デジタルドキュメント（6）

高久雅生

2015年5月21日（木）3・4時限

1

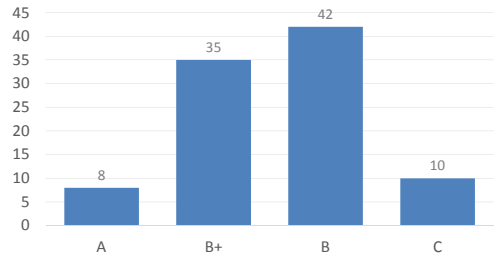
本日のお品書き

- （第2回レポートの返却、講評）
- （前回の復習）
- マークアップ言語とデジタルドキュメント
  - ✓メタ言語SGMLとXML
    - 整形式
    - メタ言語とスキーマ
  - ✓様々な応用
- 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

2

第2回レポート：成績分布

- 第1回レポートの評定に比べると、だいぶ良くなってきた。



3

第2回レポート：内訳

- 取り上げられた論文の掲載誌（62タイトル中）
  - ✓情報処理学会研究報告 16
  - ✓情報管理 5
  - ✓ACM SIGRAPH 4
  - ✓情報処理学会論文誌 4
  - ✓情報の科学と技術 3
  - ✓Nature 3
  - ✓Science 3
  - ✓図書館界 2
- 掲載プラットフォーム（22プラットフォーム）
  - ✓J-STAGE 21
  - ✓CiNii 19
  - ✓情報処理学会電子図書館 15
  - ✓機関リポジトリ 14
  - ✓Nature 4
  - ✓ACM Digital Library 3
  - ✓Science 3
  - ✓IEEE Xplore Digital Library 2

4

第2回レポート：講評

- 採点基準に示したとおり、該当論文の提供形態や構成要素、プラットフォームの特徴等を分かりやすく説明したものに加点
- 一部のプラットフォームは複合的なワークフローを用いている点に注意
  - ✓J-STAGEにおける電子化の担い手
  - ✓CiNii・機関リポジトリ・J-STAGE等における論文書誌情報（メタデータ）の情報流通など

5

（前回の復習 = ふりかえり）

- 電子書籍
  - ✓事例とともに：利用と閲覧環境、コンテンツ
  - ✓今後の課題
- デジタルドキュメントとしてのウェブ
  - ✓WWW（World Wide Web）
  - ✓基本設計とその成立
  - ✓ハイパーテキストシステム
  - ✓ウェブにおける応用

6

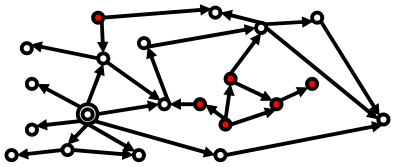
ワールド・ワイド・ウェブ

WORLD WIDE WEB (2)

7

WebとHTML

- HTTP, URI, HTMLの3つの枠組みを用いた**Web**の実現！
- HTTPレイヤでのデータ転送とドキュメントフォーマット指定
- URI指定によるリンクを通じたハイパーメディアの実現



8

HTML (Hypertext Markup Language)

- ウェブ上でのコンテンツ記述用言語
- テキストデータのやり取り
- ウェブブラウザによる解釈と描画
- タグ；メタデータ；構造
- ハイパーメディア；ハイパーテキスト
  - ✓リンク
  - ✓ページ埋め込み
- マルチメディア（画像、音声、動画）
- ダイナミックな表現；インタラクション

```
<html>
<head>
<title>タイトル</title>
</head>
<body>
<h1>見出し</h1>
<p>段落</p>
</body>
</html>
```

9

メタ言語としてのXML

10

メタ言語SGMLとXML

- HTML
  - ✓文書内に出現するタグ <a>, <p>, <title>, <h1> は、あらかじめ定められている（限定がある）
  - ✓どのような要素が出現するかを定めるメタ言語が必要
- SGML (Standard Generalized Markup Language)
  - ✓1986年に国際標準として制定（ISO）
  - ✓電子文書のパイオニア
  - ✓メタ言語：DTD (Document Type Definition)
- XML (Extensible Markup Language)
  - ✓Webでの経験を元に、SGMLをより扱いやすくし、汎用言語として整備

11

XML (Extensible Markup Language)

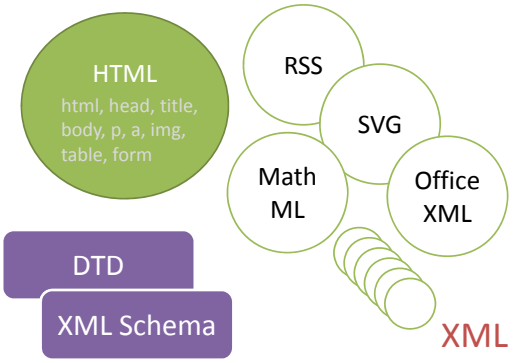
- 順序付き木構造による表現
- 整形式（Well-formedness）
- 文字コード：Unicode
- スキーマ（文書内容の定義）に則った、個別のインスタンス言語
  - ✓XHTML
  - ✓SVG（図版）, MathML（数式）
  - ✓その他の文書形式
- 参照関係
- 元情報としてのXML文書
  - ✓変換することにより、他のデータ形式へ
  - ✓One source multiple use…

12

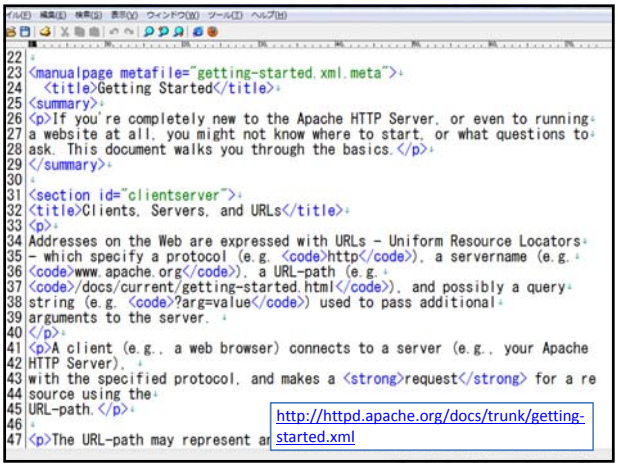
XMLにおける改善

- 文字コード指定の枠組み
  - ✓XMLの場合：ファイル冒頭での文字コード指定
    - <?xml version="1.0" encoding="utf-8"?>
- 整形式: 終了タグの省略を許さないなど
  - ✓HTMLの場合：
    - 第1段落…
    - 第2段落…
  - ✓XMLの場合：
    - 第1段落…
    - 第2段落…
  - ✓※XML（メタ言語）vs HTML（インスタンス言語）という立場の違いが厳然としてあることに注意

13

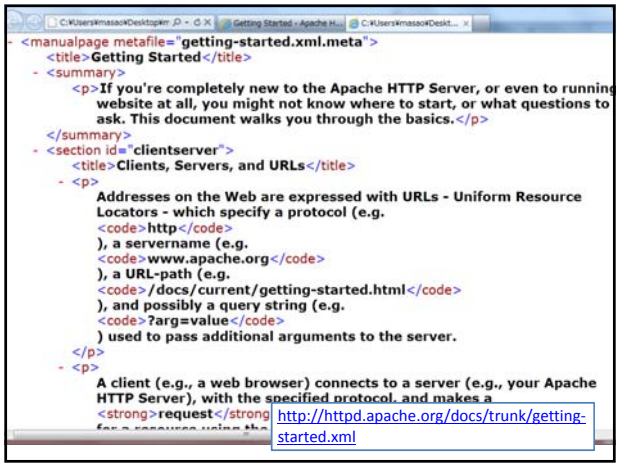


14

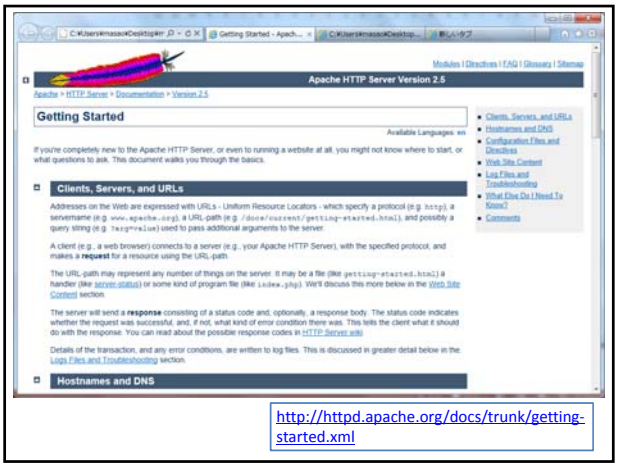


```
<manualpage metafile="getting-started.xml.meta">
  <title>Getting Started</title>
  <summary>
    <p>If you're completely new to the Apache HTTP Server, or even to running a website at all, you might not know where to start, or what questions to ask. This document walks you through the basics.</p>
  </summary>
  <section id="clientserver">
    <title>Clients, Servers, and URLs</title>
    <p>
      Addresses on the Web are expressed with URLs - Uniform Resource Locators - which specify a protocol (e.g. <code>http</code>), a servername (e.g. <code>www.apache.org</code>), a URL-path (e.g. <code>/docs/current/getting-started.html</code>), and possibly a query string (e.g. <code>?arg=value</code>) used to pass additional arguments to the server.
    </p>
    <p>
      A client (e.g., a web browser) connects to a server (e.g., your Apache HTTP Server), with the specified protocol, and makes a <strong>request</strong> for a resource using the URL-path.
    </p>
    <p>
      The URL-path may represent any number of things on the server. It may be a file (like getting-started.html) a handler (like server-status) or some kind of program file (like index.php). We'll discuss this more below in the Web Site Content section.
    </p>
    <p>
      The server will send a <strong>response</strong> consisting of a status code and, optionally, a response body. The status code indicates whether the request was successful, and, if not, what kind of error condition there was. This tells the client what it should do with the response. You can read about the possible response codes in HTTP Server wiki.
    </p>
    <p>
      Details of the transaction, and any error conditions, are written to log files. This is discussed in greater detail below in the Log Files and Troubleshooting section.
    </p>
  </section>
</manualpage>
```

http://httpd.apache.org/docs/trunk/getting-started.xml



http://httpd.apache.org/docs/trunk/getting-started.xml



http://httpd.apache.org/docs/trunk/getting-started.xml

（参考）検証サービス

http://validator.w3.org/

What is Semantic Web?

セマンティック・ウェブの世界

19

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

21

Semantic Webアプリケーション (1)

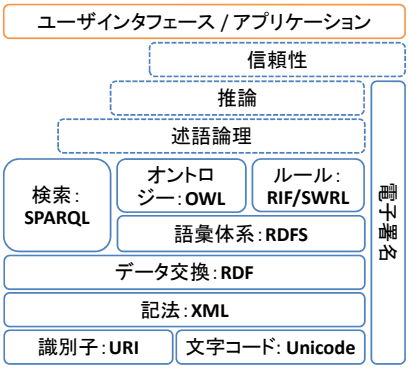
- 例:「仕事帰りに寄れる歯医者さんを調べたい」
  - ✓仕事: 平日9:00-18:00
  - ✓仕事帰りに寄る:つくばエクスプレス(TX)沿線
    - 18時以降に診察可能
    - TX沿線の駅名:つくば, 研究学園, ..., 南流山, 北千住, ...
    - 駅から徒歩 500m 圏内

22

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

Semantic Webの構成要素



24

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えない
  - ✓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

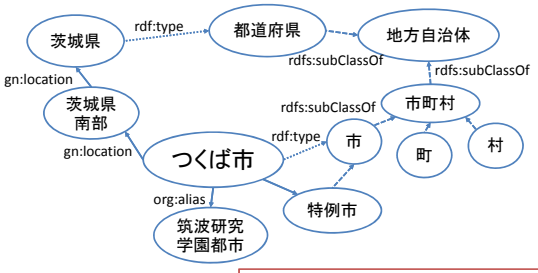
25

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

26

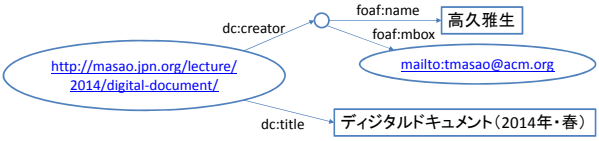
RDFグラフモデルの一例



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

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「ディ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：



28

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 “ディ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>
        ] .
```

29

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>ディ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>
```

30

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)（ー雑誌情報）
  - ✓FOAF (The Friend of a Friend)（ー著者情報）

31

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

32

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

33

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

34

（余談）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

35

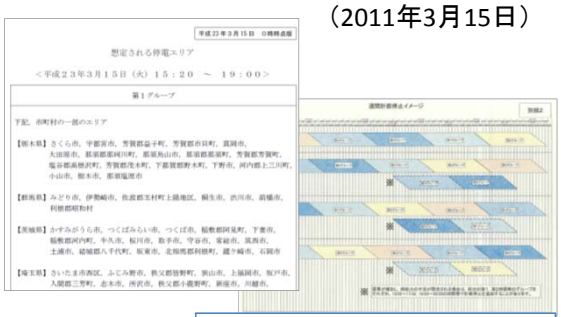
オープンデータとデジタルドキュメント

- 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

36

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）



http://www.tepco.co.jp/cc/press/11031414-j.html

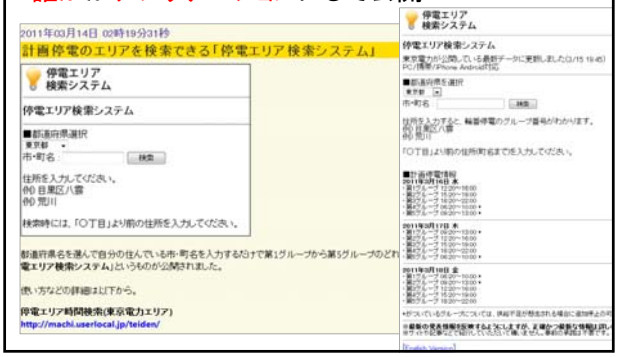
東京電力計画停電の例から (2)

誰かがCSV形式に変換して公開

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

東京電力計画停電の例から (3)

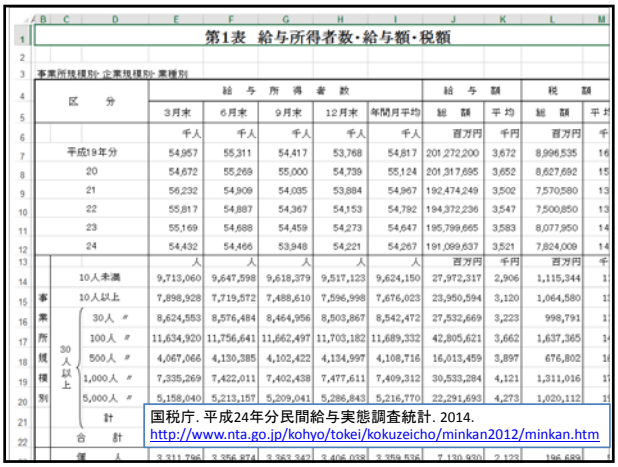
誰かがアプリケーションにして公開



ネ申Excel問題

- ※データとして扱うことを前提とせず、紙に印刷することを前提としたExcelフォーマットにおける表現の総称。
  - ✓（紙 → 神 → ネ申
- 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、計算機理解は困難となってしまう

40



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 3月末	6月末	9月末	12月末	年間月平均	給与額 総額	平均	税額 総額	平均
	千人	千人	千人	千人	千人	百万円	千円	百万円	千
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
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	1
事業所規模別 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	1
30人以上 30人〃	8,624,553	8,576,484	8,464,956	8,503,867	8,542,472	27,532,669	3,223	998,791	1
30人以上 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	1
30人以上 500人〃	4,067,066	4,130,385	4,102,422	4,134,997	4,108,716	16,013,459	3,897	676,802	1
30人以上 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	1
30人以上 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	1
30人以上 計									
合計									

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)

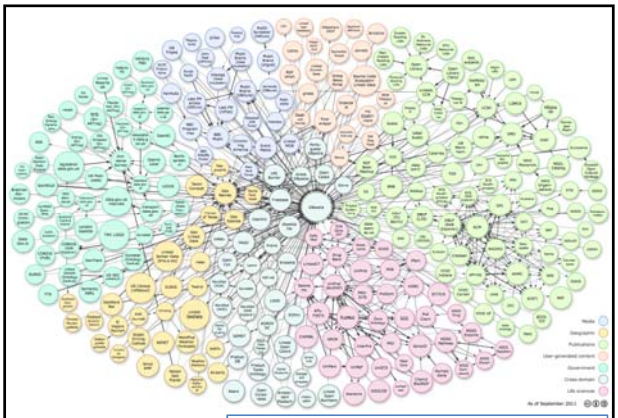
- オープンデータ + 構造化データ
- オープンでリンクできるデータ
  - ✓“ドキュメントとしてのWeb” → “データのWeb”
- Linked Dataの4原則
  - ✓事物をURIを使って名前付ける
  - ✓事物をHTTP + URIで参照する
  - ✓URIを参照したときに関連情報を表示する
  - ✓外部のリソースへのリンクも含める

42

LODへの5つのステップ

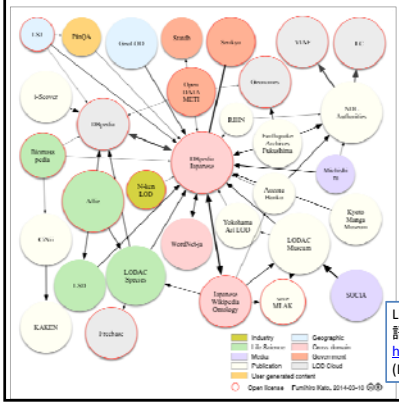
Semantic Webへのアプローチ

http://www.ted.com/talks/tim_berners_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)

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）

49

まとめ

- マークアップ言語とデジタルドキュメント
  - ✓メタ言語SGMLとXML
    - 整形式
    - メタ言語とスキーマ
  - ✓様々な応用
- 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
- （次回はファイル形式の役割について話します）

50