

デジタルドキュメント (9)

高久雅生

2014年6月12日(木)3・4時限

本日のお品書き

- (前回の振り返り)
- (前回講義での質問・要望への回答: 2件)
- セマンティックウェブとデジタルドキュメント
 - Semantic Webの基盤技術
 - オープンデータとメタデータ, ライセンス
 - Linked Open Data (LOD)
 - デジタルドキュメントのワークフロー

質問・要望へのお返事 (1)

- (第3回レポート返却分について)PCやブラウザで日本のKindleサイトで買った電子書籍を見ることが出来ると書いてあったのですが、どれだけ調べても日本のサイトで買った電子書籍をPCやブラウザで読む方法が分かりません。たぶんできないと思います。

はい、これはわたしの勘違いでした。日本版Kindle (Amazon.co.jp)で購入分のコンテンツには、PC版Kindleはまだ未対応の状態が続いているようです。

質問・要望へのお返事 (2)

- 暑いです。エアコンってまだつきませんか。

筑波大学公式の冷房稼働期間は6月25日(水)から9月5日(金)です...

質問・要望へのお返事 (3)

- EPS形式もPostScriptに関連したものか？

EPS (Encapsulated PostScript) 形式はPostScriptフォーマットの亜種です。

通常のPostScriptがページ単位の出力を前提にしているのに対して、EPSは図画像用途に用いられるため、図のサイズ情報(縦横)を持つことを前提にした描画フォーマットに特化しています。

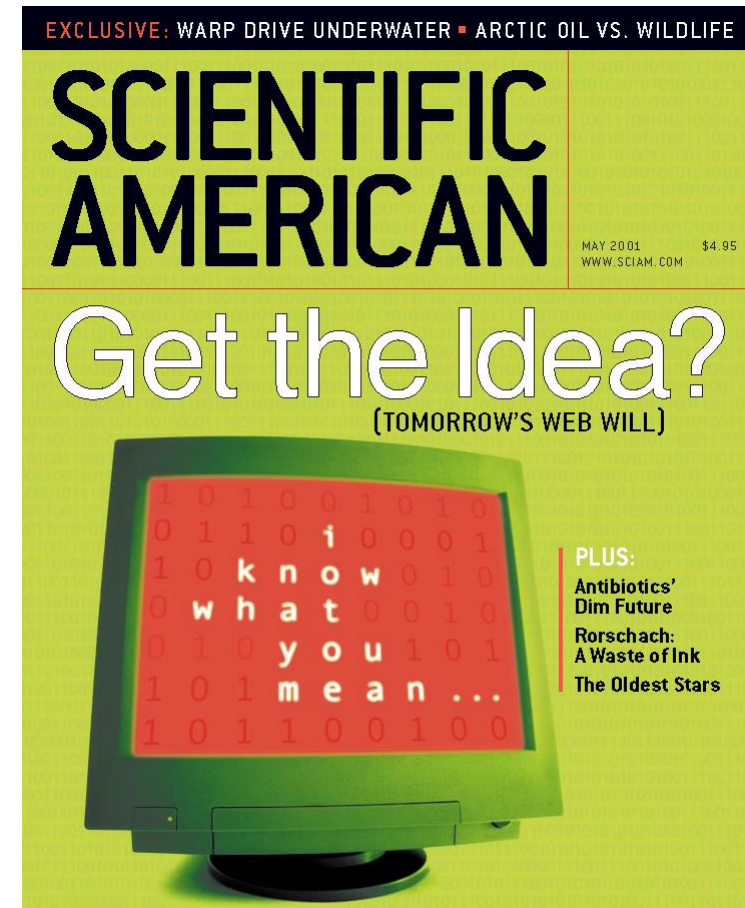
(前回の振り返り)

- LaTeX
 - 文書例
 - LaTeX文書の構成要素
 - 利用例 / 組版
- PDF
 - 歴史, PostScript
 - 構成要素
 - 文書フォーマットの構造
- オフィス文書とデジタルドキュメント
 - ワークフロー
 - 処理モデル
- OOXMLとODF
 - 標準化(独占規格から国際標準へ)
 - パッケージ形式による複合オブジェクトの埋め込み
 - 文書内容にあわせた文書要素

Semantic Web (1)

Tim Berners-Lee, James Hendler, Ora Lassila. The Semantic Web. Scientific American, 2001, Vol.284, No.5, pp.35-43.

- WebからSemantic Webへ
- 意味的記述と機械的理解を可能とするWebマークアップ



Semantic Web (2)

- Semantic Web
 - 意味的に解釈可能なWeb
 - 誰に?
 - 機械[可読]と機械[理解]
- ツール
 - Web
 - Semantics
 - Metadata
 - Ontology

Semantic Webアプリケーション (1)

- 例:「仕事帰りに寄れる歯医者さんを調べたい」
 - 仕事: 平日9:00-18:00
 - 仕事帰りに寄る: つくばエクスプレス(TX)沿線
 - 18時以降に診察可能
 - TX沿線の駅名: つくば, 研究学園, ..., 南流山, 北千住, ...
 - 駅から徒歩 500m 圏内

Semantic Webアプリケーション (2)

- 曖昧性の解消
 - 月=月曜日 = Monday = Mon.
 - 「9:00-13:00・15:00-19:00」
 - 休診日, 診療時間
 - 祝日, 祝祭日, 年中無休
- 常識の理解
 - 1週間 = 月火水木金**土****日**
 - 平日=月～金
- Webマークアップを通じた情報抽出

医療法人〇〇会
歯科クリニック
〒 305-0822 つくば市 〇〇112
TEL : 029-800-0000
診療時間

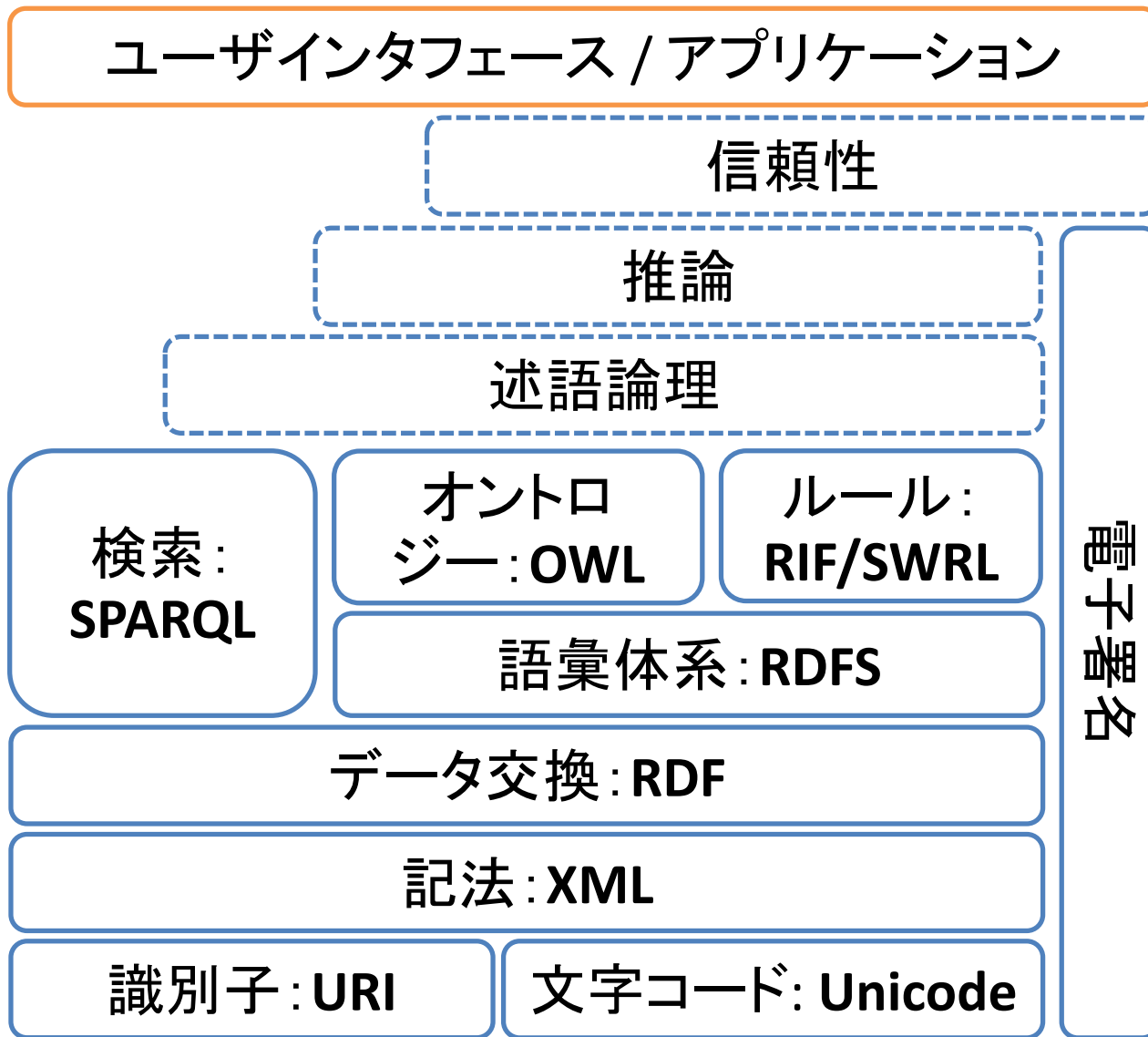
==平日==
9:00-13:00・15:00-19:00

==土曜==
9:00-13:00

月火木金 (水曜のぞく)	10:00～13:30 15:30～20:30
土曜日	10:00～13:30 15:00～18:00
日曜日	9:30～13:30
休診日	水曜日 祝祭日

セブンイレブン
りそな ガスト
千住宿場町通り歯科
北千住西口 徒歩5分
西口
北千住駅

Semantic Webの構成要素

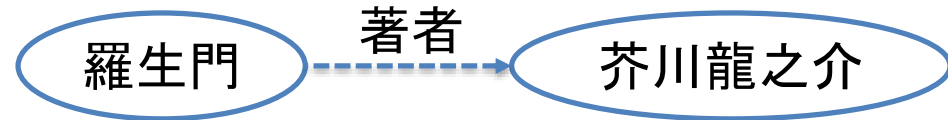


Semantic Webの課題

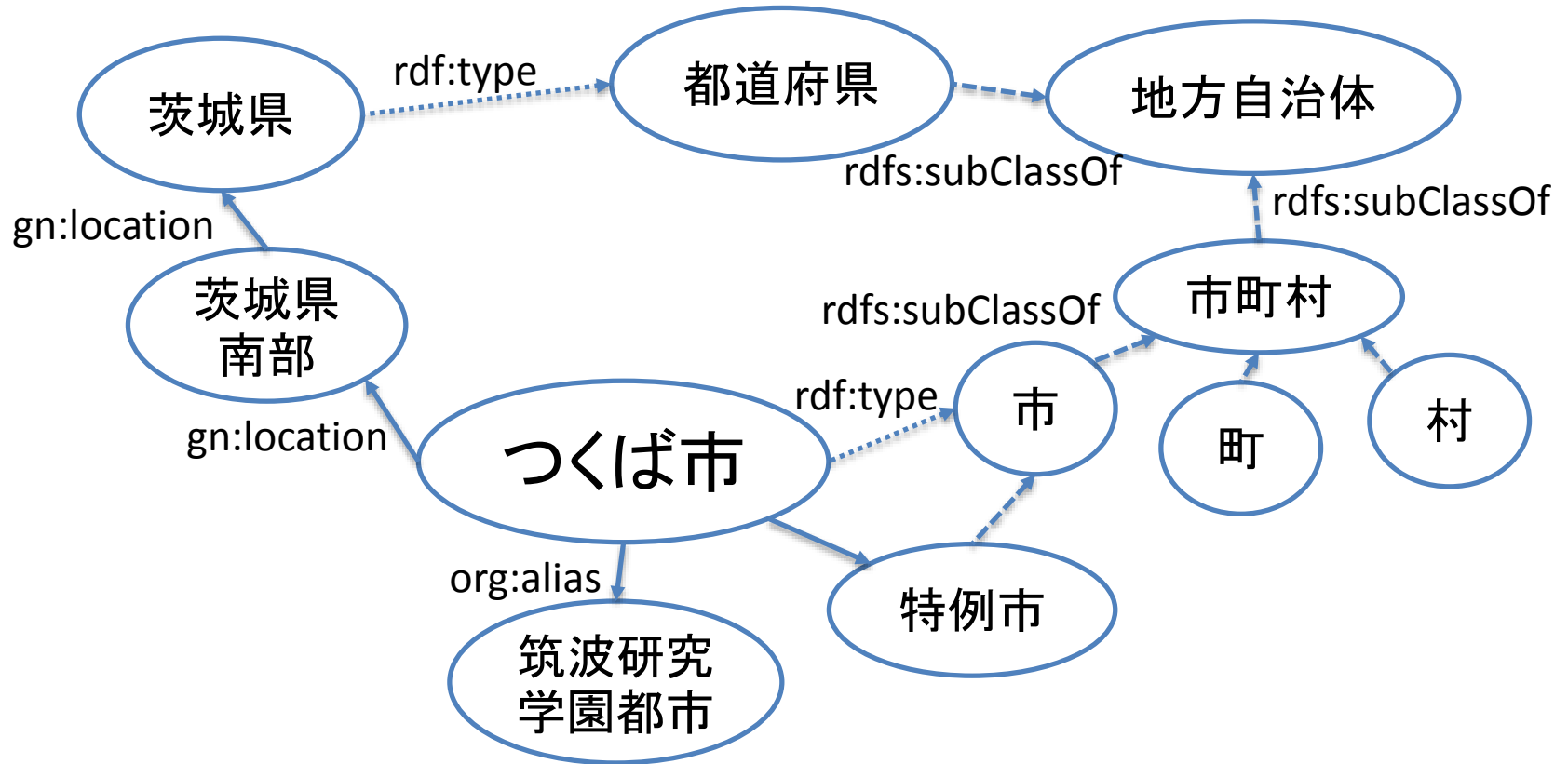
- Webの分散性 + 膨大な情報
 - 数千億 - 1兆ページを超えるWeb空間
 - 多様な情報発信ルート
 - 多言語、多文化
 - 統制された語彙使用や慣習を前提としない
 - 多様な概念や記述によるビッグデータが取得可能
- 汎用モデルの困難さ
 - 計算機アプリケーションによる意味理解は困難

RDFデータモデル

- RDF (Resource Description Framework)
- グラフデータモデル
 - ラベル付き有向グラフ
 - 3つ組 (Triple) による表現
- 特徴
 - シンプルで強力なデータ表現
 - 記述規則が複雑となりがち
 - 処理演算に時間がかかる



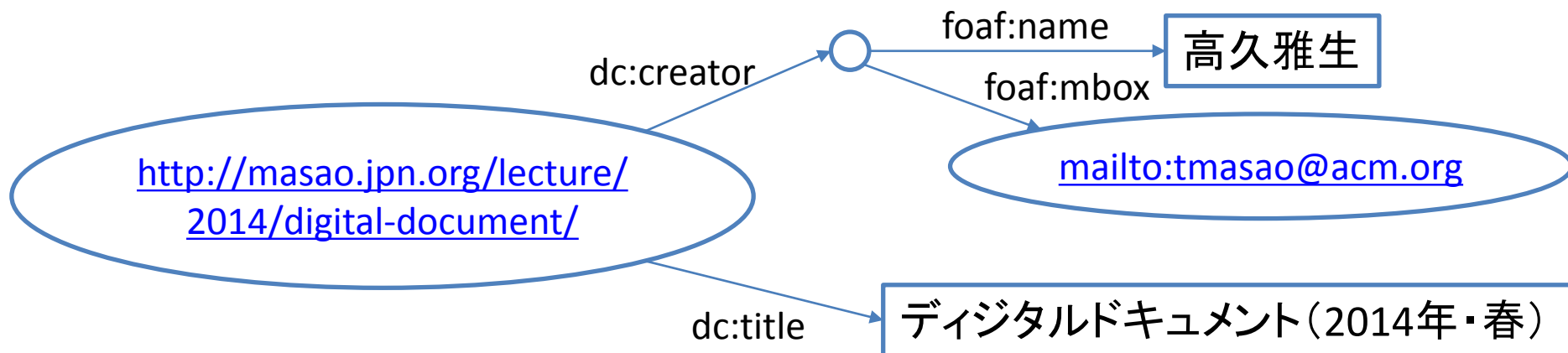
RDFグラフモデルの一例



つくば市(つくばし)は、[茨城県南部](#)に位置する[市](#)である。[学術・研究都市](#)としての[筑波研究学園都市](#)はつくば市全域を区域とする。[特例市](#)、[業務核都市](#)、[国際会議観光都市](#)に指定されている。(Wikipedia日本語版)

RDFデータモデルの記述例

- あるリソース(URL)のタイトル(title)は「デジタルドキュメント(2014年・春)」で、著者(creator)が氏名「高久雅生」であるという関係は、下記のようなモデルとして表現:



RDFデータモデルの記述例 (Turtle)

```
@prefix dc: <http://purl.org/dc/elements/1.1/>
@prefix foaf: <http://xmlns.com/foaf/0.1/>
<http://masao.jpn.org/lecture/2014/digital-
document/>
    dc:title “デジタルドキュメント(2014年・春)”
    dc:creator [
        foaf:name “高久雅生”
        foaf:mbox <mailto:tmasao@acm.org>
    ] .
```


RDFデータモデルの記述例 (XML)

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <rdf:Description rdf:about="http://masao.jpn.org/lecture/2014/digital-
document/">
    <dc:title>デジタルドキュメント(2014年・春)</dc:title>
    <dc:creator>
      <rdf:Description>
        <foaf:name>高久雅生</foaf:name>
        <foaf:mbox rdf:resource="mailto:tmasao@acm.org" />
      </rdf:Description>
    </dc:creator>
  </rdf:Description>
</rdf:RDF>
```

RDFの提供例 (CiNii)

- 通常のWebページ (HTML)
 - <http://ci.nii.ac.jp/naid/130003392336>
- RDF表現 (XML)
 - <http://ci.nii.ac.jp/naid/130003392336.rdf>
- メタデータ
 - Dublin Core
 - Prism (Publishing Requirements for Industry Standard Metadata) (—雑誌情報)
 - FOAF (The Friend of a Friend) (—著者情報)

RDFの提供例 (DBPedia)

- 例:

<http://ja.dbpedia.org/page/%E3%81%A4%E3%81%8F%E3%81%B0%E5%B8%82>

- 百科事典サービスWikipedia上からデータ抽出したもの(＋人手による属性同定とマッピング)
 - http://mappings.dbpedia.org/index.php/Mapping_ja

よりシンプルなWebデータの構造化

- Microformat, Microdata

`高久雅生`



`高久雅生`

- ハイパーテキストの構造に意味を付与
 - ドメイン(分野)毎に合意された意味関係を付与することにより、アプリケーション利用、再利用を促進
 - サーチエンジンにおける利用: Schema.org

オープンデータ

- データ
 - 学術研究データ
 - 統計データ
 - その他
- 「オープン(Open)」の意味
 - 無償利用
 - 非営利使用
 - 加工、再利用

ライセンスと著作権

- 著作権: もっとも基本的な知的財産権
 - 著作財産権と著作人格権から構成される
 - 無許諾で生成される
- ソフトウェアライセンス
 - コピーレフト (Copyleft)
 - オープンソース (Open Source)
- Creative Commons (CC) ライセンス
 - 権利者表示 (BY)
 - (+継承) - Share Alike (SA)
 - (+非商用) - Non Commercial (NC)
 - (+改変禁止) - No Derivatives (ND)
- オープンデータライセンス

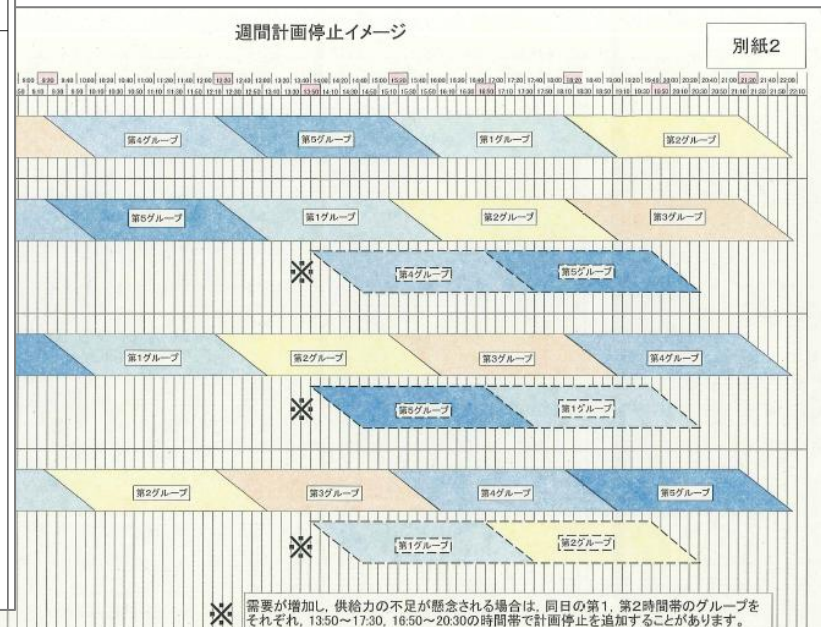
オープンデータとデジタルドキュメント

- 何のためのドキュメント?
- 誰のためのドキュメント?
 - データとメタデータ
 - データの交換可能性, 利用可能性

東京電力計画停電の例から (1)

東京電力が計画停電の予定をPDFで公開 (2011年3月15日)

平成 23 年 3 月 15 日 0 時時点版
想定される停電エリア
<平成 23 年 3 月 15 日 (火) 15:20 ~ 19:00>
第 1 グループ
下記、市町村の一部のエリア
【栃木県】さくら市、宇都宮市、芳賀郡益子町、芳賀郡市貝町、真岡市、大田原市、那須郡那珂川町、那須烏山市、那須郡那須町、芳賀郡芳賀町、塩谷郡高根沢町、芳賀郡茂木町、下都賀郡野木町、下野市、河内郡上三川町、小山市、栃木市、那須塩原市
【群馬県】みどり市、伊勢崎市、佐波郡玉村町上陽地区、桐生市、渋川市、前橋市、利根郡昭和村
【茨城県】かすみがうら市、つくばみらい市、つくば市、稲敷郡阿見町、下妻市、稲敷郡河内町、牛久市、桜川市、取手市、守谷市、常総市、筑西市、土浦市、結城郡八千代町、坂東市、北相馬郡利根町、龍ヶ崎市、石岡市
【埼玉県】さいたま市西区、ふじみ野市、秩父郡皆野町、狭山市、上福岡市、坂戸市、入間郡三芳町、志木市、所沢市、秩父郡小鹿野町、新座市、川越市、



東京電力計画停電の例から (2)

誰かがcsvに変換して公開

東京電力:計画停電(輪番停電)グループCSV【3月14日15時現在】

Posted by: Kotsutsumi

2011/3/14 月曜日 21:17:37

記事更新お知らせ

東京電力のPDFがコピーできなく、プロテクトかかっているなど「腹立つなー」ということで、全部CSV化しました。
システムなどにご利用ください。

東京電力:計画停電(輪番停電)グループCSV【3月14日15時現在】CSVダウンロード一覧

“

CSV形式/エンコード:ShiftJIS

- 栃木
- 茨城
- 群馬
- 千葉
- 神奈川
- 東京
- 埼玉
- 山梨
- 静岡
- [すべてのエリア](#)

<http://code.xenophy.com/?p=1323>

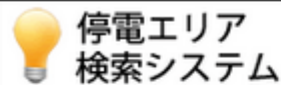
	A	B	C	D	E
1	東京都	あきる野市 網代		2	
2	東京都	あきる野市 洲上		3	
3	東京都	あきる野市 洲上		3	
4	東京都	あきる野市 養原		3	
5	東京都	あきる野市 養沢		3	
6	東京都	あきる野市 油平		3	
7	東京都	あきる野市 野辺		3	
8	東京都	あきる野市 網代		3	
9	東京都	あきる野市 平沢東1丁		3	
10	東京都	あきる野市 平沢西1丁		3	
11	東京都	あきる野市 平沢		3	
12	東京都	あきる野市 入野		3	
13	東京都	あきる野市 二宮東3丁		3	
14	東京都	あきる野市 二宮東2丁		3	
15	東京都	あきる野市 二宮東1丁		3	
16	東京都	あきる野市 二宮		3	
17	東京都	あきる野市 草花		3	
18	東京都	あきる野市 切欠		3	
19	東京都	あきる野市 瀬戸岡		3	
20	東京都	あきる野市 菅生		3	
21	東京都	あきる野市 深沢		3	
22	東京都	あきる野市 深沢		3	
23	東京都	あきる野市 上代継		3	
24	東京都	あきる野市 上代継		3	
25	東京都	あきる野市 上ノ台		3	
26	東京都	あきる野市 上ノ台		3	
27	東京都	あきる野市 小和田		3	
28	東京都	あきる野市 小和田		3	
29	東京都	あきる野市 小峰台		3	
30	東京都	あきる野市 小峰台		3	
31	東京都	あきる野市 小中野		3	
32	東京都	あきる野市 小中野		3	
33	東京都	あきる野市 小川東3丁		3	
34	東京都	あきる野市 小川東3丁		3	
35	東京都	あきる野市 小川東2丁		3	
36	東京都	あきる野市 小川東2丁		3	
37	東京都	あきる野市 小川東1丁		3	
38	東京都	あきる野市 小川東1丁		3	
39	東京都	あきる野市 小川		3	
40	東京都	あきる野市 小川		3	

東京電力計画停電の例から (3)

誰かがアプリケーションにして公開

2011年03月14日 02時19分31秒

計画停電のエリアを検索できる「停電エリア検索システム」



停電エリア検索システム

■都道府県選択

東京都 ▼

市・町名:

住所を入力してください。

例) 目黒区八雲

例) 荒川

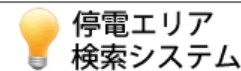
検索時には、「〇丁目」より前の住所を入力してください。

都道府県名を選んで自分の住んでいる市・町名を入力するだけで第1グループから第5グループのどれ
電エリア検索システム」というものが公開されました。

使い方などの詳細は以下から。

停電エリア時間検索(東京電力エリア)

<http://machi.userlocal.jp/teiden/>



停電エリア検索システム

東京電力が公開している最新データに更新しました(3/15 19:45)
PC/携帯/iPhone Android対応

■都道府県を選択

東京都 ▼

市・町名:

住所を入力すると、輪番停電のグループ番号がわかります。

例) 目黒区八雲

例) 荒川

「〇丁目」より前の住所(町名まで)を入力してください。

■計画停電情報

2011年3月16日 水

- ・第1グループ 12:20~16:00
- ・第2グループ 15:20~19:00
- ・第3グループ 18:20~22:00
- ・第4グループ 06:20~10:00 *
- ・第5グループ 09:20~13:00 *

2011年3月17日 木

- ・第1グループ 09:20~13:00 *
- ・第2グループ 12:20~16:00
- ・第3グループ 15:20~19:00
- ・第4グループ 18:20~22:00
- ・第5グループ 06:20~10:00 *

2011年3月18日 金

- ・第1グループ 06:20~10:00 *
- ・第2グループ 09:20~13:00 *
- ・第3グループ 12:20~16:00
- ・第4グループ 15:20~19:00
- ・第5グループ 18:20~22:00

*がついているグループについては、供給不足が懸念される場合に追加停止の可

※最新の発表情報を反映するようにしますが、正確かつ最新の情報は詳しく
※サイトや記事などで紹介していただいて構いません。事前の承諾は不要です。

[English Version]

ネ申Excel問題

- 人間のための構造化と計算機処理のための構造化のアプローチが異なることにより発生 (?)
 - A4一枚に収めるための努力が裏目に...
- 人間による処理: 見た目による構造化
 - 罫線とセルをまたいだ情報の入れ子
- 計算機処理: 表形式による構造化
 - 行・列とセルによる区分
- 計算機可読フォーマットだが、計算機理解は困難になってしまう

第1表 給与所得者数・給与額・税額

事業所規模別・企業規模別・業種別

区分	給与所得者数					給与額		税額	
	3月末	6月末	9月末	12月末	年間月平均	総額	平均	総額	平均
	千人	千人	千人	千人	千人	百万円	千円	百万円	千円
平成19年分	54,957	55,311	54,417	53,768	54,817	201,272,200	3,672	8,996,535	16
20	54,672	55,269	55,000	54,739	55,124	201,317,695	3,652	8,627,692	15
21	56,232	54,909	54,035	53,884	54,967	192,474,249	3,502	7,570,580	13
22	55,817	54,887	54,367	54,153	54,792	194,372,236	3,547	7,500,850	13
23	55,169	54,688	54,459	54,273	54,647	195,799,665	3,583	8,077,950	14
24	54,432	54,466	53,948	54,221	54,267	191,099,637	3,521	7,824,009	14
	人	人	人	人	人	百万円	千円	百万円	千円
10人未満	9,713,060	9,647,598	9,618,379	9,517,123	9,624,150	27,972,317	2,906	1,115,344	11
10人以上	7,898,928	7,719,572	7,488,610	7,596,998	7,676,023	23,950,594	3,120	1,064,580	13
30人以上	8,624,553	8,576,484	8,464,956	8,503,867	8,542,472	27,532,669	3,223	998,791	11
100人 "	11,634,920	11,756,641	11,662,497	11,703,182	11,689,332	42,805,621	3,662	1,637,365	14
500人 "	4,067,066	4,130,385	4,102,422	4,134,997	4,108,716	16,013,459	3,897	676,802	16
1,000人 "	7,335,269	7,422,011	7,402,438	7,477,611	7,409,312	30,533,284	4,121	1,311,016	17
5,000人 "	5,158,040	5,213,157	5,209,041	5,286,843	5,216,770	22,291,693	4,273	1,020,112	19
計									
合計									
個人	3,311,796	3,356,874	3,363,342	3,406,038	3,359,536	7,130,930	2,123	196,689	5

国税庁. 平成24年分民間給与実態調査統計. 2014.

<http://www.nta.go.jp/kohyo/tokei/kokuzeicho/minkan2012/minkan.htm>

Linked Open Data (LOD)

- オープンデータ + 構造化データ
- オープンでリンクできるデータ
 - “ドキュメントとしてのWeb” → “**データのWeb**”
- Linked Dataの4原則
 - 事物をURIを使って名前付ける
 - 事物をHTTP + URIで参照する
 - URIを参照したときに関連情報を表示する
 - 外部のリソースへのリンクも含める

LODへの5つのステップ



Semantic Webへのアプローチ



TED



Tim Berners-Lee:

The year open data went worldwide

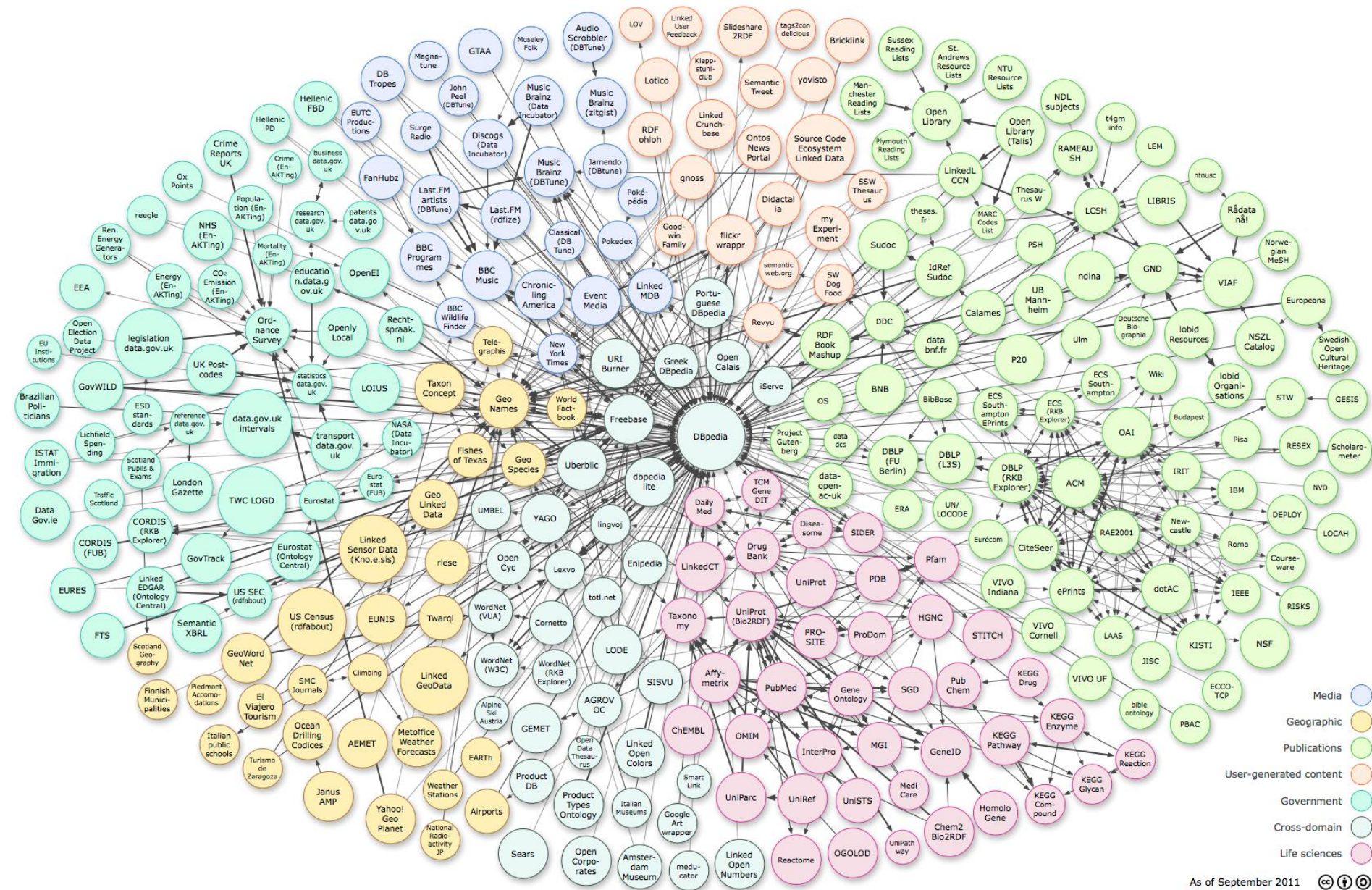
TED2010 · 5:33 · Filmed Feb 2010

Subtitles available in 32 languages



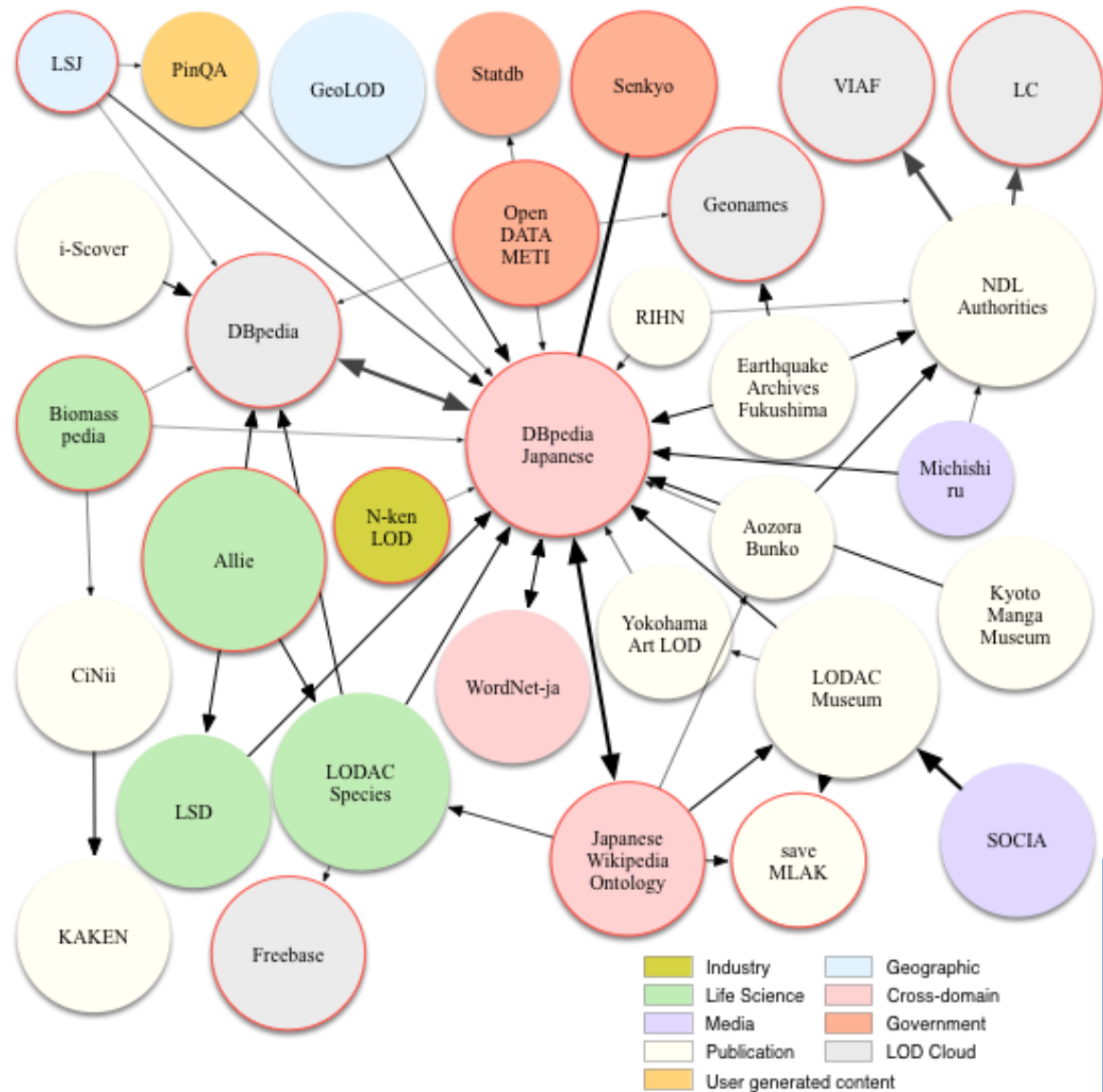
View interactive transcript

http://www.ted.com/talks/tim_berniers_lee_the_year_open_data_went_worldwide



Richard Cyganiak, Anja Jentzsch. "Linking Open Data cloud diagram". <http://lod-cloud.net/> (Last updated: 2011-09-19)

日本におけるLODデータセット



Linked Open Data Initiative. 日本語版Linked Dataクラウド図.
<http://linkedopendata.jp/?p=486>
(Last updated 2014-03-10)

本名	小野寺 章太郎 (おのでら しょうたろう)
生誕	1938年1月25日 ● 日本・宮城県登米郡石森町 (現・宮城県登米市)
死没	1998年1月28日(満60歳没) ● 日本・東京都
国籍	● 日本
職業	漫画家 特撮原作者
活動期間	1954年 - 1998年
代表作	『サイボーグ009』 『仮面ライダー』 『人造人間キカイダー』 『さるとびエッチちゃん』 『マンガ日本経済入門』 『HOTEL』
受賞	第7回講談社児童まんが賞 (『サイボーグ009』ミュータント・サブ) 第13回小学館漫画賞 (『ジュン』佐武と市捕物控) 第33回小学館漫画賞 (『HOTEL』『マンガ日本経済入門』) 第17回日本漫画家協会賞大賞 (『マンガ日本経済入門』) 日本漫画家協会賞文部大臣賞

About: 石ノ森章太郎

An Entity of Type : 漫画家, from Named Graph : http://ja.dbpedia.org, within Data Space : ja.dbpedia.org

石ノ森 章太郎(いしのもりしょうたろう、1938年〈昭和13年〉 1月25日 - 1998年〈平成10年〉 1月28日)は、日本の漫画家、特(おのでらしょうたろう)。1984年までは石森 章太郎の表記を用いたが、この時期には「いしもり〜」と呼ばれるのが通例だった。

Data Table	
Property	Value
dbpedia-owl:abstract	石ノ森 章太郎(いしのもりしょうたろう、1938年 〈昭和13年〉 1月25日 - 1998年 〈平成10者。本名、小野寺章太郎(おのでらしょうたろう)。1984年までは石森 章太郎の表記を用いたが通例だった。代表作は『サイボーグ009』、『ロボット 刑事』、『さるとびエッチちゃん』、『マンシリーズを始め特撮作品の原作者としても活躍。SF漫画から学習漫画まで幅広い分野にされた。1989年、漫画には「面白い、おかしい」だけではなく多数の表現が可能になった画宣言」を発表。以降は自らの職業を「漫画家」ではなく「萬画家」と称した。没後の2007森章太郎萬画大全集』(角川書店)が、一人の著者による最も多い漫画の出版の記録となお、テレビ原作者(アニメ、実写)としてのクレジットは放映期間のベ六十数年分に及びある。手塚治虫の衣鉢を継ぐSFマンガの第一人者であったが、熱心なSFの読者として、ことも指摘されている。息子に俳優の小野寺丈(長男)と石森プロ社長の小野寺章(次男)はAB型。
dbpedia-owl:award	<div><div></div><div>dbpedia-ja:手塚治虫文化賞</div><div></div></div> <div><div></div><div>dbpedia-ja:小学館漫画賞</div><div></div></div> <div><div></div><div>dbpedia-ja:日本漫画家協会賞</div><div></div></div> <div><div></div><div>dbpedia-ja:勲四等</div><div></div></div> <div><div></div><div>dbpedia-ja:映画特別功労賞</div><div></div></div>
dbpedia-owl:birthPlace	<div><div></div><div>dbpedia-ja:登米市</div><div></div></div> <div><div></div><div>dbpedia-ja:宮城県</div><div></div></div>
dbpedia-owl:birthYear	<div><div></div><div>0025-01-01 00:00:00 (xsd:date)</div><div></div></div>
dbpedia-owl:nationality	<div><div></div><div>dbpedia-ja:Japan</div><div></div></div>
dbpedia-owl:notableWork	<div><div></div><div>dbpedia-ja:人造人間キカイダー</div><div></div></div> <div><div></div><div>dbpedia-ja:仮面ライダーシリーズ</div><div></div></div> <div><div></div><div>dbpedia-ja:HOTEL</div><div></div></div>

prop-ja:代表作

『HOTEL』
『仮面ライダー』
『サイボーグ009』
『人造人間キカイダー』
『さるとびエッチちゃん』
『マンガ日本経済入門』

prop-ja:公式サイト

http://www.ishimoripro.com/

dbpedia-ja:手塚治虫文化賞
dbpedia-ja:映画特別功労賞
勲四等旭日小綬章
第33回小学館漫画賞
第 7回講談社児童まんが賞
《『サイボーグ009』『ミュータント・サブ』
第13回小学館漫画賞
《『ジュン』『佐武と市捕物控』
《『HOTEL』『マンガ日本経済入門』
第17回日本漫画家協会賞大賞
《『マンガ日本経済入門』
日本漫画家協会賞文部大臣賞

prop-ja:名前

石ノ森 章太郎

prop-ja:本名
小野寺 章太郎
《おのでらしょうたろう》

prop-ja:没地

・東京都

prop-ja:職業

dbpedia-ja:特撮

デジタルドキュメントの視点から

- ワークフローを考え直す手がかりに
 1. ドキュメントの作成
 2. 流通、利用、保存
 3. 他者への頒布
 - 同僚、知人
 - 同業者
 - 市民一般
 4. 再利用
 - Webへ
 - 紙へ
 - 別用途へ
- (振り出しに戻る)

まとめ

- セマンティックウェブとデジタルドキュメント
 - Semantic Webの基盤技術
 - RDFデータモデル
 - オープンデータとメタデータ
 - オープンデータとライセンス
 - Linked Open Data (LOD)
 - デジタルドキュメントのワークフロー