RDA 7th plenary 報告

蔵川圭
国立情報学研究所

蔵川圭の自己紹介

- 蔵川圭
- 国立情報学研究所
- 学術基盤推進部学術コンテンツ課
- 特任准教授
- 博士(工学)

http://researchmap.jp/kurakawa/

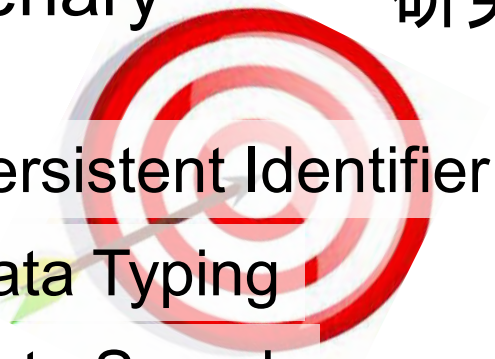
https://www.facebook.com/kei.kurakawa

http://orcid.org/0000-0002-7031-1846

- NIIで10年ほど学術情報データベースと学術情報流通関連サービスの研究開発を行ってきました。それまでは、設計工学およびソフトウェア開発関連の研究を行っていました。
- 学術情報流通まわりのコミュニティの関心の対象は、論文や本のカタログ主体から、本文主体へと変化し、ここ5年の間に研究データへと完全に変化してきています。この潮流に合わせて、3年前にできたRDAの動きを観察しつつ、今回は3度目の参加になります。
- 昨年10月にNIIで行ったSPARC Japanのイベントでは、企画メンバーとして関わり、オープンアクセスと研究データ共有をテーマに、今回のRDAへの前哨戦となるような形で企画しました。

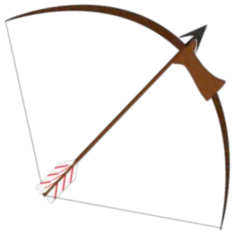
私の関心

RDA 7th plenary　　研究データ共有



1. Persistent Identifier
2. Data Typing
3. Data Search
4. Data Citation
5. Vocabulary Services
6. Text and Data Mining
7. Summer Schools in
   Data Science and Cloud Computing

情報システム開発
情報分析



IG PID

1. 学術情報リソースを有効活用するためには、リソースにPID(persistent identifier,永続識別子)を付与しよう

PIDのタイプ

デジタルオブジェクト、研究データ　研究者　組織　グラント

OpenAIRE　PANGAEA　EUDAT

ePIC (PID services for the European Research Community)

IGSN (international geoscience number)

FORCE11 data citation principles

CrossRef　DataCite

DOI

Handle System

THOR

ORCID

eduPersonOrgOrcid

eduPersonOrgDN

Federation Identity Management

BoF Initial Breakout for the Data Typing Working Group

- CNRI (Corporation for National Research Initiatives)のLarry Lannomがチェア
  - CNRI→
    - Digital Object Architecture
    - The Handle System
  - D-Lib MagazineのEditor-in-Chief
- BoFといえば、IG, WGの前段階であると思われがちだが、これはWG終了後のメンバーの同窓会の位置付けであった

2. データの型定義(data typing)は、データ共有には重要である

- 暗黙的な仮定というのは、問題がおこる。
- データタイプは、メール添付ファイルで言うところのMIME(Multipurpose Internet Mail Extensions) type。これがあると、ソフトウェアによる自動処理が可能となる。
- 研究データ特有のデータタイプは、CSV, NetCDFがあげられる。

WGのアウトプット

目的の明確化：
データタイプと
データタイプレジストリの共有

レジストリの構成：
プロトタイプシステム
http://typeregistry.org/

IDとの統合

BoF on Data Search

3. 研究データに特有の検索ユーザーインターフェースがある

熱気のあるBoFで、検索インターフェースのデモが行われた。会場参加の飛び入りのデモもあり

- デモリスト
  - ANDS: RD Switchboard
    - It spells out Research data switchboard.
  - NIH: Biocaddie
    - This project aims at “Pubmed” for data.
    - http://datamed.biocaddieorg
  - NSIDC: bCube
    - BCube produces geoscience data. This system consists of Solr and Nutch elastic map reduce hosted on the amazon.
  - Pangaea
    - International project since 1995.
    - This handles several kinds of catalogue metadata, i.e. ISO19115, Dublin core etc., which are converged into PANGAEA metadata.
  - CoS: SHARE
    - SHARE have any kind of providers, e.g. VIVO, ST, VVT, crossref, DataONE, then gather all information to share. It uses lucene as a search engine. The presenter pointed out it needs de-duplication effort.
  - EarthChem: earthchem.org
    - She shows demo to search research data in several ways. Locating regions by pointing out bounding box, spread sheet like of chemical materials, LEPR compositional selection.
  - USGIN
    - U.S. national geothermal data system.
  - Elsevier Datasearch
    - http://datasearchdemo.elsevier.com
    - It crawles arXive repository, and characteristically gives data preview feature.

WG Data Citation

Making Dynamic Data citable: Adoption of the Recommendations

- Andreas Rauber, Associate professor at the Vienna University of Technology & SBA, Austriaがチェア
- WGは、2014.03 – 2015.09までおこなわれ、現在は適用支援の段階にある

4. 論文引用とデータ引用は異なる。データの特徴を捉えて、引用する仕組みを整えよう

- データというものの特徴
  - 動的(dynamic)
    - データの追加
    - データの修正
    - データの品質向上
- データが引用されるとき
  - データの属性
    - タイムスタンプ
    - バージョン

- 提案されたデータ引用フレームワーク
  - タイムスタンプとバージョン管理されたデータベース
  - データの問い合わせにPIDを付与
- 適用例
  - VMADC (the virtual atomic and molecular data center), EU
  - CBMI (center for biomedical informatics) @ WUSTL(Washington University in St. Louis)
  - CCCA (climate change center Austria) pilot (www.ccca.ac.at/de/home)
  - ENVRIPlus (cluster of environmental research infrastructures), a Horizon 2020 project
  - Data citation for ARGO (The broad-scale global array of temperature/salinity profiling floats on the ocean)

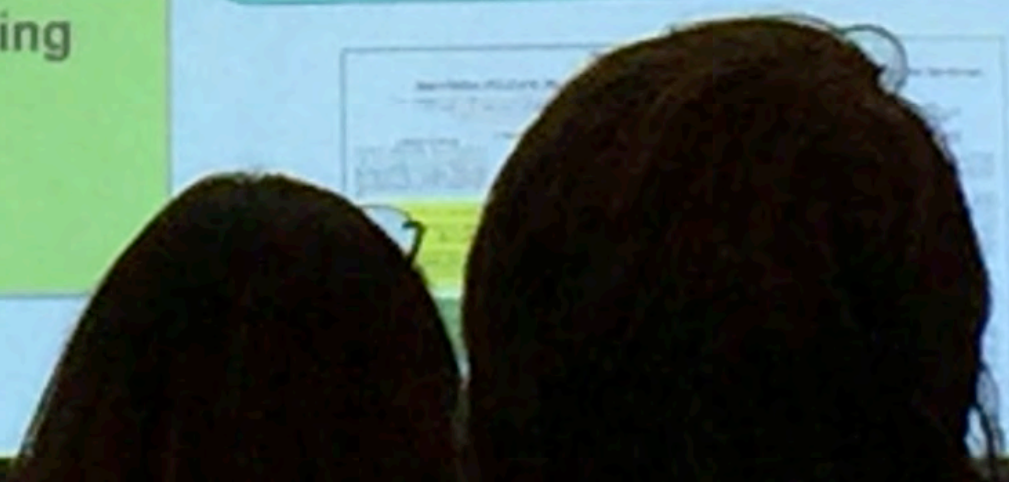
Data Citation – Recommendations

Preparing Data & Query Store
- R1 – Data Versioning
- R2 – Timestamping
- R3 – Query Store

When Data should be persisted
- R4 – Query Uniqueness
- R5 – Stable Sorting
- R6 – Result Set Verification
- R7 – Query Timestamping
- R8 – Query PID
- R9 – Store Query
- R10 – Citation Text

When Resolving a PID
- R11 – Landing Page
- R12 – Machine Actionability

Upon Modifications to the Data Infrastructure
- R13 – Technology Migration
- R14 – Migration Verification



research data sharing without barriers
rd-alliance.org

IG Vocabulary Services

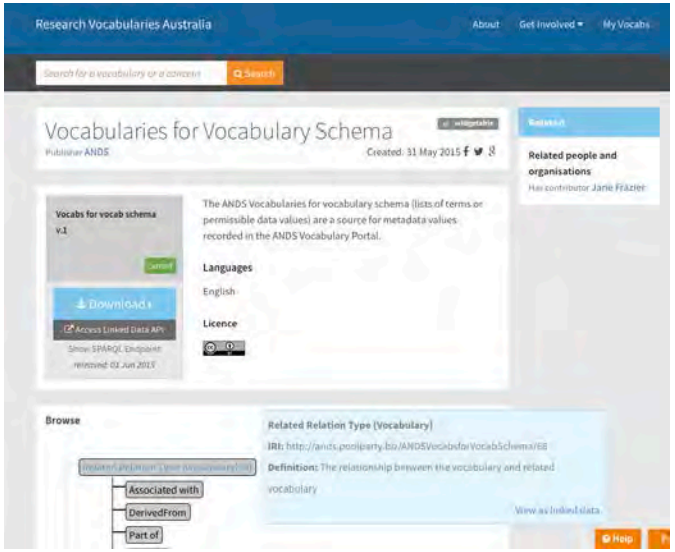
Access Methods Review of Existing Vocabulary Services

5. 研究データ共有のメタデータ管理にかかわる研究用語彙を収集して公開、共有しよう

- ANDS (Australian National Data Service)というプロジェクトにおいて、研究用の語彙の収集と共有を行っている
- データライブラリアンである、Jane Frazierがチェア

当日のデモの内容

https://vocabs.ands.org.au/



- “Pool Party” という、セマンティックウェブ対応の知識管理システムをベースに構築
- SKOS conceptのコレクションであり、スキーマ
- IGでは、既存のサービスのセマンティックウェブ語彙についてサーベイをした

BoF on Text and Data Mining

Text and Data Mining: Defining the Challenges and Actions

- このBoFは、今回が初めての集まり

6. テキストおよびデータマイニングのためにデータを公開したい機関が集まって情報共有しよう

すでに、EUではHorizon 2020のもと、オープンサイエンスを前提にした
テキストマイニング基盤のプロジェクトがある

openMINTED
Open Mining Infrastructure for Text & Data

http://openminted.eu

- 言語リソースのメタデータやテキストマイニングサービスの標準化
- 言語リソースの仕様や、異なるリソースおよびツール間の相互運用
- ライセンス

ATHENA RC, UNIV. OF MANCHESTER (NACTEM), LIBER, UNIV. OF AMSTERDAM, GESIS, GRNET, TU DARMSTADT, INRA, OPEN UNIVERSITY UK, EPFL, FRONTIERS, UNIV. OF STIRLING, EMBL-EBI, AGRO-KNOW, CNIO, UNIV. OF SHEFFIELD (GATE)

WG RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World

CODATA-RDA schools in Research Data Science

7. オープンサイエンス時代のデータサイエンスを教えよう

国境を越えて、エンジニアを対象に
こんなことを教えています

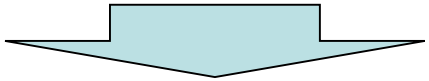
- 講義内容例の紹介

  - Open Science
  - Data carpentry
  - Visualization with R, ggplot2
  - Analysis by machine learning and statistics

- その場にいた人のやり取り
  - イベントを開催するのに、どれくらいの経費がかかるか？
  - （マネージャーレベルの責任ある人が集まっている？）
- よくあるデータサイエンスがらみのイベントとの違い
  - オープンサイエンスを前提
  - データ利用の帰属について教えることが含まれている

全体を通した雑感

- 普段、その業務についており、かつ、執行権限のあるマネージャーが参加している
- 研究データを生産している現場の研究者、エンジニアは少ないかもしれない
- 研究データそのものよりも、枠組みや運用に関する分野を超えた共通の課題が話題として成立する



- RDAは、研究データ資源のスチュワードシップを発揮する責任権限のあるマネージャー会合
- 問題提起や解決案をリードするなら、その限りではない