# When Linked Data Rules Fail

**by Mike Bergman - Monday, November 16, 2009**

http://www.mkbergman.com/846/when-linked-data-rules-fail/

## High Visibility Problems with NYT, data.gov Show Need for Better Practices

When I say, "shot", what do you think of? A flu shot? A shot of whisky? A moon shot? A gun shot? What if I add the term "bank"? Do you now think of someone being shot in an armed robbery of a local bank or similar?

And, now, what if I add a reference to say, The Hustler, or Minnesota Fats, or "Fast Eddie" Felson? Do you now see the connection to a pressure-packed banked pool shot in some smoky bar room?

As humans we need context to make connections and remove ambiguity. For machines, with their limited reasoning and inference engines, context and accurate connections are even more important.

Over the past few weeks we have seen announcements of two large and high-visibility linked data projects: One, a first release of references for articles concerning about 5,000 people from the New York Times at data.nytimes.com; and Two, a massive exposure of 5 billion triples from data.gov datasets provided by the Tetherless World Constellation (TWC) at Rennselaer Polytechnic Institute (RPI).

On various grounds from licensing to data characterization and to creating linked data for its own sake, some prominent commentators have weighed in on what is good and what is not so good with these datasets. One of us, Mike, commented about a week ago that "we have now moved beyond 'proof of concept' to the need for actual useful data of trustworthy provenance and proper mapping and characterization. Recent efforts are a disappointment that no enterprise would or could rely upon."

Reactions to that posting and continued discussion on various mailing lists warrant a more precise dissection of what is wrong and still needs to be done with these datasets [1].

## Berners-Lee's Four Linked Data "Rules"

It is useful, then, to return to first principles, namely the original four "rules" posed by Tim Berners-Lee in his design note on linked data [2]:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs so that they can discover more things.

The first two rules are definitional to the idea of linked data. They cement the basis of linked data in the Web, and are not at issue with either of the two linked data projects that are the subject of this posting.

However, it is the lack of specifics and guidance in the last two rules where the breakdowns occur. Both the NYT and the RPI datasets suffer from a lack of "providing useful information" (Rule #3). And, the nature of the links in Rule #4 is a real problem for the NYT dataset.

## What Constitutes "Useful Information"?

The Wikipedia entry on linked data expands on "useful information" by augmenting the original rule with the parenthetical clause, " (*i.e.*, a structured description — metadata)." But even that expansion is insufficient.

Fundamentally, what are we talking about with linked data? Well, we are talking about instances that are characterized by one or more attributes. Those instances exist within contexts of various natures. And, those contexts may relate to other existing contexts.

We can break this problem description down into three parts:

- A ***vocabulary*** that defines the nature of the instances and their descriptive attributes
- A ***schema*** of some nature that describes the structural relationships amongst instances and their characteristics, and, optimally,
- A ***mapping*** to existing external schema or constructs that help place the data into context.

At minimum, ANY dataset exposed as linked data needs to be described by a ***vocabulary***. Both the NYT and RPI datasets fail on this score, as we elaborate below. Better practice is to also provide a ***schema*** of relationships in which to embed each instance record. And, best practice is to also ***map*** those structures to external schema.

Lacking this "useful information", especially a defining vocabulary, we cannot begin to understand whether our instances deal with drinks, bank robberies or pool shots. This lack, in essence, makes the information worthless, even though available via URL.

### The data.gov (RPI) Case

With the support of NSF and various grant funding, RPI has set up the Data-Gov Wiki [3], which is in the process of converting the datasets on data.gov to RDF, placing them into a semantic wiki to enable comment and annotation, and providing that data as RSS feeds. Other demos are also being placed on the site.

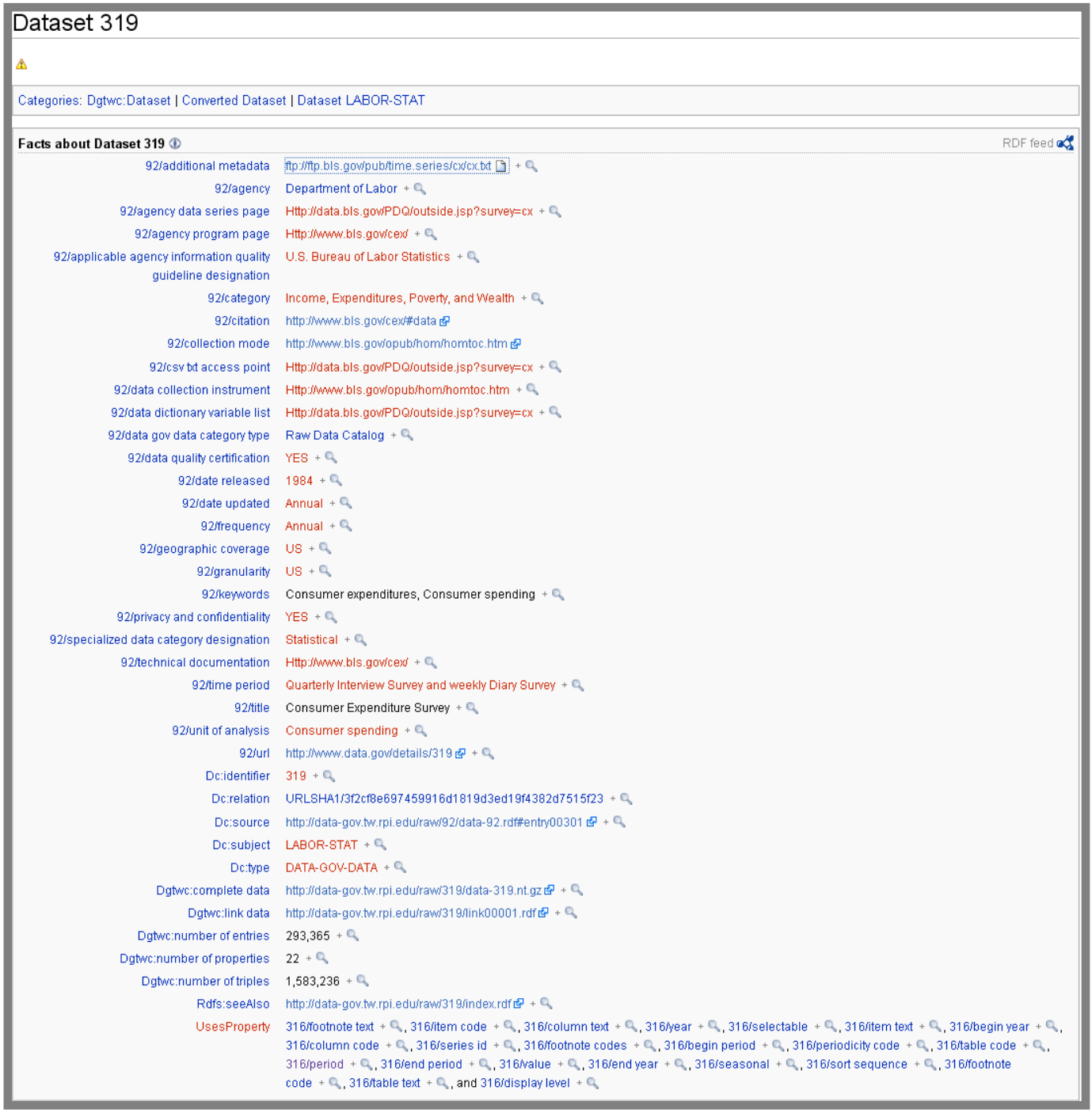
As of the date of this posting, the site had a catalog of 116 datasets from the 800 or so available on data.gov, leading to these statistics:

- 459,412,419 table entries
- 5,074,932,510 triples, and
- 7,564 properties (or attributes).

We'll take one of these datasets, #319, and look a bit closer at it:

| Wiki | Title | Agency | Name | data.gov Link | No Properties | No Triples | RDF File |
|---|---|---|---|---|---|---|---|
| Dataset 319 | Consumer Expenditure Survey | Department of Labor | LABOR-STAT | http://www.data.gov/details/319 | 22 | 1,583,236 | http://data-gov.tw.rpi.edu/raw/319/index.rdf |

This report was picked solely because it had a small number of attributes (properties), and is thus easier to screen capture. The summary report on the wiki is shown by this page:



Dataset 319

Categories: Dgtwc:Dataset | Converted Dataset | Dataset LABOR-STAT

**Facts about Dataset 319** RDF feed

| | |
|---|---|
| 92/additional metadata | ftp://ftp.bls.gov/pub/time.series/cx/cx.txt |
| 92/agency | Department of Labor |
| 92/agency data series page | Http://data.bls.gov/PDQ/outside.jsp?survey=cx |
| 92/agency program page | Http://www.bls.gov/cex/ |
| 92/applicable agency information quality guideline designation | U.S. Bureau of Labor Statistics |
| 92/category | Income, Expenditures, Poverty, and Wealth |
| 92/citation | http://www.bls.gov/cex/#data |
| 92/collection mode | http://www.bls.gov/opub/hom/homtoc.htm |
| 92/csv txt access point | Http://data.bls.gov/PDQ/outside.jsp?survey=cx |
| 92/data collection instrument | Http://www.bls.gov/opub/hom/homtoc.htm |
| 92/data dictionary variable list | Http://data.bls.gov/PDQ/outside.jsp?survey=cx |
| 92/data gov data category type | Raw Data Catalog |
| 92/data quality certification | YES |
| 92/date released | 1984 |
| 92/date updated | Annual |
| 92/frequency | Annual |
| 92/geographic coverage | US |
| 92/granularity | US |
| 92/keywords | Consumer expenditures, Consumer spending |
| 92/privacy and confidentiality | YES |
| 92/specialized data category designation | Statistical |
| 92/technical documentation | Http://www.bls.gov/cex/ |
| 92/time period | Quarterly Interview Survey and weekly Diary Survey |
| 92/title | Consumer Expenditure Survey |
| 92/unit of analysis | Consumer spending |
| 92/url | http://www.data.gov/details/319 |
| Dc:identifier | 319 |
| Dc:relation | URLSHA1/3f2cf8e697459916d1819d3ed19f4382d7515f23 |
| Dc:source | http://data-gov.tw.rpi.edu/raw/92/data-92.rdf#entry00301 |
| Dc:subject | LABOR-STAT |
| Dc:type | DATA-GOV-DATA |
| Dgtwc:complete data | http://data-gov.tw.rpi.edu/raw/319/data-319.nt.gz |
| Dgtwc:link data | http://data-gov.tw.rpi.edu/raw/319/link00001.rdf |
| Dgtwc:number of entries | 293,365 |
| Dgtwc:number of properties | 22 |
| Dgtwc:number of triples | 1,583,236 |
| Rdfs:seeAlso | http://data-gov.tw.rpi.edu/raw/319/index.rdf |
| UsesProperty | 316/footnote text, 316/item code, 316/column text, 316/year, 316/selectable, 316/item text, 316/begin year, 316/column code, 316/series id, 316/footnote codes, 316/begin period, 316/periodicity code, 316/table code, 316/period, 316/end period, 316/value, 316/end year, 316/seasonal, 316/sort sequence, 316/footnote code, 316/table text, and 316/display level |

*(click to expand)*

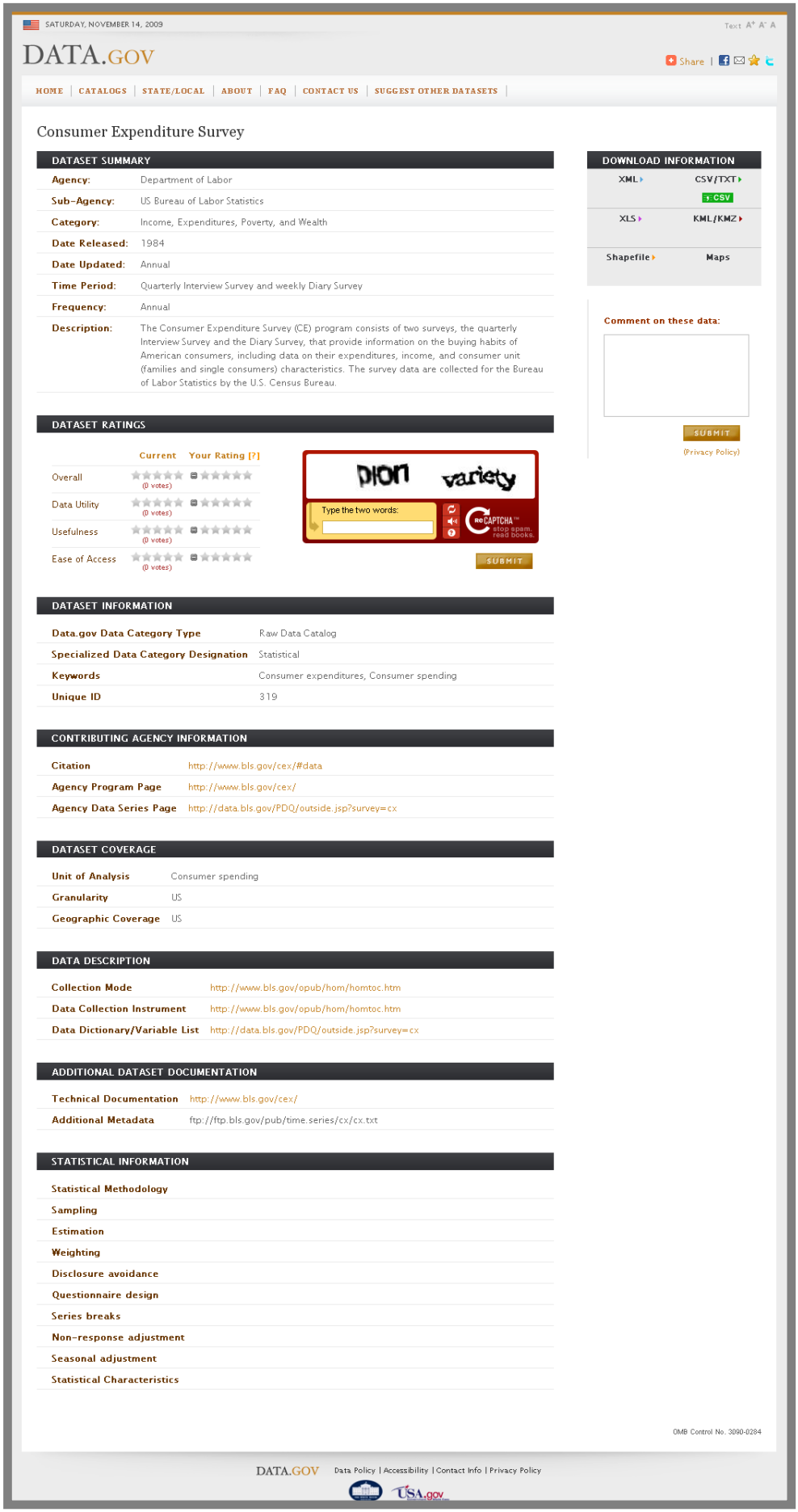
So, we see that this specific dataset contains about 22 of the nearly 8,000 attributes across all datasets.

When we click on one of these attribute names, we are then taken to a specific wiki page that only reiterates its label. There is no definition or explanation.

When we inspect this page further we see that, other than the broad characterization of the dataset itself (the bulk of the page), we see at the bottom 22 undefined attributes with labels such as *item code*, *periodicity code*, *seasonal*, and the like. These attributes are the real structural basis for the data in this dataset.

But, what does all of this mean???

To gain a clue, now let's go to the source data.gov site for this dataset (#319). Here is how that report looks:



SATURDAY, NOVEMBER 14, 2009

Text A+ A- A

DATA.GOV

Share | | | | |

HOME | CATALOGS | STATE/LOCAL | ABOUT | FAQ | CONTACT US | SUGGEST OTHER DATASETS |

Consumer Expenditure Survey

**DATASET SUMMARY**

| | |
|---|---|
| **Agency:** | Department of Labor |
| **Sub-Agency:** | US Bureau of Labor Statistics |
| **Category:** | Income, Expenditures, Poverty, and Wealth |
| **Date Released:** | 1984 |
| **Date Updated:** | Annual |
| **Time Period:** | Quarterly Interview Survey and weekly Diary Survey |
| **Frequency:** | Annual |
| **Description:** | The Consumer Expenditure Survey (CE) program consists of two surveys, the quarterly Interview Survey and the Diary Survey, that provide information on the buying habits of American consumers, including data on their expenditures, income, and consumer unit (families and single consumers) characteristics. The survey data are collected for the Bureau of Labor Statistics by the U.S. Census Bureau. |

**DOWNLOAD INFORMATION**

XML | CSV/TXT | CSV | XLS | KML/KMZ | Shapefile | Maps

**Comment on these data:**

SUBMIT

(Privacy Policy)

**DATASET RATINGS**

| | Current | Your Rating [?] |
|---|---|---|
| Overall | (0 votes) | |
| Data Utility | (0 votes) | |
| Usefulness | (0 votes) | |
| Ease of Access | (0 votes) | |

Type the two words:

reCAPTCHA stop spam. read books.

SUBMIT

**DATASET INFORMATION**

| | |
|---|---|
| **Data.gov Data Category Type** | Raw Data Catalog |
| **Specialized Data Category Designation** | Statistical |
| **Keywords** | Consumer expenditures, Consumer spending |
| **Unique ID** | 319 |

**CONTRIBUTING AGENCY INFORMATION**

| | |
|---|---|
| **Citation** | http://www.bls.gov/cex/#data |
| **Agency Program Page** | http://www.bls.gov/cex/ |
| **Agency Data Series Page** | http://data.bls.gov/PDQ/outside.jsp?survey=cx |

**DATASET COVERAGE**

| | |
|---|---|
| **Unit of Analysis** | Consumer spending |
| **Granularity** | US |
| **Geographic Coverage** | US |

**DATA DESCRIPTION**

| | |
|---|---|
| **Collection Mode** | http://www.bls.gov/opub/hom/homtoc.htm |
| **Data Collection Instrument** | http://www.bls.gov/opub/hom/homtoc.htm |
| **Data Dictionary/Variable List** | http://data.bls.gov/PDQ/outside.jsp?survey=cx |

**ADDITIONAL DATASET DOCUMENTATION**

| | |
|---|---|
| **Technical Documentation** | http://www.bls.gov/cex/ |
| **Additional Metadata** | ftp://ftp.bls.gov/pub/time.series/cx/cx.txt |

**STATISTICAL INFORMATION**

**Statistical Methodology**
**Sampling**
**Estimation**
**Weighting**
**Disclosure avoidance**
**Questionnaire design**
**Series breaks**
**Non-response adjustment**
**Seasonal adjustment**
**Statistical Characteristics**

OMB Control No. 3090-0284

DATA.GOV Data Policy | Accessibility | Contact Info | Privacy Policy

USA.gov

*(click to expand)*

Contained within this report we see a listing for additional metadata. This link tells us about the various data fields contained in this dataset; we see many of these attributes are "codes" to various data categories.

Probing further into the dataset's technical documentation, we see that there is indeed a rich structure underneath this report, again provided via various code lookups. There are codes for geography, seasonality (adjusted or not), consumer demographic profiles and a variety of consumption categories. (See, for example, the link to this glossary page.) These are the keys to understanding the actual values within this dataset.

For example, one major dimension of the data is captured by the attribute *item_code*. The survey breaks down consumption expenditures within the broad categories of Food, Housing, Apparel and Services, Transportation, Health Care, Entertainment, and Other. Within a category, there is also a rich structural breakdown. For example, expenditures for Bakery Products within Food is given a code of FHC2.

But, nowhere are these codes defined or unlocked in the RDF datasets. This absence is true for virtually all of the datasets exposed on this wiki.

So, for literally billions of triples, and 8,000 attributes, we have **ABSOLUTELY NO INFORMATION ABOUT WHAT THE DATA CONTAINS OTHER THAN A PROPERTY LABEL**. There is much, much rich value here in data.gov, but all of it remains locked up and hidden.

The sad truth about this data release is that it provides absolutely no value in its current form. We lack the keys to unlock the value.

To be sure, early essential spade work has been done here to begin putting in place the conversion infrastructure for moving text files, spreadsheets and the like to an RDF form. This is yeoman work important to ultimate access. But, until a ***vocabulary*** is published that defines the attributes and their codes so we can unlock this value, it will remain hidden. And only when its further value (by connecting attributes and relations across datasets) through a ***schema*** of some nature is also published, the real value from connecting the dots will also remain hidden.

These datasets may meet the partial conditions of providing clickable URLs, but the crucial "useful information" as to what any of this data means is absent.

Every single dataset on data.gov has supporting references to text files, PDFs, Web pages or the like that describe the nature of the data within each dataset. Until that information is exposed and made usable, we have no linked data.

Until ontologies get created from these technical documents, the value of these data instances remain locked up, and no value can be created from having these datasets expressed in RDF.

The devil lies in the details. The essential hard work has not yet begun.

**The NYT Case**

Though at a much smaller scale with many fewer attributes, the NYT dataset suffers from the same failing: it too lacks a ***vocabulary***.

So, let's take the case of one of the lead actors in The Hustler, Paul Newman, who played the role of "Fast Eddie" Felson. Here is the NYT record for the "person" *Paul Newman* (which they also refer to as http://data.nytimes.com/newman_paul_per). Note the header title of **Newman, Paul**:

The New York Times **Developer Network** BETA

**Newman, Paul**

| | |
|---|---|
| nyt:associated_article_count | 48 |
| nyt:first_use | 2001-03-18 |
| nyt:latest_use | 2009-04-05 |
| nyt:number_of_variants | 1 |
| nyt:search_api_query | http://api.nytimes.com/svc/search/v1/article ?query=+nytd_per_facet%3A%5BNewman%2C+Paul%5D &rank=newest &fields=abstract, author, body, byline, classifiers_facet, column_facet, date, day_of_week_facet, des_facet, desk_facet, fee, geo_facet, lead_paragraph, material_type_facet, multimedia, nytd_byline, nytd_des_facet, nytd_geo_facet, nytd_lead_paragraph, nytd_org_facet, nytd_per_facet, nytd_section_facet, nytd_title, nytd_works_mentioned_facet, org_facet, page_facet, per_facet, publication_day, publication_month, publication_year, section_page_facet, small_image_height, small_image_url, small_image_width, source_facet, title, url, word_count, works_mentioned_facet |
| nyt:topicPage | http://topics.nytimes.com/top/reference/timestopics/people/n/paul_newman/index.html |
| rdf:type | http://www.w3.org/2004/02/skos/core#Concept |
| owl:sameAs | http://rdf.freebase.com/ns/en.paul_newman |
| owl:sameAs | http://dbpedia.org/resource/Paul_Newman |
| owl:sameAs | http://data.nytimes.com/newman_paul_per |
| skos:inScheme | http://data.nytimes.com/elements/nytd_per |
| skos:prefLabel | Newman, Paul |

**http://data.nytimes.com/N31738445835662083893.rdf**

| | |
|---|---|
| cc:attributionName | The New York Times Company |
| cc:attributionURL | http://data.nytimes.com/N31738445835662083893 |
| cc:license | http://creativecommons.org/licenses/by/3.0/us/ |
| dc:creator | The New York Times Company |
| dcterms:modified | 2009-11-11 |
| dcterms:rightsHolder | The New York Times Company |
| foaf:primaryTopic | http://data.nytimes.com/N31738445835662083893 |

New York Times Linked Open Data by The New York Times Company is licensed under a Creative Commons Attribution 3.0 United States License.

RDF Copyright 2009 The New York Times Company Terms of Service Privacy Policy Work for Us

*(click to expand)*

Click on any of the internal labels used by the NYT for its own attributes (such as nyt:first_use), and you will be given this message:

*"An RDFS description and English language documentation for the NYT namespace will be provided soon. Thanks for your patience."*

We again have no idea what is meant by all of this data except for the labels used for its attributes. In this case for nyt:first_use we have a value of "2001-03-18".

Hello? What? What is a "first use" for a "Paul Newman" of "2001-03-18"???

The NYT put the cart before the horse: even if minimal, they should have released their ontology first — or at least at the same time — as they released their data instances. (See further this discussion about how an ontology creation workflow can be incremental by starting simple and then upgrading as needed.)

## Links to Other Things

Since there really are no links to other things on the Data-Gov Wiki, our focus in this section continues with the NYT dataset using our same example.

We now are in the territory of the fourth "rule" of linked data: *4. Include links to other URIs so that they can discover more things.*

This will seem a bit basic at first, but before we can talk about linking to other things, we first need to understand and define the starting "thing" to which we are linking.

**What is a "Newman, Paul" Thing?**

Of course, without its own vocabulary, we are left to deduce what this thing "**Newman, Paul**" is that is shown in the previous screen shot. Our first clue comes from the statement that it is of *rdf:type* SKOS *concept*. By looking to the SKOS vocabulary, we see that *concept* is a class and is defined as:

*A SKOS concept can be viewed as an idea or notion; a unit of thought. However, what constitutes a unit of thought is subjective, and this definition is meant to be suggestive, rather than restrictive. The notion of a SKOS concept is useful when describing the conceptual or intellectual structure of a knowledge organization system, and when referring to specific ideas or meanings established within a KOS.*

We also see that this instance is given a foaf:primaryTopic of *Paul Newman*.

So, we can deduce so far that this instance is about the concept or idea of *Paul Newman*. Now, looking to the attributes of this instance — that is the defining properties provided by the NYT — we see the properties of nyt:associated_article_count, nyt:first_use, nyt:last_use and nyt:topicPage. Completing our deductions, and in the absence of its own vocabulary, we can now define this concept instance somewhat as follows:

*New York Times articles in the period 2001 to 2009 having as their primary topic the actor Paul Newman*

(BTW, across all records in this dataset, we could see what the earliest first use was to better deduce the time period over which these articles have been assembled, but that has not been done.)

We also would re-title this instance more akin to "2001-2009 NYT Articles with a Primary Topic of Paul

Newman" or some such and use URIs more akin to this usage.

**sameAs Woes**

Thus, in order to make links or connections with other data, it is essential to understand what the nature is of the subject "thing" at hand. There is much confusion about actual "things" and the references to "things" and what is the nature of a "thing" within the literature and on mailing lists.

Our belief and usage in matters of the semantic Web is that all "things" we deal with are a reference to whatever the "true", actual thing is. The question then becomes:  What is the nature (or scope) of this referent?

There are actually quite easy ways to determine this nature. First, look to one or more instance examples of the "thing" being referred to. In our case above, we have the "**Newman, Paul**" instance record. Then, look to the properties (or attributes) the publisher of that record has used to describe that thing. Again, in the case above, we have nyt:associated_article_count, nyt:first_use, nyt:last_use and nyt:topicPage.

Clearly, this instance record — that is, its nature — deals with articles or groups of articles. The relation to *Paul Newman* occurs as a basis of the primary topic of these articles, and not a person basis for which to describe the instance. If the nature of the instance was indeed the person *Paul Newman*, then the attributes of the record would more properly be related to "person" properties such as age, sex, birthdate, death date, marital status, etc.

This confusion by NYT as to the nature of the "things" they are describing then leads to some very serious errors. By confusing the topic (*Paul Newman*) of a record with the nature of that record (articles about topics), NYT next misuses one of the most powerful semantic Web predicates available, **owl:sameAs**.

By asserting in the "**Newman, Paul**" record that the instance has a **sameAs** relationship with external records in Freebase and DBpedia, the NYT both entails that properties from any of the associated records are shared and infers a chain of other types to describe the record. More precisely, the NYT is asserting that the "thing" referred to by these instances are **identical** resources.

Thus, by the **sameA**s statements in the "**Newman, Paul**" record, the NYT is also asserting that that record is an instance of all these things [5]:

- owl:Thing
- foaf:Agent
- foaf:Person
- dbpedia-owl:Actor
- http://dbpedia.org/class/yago/JewishActors
- http://dbpedia.org/class/yago/PeopleFromCleveland,Ohio
- dbpedia-owl:Artist
- dbpedia-owl:Person
- http://dbpedia.org/class/yago/Person1000078

46
- http://dbpedia.org/class/yago/AmericanFilmDirectors
- http://dbpedia.org/class/yago/YaleUniversityAlumni
- http://dbpedia.org/class/yago/OhioUniversityAlumni
- opencyc:en/MaleHuman
- http://dbpedia.org/class/yago/AmericanFilmActors
- http://dbpedia.org/class/yago/Liberals
- http://dbpedia.org/class/yago/OhioActors
- http://dbpedia.org/class/yago/UnitedStatesNavySailors
- http://dbpedia.org/class/yago/PeopleFromWestport,Connecticut
- opencyc:en/JewishPerson
- opencyc:en/ActorInMovies
- http://dbpedia.org/class/yago/LivingPeople
- http://dbpedia.org/class/yago/Actor109765278
- http://dbpedia.org/class/yago/AmericanVegetarians
- http://dbpedia.org/class/yago/AmericanPhilanthropists
- http://dbpedia.org/class/yago/KenyonCollegeAlumni
- http://dbpedia.org/class/yago/WesternFilmActors
- http://dbpedia.org/class/yago/ActorsStudioAlumni
- and, a hundred other dbpedia_yago superClasses.

Furthermore, because of its strong, reciprocal entailments, the **owl:sameAs** assertion would also now entail that the person *Paul Newman* has the nyt:first_use and nyt:last_use attributes, clearly illogical for a "person" thing.

This connection is clearly wrong in both directions. *Articles* are not *persons* and don't have *marital status*; and *persons* do not have *first_uses*. By misapplying this **sameAs** linkage relationship, we have screwed things up in every which way. And the error began with misunderstanding what kinds of "things" our data is about.

**Some Options**

However, there are solutions. First, the **sameAs** assertions, at least involving these external resources,

should be dropped.

Second, if linkages are still desired, a vocabulary such as UMBEL [4] could be used to make an assertion between such a concept, and these other related resources. So, even though these resources are not the same, they are **closely** related. The UMBEL ontology helps us to define this kind of relation between related, but non-identical, resources.

Instead of using the **owl:sameAs** property, we would suggest the usage of the **umbel:linksEntity**, which links a **skos:Concept** to related named entities resources. Additionally, Freebase, which also currently asserts a **sameAs** relationship to the NYT resource, could use the **umbel:isAbout** relationship to assert that their resource "is about" a certain concept, which is the one defined by the NYT.

Alternatively, still other external vocabularies that more precisely capture the intent of the NYT publishers could be found, or the NYT editors could define their own properties specifically addressing their unique linkage interests.

**Other Minor Issues**

As a couple of additional, minor suggestions for the NYT dataset, we would suggest:

- Create a **foaf:Organization** description of the NYT organization, then use it with **dc:creator** and **dcterms:rightsHolder** rather than using a literal, and
- The dual URIs such as "http://data.nytimes.com/N31738445835662083893" and "http://data.nytimes.com/newman_paul_per" are not wrong in themselves, but the purpose is hard to understand. Why does a single organization need to create multiple resources for the **identical resource,** when it comes from the same system and has the same purpose?

**Re-visiting the Linkage "Rule"**

There are very valuable benefits from entailment, inference and logic to be gained from linking resources. However, if the nature of the "things" being linked — or the properties that define these linkages — are incorrect, then very wrong logical implications result. Great care and understanding should be applied to linkage assertions.

## In the End, the Challenge is Not Linked Data, but *Connected* Data

Our critical comments are not meant to be disrespectful and are not being picky. The NYT and TWC are prominent institutions for which we should expect leadership on these issues. Our criticisms (and we believe those of others) are also not an expression of a "trough of disillusionment" as some have been pointing out.

This posting has been jointly authored by Mike Bergman and Fred Giasson and simultaneously published on both of their blogs, hoping to draw more attention to the need for better practices in publishing linked data.

This posting is about poor practices, pure and simple. The time to correct them is now. If asked, we would be pleased to help either institution establish exemplar practices. This is not automatic, and it is not always easy. The data.gov datasets, in particular, will require much time and effort to get right. There is much documentation that needs to be transitioned and expressed in semantic Web formats.

In a broader sense, we also seem to lack a definition of best practices related to **vocabularies**, **schema** and **mappings**. The Berners-Lee rules are imprecise and insufficient as is. Prior best guidance documents tend to be more how to publish and make URIs linkable, than to properly characterize, describe and connect the data.

Perhaps, in part, this is a bit of a semantics issue. The challenge is not the mechanics of *linking data*, but the meaning and basis for connecting that data. Connections require logic and rationality sufficient to reliably inform inference and rule-based engines. It also needs to pass the sniff test as we "follow our nose" by clicking the links exposed by the data.

It is exciting to see high-quality content such as from national governments and major publishers like the New York Times begin to be exposed as linked data. When this content finally gets embedded into usable contexts, we should see manifest uses and benefits emerge. We hope both institutions take our criticisms in that spirit.

---

[1] The NYT has been updated with improvements and they fixed multiple issues from the first release. The problems listed herein, however, still pertain after these improvements.

[2] Tim Berners-Lee, 2006. Linked Data (Design Issues), first posted on 2006-07-27; last updated on 2009-06-18. See http://www.w3.org/DesignIssues/LinkedData.html. Berners-Lee refers to the steps above as "rules," but he elaborates they are expectations of behavior. Most later citations refer to these as "principles."

[3] Li Ding, Dominic DiFranzo, Sarah Magidson, Deborah L. McGuinness and Jim Hendler, 2009. Data-GovWiki: Towards Linked Government Data. See http://www.cs.vu.nl/~pmika/swc/documents/Data-gov%20Wiki-data-gov-wiki-v1.pdf.

[4] UMBEL *(Upper Mapping and Binding Exchange Layer)* is a lightweight ontology structure in development for relating Web content and data to a standard set of subject concepts. It purpose has resulted in its creation of an associated vocabulary geared to both class-instance and reciprocal relationships, as well as partial or likelihood relationships. See http://umbel.org/technical_documentation.html#vocabulary.

[5] We'd like to thank Denny Vrandecic (see comments) for pointing out an imprecision in our original wording. This phrase was originally stated as, "Thus, by the sameAs statements in the 'Newman, Paul' record, the NYT is also asserting that that record is the same as these other things."

---