

国立情報学研究所


Linked Open Dataと 学術・公共情報流通

国立情報学研究所
大向 一輝
@i2k

国立情報学研究所

自己紹介


- ・ コンテンツ科学研究系・准教授
 - ・ セマンティックウェブ・ソーシャルメディア
 - ・ オープンデータ
- ・ コンテンツシステム開発室長
 - ・ CiNii Articles / Books
- ・ 株式会社グルコース
 - ・ RSSリーダー・Twitterクライアント
- ・ 「ウェブらしさを考える本」
 - ・ 全文公開中



国立情報学研究所

知識は失われる


- ・ 壊れる
 - ・ 物理的損壊・エラー
- ・ 読めなくなる
 - ・ 記録メディアのライフサイクル
 - ・ ソフトウェアのライフサイクル
- ・ 意味がわからなくなる
 - ・ 社会のライフサイクル
- ・ 「データの保存」とは
 - ・ コピー?
 - ・ バックアップ?
 - ・ 使う



国立情報学研究所

知識インフラ

- ・ 収集する
 - ・ コピー・アクセス
- ・ 活用・創造する
 - ・ 組み合わせ・変換
- ・ 公開・発信する
 - ・ 次の収集プロセスのために
- ・ 知識インフラの役割
 - ・ 時間・空間を超えて使われ続けるようにすること
 - ・ 未知の誰か・未知の用途のため
 - ・ **オープンかつ共通の基盤が必要**



国立情報学研究所

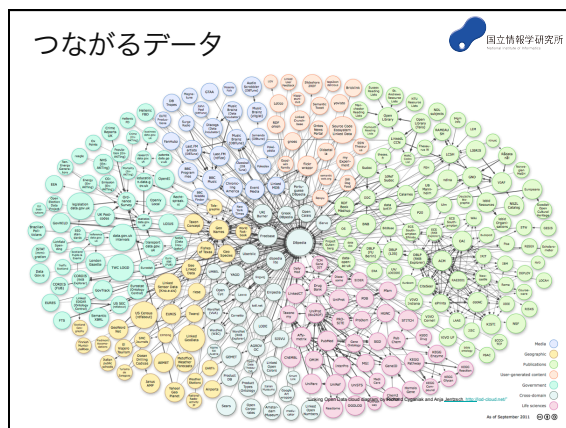
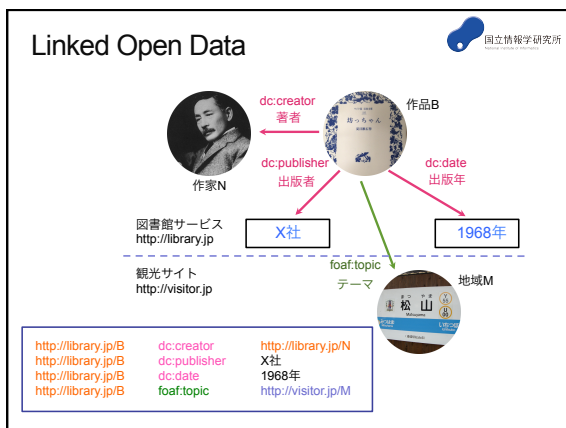
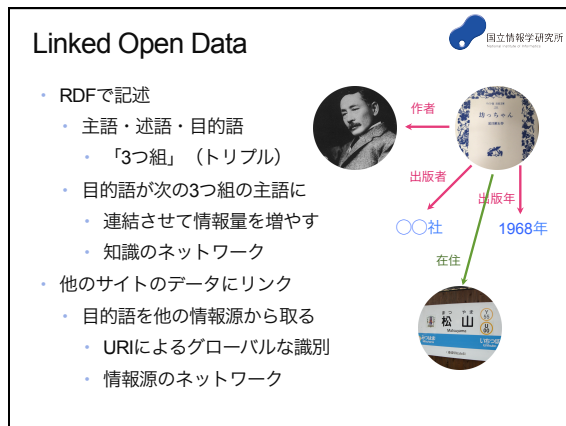
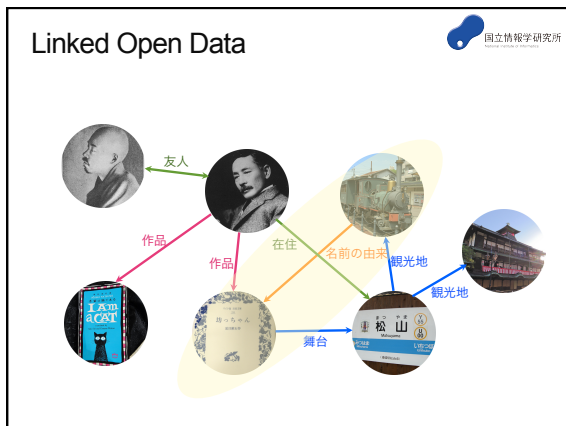
データの価値を高める

- ・ 他分野の知識とつなげる
 - ・ 相手の知識を豊穡に
 - ・ 自身を豊穡に
 - ・ 新たなつながりを見出す
- ・ 新たなユーザを発見する
 - ・ すぐに使える形で提供
 - ・ 理解しやすく使いやすい
- ・ **Linked Open Data (LOD)**

国立情報学研究所

Linked Open Data

- ・ セマンティックウェブ
 - ・ ティム・バーナーズ＝リーが提唱
 - ・ 機械可読なフォーマット
 - ・ 厳密な知識表現とオントロジー（辞書）
- ・ Linked Open Data
 - ・ セマンティックウェブの実践
 - ・ データのオープン化・構造化を中心に
 - ・ 多様な情報源・データベースが存在
 - ・ API・マッシュアップに親しんだ開発者の存在
 - ・ **文書のウェブからデータのウェブへ**




- ### LODの利点
- 標準化による分業体制の確立
 - データの書き方: RDF
 - 分野のスペシャリストはデータの記述に集中
 - データの取り出し方: SPARQL
 - RDFストアのクエリ言語
 - アプリケーション開発者はユーザインターフェイスに集中
 - データの取り出しに独自プログラムを必要としない
 - TXT・CSV・XLS・XMLとの相違点
- 国立情報学研究所

- ### 学術情報とLOD
- 学術情報分野の特徴
 - 情報の構造化を生業とする職業集団・組織がある
 - 研究者(大学・研究機関)・学会・図書館...
 - 情報の構造化フォーマットが共有されている
 - タイトル・著者名・抄録・本文・参考文献...
 - 「何を」「どう作る」は解決済み
 - フォーマット変換のみ
- Weblogの現在と展望: セマンティックWebおよびソーシャルネットワークの基礎として
The State-of-Art and Prospects of Weblog: An Infrastructure of Semantic Web and Social Networking
- 国立情報学研究所
 国立情報学研究所
 国立情報学研究所
 国立情報学研究所
- ▼ 巻末文庫: 21冊 ▼ 雑誌文庫: 15冊
- 国立情報学研究所

CiNii

- 国内最大級の学術情報サービス
 - 論文データ
 - 本文400万件
 - 書誌1500万件
 - 図書・雑誌データ
 - 書誌1100万件
 - 所蔵1億1000万件
 - 大学図書館・学協会によるデータ作成/維持管理
 - CiNiiリニューアル ('09~'11)
 - オープン・コネクタ戦略



CiNiiの書誌メタデータ

```

<rdf:Description rdf:about="http://ci.nii.ac.jp/ncid/BB02488158#entity">
  <foaf:isPrimaryTopicOf rdf:resource="http://ci.nii.ac.jp/ncid/BB02488158.rdf"/>

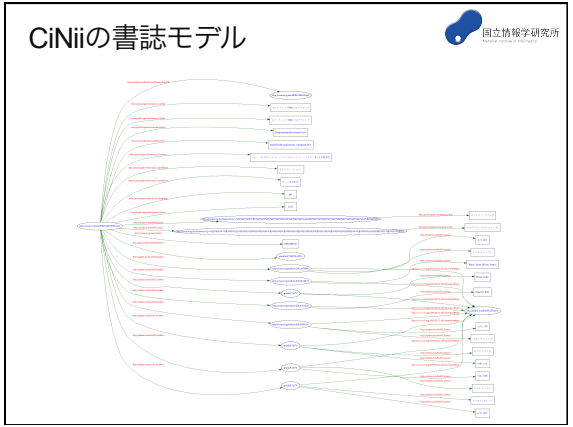
  <dc:title>セマンティックWebプログラミング</dc:title>
  <dc:title xml:lang="ja-hrkt">セマンティック Web プログラミング</dc:title>
  <dcterms:alternative>Programming the semantic web</dcterms:alternative>
  <dc:creator>トビー・セガラン著; 玉川竜司訳</dc:creator>
  <dc:publisher>オライリー・ジャパン</dc:publisher>
  <dc:language>jpn</dc:language>
  <dc:date>2010</dc:date>
  
```

CiNiiの書誌メタデータ

```

<foaf:topic rdf:resource="http://ci.nii.ac.jp/books/search?q=セマンティックウェブ" dc:title="セマンティックウェブ"/>
<ciinii:ncid>BB02488158</ciinii:ncid>
<dcterms:hasPart rdf:resource="urn:isbn:9784873114521"/>
</rdf:Description>

<rdf:Description rdf:about="http://ci.nii.ac.jp/ncid/BB02488158#entity">
  <foaf:maker>
    <foaf:Person rdf:about="http://ci.nii.ac.jp/author/DA15839119">
      <foaf:name>大向, 一輝</foaf:name>
      <foaf:name xml:lang="ja-hrkt">オオムカイ, イツキ</foaf:name>
    </foaf:Person>
  </foaf:maker>
</rdf:Description>
  
```



CiNiiのメタデータデザイン

- 語彙と構造をどのように決めるか
 - 標準化された構造はない
 - まったく同じサービスは存在しない
 - 日本独自の事情 (よみ)
 - 語彙の重複・独自語彙の必要性
 - 厳密性と利便性のトレードオフ
- 方針
 - シンプルなデータ構造
 - ライブラリの普及状況を念頭に
 - 世界標準に準拠

メタデータを作る

- 構造化されてこなかった情報への対応
- 既存データに手をつけるコスト
 - 分量・ワークフロー
- 完全性の保証
- 代表的な例：著者ID
 - 高まる重要性
 - 個人の業績管理
 - 国際競争 (ResearcherID・ORCID)
 - 著者名典拠がない
 - 論文の著者名は膨大かつロングテール
 - 同姓同名・旧姓・タイプミス...

CiNii著者検索

国立情報学研究所

- NII著者ID (NRID) の導入
 - 科研費番号+機械処理による著者へのID付与
 - 著者ごとのページを生成
- NRIDベースの論文検索機能
 - 著者名→IDリスト→論文リスト
 - APIの提供
- 新たなデータ生成・管理モデル
 - 研究成果の活用
 - ユーザーフィードバック

CiNii著者検索の概要

CiNii著者検索

国立情報学研究所

- ALS (Author Linking System)
 - i-Linkage (NII相澤教授) の大規模・実運用システム
 - CPU32コア・メモリ320GB・計算時間5日 (全件処理)
- フィードバック (同一人物の報告)
 - 機械処理だけで100%の精度を得ることは不可能
 - あらかじめフィードバックを織り込んだシステム・アルゴリズム設計
 - 例: 過統合より未統合を指摘する方が簡単
- 実績: 6217件 (4月1日~7月15日)
 - Researchmap経由で研究者本人からのフィードバックも可能に

CiNiiのデータ共有

国立情報学研究所

- ウェブAPIコンテスト ('09・'10)
 - Twitter・地図・学術DBとの連携
 - 専門家を探す・志望校を選ぶ...
- アクセス数

LODをつくる

国立情報学研究所

- LODAC (Linked Open Data for ACademia)
- 国内の学術情報・公共情報をLODで公開し、共有を促進
 - 学術分野のみならず、広く情報を共有するための情報流通基盤の構築
- 複数の情報源・分野にまたがる情報を共有するためのモデル構築
 - データ構造・スキーマの違い
 - 情報の同一性
- 現在の活動
 - Museum: 美術館・博物館情報 (人文科学)
 - Location: 地図・地名情報 (公共・公的情報)
 - Species: 生物情報 (自然科学)

<http://lod.ac>

LODAC Museum

国立情報学研究所

- 美術館・博物館情報の統合と共有
 - 日本国内に6000館以上
 - 資料情報は個別管理
 - 網羅的な検索・調査ができない
 - 資料間の関連が不明
 - 集中管理は可能か?
 - 決められた枠内のメタデータでは資料情報記述に対応できない
 - 情報が欠落する可能性
 - 細かすぎると使われない
 - そもそもどのような属性項目があるのか不明

LODAC Museumの情報源

- 美術館・博物館情報の統合と共有
 - 提供
 - 日本美術シソーラス [福田97]
 - 取得
 - 収蔵品資料 (80館)
 - 国指定文化財データベース
 - 文化遺産オンライン
 - API経由
 - 日本語版DBpedia
 - 約100,000項目+DBpedia

美術館・博物館	
(1) 東京国立近代美術館	
(2) 国立西洋美術館	
(3) 京都国立近代美術館	
(4) 国立国際美術館	
(5) 京都国立博物館	
(6) 奈良国立博物館	
(7) 福島県立美術館	
(8) 栃木県立美術館	
(9) 秋田県立近代美術館	
(10) 岩手県立美術館	
(11) 徳島県立近代美術館	
(12) 山梨県立美術館	
(13) 東京都現代美術館	
(14) 香川県立栗山魁夷せとうち美術館	
(15) 横浜美術館	

データの標準化

- スキーマの観察・分類→簡易スキーマを定義
 - 標準的な語彙を優先的に使用
 - 既存の名前空間・プロパティ
 - 独自語彙は最小限に
 - 各情報源のスキーマを簡易スキーマにマッピング
 - 「メタデータ情報共有のためのガイドライン」の議論

Property (一部項目省略)	PREFIX	URI
資料分類	lodac:genre	crm
文化財	lodac:culturalAssets	http://purl.org/NET/oidoc:crm/core#
制作者	dc:creator / dc:l1-creator	determs
制作	crm:PT_look_place_at	http://purl.org/dc/terms/
作品名	dc:title / skos:prefLabel	dc
作品名読み	dc:title @ja / rdfs:skos:altLabel	foaf
作品名英語	dc:title @en / skos:altLabel	http://www.w3.org/2004/02/skos/core#
図文	crm:P621_is_depicted_by	rdfs
図像	crm:P625_showing_visual_item	http://www.w3.org/2000/01/rdf-schema#
異数	crm:P57_has_number_of_parts	ical
コレクション	dc:isPartOf	rd2
制作年	dc:created	http://rdvocab.info/ElementsGr2
推定始年	lodac:estimatedStartYear	lodac
材質	dc:medium / crm:P45_consists_of	http://lod.ac/ns/lodac#

情報の統合

- 日本美術シソーラスを中心とした「名寄せ」
 - メンテナンスされている知識体系
- 文字列マッチによる統合
 - 作者名はユニークネスが高い
 - 論文・書籍では同姓同名が多い
 - 組織・機械処理による名寄せが必要

LODACデータベースの公開

SPARQL Endpointの提供

```

Query
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX lodac: <http://lod.ac/ns/lodac#>
PREFIX lodocid: <http://lod.oc/id/>
SELECT ?work ?title
WHERE {
  ?work lodocid:359 lodac:creates ?work .
  ?work dc:title ?title .
}
LIMIT 100
  
```

実行 (リセット) Options Output Format (html) Show query URI Sample queries

地理・地名情報のLODAC化

- 情報源
 - 大字・町丁目レベル位置参照情報 (国交省)
 - 郵便番号データ (日本郵便)
 - 国土数値情報 (国交省)
 - 駅データ
- 情報
 - 住所
 - 緯度・経度
 - 施設名
 - 事業所名

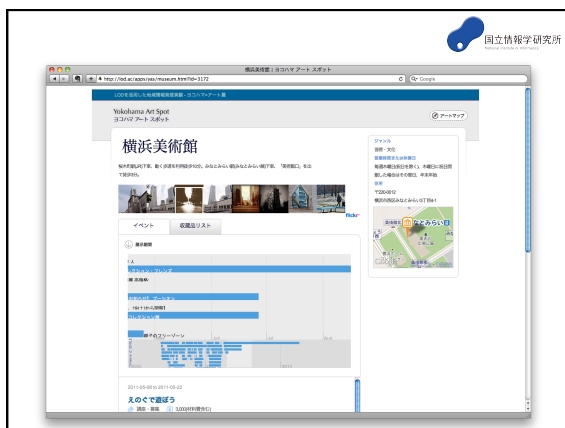
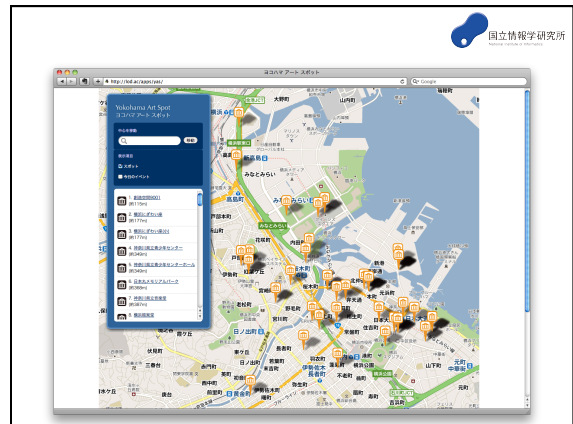
LODAC Maps

- 複数のSPARQLエンドポイントを統合・差し替え可能
 - LODAC DB / DBpedia / LinkedGeoData
- 標準的なデータアクセスの意義
 - データを再利用可能なマイマップ

ヨコハマアートスポット



- ・ 地域コミュニティとの連携
 - ・ 横浜市文化芸術振興財団 / 横浜LODプロジェクト
- ・ アートイベントのLinked Data化
 - ・ 施設情報をキーにした位置情報・イベント情報・作品情報のリンク
 - ・ フロー情報からアーカイブを作る
 - ・ SPARQLでアクセス



その他のアプリ



- ・ CiNii x BDLS
 - ・ 生物学辞書によるキーワード拡張
- ・ DashSearch for SPARQL
 - ・ 試行錯誤を許す検索インターフェイス
- ・ オープンハウスつながりマップ
 - ・ 論文・研究報告書から見るネットワーク
- ・ MMap
 - ・ 国際会議支援サービス
- ・ SPARQL+UI

今後の課題



- ・ 「生データとの格闘」
- ・ データ量の拡大
- ・ 対象の拡大
 - ・ DBpedia日本語版 (Wikipedia)
 - ・ Wordnet日本語版 (シソーラス)
- ・ メタデータをつくるためのワークフロー
 - ・ 1次情報源に委ねるために
 - ・ メリット・インセンティブの設計
- ・ メタデータの流通・2次利用
 - ・ ライセンス

オープンデータの潮流



- ・ オープンガバメントの潮流
- ・ 政府・行政機関のデータ提供
 - ・ アメリカ・イギリス・
 - ・ 鯖江市・会津若松市・流山市...
 - ・ 総務省・経済産業省...
- ・ サービス向上・透明性・コミュニケーション



データシティ鯖江 (XMLRDFによるオープンデータ化の推進)
 情報統計課 情報統計グループ SC-JohoTokai@city.sabae.lg.jp
 電話番号: 情報統計グループ (情報) 0778-53-2213
 情報統計グループ (統計) 0778-53-2212



国立情報学研究所

5 ★ Open Data

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5 star deployment schema for Open Data. Here, we give examples for each step of the stars and explain costs and benefits that come along with it.

LINKED OPEN DATA
 ★ On the web
 ★ Machine-readable
 ★ Non-proprietary format
 ★ RDF standards
 ★ Linked RDF
 ★ YOUR DATA 5★

XML CSV JSON JSON-LD

http://data...

CC BY NC ND

国立情報学研究所

Linked Open Data Challenge Japan 2012

開催情報 エントリー スポンサー イベント LODとは 公式ブログ

「Linked Open Data チャレンジ Japan 2012」では、様々な分野で Linked Open Data (LOD) の仕組み作りやアプリケーションにチャレンジされている方々による活動の発表の場を提供します。

Open = つながる。

4つの部門で開催中!!

データセット Z11Z2 Z22Z1Z2 Z1Z2Z1Z2

開催日程
 応募期間 2012年10月1日～2013年1月31日
 審査結果発表会実施 2013年3月7日 (木)

News

長瀬コラム 第4回「第3回 LODチャレンジアー「オープンデータハッカソン」の経過」
 更新日: 2012-11-09

「働きしよしよ」第5回LODチャレンジアー「アイデアソン」を開催します
 更新日: 2012-10-22

第3回 第4回LODチャレンジアーを開催します
 更新日: 2012-10-04

第6回LODチャレンジアーをクリエイティブライセンス・コンソーシアム 第5回発表会で発表
 スポンサー募集中です!

国立情報学研究所

まとめ

- 何のためのLOD?
 - 公開すること自体の重要性
 - ドメイン内の利用
 - ドメインを超えた相互利用・共有
 - 自身が持つデータの新たな価値を知る
- 「Information wants to be linked」