Journalism

and the

Semantic Web

Kurt Cagle

XML Industry Analyst
Managing Editor,
XMLToday.org
O'Reilly Media Contributing Editor

kurt.cagle@gmail.com
Twitter: @kurt_cagle

Is Journalism Dead? No.

The role of journalist has always been evolving – it is only the hubris of the established generation that has remained constant.

Journalist as Analyst

The significant *journalists* of today are analysts

Their role today is to discern meaning & validity in a rushing tide of assertions.

Increasingly less important is their role as reporters.

However, their role as raconteur – the ability to create a meaningful (and entertaining) narrative on a set of related events – is crucial.

Finally, their authority rests on their authenticity, their veracity, and their insight.

Analyst and Programmers

- The programmer creates sets of assertions that, when compiled, dereferenced and validated, builds a program.
- The analyst creates sets of assertions that, when compiled, dereferenced and validated, builds a narrative
- Welcome to Programming!!

What Is Semantics?

Semantics in computer science is the study of assertions, and the relationships that assertions have with respect to one another.

Informally, semantics touches on the nature of objects, classes and classification.

Finally, computational semantics is tied into the notion of abstraction, of creating simpler models that embody the relevant aspects of the larger system.

Operational Semantics

- Classification
  - What is this like?
- Abstraction
  - What is this about?
- Objectivization
  - What is the structure of this?
- Correlation
  - How much does this relate to that?

Operational Semantics II

- Inference
  - What does this imply?
- Context
  - How is this affected by what's around it?
- Identity
  - How do I (or can I) uniquely identify this?
- Authentication
  - How trustworthy is this assertion?

Resources and Collections

A ***resource*** is a noun – a person, place or thing – that can be represented in the virtual world with a record.

Resources have resource ***keys*** or **addresses** which uniquely define each resource (for some arbitrary definition of uniqueness).

A ***collection*** is a set of resources.

Collections play a huge part in the Semantic Web.

Classification

Classification is the process of establishing categories in a taxonomy and then assigning resources to one or more of those categories.

A ***taxonomy***, or classification schema, is a set of terms, sometimes with an underlying relationship between these terms.

Any resource can be thought of belonging to one or more collections named by a two part moniker - category:term.

Classification is "*difficulty:hard*".

Digital Orienteering

The map is not the territory.
Korsybski, 1946

On the web, the map IS the territory.
Cagle, 2009

Information Navigation

Put another way, everything on the web is a model – a model of an article, a person, a page

Hyperlinks are *assertions* of relationships.

Most web navigation, most links, point to:

- A resource (a web page)
- A collection of resource links (a feed)
- A collection of collections (portals)

Collections & Feeds

- A ***feed*** is a collection of links with some metadata that provides context for each link ***entry***
- Syndication formats (RSS, Atom, json) are feed formats
- HTML structures, such as tables or HTML lists, with one or more links per entry, are also feeds
- Apps like Drupal are fundamentally resource feed providers and consumers.

Categories and Navigation

- Categorization creates collections of related resources from the domain of all possible resources (i.e., a category is a feed)
- Most web links point to either resources, feeds, or portals.
- Implication: web navigation IS categorization – categories provide the navigational structure of the web.
- Look at *Drupal* for a good example of this.

Categorization, Query and Search

A category is a form of query, typically on some set of category terms in the resource.

Web Search is also a form of query – it uses an algorithm (and dynamic parameters to return a feed).

Relevance is the degree to which each resource satisfies this query algorithm.

At this point in time, the most relevant aspect of Semantics is search.

Relevance

- Why is all this theory relevant to you?
- Broad category silos are being replace by an explosive number of microcategories
- The Long Tail (Chris Anderson) has become a fractal forest of short tails
- Each microcategory is a micro-market with a limited market size.
- This is the essence of power laws.

Relevance and Audience Size

$$Aud = \sum_{n \in \Omega} Rel(n) MktSz(n)$$

Aut = Total Audience for a given media piece
N = Index of categorization partitition
Rel(n) = Relevance of nth partition
MktSz(n) = Size of market in each partition

In English

Your total readership for any given piece of media is proportional to how relevant it is for the broadest number of micromarkets.*

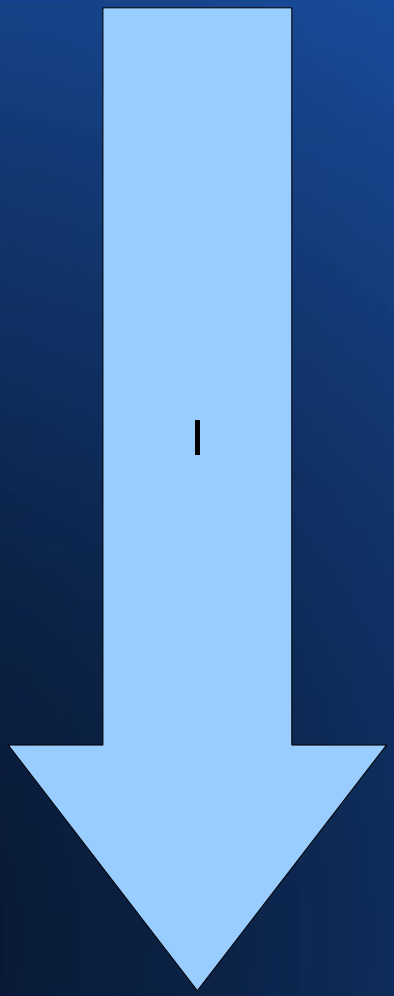
* One caveat – as size grows, relevance drop

Human vs Machine Relevance

- Machine Relevance is Search Engine Optimizations (SEO)– Gaming the System.
- Human Relevance is referential – how many people *who are themselves seen as relevant* (micromarkets) link to you.
  - This the premise of Twitter and Facebook
- Put another way – quality matters.
- Semantic Web Tech can increase machine relevance, but can destroy human relevance.

Semantic Web Technologies

Increasing Power and Increasing Complexity



Taxonomies and Folksonomies

Search, Query and Feeds

Widgets

Microformats & RDFa

Document Enrichment

XML Tools (Xquery!)

XML Ontologies

RDF and OWL, SPARQL, GRDDL

Semantic Rules Architectures

Widgets

- Widgets provide a visual interface for resources and feeds
- Widgets create visual semantics – associate “media” with content
- Widgets represent the componentization of the web, separating content from presentation

Inline Semantics

- *Inline semantics* create one or more additional layers of meaning in a document
  - Use attributes to add inline categorization
  - *Microformats* use fixed ontologies (vCard, Dublin Core, geoformats) ... fading
  - *Document Enrichment*
  - *RDFa* (Resource Description Frameworks for Attributes)

Document Enrichment

- Takes resource content and passes it through a *web service* to create categorization of names, events, scientific terms and so forth.
- These categories are embedded as XML elements or attributes.
- It uses semantic tools to disambiguate terms.
- Good starting point: *OpenCalais.com (Reuters)*

XML

- XML is moving into big data – most organizations now use XML both as message stores and messaging formats for documents AND data.
- XQuery is relatively new (2007) standard for query of XML data stores.
- XQuery provides distributed queries, development of template outputs, and a full

RFDa

- RDFa provides a way to make inline assertions about blocks of text.
- RDFa can be hand entered, or can be added via document enrichment.
- GRDDL can then read RDFa enriched documents and generate RDF.
- RDFa/GRDDL represents the bridge between text indexing and the Semantic Web

Data Stores & Modeling

- Relational Data Model
  - Data as Tables
  - Query Modeler: SQL
  - Local, Static, Bounded
- XML Data Model
  - Data as Documents
  - Query Modeler: XQuery
  - Distributed, Flexible, but still Bounded

Data Stores & Modeling II

- RDF Data Model
  - Data as Assertions
  - Query Modeler: SPARQL
  - Distributed, Flexible and Unbounded
- Bound data models
  - All data models maps to a defined schema
- Unbound data models
  - Data models may add or remove arbitrary attributes.

Linked Data

- Because RDF/OWL data models are dynamic, queries can search on multiple distributed RDF stores at once.
- This principle is known as *Linked Data*.
- RDF provides links (or acts as payloads) to resources and can also abstract resource content.
- Linked Data is distributed – silos disappear.

Query Unification & XProc

XProc is a W3C pipeline language standard

Each pipe in the pipeline is a specific type of XML operation, from counting nodes to performing xqueries

SPARQL queries could be used to extract RDF from distributed Linked Data repositories as a pipe.

This would unify SPARQL and XQuery, making both semantic and syntactic queries possible.

Future of Journalistic Semantics

- Sophisticated inferential analysis
- More effective user agents and avatars
- Automated production of intelligent abstracts
- Semantic rules can launch nuanced applications based upon meaning matching
- Semantic rules engines + inferential analysis = sophisticated composition engines
- Pulitzer Prize by an A.I. by 2030?