LOD challenge 2019

RDFのチェックツール「rdflint」とコミュニティによるオープンデータの作成

三上 威 - @takemikami
アーリース情報技術株式会社 代表取締役 社長

LODチャレンジ2019ミートアップ （キックオフイベント）
2019.7.5 @国立情報学研究所 12階 1208・1210会議室

自己紹介

発表者のプロフィールを紹介します

- 三上威 (@takemikami)
- データエンジニア・サイエンティスト
  - タスク: 分析・予測モデル開発・基盤構築 etc
  - 対象: マーケティングデータ etc
- 略歴
  - 甲南大学理学部応用数学科 卒
  - EC, CRM等のシステム構築 @ NEC系SIer
  - ECサイトのマーケティングデータ分析 @ DeNA
  - データ分析・予測モデル開発・基盤構築 @アーリース情報技術(株)

※フリーランスの法人成り

im@sparqlとそのデータ作成の運用

im@sparqlとそのデータ作成の運用について紹介をします

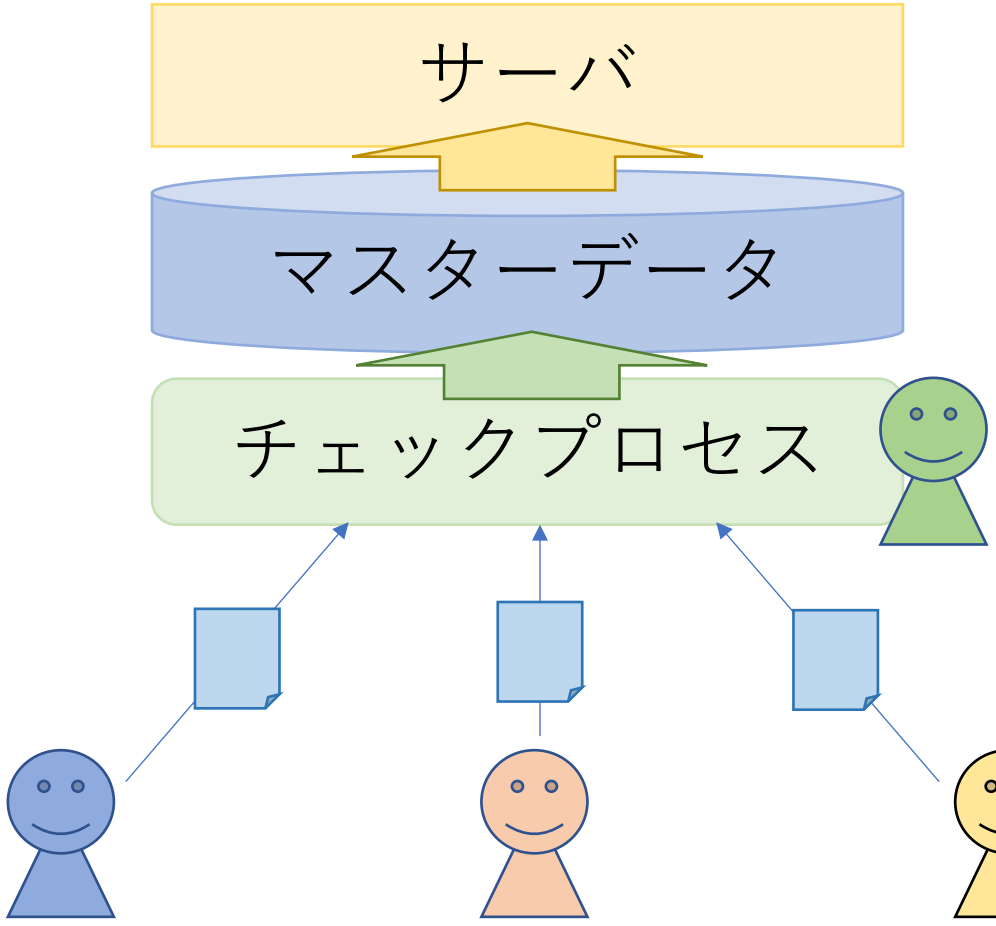
im@s + sparql = im@sparql

- 「アイドルマスター」作品世界のデータセットを持つ、SPARQLでアクセス可能なエンドポイント
- GitHubというソースコードの管理システムを使用、**誰でもデータの追加・修正のリクエストをすることが出来る**

※im@sparqlは、
LODチャレンジ2018 データセット部門 受賞作品です

コミュニティによるデータ作成の全体像

コミュニティによるデータ作成の全体像を示します



④サーバに反映

③マスターに反映

②依頼内容をチェック

チェックに手間のかかる
＆システムの知識も必要

①修正依頼を送る

コミュニティによるデータ作成の課題意識

コミュニティによるデータ作成の課題意識を説明します

- 目標：
  誰でもデータの追加・修正のリクエストをすることが出来る
- 障害：
  確認作業（サーバを立ち上げて、修正データをロード）
  に手間と知識が必要になる
  ＝「誰でも」リクエスト出来る状態ではない

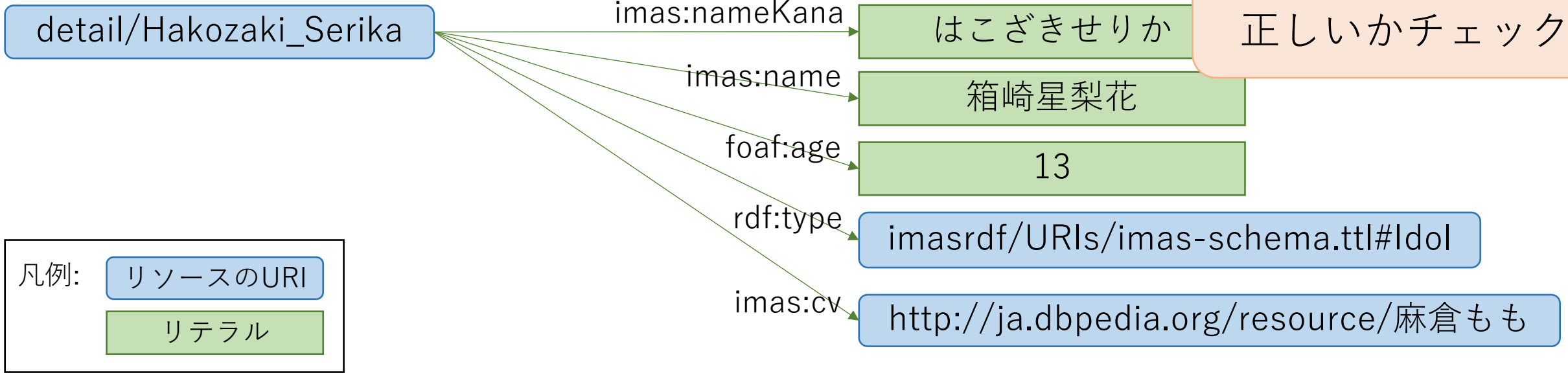
→確認作業を簡単にできるチェックツールを整備する
RDFチェックツール「rdflint」の開発・導入

rdflintで出来ること① ～RDFファイルとして正しいか～

rdflintで出来ること: RDFファイルとして正しいかのチェック のイメージを示します

```
<rdf:Description rdf:about="detail/Hakozaki_Serika">
  <imas:nameKana xml:lang="ja">はこざきせりか</imas:nameKana>
  <schema:name xml:lang="ja">箱崎星梨花</schema:name>
  <foaf:age rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">13</foaf:age>
  <rdf:type rdf:resource="https://sparql.crssnky.xyz/imasrdf/URIs/imas-schema.ttl#Idol"/>
  <imas:cv rdf:resource="http://ja.dbpedia.org/resource/麻倉もも"/>
</rdf:Description>
```

RDFファイルとして正しいかチェック

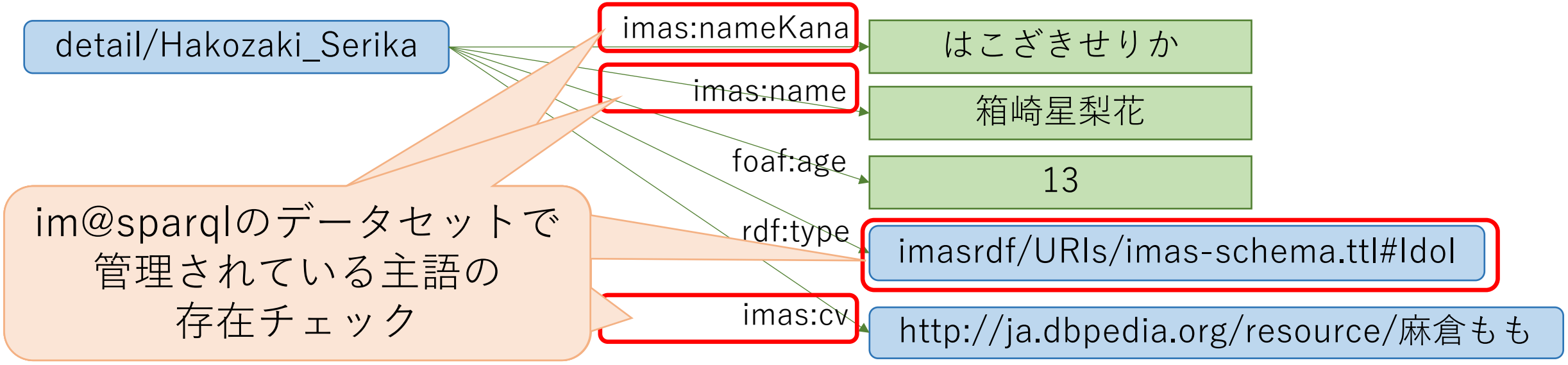


rdflintで出来ること② ～主語の存在チェック～

要するに
リンク切れチェック

rdflintで出来ること: 主語の存在チェック のイメージを示します

```
<rdf:Description rdf:about="detail/Hakozaki_Serika">
  <imas:nameKana xml:lang="ja">はこざきせりか</imas:nameKana>
  <schema:name xml:lang="ja">箱崎星梨花</schema:name>
  <foaf:age rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">13</foaf:age>
  <rdf:type rdf:resource="https://sparql.crssnky.xyz/imasrdf/URIs/imas-schema.ttl#Idol"/>
  <imas:cv rdf:resource="http://ja.dbpedia.org/resource/麻倉もも"/>
</rdf:Description>
```



検査の実行イメージ

検査の実行イメージを示します

rdflintのダウンロード

```
$ wget https://jitpack.io/com/github/imas/rdflint/0.0.6/rdflint-0.0.6-all.jar
```

rdflint検査の実行

```
$ java -jar rdflint-0.0.6-all.jar -config .circleci/rdflint-config.yml
RDFs/Event.rdf
  warn   Undefined URI: https://sparql.crssnky.xyz/imasrdf/RDFs/detail/Hakozaki_serika (Triple:
http://sparql.crssnky.xyz/imasrdf/RDFs/detail/765Caravan_2 - http://schema.org/actor -
http://sparql.crssnky.xyz/imasrdf/RDFs/detail/Hakozaki_serika)
```

エラー理由

対象ファイル

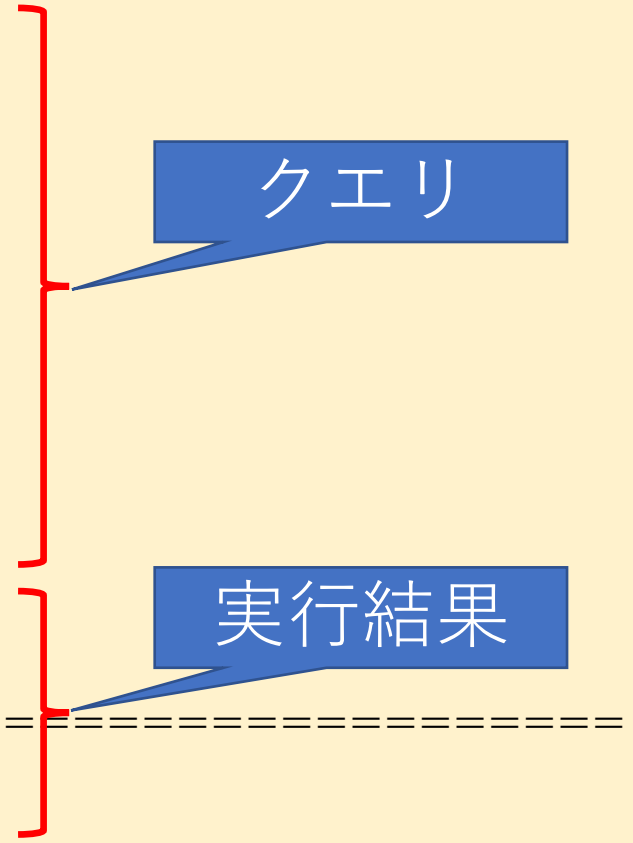
※Hakozaki_serika → Hakozaki_Serika (大文字・小文字の誤り)

SPARQLのテスト実行環境 実行イメージ

SPARQLのテスト実行環境(インタラクティブモード)の実行イメージを示す

rdflintインタラクティブモードでのクエリ実行

```
$ java -jar rdflint-0.0.6-all.jar -i -config .circleci/rdflint-config.yml
sparql > PREFIX schema: <http://schema.org/>
> PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
> PREFIX imas: <https://sparql.crssnky.xyz/imasrdf/URIs/imas-schema.ttl#>
> SELECT ?星梨花の主語
> WHERE {
>   ?星梨花の主語 rdf:type imas:Idol;
>             schema:name ?アイドル名.
>   filter(contains(?アイドル名,"箱崎星梨花"))
> }
>
-----------------------------------------------------------
| 星梨花の主語                                            |
===========================================================
| <https://sparql.crssnky.xyz/imasrdf/RDFs/detail/Hakozaki_Serika> |
-----------------------------------------------------------
```



rdflintで出来ること ～その他の機能と実装予定～

rdflintで出来る紹介した以外の機能、今後実装予定の機能を紹介します

- RDFファイルとして正しいか
- 主語の存在チェック
- SPARQLのテスト実行機能

紹介済み

- SPARQLクエリによるカスタムチェック
- 主語・トリプルが削除されていないかチェック
- RDFファイルの生成機能
- 文字コード、改行コードのチェック
- 数値・文字列などのデータ型チェック
- 外れ値のチェック

実装予定

まとめ

本Lightning Talkのまとめです

- コミュニティでのデータセット作成を運用すると、
  データ作成者・チェック者に、手間のかかる確認作業が発生
  - 手間がかかる上に、システムの知識も必要

- 確認作業は「rdflint」で支援・自動化出来る
  - 機械的にチェック出来ることはたくさんあるので機能強化中
  - まだまだ成長途中なんです！です！

- 「rdflint」は「LOCチャレンジ2019 基盤部門」に応募