するボキャブラリ

|  |  |  |
| --- | --- | --- |
| **属性** | **値域** | **説明** |
| rdf:type | rdfs:Class | 基準項目軸skos:Conceptクラスを指定する |
| skos:prefLabel | rdfs:Literal | 項目のタイトルを表す |
| skos:topConceptOf | skos:Concept | コードリストを指定する |
| skos:inScheme | skos:Concept | 項目を構成する基準項目次元の基準項目コードリストを指定する |
| odstat:min | xsd:decimal | 主語が数値の範囲を表現する場合、数値の範囲の下限を表す |
| odstat:max | xsd:decimal | 主語が数値の範囲を表現する場合、数値の範囲の上限を表す |

#### 基準項目コードリスト

##### 概要

調査票などから抽出した、基準項目のリストを記述する。

統計実施年に依らず一意である。

##### 名前空間の接頭辞

名前空間の接頭辞は以下の通りである。

* odstat-ref: [http://odstat.jp/ref#](http://odstat.jp/codes)

##### 属性

データの属性は表 15の通りである。

表 15 基準項目コードリストを記述するボキャブラリ

|  |  |  |
| --- | --- | --- |
| **属性** | **値域** | **説明** |
| rdf:type | rdfs:Class | skos:ConceptSchemeクラスを 指定する |
| skos:prefLabel | rdfs:Literal | コードリストの名称を指定する |
| skos:hasTopConcept | skos:Concept | コードリストに含まれる軸の項目を 指定する |

### 統計情報のRDF化と登録方法

#### RDF化の基本方針

統計情報は、データセット、データセット構造定義、統計軸、統計データ、基準項目から構成されている。統計軸は、統計実施年毎に定義されており、統計データの集計項目を表現する。さらに、統計軸はさらに、統計データの集計単位を表す統計軸項目、および統計軸項目の集合を表す軸の項目リストから構成される。

これらのことから、統計軸を図 4のように統計実施年ごとにクラス定義し、実施年に依存しない軸を表現するクラスのサブクラスとして、実施年ごとのクラスを定義する。



図 4 統計軸のクラス定義

また、統計データは、図 5のように、統計データの識別子を中心としてさまざまな情報を紐付ける形で表現する。

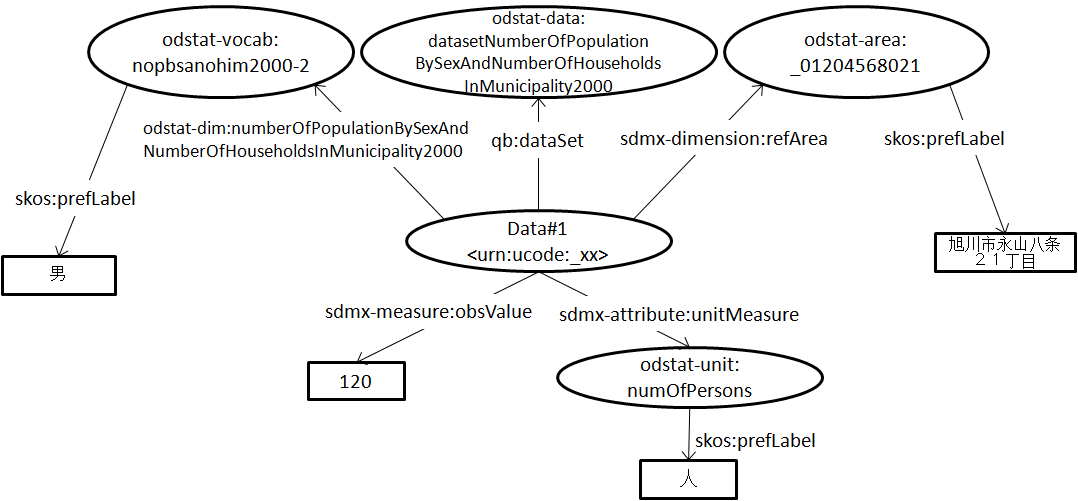


図 5 統計データのRDFグラフ

統計軸項目は、調査票や日本標準産業分類、日本標準職業分類等から作成した基準項目を組み合わせて記述する。

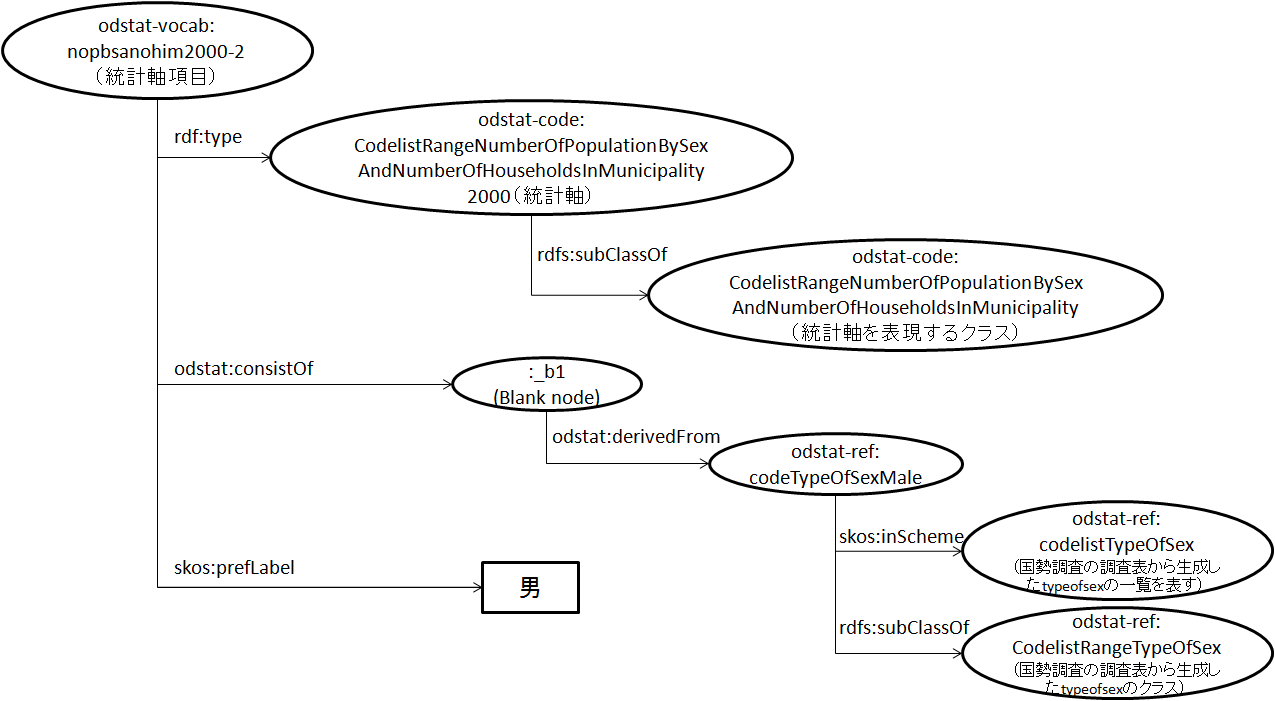


図 6 集計表から生成した基準項目と、統計軸との関係

#### 統計データのRDF化・登録手順の概要

以下の手順で統計データをRDF化し、本システムにキャッシュする。

1. 対象とする統計表を、次世代統計利用システムからダウンロードする。
2. 取得した統計表に含まれる統計データを、前節の方針に基づき、統計表を表すデータセットや軸に対応したボキャブラリを利用してRDF化し、Turtle形式で出力する。本実証においては、これを以下の手順で行った。
   1. 統計表ごとに、ボキャブラリの定義を行うTurtleファイルを作成した。これは、図 6に記した基準項目と統計軸との関係を記したRDFデータに、Turtleのコメントとして対象の統計表IDを付与したものである。
   2. 次世代統計システム利用APIを利用して入手したXML形式の統計データは、多次元の表を書き下している。これと、(A)のボキャブラリ定義ファイルを入力として、統計データをRDF形式に変換するスクリプトにかけることにより、入手した統計データをRDF化した。  
      スクリプトの挙動は以下の通りである。XML形式の統計データをパースすると、多次元表のセル集合が得られる。それぞれのセルにucodeを付与し、それを主語として、所属するデータセットや統計軸などの情報を、図 5に示すようなRDFデータとして生成する。この際に利用するタームは、(A)のボキャブラリ定義ファイルから統計表IDと統計軸を検索して抽出する。
3. 前工程において出力されたTurtle形式のファイルを、本システムのGraph Database Serverに登録する。

## 「政府データカタログサイト」メタデータ記述ボキャブラリ

### ボキャブラリの設計方針

本システムでは、W3C勧告として標準化されているData Catalog Vocabulary（DCAT）を中心として政府データカタログサイト試行版のデータセットに関するメタデータを記述している。

政府データカタログサイト試行版のシステムであるCKANは、諸外国など、他組織でも広く利用されている。このため、他組織のデータカタログサイトとの相互接続性を意識して、本実証独自の語彙は定義していない。また、DCATとその他の既存のボキャブラリを組み合わせ、CKAN推奨のボキャブラリセットを利用してデータセットを記述した。

### ボキャブラリ定義

政府データカタログサイトのメタデータを記述するためのボキャブラリを表 16に示す。

表 16 政府データカタログサイトメタデータのRDF化のためのボキャブラリ

| **メタデータ 付与対象** | **属性名** | **説明** |
| --- | --- | --- |
| データセット | rdf:type | データセットの型。dcat:Datasetを格納する。 |
| データセット | dct:creator | データセットの作成者を表すノードのURI |
| データセット | dct:description | データセットの説明（notes） |
| データセット | dct:identifier | データセットの識別用（一意であること） |
| データセット | dct:title | データセットのタイトル |
| データセット | rdfs:label | データセットの可読用名称（一意であること） |
| データセット | owl:sameAs | システム内部で利用しているUUID |
| データセット | dcat:distribution | リソース（参照先情報）を表すノードのURI |
| データセット | dcat:keyword | リソース（参照先情報）を表すノードのタグ |
| データセット | foaf:homepage | データセットへアクセスするためのCKANのURL |
| リソース（参照先情報） | rdf:type | リソース（参照先情報）の型。dcat:Distributionを格納する。 |
| リソース（参照先情報） | dct:title | リソースの名前 |
| リソース（参照先情報） | dcat:accessURL | リソース（ファイル）へアクセスするためのURL |
| リソース（参照先情報） | dct:format | リソースのフォーマットを表すノードのURI |
| ファイルフォーマット | rdf:type | ファイルフォーマットの型。dct:IMTを格納する。 |
| ファイルフォーマット | rdf:value, rdfs:label | フォーマット名 |
| データセットの作成者 | foaf:name | データセットの作者名 |
| data.go.jp拡張メタデータ | rdfs:label | release\_day, frequency\_of\_update, publisher などのメタデータ項目名 |
| data.go.jp拡張メタデータ | rdf:value | メタデータ項目名に対応する値 |

### 「政府データカタログサイト」メタデータの登録方法

政府データカタログサイト試行版のシステムであるCKANが、データカタログに格納されているデータのメタデータをRDF形式で提供する機能を提供している。また、CKANは、登録されているメタデータを一括出力するAPIを持っている。

これらを利用して、CKANに登録されているメタデータを以下の手順で本システムにキャッシュする。

1. 以下のURLにアクセスし、CKANに登録されているデータの識別子一覧を取得する。  
   http://www.data.go.jp/data/api/action/package\_list
2. 得られた識別子をキーとして、登録されているデータのメタデータをRDF形式で取得し、Turtle形式に変換する。
3. 得られたTurtleファイルを、本システムのGraph Database Serverに登録する。

# 実装の詳細

本章では、API機能と開発者サイト機能の詳細実装について示す。

## API機能

API機能に携わるサーバは、図 2のうち、Frontend Server、User Information Server、Graph Database Serverの3つである。

Frontend Serverでは、情報流通連携基盤外部仕様書に掲載されている標準APIのうち、SPARQL-Based Commandに準拠したAPIを提供した。一般ユーザ向けには、以下のAPIを提供した。

* 3.1.1 SPARQL1.1準拠のクエリ発行（GETメソッド）
* 3.1.2 SPARQL1.1準拠のクエリ発行（POSTメソッド）

また、開発者サイト向けにSecurity Management Commandのうち下記を提供した。

* 3.4.1 ロールの検索

### SPARQL-Based Command

SPARQL1.1が規定するクエリのうち、データの検索、取得のために一般的に利用されているSELECT、CONSTRUCT、ASKの3種類を提供した。

SPARQLのエンドポイントは、https://api.odstat.jp/api/v1/sparqlとした。

リクエストは全てHTTP GETまたはHTTP POSTリクエストとした。GET時、POST時のヘッダなどパラメータの詳細は、SPARQL1.1プロトコルに準拠している。

クライアントは、GETリクエストを行う際のクエリ文字列に対して、必ずURIエンコードを行わなければならない。また、リクエスト時には必ずパラメータ、またはリクエストヘッダのacl:consumerKeyにユーザを識別するConsumerKeyを付与してリクエストしなければならない。

Frontend Serverは、リクエストを受け取ると、ユーザがシステムに登録されているか、Security Management Commandの/api/v2/rolesを利用してUser Information Serverに問い合わせることにより、登録ユーザであるか否かを判断する。登録ユーザであれば、Graph Database Serverにリクエストを転送してGraph Database Serverの結果を返す。もし登録ユーザでは無い場合には、ステータスコード403（Forbidden）を返す。

Graph Database ServerのSPARQLエンドポイントは、RDFデータベースとして稼働させているVirtuosoのエンドポイントそのものである。Graph Database Serverは、Virtuosoが提供するSPARQL検索機能を利用してRDFデータベースの検索を行い、その結果をFrontend Serverに返す。

### Security Management Comamnd

User Information Serverは、MongoDBの中に、登録されているユーザのConsumerKeyに対して、統計データのすべてを示すodstat:aclAllを結びつけて管理している。すなわち、あるユーザXのロールの識別子がurn:ucode:\_00001C00000000000001000000012345、このユーザのConsumerKeyが”ABCDEFG”である場合，MongoDBには以下のようなデータが存在している。

|  |
| --- |
| {  @id: "urn:ucode:\_00001C00000000000001000000012345",  @type: "acl:RightsStatement",  acl:accessTarget: "odstat:aclAll",  acl:consumerKey: "ABCDEFG",  acl:isActive: true,  acl:hasReadPermission: true  } |

User Information Serverは、Frontend Serverから問い合わせを受けると、パラメータとして渡されたConsumerKeyをキーとしてデータベースを検索する。ConsumerKeyが見つかると、ステータスコード200（OK）とロールurn:ucode:\_00001C00000000000001000000012345の情報を返す。見つからない場合は、ステータスコード404（Not Found）を返す。

## 開発者サイト機能

開発者サイト機能に携わるサーバは、図 2のうち、Developer Site ServerとUser Information Serverの2つである。

開発者サイト機能は、API仕様、サンプルコード、利用規約、データ一覧のスタティックページ表示部と、ニュースを扱うCMS部と、フォーラムを提供するフォーラム部と、ユーザ管理機能を提供するユーザ管理部からなり、それぞれRuby on Rails 3.2.15を利用して実装した。

また、APIの利用傾向を把握するため、ユーザにはそれぞれ固有のConsumerKeyを発行し、APIへのリクエスト時にこのConsumerKeyを付与する仕様とした。このため、新規ユーザ登録時には、外部仕様書のSecurity Management Commandを利用してユーザが統計情報・データカタログ情報にアクセスできるよう設定を行った。

API仕様、サンプルコード、利用規約、データ一覧、ニュース、フォーラムの各機能への遷移は、各ページの一番上にあるメニューにより行う。開発者サイトは、すべてのページをメニューから遷移できるようにしている。

### スタティックページ表示部

スタティックページは、トップページ、API仕様、サンプルコード、利用規約、データ一覧を掲載する部分である。これらのページは、Ruby on Rails向けにCMS機能を提供するライブラリであるRefinery CMSを利用して作成した。

それぞれの画面イメージを以下に記す。

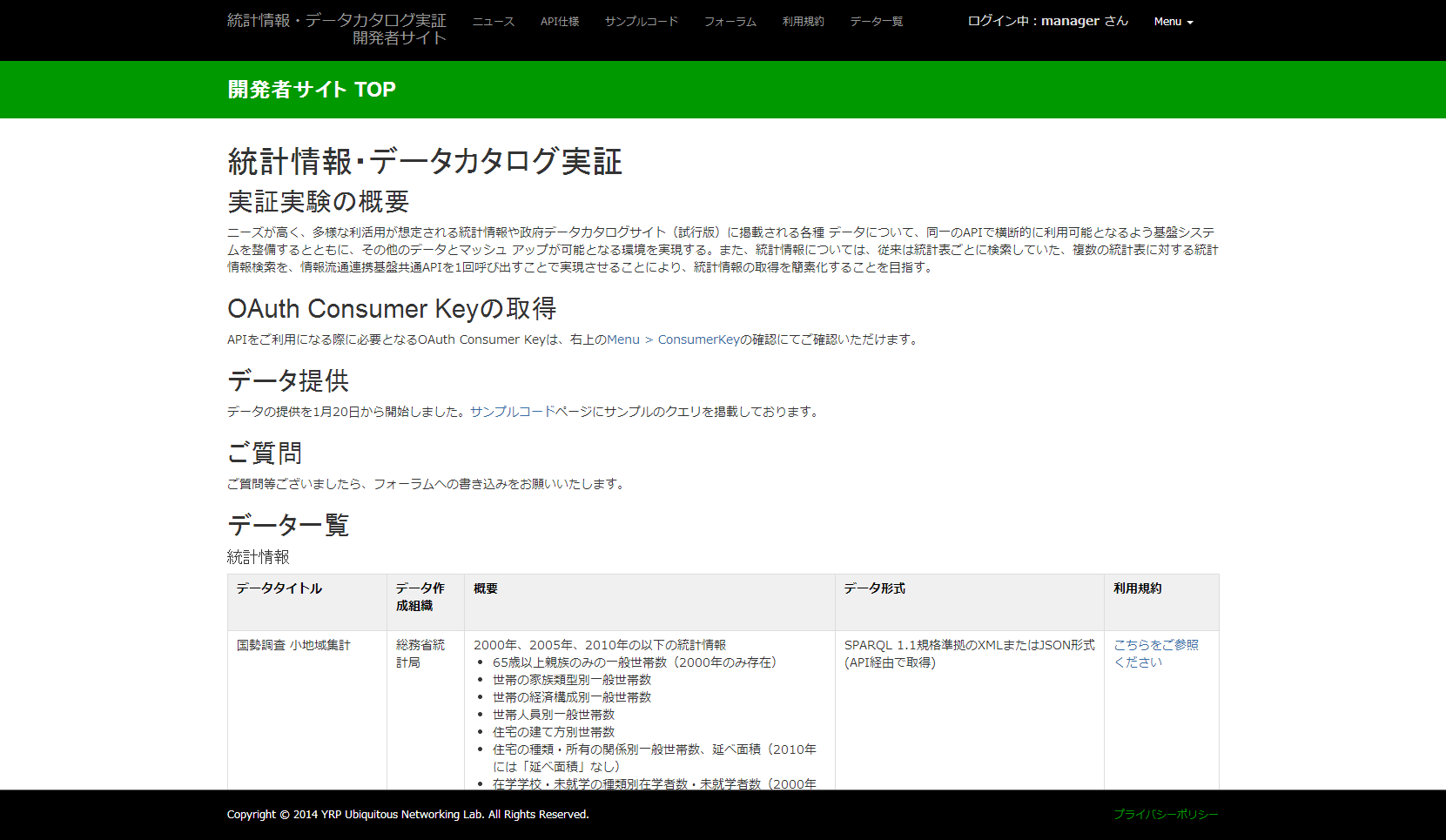


図 7 開発者サイトトップ画面

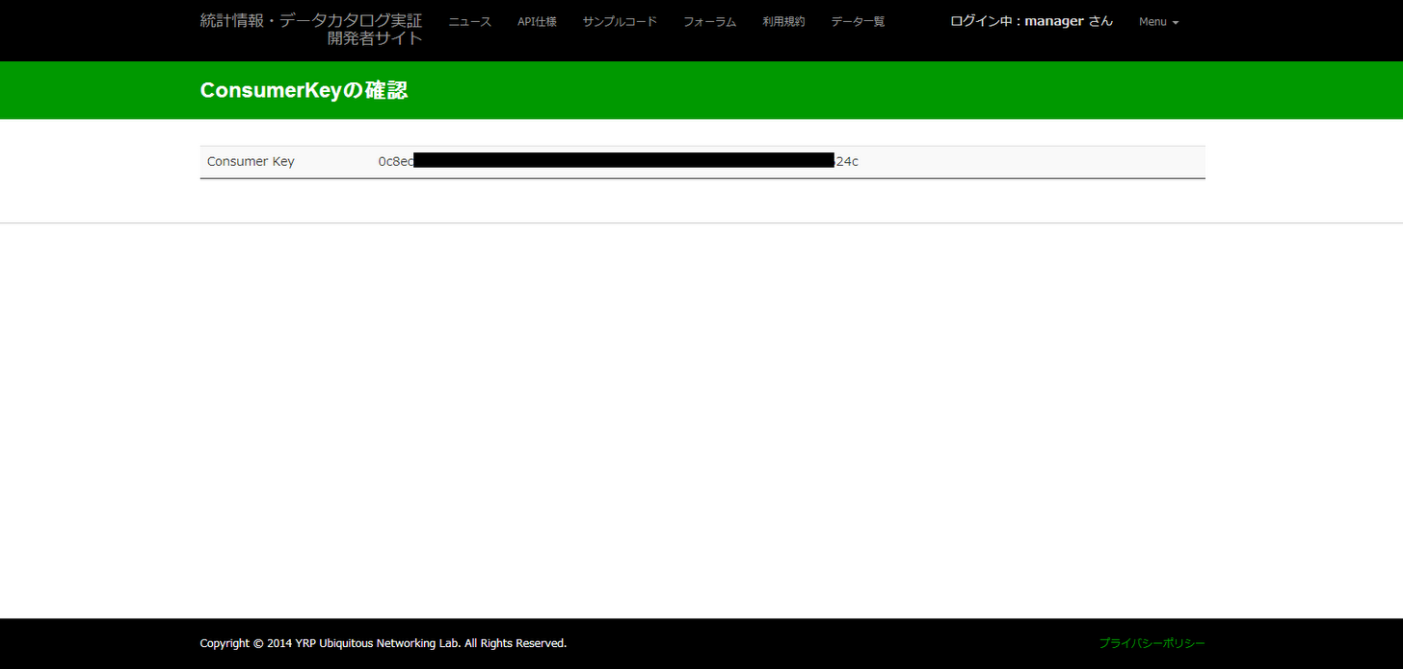


図 8 ConsumerKeyの確認画面



図 9 API仕様画面



図 10 サンプルコード画面



図 11 データ一覧画面

### ニュースページ

ニュースページは、開発者サイトの管理者がリリースノートなど更新情報や変更情報などを掲載するページである。

本機能は、Ruby on Rails向けにCMS機能を提供するライブラリであるRefinery CMSと、ニュースページを提供するためのRefinery CMSのプラグインであるNews Plugin[[6]](#footnote-6)を利用して実装した。



図 12 ニュース画面

### フォーラムページ

フォーラムページは、参加者同士が意見交換をできるような掲示板である。本機能は、Ruby on Railsにフォーラム機能を提供するライブラリであるForem[[7]](#footnote-7)を用いて実装した。

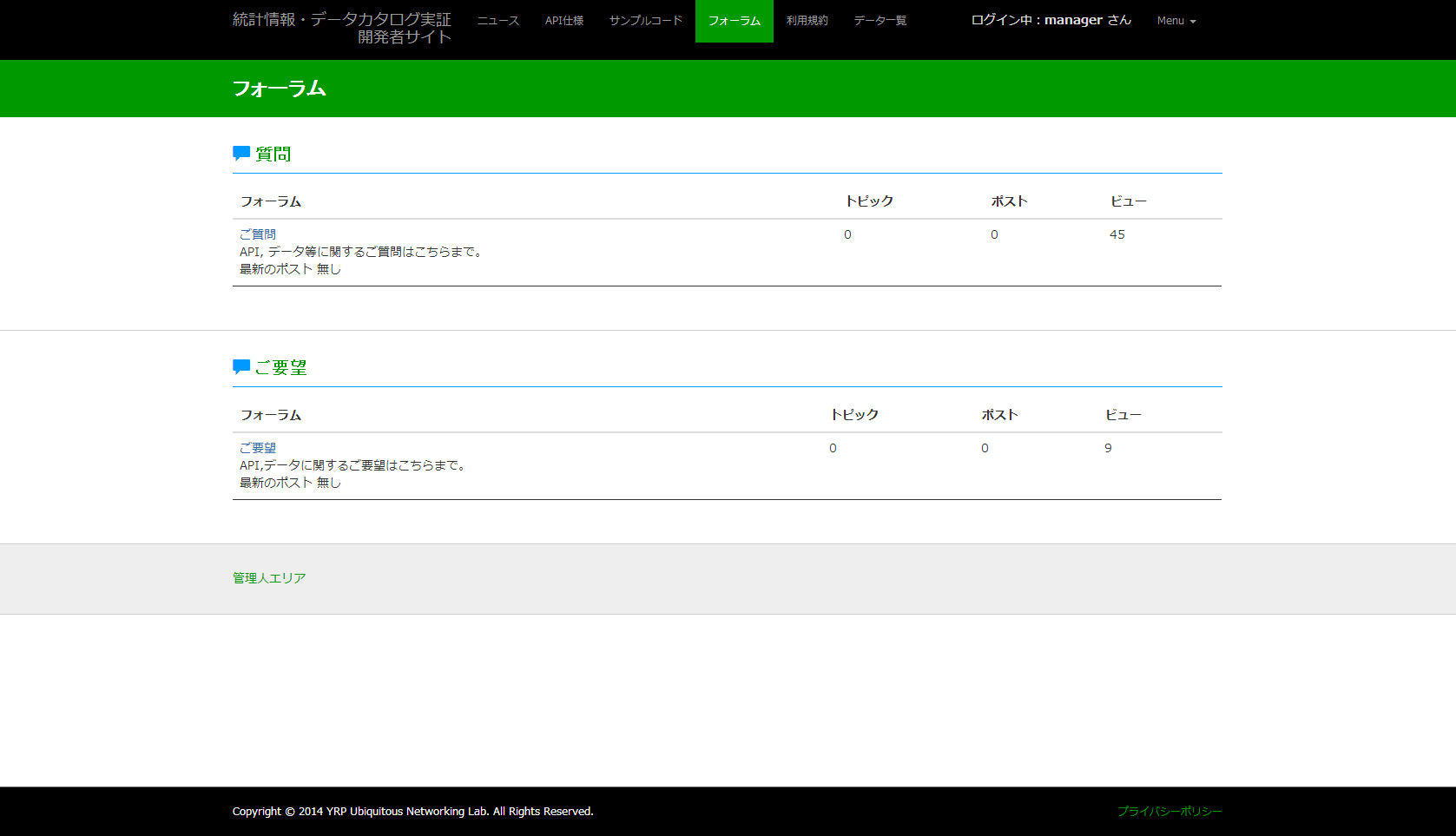


図 13 フォーラム画面

1. http://www.opendata.gr.jp/cfc/。ただし、SPARQL-Based Commandの内部で、平成25年度版仕様にあるSecurity Management Commandを利用している。 [↑](#footnote-ref-1)
2. http://statdb.nstac.go.jp/system-info/api/ [↑](#footnote-ref-2)
3. http://docs.ckan.org/en/latest/api/ [↑](#footnote-ref-3)
4. http://statdb.nstac.go.jp/ [↑](#footnote-ref-4)
5. http://data.go.jp/ [↑](#footnote-ref-5)
6. https://github.com/refinery/refinerycms-news [↑](#footnote-ref-6)
7. https://github.com/radar/forem [↑](#footnote-ref-7)