Role of Metadata in Data Programs

Joseph A Busch

Agenda

- Metadata and interoperability
- Data management challenges

What is metadata

- Metadata provides enough information for any user, tool, or program to find and use any piece of content.

- Metadata is also used to drive business processes
  - A simple example – publication and expiration dates.

… and business processes may be used to generate metadata

Types of metadata

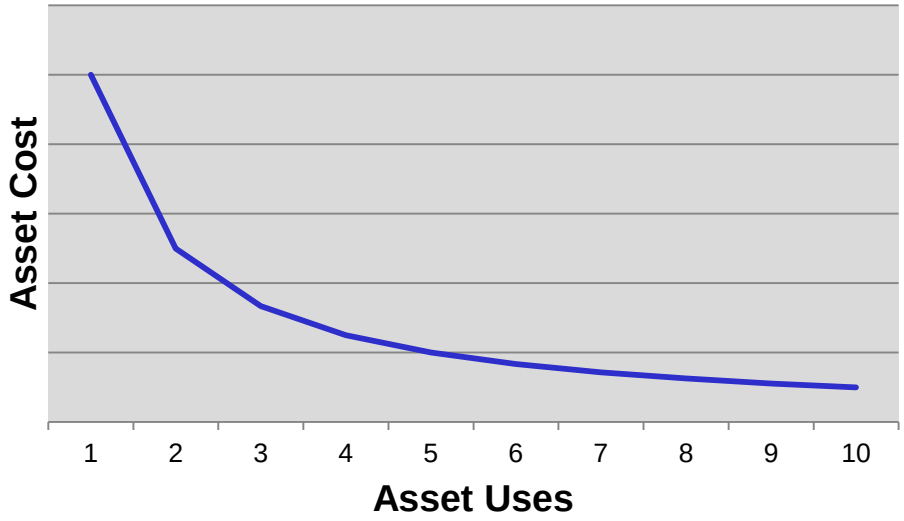
- Asset metadata
  - Identifier, Creator, Owner, Title, Description, Format, Type, Size, Date, etc.
- Subject metadata
  - Names of People, Names of Organizations, Names of Events, Names of Assets, Topics, Purpose, Location, etc.
- Use metadata
  - Audience, Language, Channel, Rights, Role, Expertise, etc.
- Relational metadata
  - Source, Collection, Parts, Related to, etc.

Interoperability

- **The ability of diverse systems and organizations to work together by exchanging information.**
- Semantic interoperability is the ability to automatically interpret the information exchanged meaningfully and accurately.

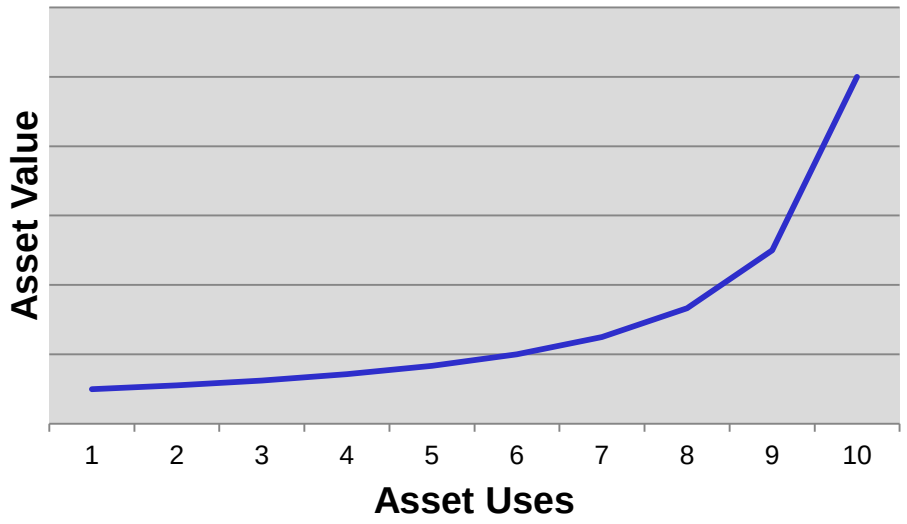
Interoperability ROI

- Content assets are expensive to create so it's critical that they can be found, so they can be used and re-used.
- Every re-use decreases the asset creation cost and increases the content asset value.



Metadata capital*

- Content asset reuse is contingent on the creation and accessibility of complete and consistent metadata.
- Every re-use increases the content asset value.



* "Metadata capital" is a term recently coined by Dr. Jane Greenberg, Director of the Metadata Research Center at the University of North Carolina at Chapel Hill.

Interoperability barriers

- If content assets are so important, why can't they be found?
  - **They are named in different ways.**
  - There is no metadata, or the metadata is incomplete and inconsistent.
  - There is no searchable text (data, graphics, visualizations, etc.)
  - They exist in different applications, file shares and/or desktops.
  - They are not in the search engine index.
  - You do not have proper entitlements to access the content.
  - They have been discarded or lost.
- When they are found why can't content assets be reused?
  - **There are no authoritative sources.**
  - When there are multiple versions, it's difficult to choose which one to use.
  - The source, accuracy and/or authority are unclear.
  - The usage rights may not be clear.

Interoperability vision

- I want to easily find any information assets in a particular format that can be used for a specific purpose regardless of where they are located.
- I want an authoritative source for key named entity* data such as “organization”, “project”, “program”, or “facility”.
- I want to analyze my collection to …
  - Determine strengths and weaknesses.
  - Evaluate types of content, it’s quality, etc.
  - Validate compliance with regulations.
  - Aggregate and report information.
  - Develop new information products and services.
  - etc.

---

* Named entities are people, organizations, locations, events and things that have proper names with which they are typically or legally labelled..

Agenda

- Metadata and interoperability
- Data management challenges

Data management challenge

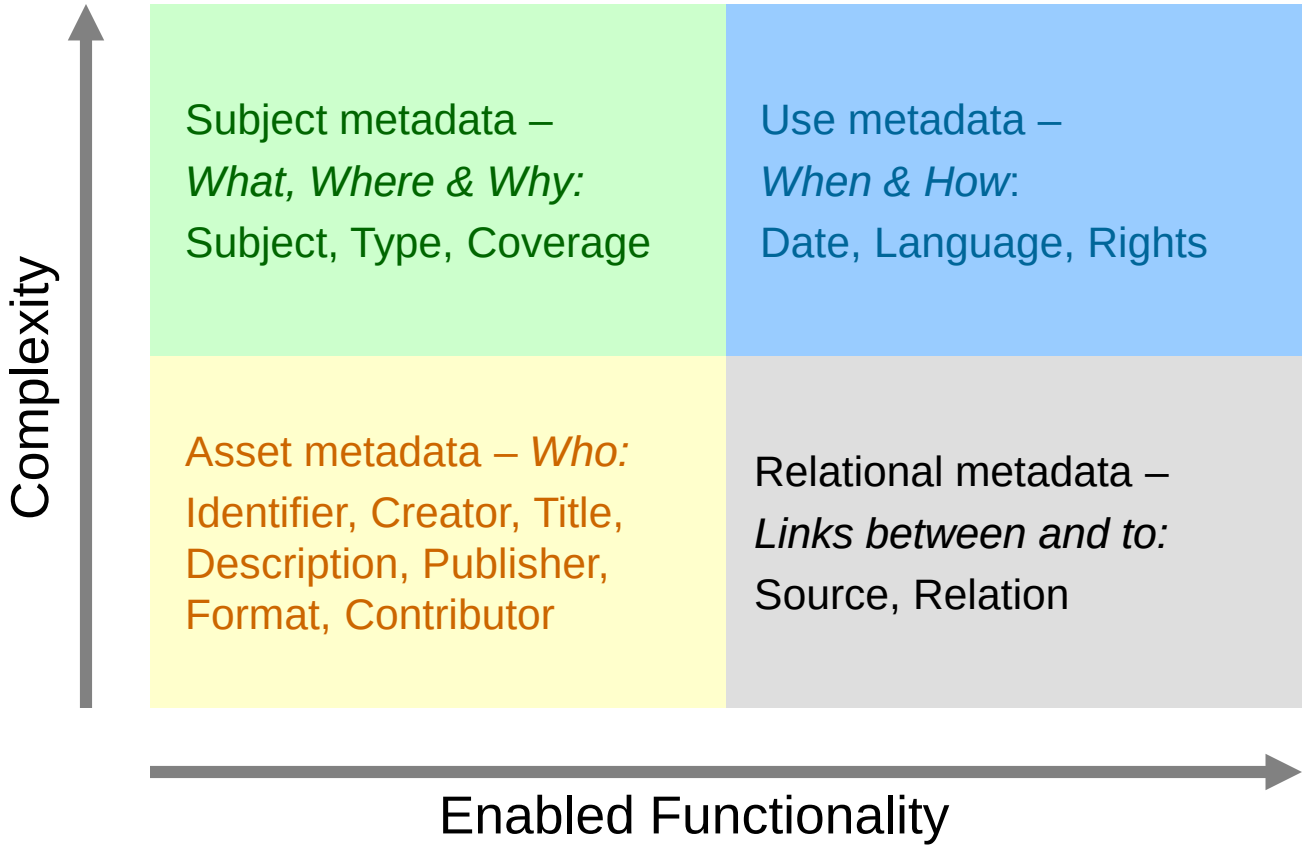
- How to align different metadata properties
  - E.g., *Originator* and *Creator*; *Bounding Coordinates* and *Coverage*; etc.
- How to align different vocabularies
  - E.g., *CA* and *California*; *FHA* and *Federal Highway Administration*; *P&R* and *Park and Ride*; etc.
- Master Data Management (MDM) aims to normalize metadata schemas and valid values across heterogeneous data management systems.

MDM is concerned with at least two types of vocabularies

- Metadata schemes like Dublin Core (ISO 15836), Geographic Information Metadata (ISO 19115) and XML Schema (ISO 19139) for content description.
- Semantic schemes like SKOS (W3C Recommendation 18), OWL (W3C Recommendation 11) and Topic Maps (ISO 13250) for value vocabularies like the Transportation Research Thesaurus (TRT) and Federal Geographic Data Committee (FGDC) Topic Categories.
- XML namespaces (a.k.a. vocabularies) need to be combined to represent content on the web.

What is Dublin Core?

- Provides the basis for any user, tool, or program to find and use any information asset.



http://dublincore.org/

Why Dublin Core?
According to R. Todd Stephens*

- Dublin Core is a de-facto standard across many other systems and standards
  - RSS (1.0), OAI (Open Archives Initiative), etc.
  - Inside organizations – SharePoint, ECMS, etc.
  - Federal public websites (to comply with OMB Circular A–130, http://www.howto.gov/web-content/manage/categorize/meta-data)
- Mapping to DC elements from most existing schemes is simple.
- Metadata already exists in enterprise applications
  - ArcGIS, GeoMedia, MapInfo, Windchill, SAP, Documentum, MS Office, SharePoint, Drupal, OpenText, MarkLogic, etc.

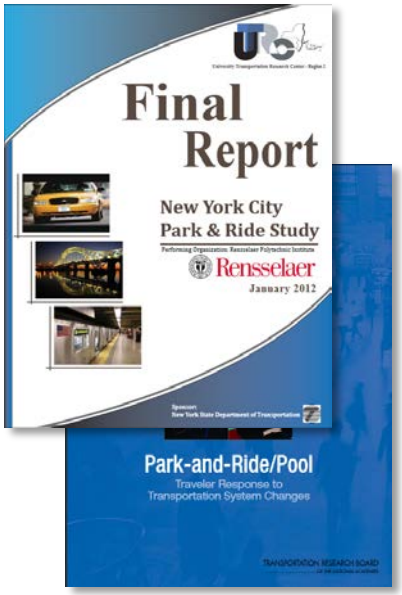
---

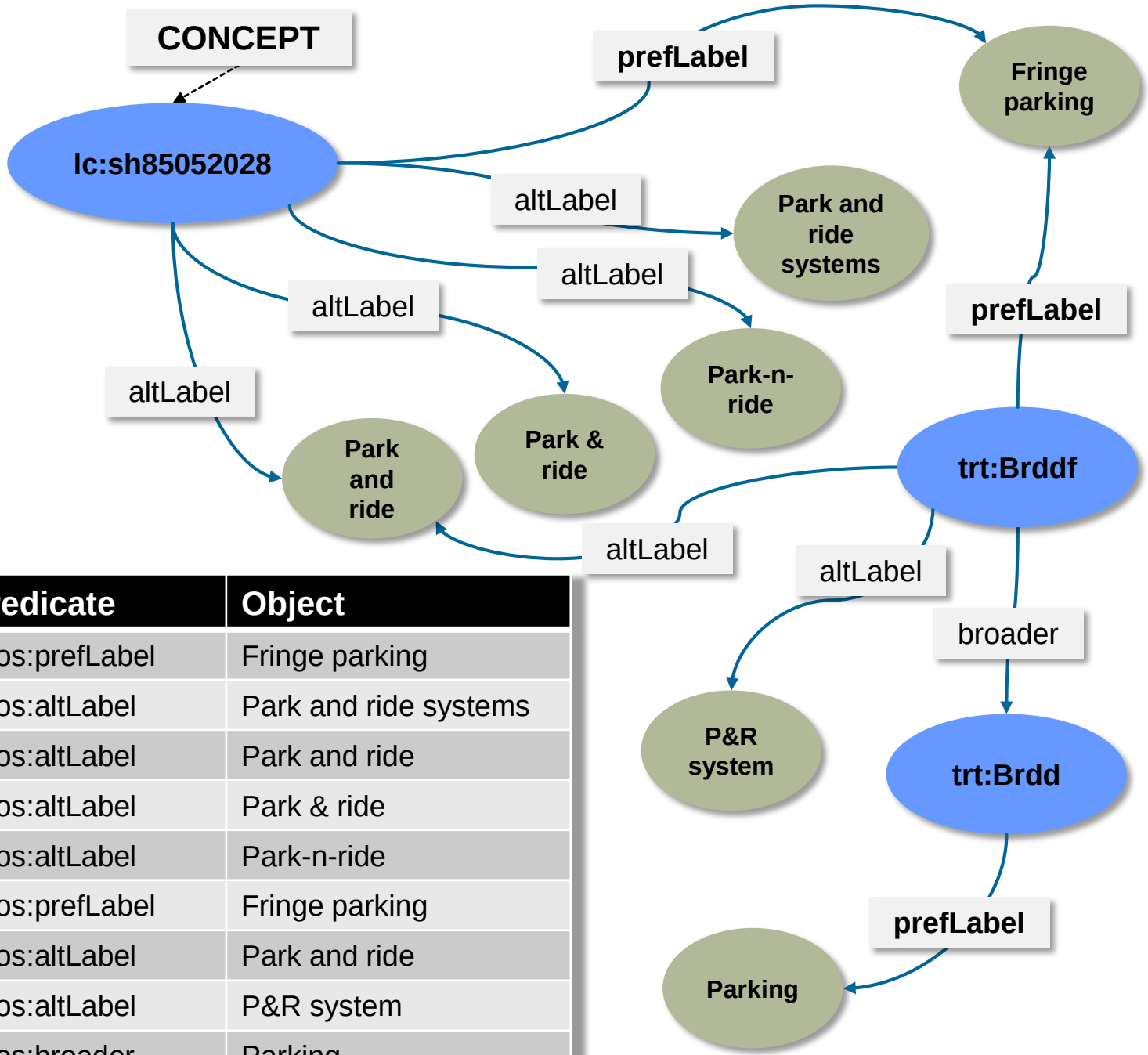
* Sr. Technical Architect (Collaboration and Online Services) at AT&T

What is SKOS?

- Provides the basis for any user, tool, or program to identify, define and link concept vocabularies.

Relationship	Definition
Concept	A unit of thought, an idea, meaning, or category of objects or events. A Concept is independent of the terms used to label it.
Preferred Label	A preferred lexical label for the resource such as a term used in a digital asset management system.
Alternate Label	An alternative label for the resource such as a synonym or quasi-synonym.
Broader Concept	Hierarchical link between two Concepts where one Concept is more general than the other.
Narrower Concept	Hierarchical link between two Concepts where one Concept is more specific than the other.
Related Concept	Link between two Concepts where the two are inherently "related", but that one is not in any way more general than the other.





Subject	Predicate	Object
lc:sh85052028	skos:prefLabel	Fringe parking
lc:sh85052028	skos:altLabel	Park and ride systems
lc:sh85052028	skos:altLabel	Park and ride
lc:sh85052028	skos:altLabel	Park & ride
lc:sh85052028	skos:altLabel	Park-n-ride
trt:Brddf	skos:prefLabel	Fringe parking
trt:Brddf	skos:altLabel	Park and ride
trt:Brddf	skos:altLabel	P&R system
Trt:Brdd	skos:broader	Parking

Why SKOS?
According to Alistair Miles* (SKOS co-author)

- **Ease of combination** with other standards
  - Vocabularies are used in great variety of contexts.
    - E.g., databases, faceted navigation, website browsing, linked open data, spellcheckers, etc.
  - Vocabularies are re-used in combination with other vocabularies.
    - E.g., Library of Congress Subject Headings + Transportation Research Thesaurus; USPS states + Federal Lands Highway Division Offices; USPS zip codes + US Congressional districts; etc.
- **Flexibility and extensibility** to cope with variations in structure and style
  - Variations between types of vocabularies
    - E.g., list vs. classification scheme
  - Variations within types of vocabularies
    - E.g., Z39.19-2005 monolingual controlled vocabularies and the Transportation Research Thesaurus

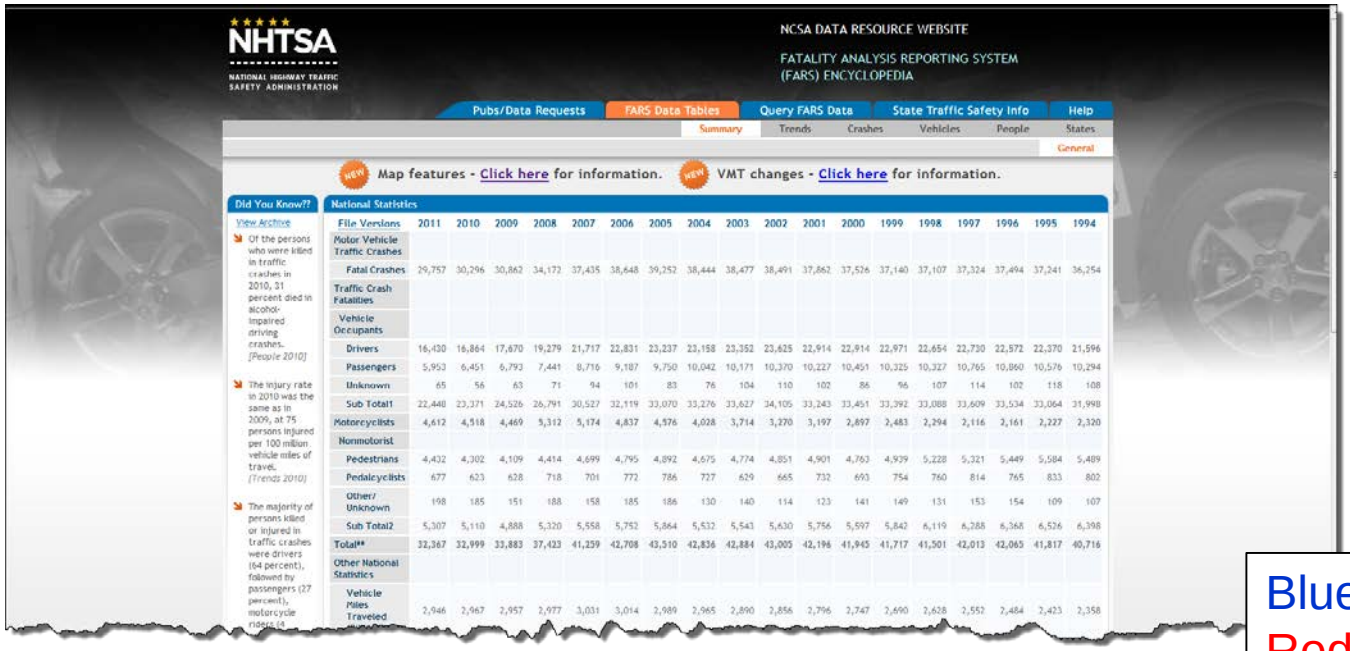
---

* Head of Epidemiological Informatics at Oxford University Wellcome Trust Centre for Human Genetics (formerly OUP Senior Computing Officer)

Why SKOS? (2)

- **Publish managed vocabularies** so they can readily be consumed by applications
  - Identify the concepts
    - What are the named entities?
  - Describe the relationships
    - Labels, definitions and other properties
  - Publish the data
    - Convert data structure to standard format
    - Put files on an http server (or load statements into an RDF server)
- **Ease of integration** with external applications
  - Use web services to use or link to a published concept, or to one or more entire vocabularies.
    - E.g., Google maps API, NY Times article search API, Linked open data; etc.
- **A W3C standard** like HTML, CSS, XML and RDF, RDFS, and OWL.

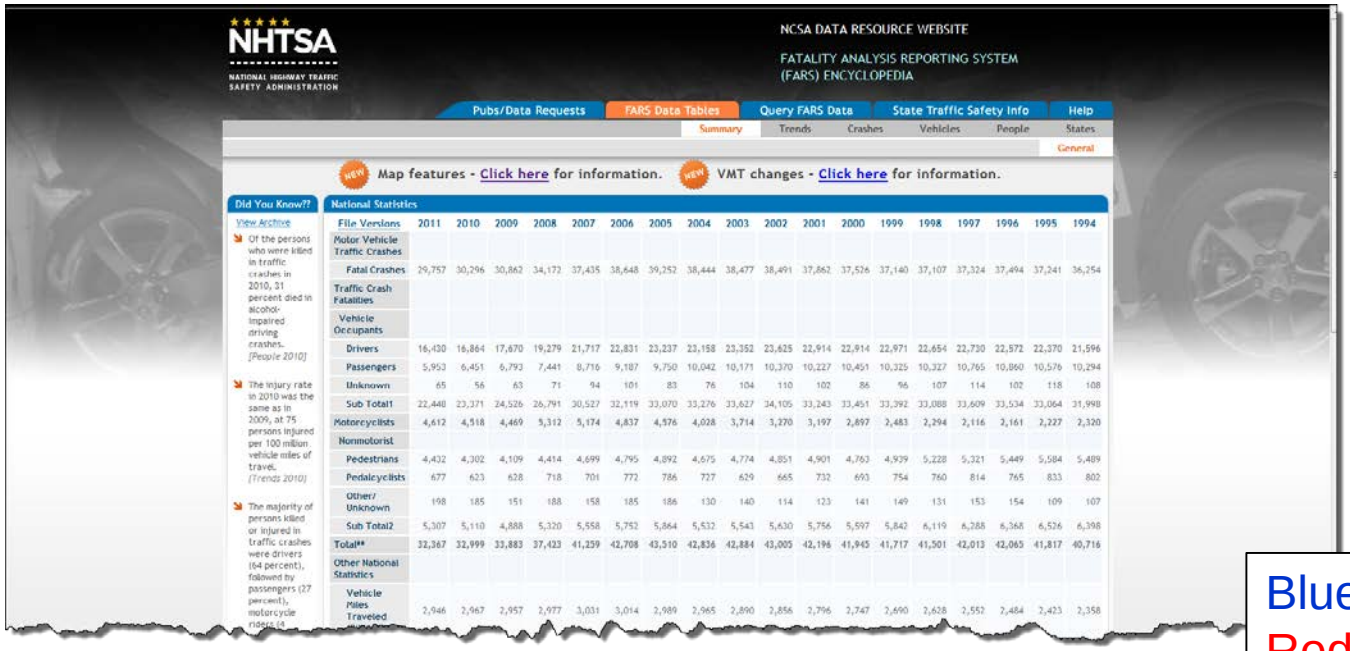
Metadata elements vs. metadata values



Blue = Dublin Core
Red = Vocabularies

Element	Scheme	Value
dc.identifier		http://www-fars.nhtsa.dot.gov/Main/index.aspx
dc.title		FAR Encyclopedia
dc.description		The FARS Encyclopedia offers an intuitive and powerful approach for retrieving fatal crash information.
dc.type	DITA	Reference resource
dc.subject	TRT	Traffic crashes, Fatalities
dc.coverage		1994-2011
dc.creator	US Govt Manual	National Highway Traffic Safety Administration
dc.date		2012

Metadata elements vs. metadata values



Blue = Dublin Core
Red = Vocabularies

<!--Each page must contain this info, per OMB-->

<meta name="dc.identifier" content="http://www-fars.nhtsa.dot.gov/Main/index.aspx" />

<meta name="dc.title" content="FAR Encyclopedia" />

<meta name="dc.description" content="The FARS Encyclopedia offers an intuitive and powerful approach for retrieving fatal crash information." />

<meta name="dc.type" scheme="DITA" content="Reference resource" />

<meta name="dc.subject" scheme="TRT" content="Traffic crashes, Fatalities" />

<meta name="dc.coverage" content="1994-2011" />

<meta name="dc.creator" scheme="US Government Manual" content="National Highway Traffic Safety Administration" />

<meta name="dc.date" content= "2012" />

<meta name="dc.format" content="text/html; charset=utf-8" />

<meta name="dc.language" scheme= "DCTERMS.RFC1766" content= "EN-US" />

<meta name="keywords" content= "FARS, Fatality Analysis Reporting System, PAR, Police Accident Reports, statistics, data, facts, car, truck, motorcycle, vehicle, pedestrian, street, road, highway, interstate, accident, injury" />

Effective data management IS the search solution

- Generate more consistent content to search on
  - Enable organic (web) and website search.
  - Support indexing, harvesting and linking to web pages.
- Correct user errors
  - Map the language of users to the language of the target content.
- Augment search results with linked data
  - Suggest related content.
  - Support mash-ups.
  - Publish vocabulary namespaces (Web enabled unique ID's)

You should not need to feel lucky to find the content you are looking for!

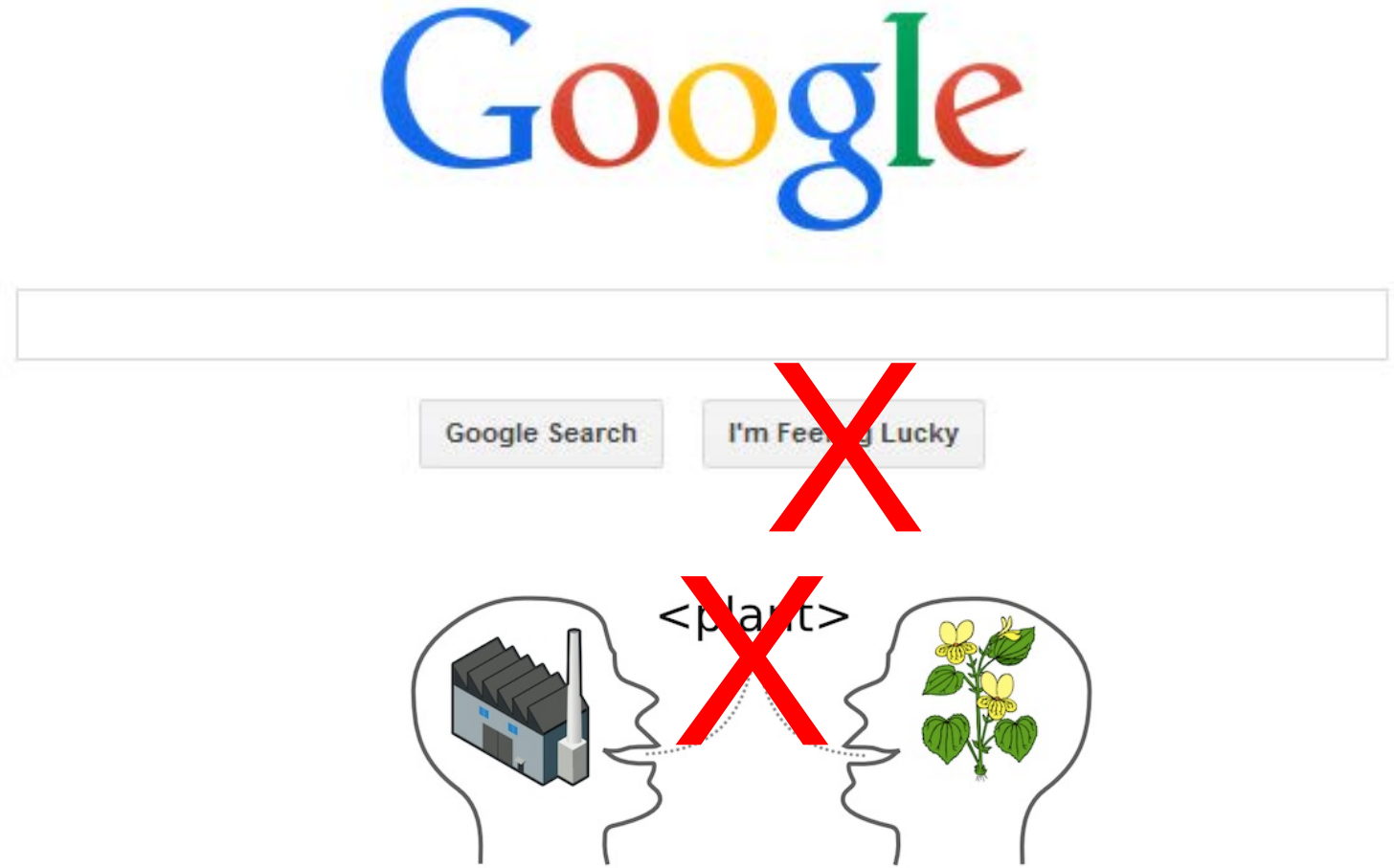


Image from Peter Krantz. http://www.peterkrantz.com/2010/semantic-interoperability/

Resources

- Cambridge Systematics. NCHRP Report 754: Improving Management of Transportation Information. 2013. http://www.trb.org/Publications/Blurbs/169522.aspx
- Dublin Core (ISO Standard 15836:2009) http://dublincore.org/documents/dces/
- J. Greenberg, S. Swauger, E.M. Feinstein. Metadata Capital in a Data Repository. Proceedings of International Conference on Dublin Core and Metadata Applications 2013. http://http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/189/172.
- Simple Knowledge Organization System (W3C Recommendation 18 August 2009) http://www.w3.org/TR/2009/REC-skos-reference-20090818/
- Transportation Research Thesaurus. http://trt.trb.org/

Joseph A Busch, Principal
jbusch@taxonomystrategies.com
twitter.com/joebusch
415-377-7912

QUESTIONS?