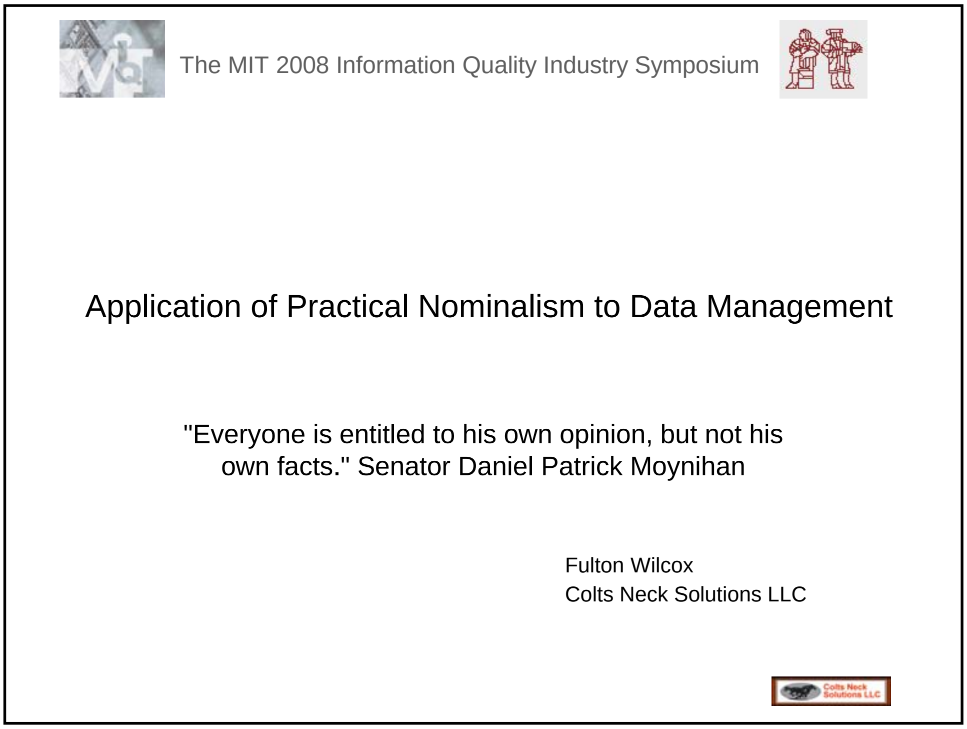
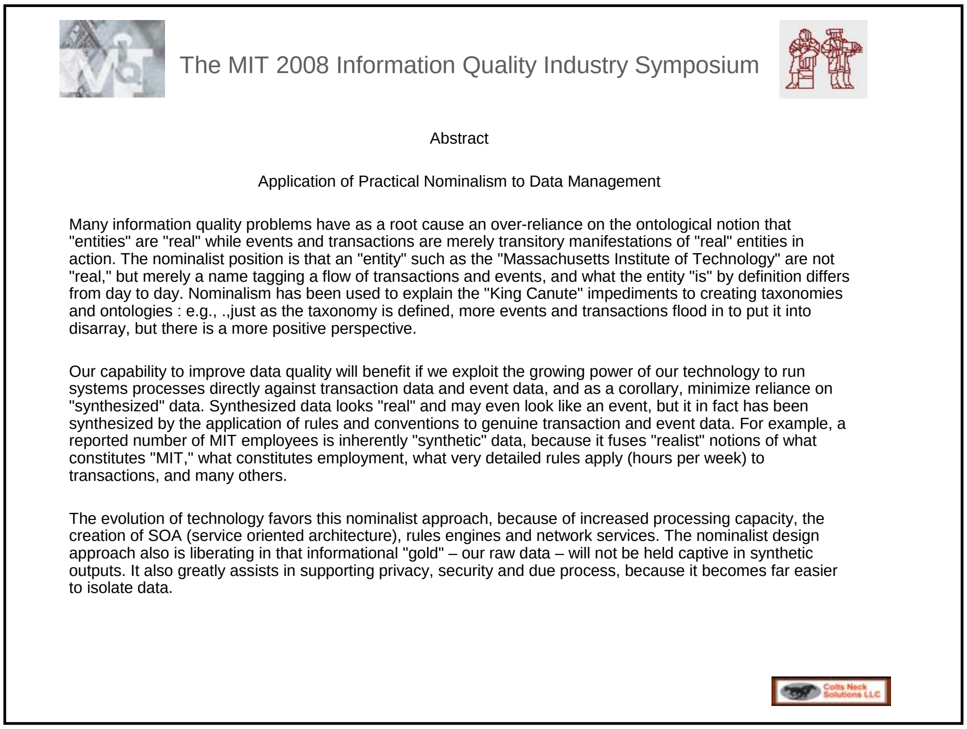


The MIT 2008 Information Quality Industry Symposium

Application of Practical Nominalism to Data Management

"Everyone is entitled to his own opinion, but not his own facts." Senator Daniel Patrick Moynihan

Fulton Wilcox
Colts Neck Solutions LLC



The MIT 2008 Information Quality Industry Symposium

Abstract

Application of Practical Nominalism to Data Management

Many information quality problems have as a root cause an over-reliance on the ontological notion that "entities" are "real" while events and transactions are merely transitory manifestations of "real" entities in action. The nominalist position is that an "entity" such as the "Massachusetts Institute of Technology" are not "real," but merely a name tagging a flow of transactions and events, and what the entity "is" by definition differs from day to day. Nominalism has been used to explain the "King Canute" impediments to creating taxonomies and ontologies : e.g., .,just as the taxonomy is defined, more events and transactions flood in to put it into disarray, but there is a more positive perspective.

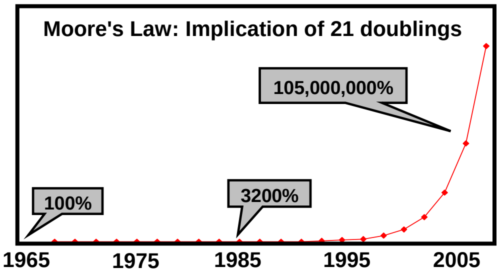
Our capability to improve data quality will benefit if we exploit the growing power of our technology to run systems processes directly against transaction data and event data, and as a corollary, minimize reliance on "synthesized" data. Synthesized data looks "real" and may even look like an event, but it in fact has been synthesized by the application of rules and conventions to genuine transaction and event data. For example, a reported number of MIT employees is inherently "synthetic" data, because it fuses "realist" notions of what constitutes "MIT," what constitutes employment, what very detailed rules apply (hours per week) to transactions, and many others.

The evolution of technology favors this nominalist approach, because of increased processing capacity, the creation of SOA (service oriented architecture), rules engines and network services. The nominalist design approach also is liberating in that informational "gold" – our raw data – will not be held captive in synthetic outputs. It also greatly assists in supporting privacy, security and due process, because it becomes far easier to isolate data.

The MIT 2008 Information Quality Industry Symposium

Technological Progress: The 900 Pound Gorilla

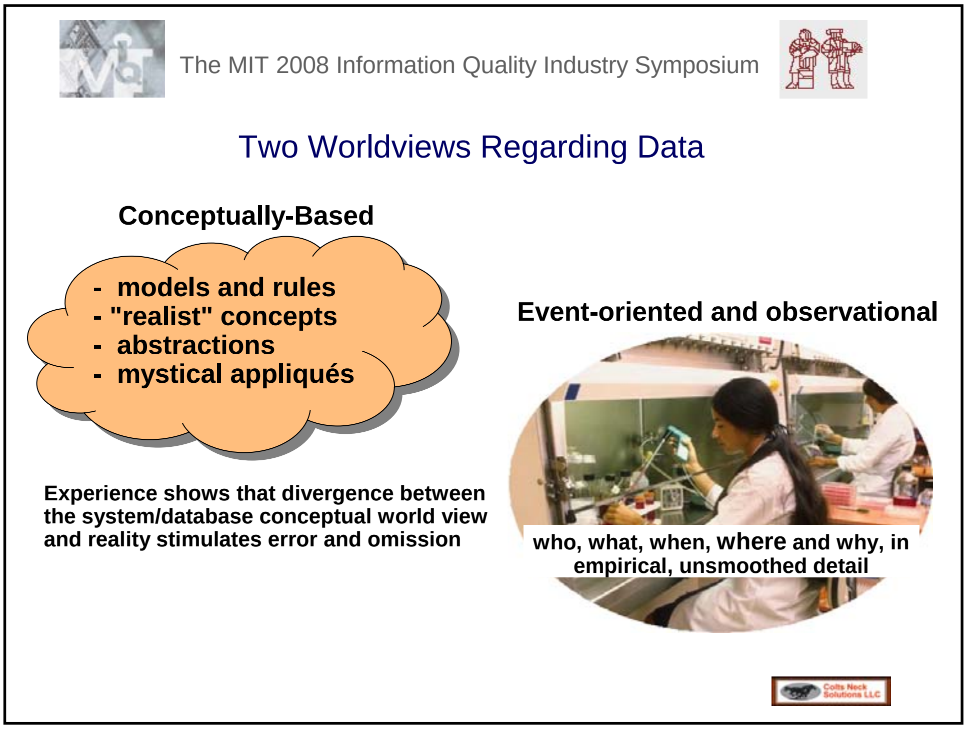
- Opens up new opportunities, e.g. Internet 2.0 and Internet 3.0
- Moore's Law is a proxy for brute force increases in capability across IT: processing, memory, storage, networking, virtualization ...
- Plotting it on a linear scale dramatizes effects
- The plot also prompts questions:
  - were IT users 150,000,000% better off in 2007 than in 1965?
  - Indeed, were they 800% better off in 2007 than in 2001? If not, why not?
  - With more doublings coming, how do we put the gorilla to work?
- What are the data management and data quality implications?



The MIT 2008 Information Quality Industry Symposium

Must the data warehouse inmates take over the asylum?

- **ETL (extract, transform and load) processes rub people's noses into the shortcomings of today's data and data management processes**
  - What looks sensible and "good enough" within a given venue breaks down when used across venues
- **As described by Claudia Imhoff of Intelligent Solutions, those implementing data warehouses and ETL identified three needs:**
  - improve the quality of the data being integrated,
  - create sets of integrated master or reference data (MDM),
  - implement repositories of current data for management and operational purposes (ODS), and so on...
  - so the data warehouse team took a much broader, more diverse set of projects. "
- **Her reaction to this expansion of ETL/DW scope was "Back off!" ... *None* of these is a data warehouse project."**
  - From a scope creep perspective, she is right, but the problem needs to be addressed, and rescue is probably not coming from BPEL, etc.



The MIT 2008 Information Quality Industry Symposium

Two Worldviews Regarding Data

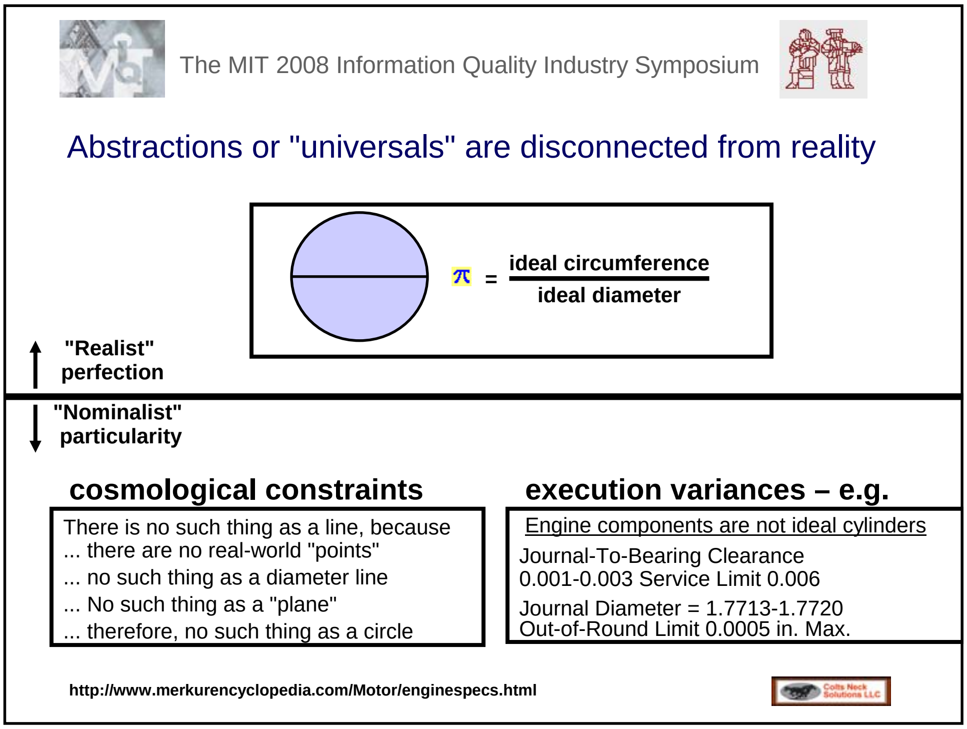
Conceptually-Based

- **models and rules**
- **"realist" concepts**
- **abstractions**
- **mystical appliqués**

Experience shows that divergence between the system/database conceptual world view and reality stimulates error and omission

Event-oriented and observational

who, what, when, where and why, in empirical, unsmoothed detail



The MIT 2008 Information Quality Industry Symposium

Abstractions or "universals" are disconnected from reality

$$\pi = \frac{\textbf{ideal circumference}}{\textbf{ideal diameter}}$$

↑ **"Realist" perfection**

↓ **"Nominalist" particularity**

cosmological constraints

There is no such thing as a line, because
... there are no real-world "points"
... no such thing as a diameter line
... No such thing as a "plane"
... therefore, no such thing as a circle

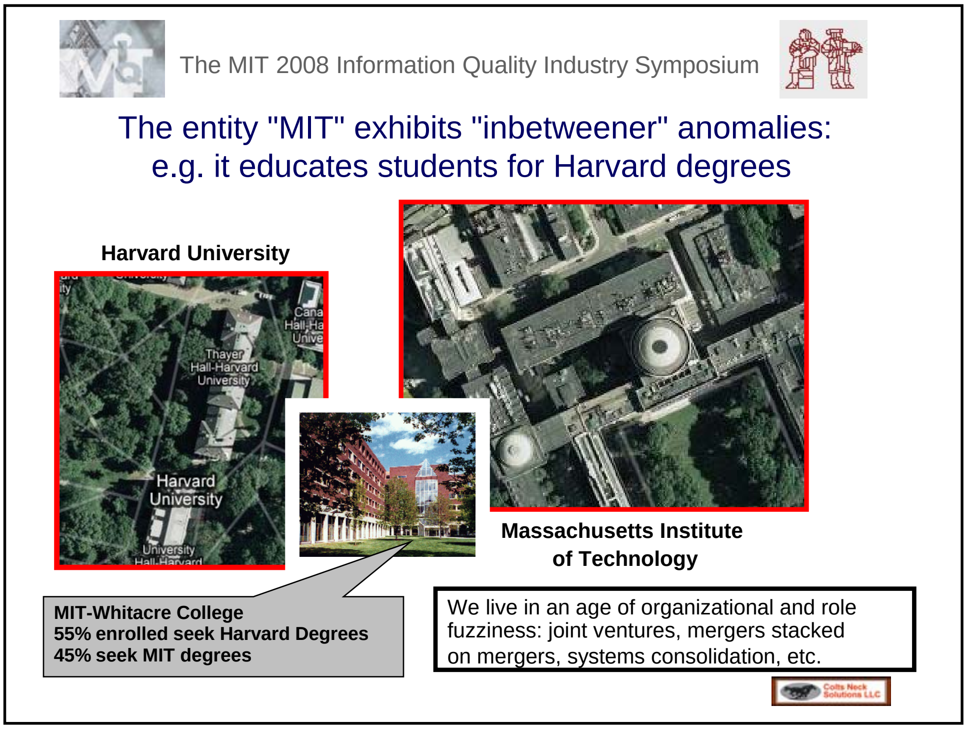
execution variances – e.g.

Engine components are not ideal cylinders

Journal-To-Bearing Clearance
0.001-0.003 Service Limit 0.006

Journal Diameter = 1.7713-1.7720
Out-of-Round Limit 0.0005 in. Max.

http://www.merkurencyclopedia.com/Motor/enginespecs.html


Cana
Hall-Ha
Unive
Thayer
Hall-Harvard
University
Harvard
University
University
Hall-Harvard
Colts Neck Solutions LLC

The MIT 2008 Information Quality Industry Symposium

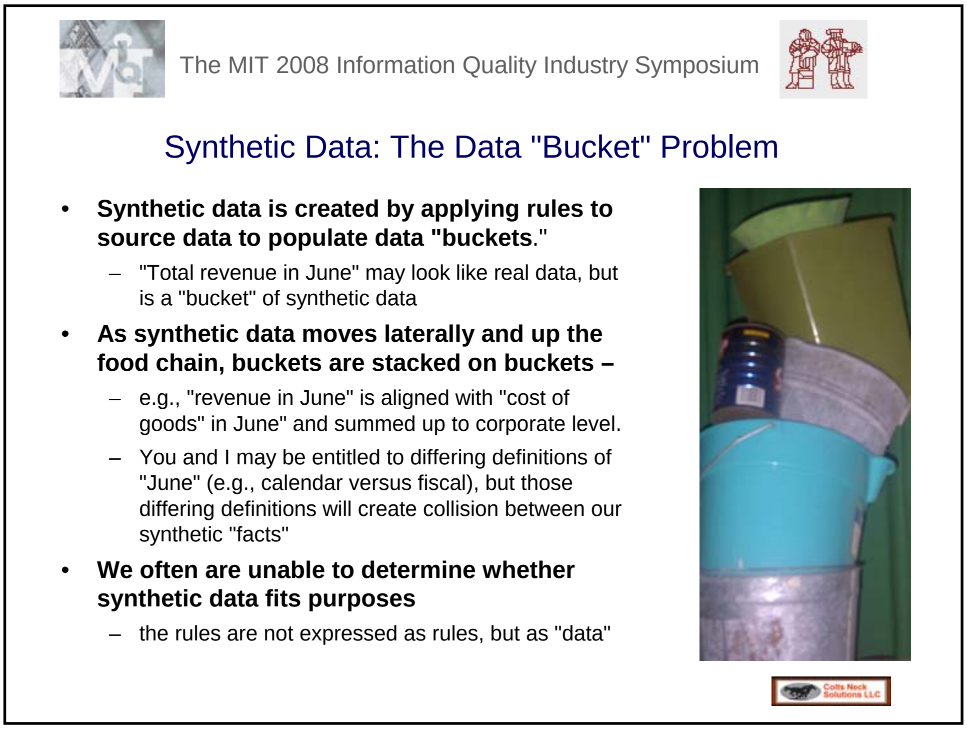
The entity "MIT" exhibits "inbetweener" anomalies: e.g. it educates students for Harvard degrees

Harvard University

Massachusetts Institute of Technology

MIT-Whitacre College
55% enrolled seek Harvard Degrees
45% seek MIT degrees

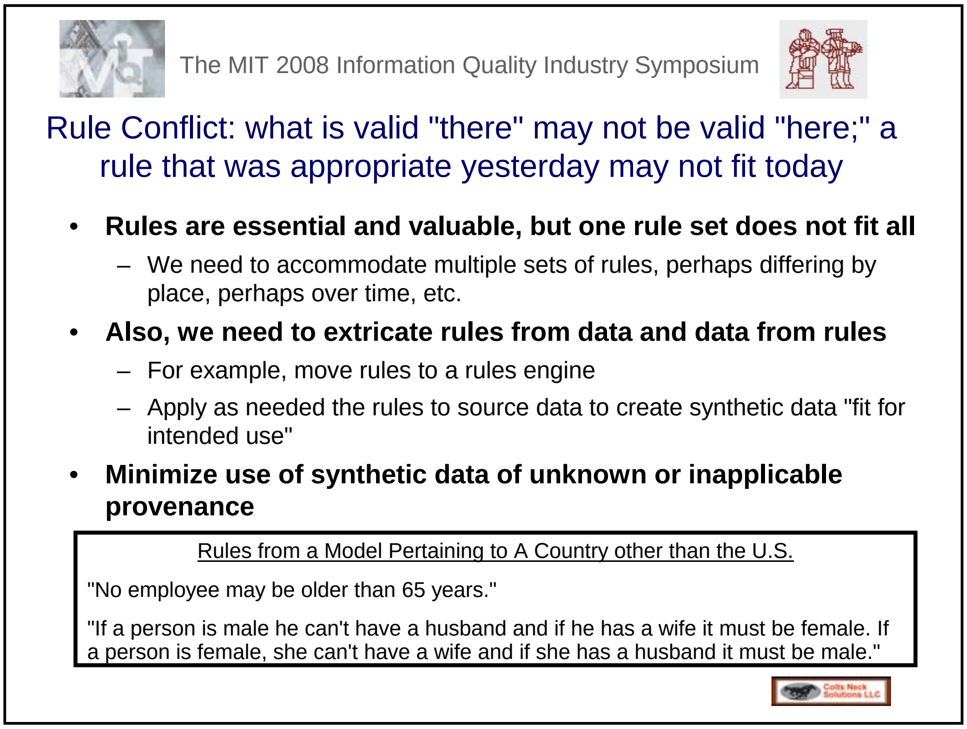
We live in an age of organizational and role fuzziness: joint ventures, mergers stacked on mergers, systems consolidation, etc.


Colts Neck Solutions LLC

The MIT 2008 Information Quality Industry Symposium

Synthetic Data: The Data "Bucket" Problem

- **Synthetic data is created by applying rules to source data to populate data "buckets**."
  - "Total revenue in June" may look like real data, but is a "bucket" of synthetic data
- **As synthetic data moves laterally and up the food chain, buckets are stacked on buckets –**
  - e.g., "revenue in June" is aligned with "cost of goods" in June" and summed up to corporate level.
  - You and I may be entitled to differing definitions of "June" (e.g., calendar versus fiscal), but those differing definitions will create collision between our synthetic "facts"
- **We often are unable to determine whether synthetic data fits purposes**
  - the rules are not expressed as rules, but as "data"



The MIT 2008 Information Quality Industry Symposium

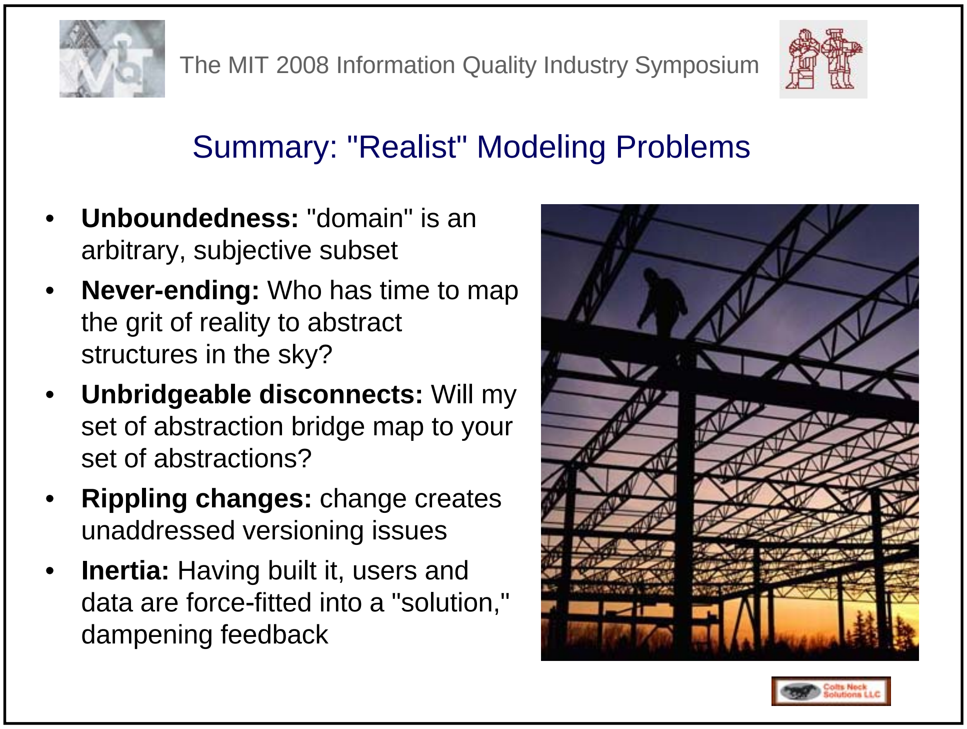
Rule Conflict: what is valid "there" may not be valid "here;" a rule that was appropriate yesterday may not fit today

- **Rules are essential and valuable, but one rule set does not fit all**
  - We need to accommodate multiple sets of rules, perhaps differing by place, perhaps over time, etc.
- **Also, we need to extricate rules from data and data from rules**
  - For example, move rules to a rules engine
  - Apply as needed the rules to source data to create synthetic data "fit for intended use"
- **Minimize use of synthetic data of unknown or inapplicable provenance**

Rules from a Model Pertaining to A Country other than the U.S.

"No employee may be older than 65 years."

"If a person is male he can't have a husband and if he has a wife it must be female. If a person is female, she can't have a wife and if she has a husband it must be male."



The MIT 2008 Information Quality Industry Symposium

Summary: "Realist" Modeling Problems

- **Unboundedness:** "domain" is an arbitrary, subjective subset
- **Never-ending:** Who has time to map the grit of reality to abstract structures in the sky?
- **Unbridgeable disconnects:** Will my set of abstraction bridge map to your set of abstractions?
- **Rippling changes:** change creates unaddressed versioning issues
- **Inertia:** Having built it, users and data are force-fitted into a "solution," dampening feedback

The MIT 2008 Information Quality Industry Symposium

Revisiting choices between "Realism" and "Nominalism"

- **Realism and its cousin, conceptualism, treat abstractions as "real"**
  - The "model" or the ideal of a given object are thought of as the highest truth, while real-world "instantiation" is viewed as annoyingly noisy
  - Real-world instances are idiosyncratic and, like snowflakes, no two are identical
- **Within narrowly bounded, disciplined abstractions are useful, while divergences from reality may be of minor importance**
- **However, as systems reach and data reuse overruns "stovepiped" domains**
  - Model design and detail cannot keep pace with expanding "footprints" of automation
  - Use of given models over extended time lead to accumulation of error
  - Tighter coupling of models to software development removes a buffering effect
- **To expand reach and increase the reuse of data we need to accommodate and exploit particularity rather than mask it**

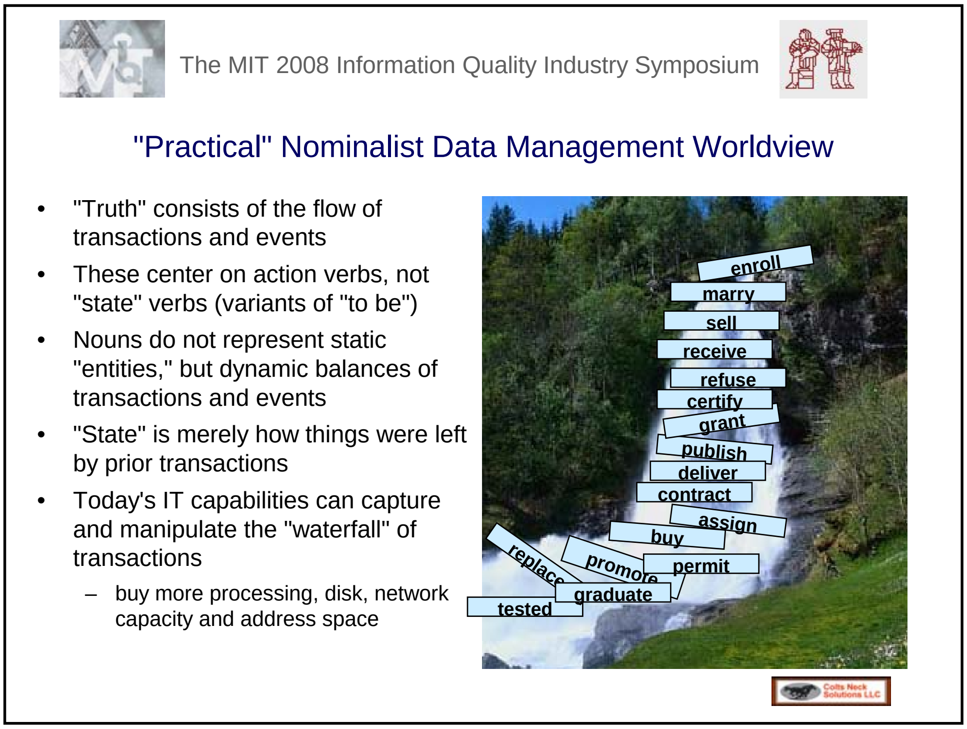
Colts Neck Solutions LLC

The MIT 2008 Information Quality Industry Symposium

Heraclitus: "We never step twice into the same river"

- To the nominalist, experience is flow
- We may name flowing water "river" or "stream" or "brook," but Heraclitus's point was that we are labeling a flow
  - Particular molecules flow by, to be replaced by other molecules
  - Words like "river" or "stream" are labels for fuzzy sets of instances
  - If it rains hard, the stream becomes a river; in drought the river becomes a riverbed
- An image of a river flowing over a waterfall is a good introduction to "transactions" and "all is flow"

Colts Neck Solutions LLC

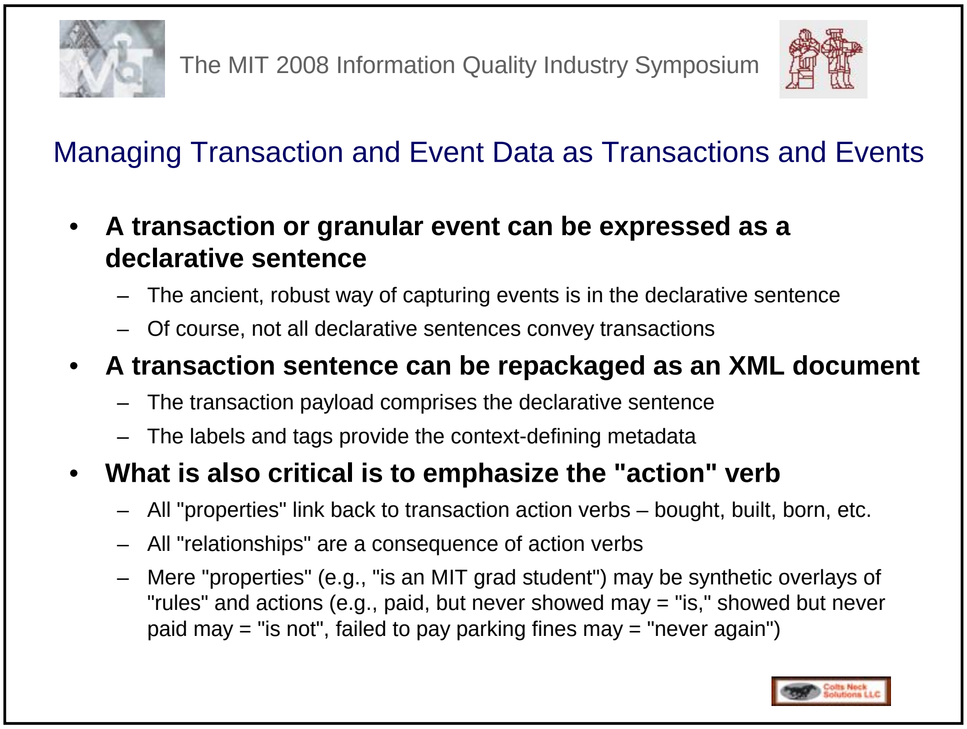


The MIT 2008 Information Quality Industry Symposium

"Practical" Nominalist Data Management Worldview

- "Truth" consists of the flow of transactions and events
- These center on action verbs, not "state" verbs (variants of "to be")
- Nouns do not represent static "entities," but dynamic balances of transactions and events
- "State" is merely how things were left by prior transactions
- Today's IT capabilities can capture and manipulate the "waterfall" of transactions
  - buy more processing, disk, network capacity and address space

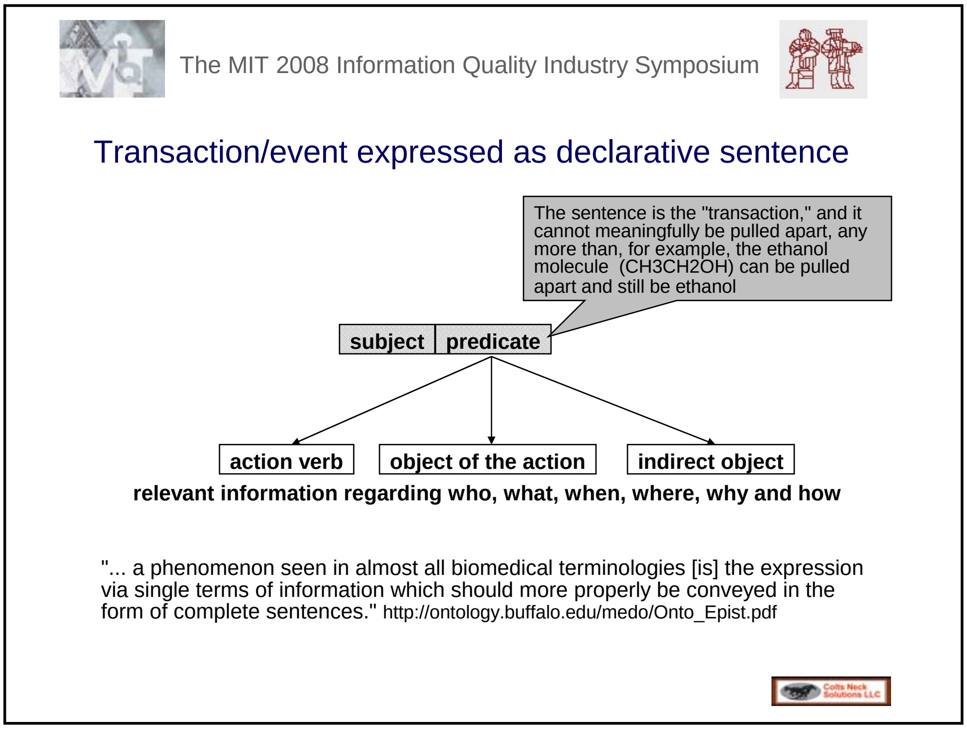
enroll
marry
sell
receive
refuse
certify
grant
publish
deliver
contract
assign
buy
replace
promote
permit
graduate
tested



The MIT 2008 Information Quality Industry Symposium

Managing Transaction and Event Data as Transactions and Events

- **A transaction or granular event can be expressed as a declarative sentence**
  - The ancient, robust way of capturing events is in the declarative sentence
  - Of course, not all declarative sentences convey transactions
- **A transaction sentence can be repackaged as an XML document**
  - The transaction payload comprises the declarative sentence
  - The labels and tags provide the context-defining metadata
- **What is also critical is to emphasize the "action" verb**
  - All "properties" link back to transaction action verbs – bought, built, born, etc.
  - All "relationships" are a consequence of action verbs
  - Mere "properties" (e.g., "is an MIT grad student") may be synthetic overlays of "rules" and actions (e.g., paid, but never showed may = "is," showed but never paid may = "is not", failed to pay parking fines may = "never again")



The MIT 2008 Information Quality Industry Symposium

Transaction/event expressed as declarative sentence

The sentence is the "transaction," and it cannot meaningfully be pulled apart, any more than, for example, the ethanol molecule (CH3CH2OH) can be pulled apart and still be ethanol

subject | **predicate**

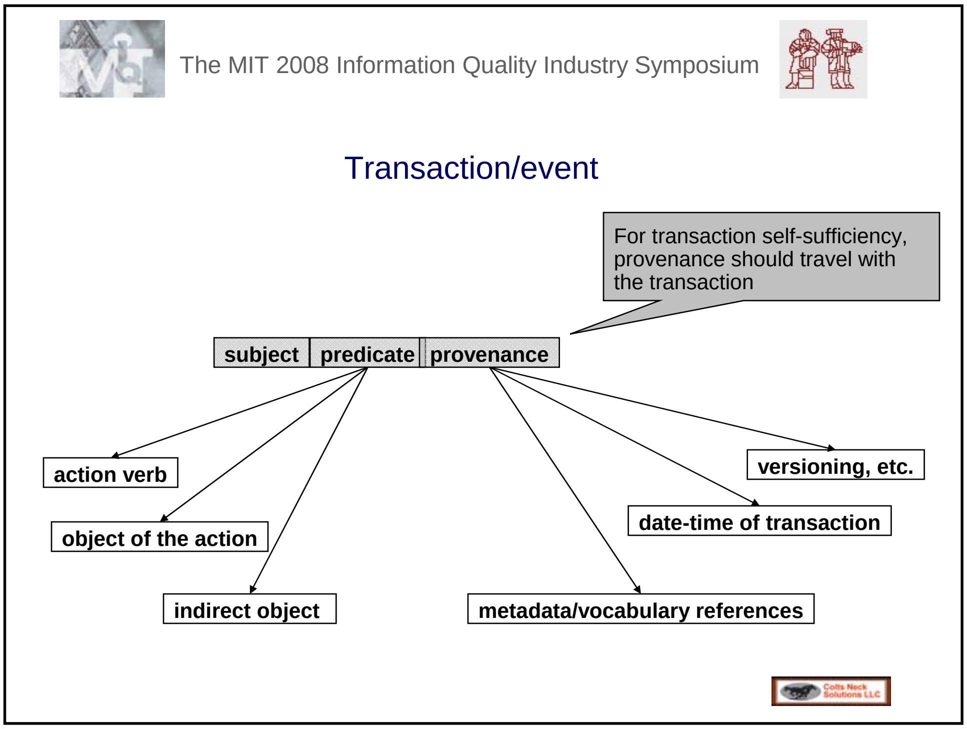
action verb

object of the action

indirect object

relevant information regarding who, what, when, where, why and how

"... a phenomenon seen in almost all biomedical terminologies [is] the expression via single terms of information which should more properly be conveyed in the form of complete sentences." http://ontology.buffalo.edu/medo/Onto_Epist.pdf



The MIT 2008 Information Quality Industry Symposium

Transaction/event

For transaction self-sufficiency, provenance should travel with the transaction

subject | **predicate** | **provenance**

action verb

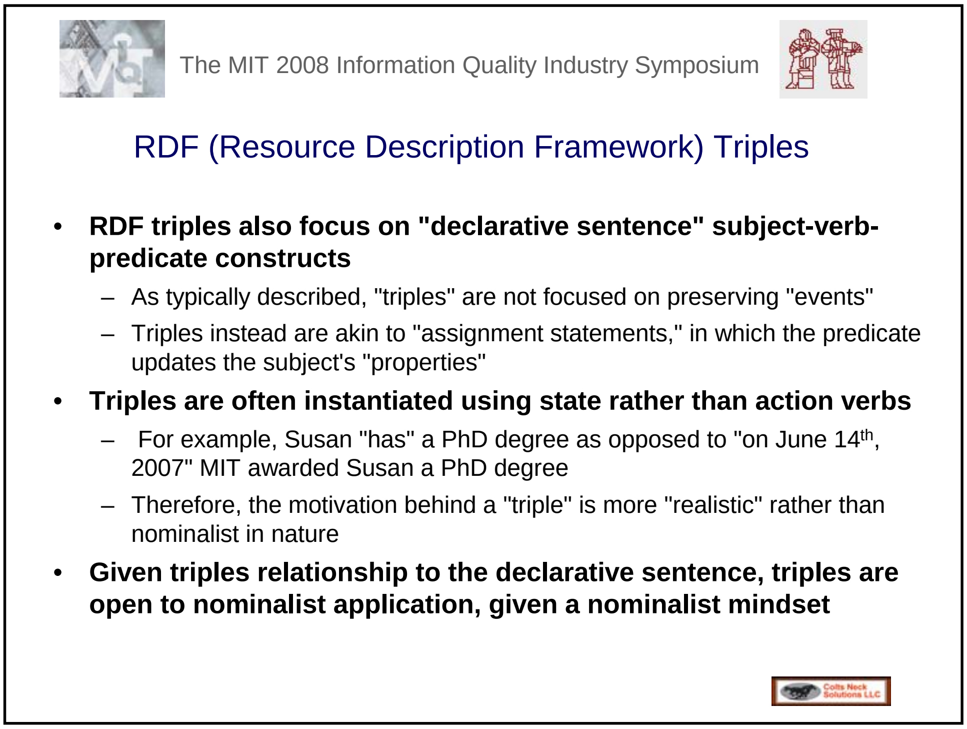
object of the action

indirect object

versioning, etc.

date-time of transaction

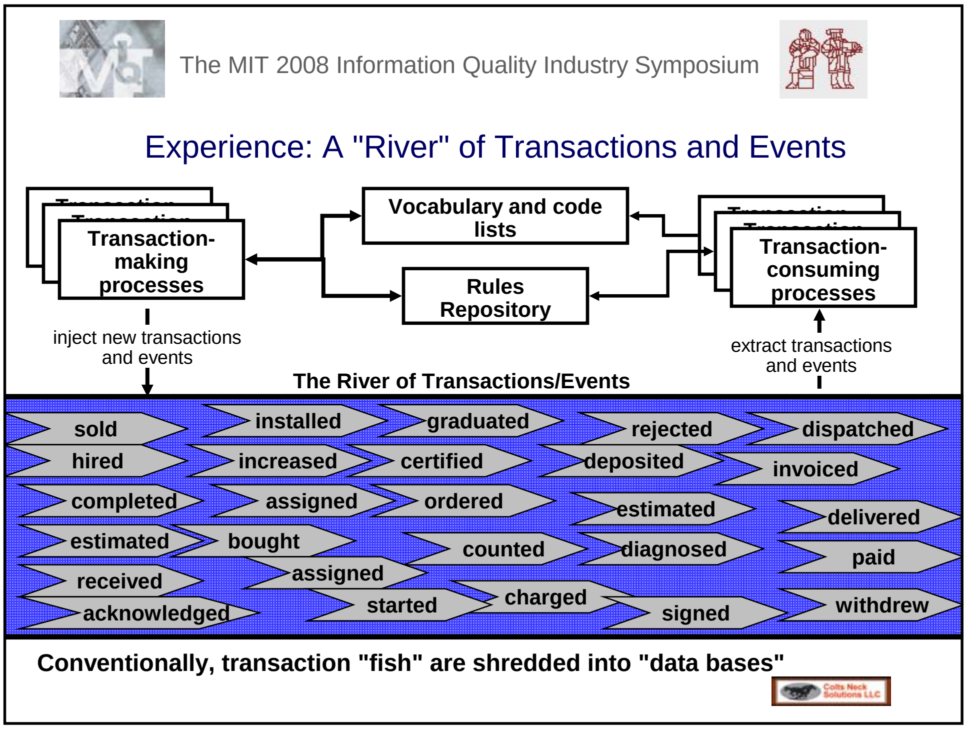
metadata/vocabulary references



The MIT 2008 Information Quality Industry Symposium

RDF (Resource Description Framework) Triples

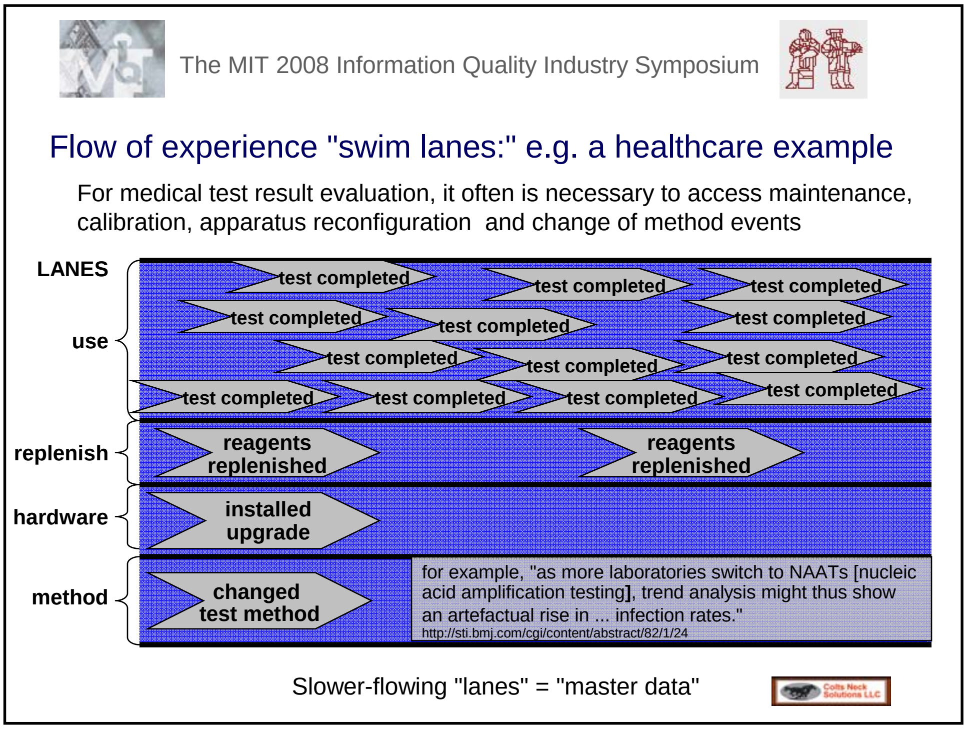
- **RDF triples also focus on "declarative sentence" subject-verb-predicate constructs**
  - As typically described, "triples" are not focused on preserving "events"
  - Triples instead are akin to "assignment statements," in which the predicate updates the subject's "properties"
- **Triples are often instantiated using state rather than action verbs**
  - For example, Susan "has" a PhD degree as opposed to "on June 14th, 2007" MIT awarded Susan a PhD degree
  - Therefore, the motivation behind a "triple" is more "realistic" rather than nominalist in nature
- **Given triples relationship to the declarative sentence, triples are open to nominalist application, given a nominalist mindset**



The MIT 2008 Information Quality Industry Symposium

Experience: A "River" of Transactions and Events

Conventionally, transaction "fish" are shredded into "data bases"



The MIT 2008 Information Quality Industry Symposium

Flow of experience "swim lanes:" e.g. a healthcare example

For medical test result evaluation, it often is necessary to access maintenance, calibration, apparatus reconfiguration and change of method events

LANES

use

test completed
test completed
test completed
test completed
test completed
test completed
test completed
test completed
test completed
test completed
test completed
test completed
test completed
test completed

replenish

reagents replenished
reagents replenished

hardware

installed upgrade

method

changed test method

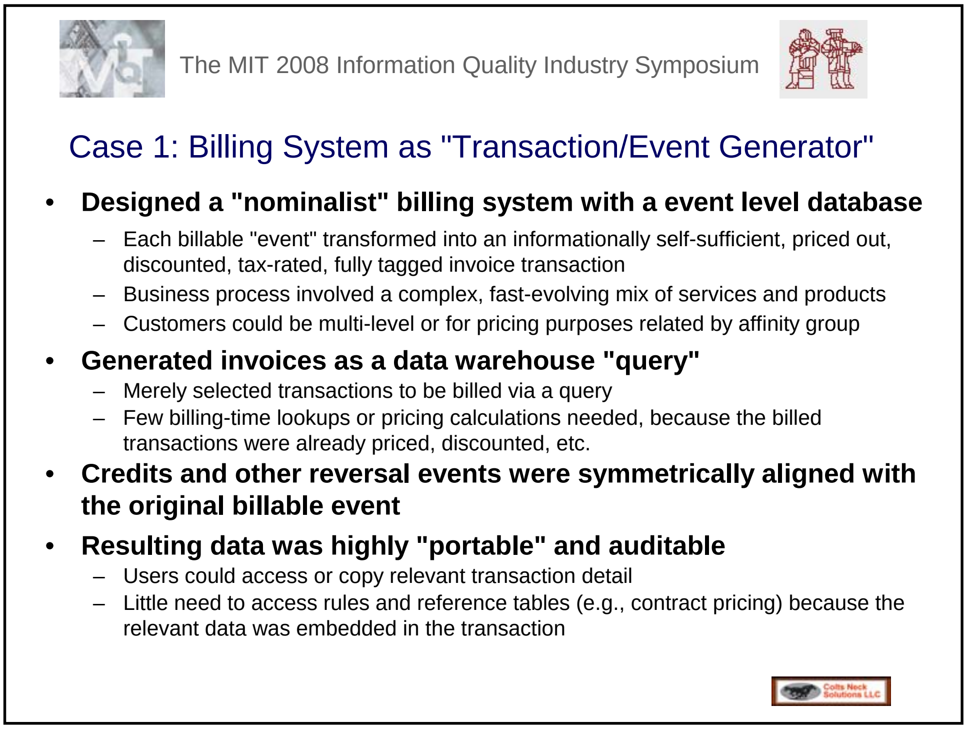
for example, "as more laboratories switch to NAATs [nucleic acid amplification testing], trend analysis might thus show an artefactual rise in ... infection rates."
http://sti.bmj.com/cgi/content/abstract/82/1/24

Slower-flowing "lanes" = "master data"

The MIT 2008 Information Quality Industry Symposium

Three Nominalist Success "Cases":

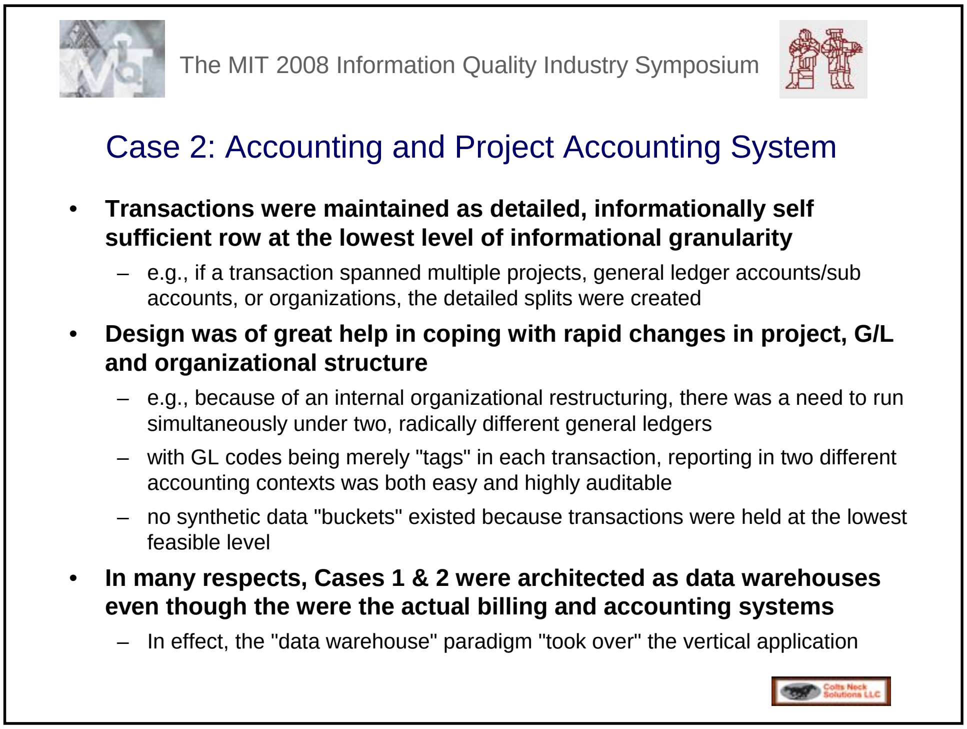
A Few Minor Adjustments in Architecture



The MIT 2008 Information Quality Industry Symposium

Case 1: Billing System as "Transaction/Event Generator"

- **Designed a "nominalist" billing system with a event level database**
  - Each billable "event" transformed into an informationally self-sufficient, priced out, discounted, tax-rated, fully tagged invoice transaction
  - Business process involved a complex, fast-evolving mix of services and products
  - Customers could be multi-level or for pricing purposes related by affinity group
- **Generated invoices as a data warehouse "query"**
  - Merely selected transactions to be billed via a query
  - Few billing-time lookups or pricing calculations needed, because the billed transactions were already priced, discounted, etc.
- **Credits and other reversal events were symmetrically aligned with the original billable event**
- **Resulting data was highly "portable" and auditable**
  - Users could access or copy relevant transaction detail
  - Little need to access rules and reference tables (e.g., contract pricing) because the relevant data was embedded in the transaction



The MIT 2008 Information Quality Industry Symposium

Case 2: Accounting and Project Accounting System

- **Transactions were maintained as detailed, informationally self sufficient row at the lowest level of informational granularity**
  - e.g., if a transaction spanned multiple projects, general ledger accounts/sub accounts, or organizations, the detailed splits were created
- **Design was of great help in coping with rapid changes in project, G/L and organizational structure**
  - e.g., because of an internal organizational restructuring, there was a need to run simultaneously under two, radically different general ledgers
  - with GL codes being merely "tags" in each transaction, reporting in two different accounting contexts was both easy and highly auditable
  - no synthetic data "buckets" existed because transactions were held at the lowest feasible level
- **In many respects, Cases 1 & 2 were architected as data warehouses even though the were the actual billing and accounting systems**
  - In effect, the "data warehouse" paradigm "took over" the vertical application

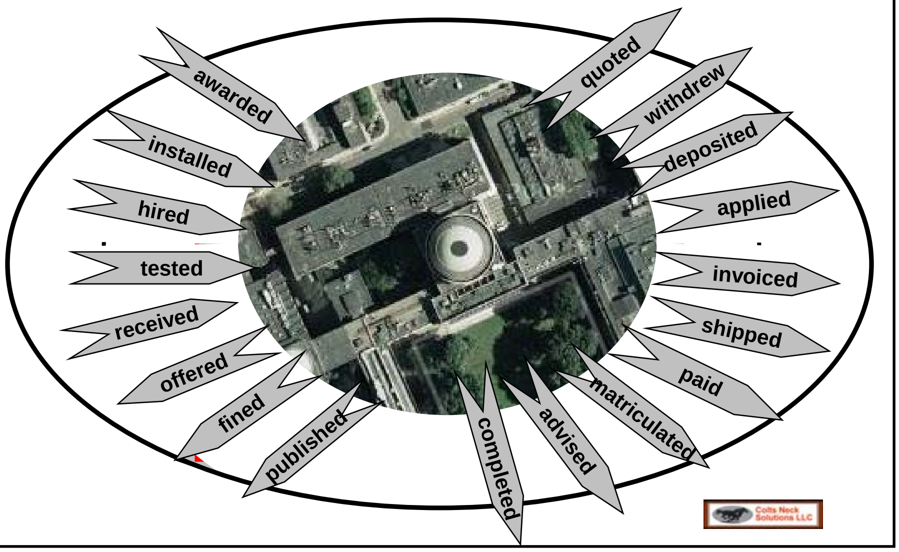
The MIT 2008 Information Quality Industry Symposium

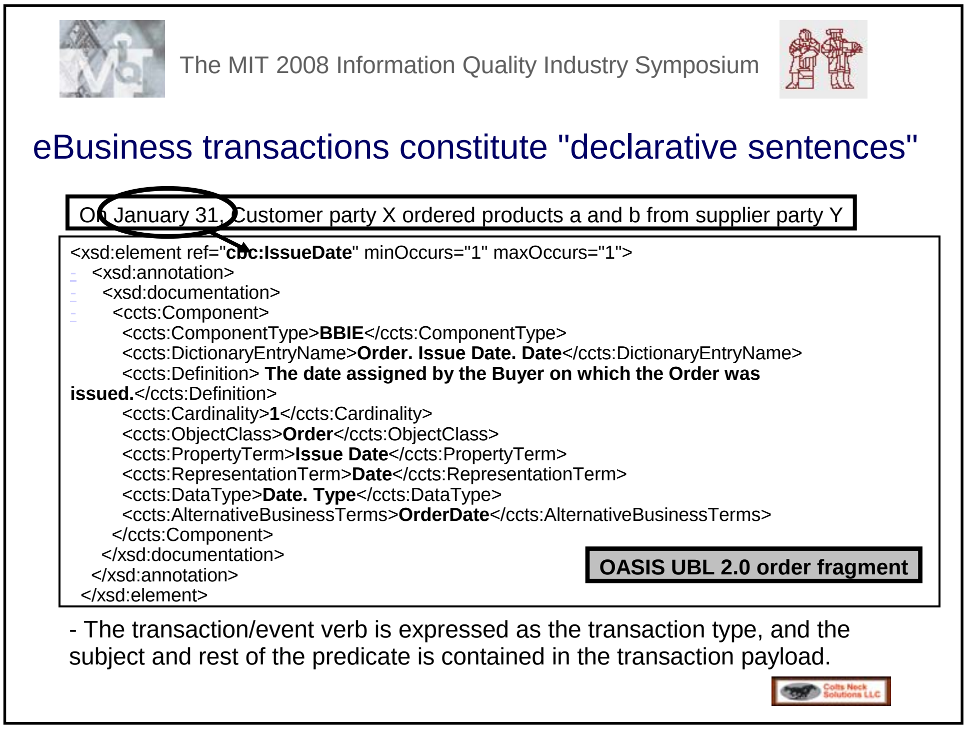
Case 3: eBusiness transaction processes

- **B2B eBusiness transaction flow is inherently "nominalistic"**
  - eBusiness transactions typically are not "native" to any of the parties
  - Each transaction has to "speak for itself" because recipients have little or no access to the sender's systems and data
- **B2B transactions comprise an event-driven "flow"**
  - The data passed is aligned with the eBusiness transaction's action "verb," (such as "ordered", "shipped", "invoiced") etc.
  - Although transactions are standardized, the flow often is not because not all the transactions are used or are used in non-standard sequences
- **An organization's eBusiness transaction "corona" becomes a potentially important source of "nominalistic" transaction data**
  - "You are what you eat:" an entity is defined by its inputs and outputs
  - However, today most eBusiness transactions are treated as perishable, initiated for a specific need and unceremoniously unpackaged on arrival at the "far end"

The MIT 2008 Information Quality Industry Symposium

MIT's nominalist electronic transaction "corona"





The MIT 2008 Information Quality Industry Symposium

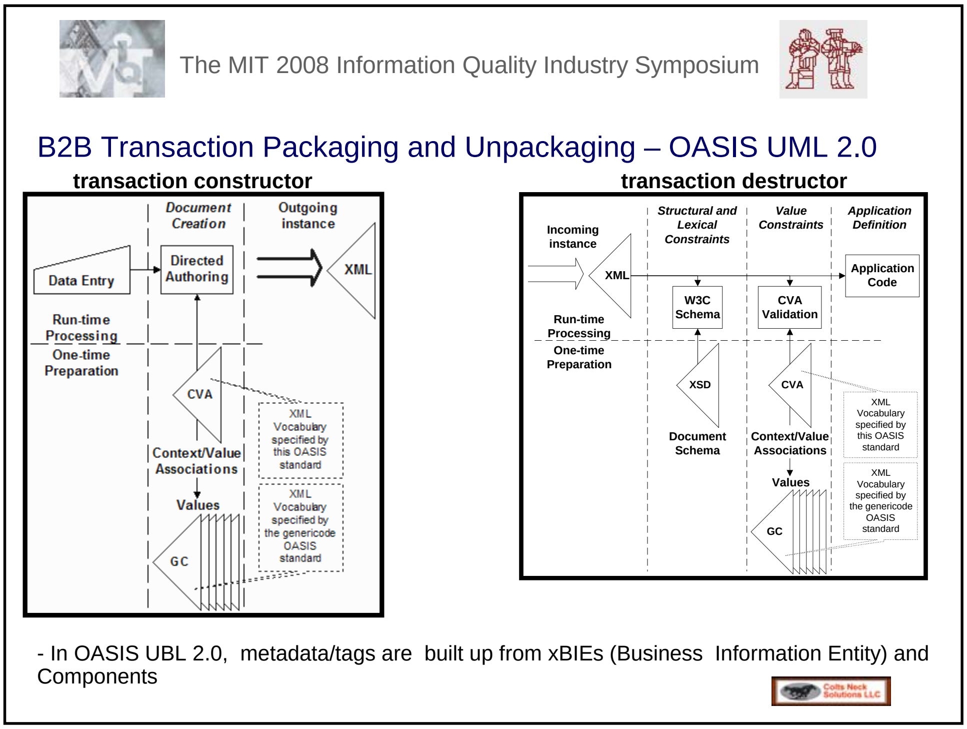
eBusiness transactions constitute "declarative sentences"

On January 31, Customer party X ordered products a and b from supplier party Y

```
<xsd:element ref="cbc:IssueDate" minOccurs="1" maxOccurs="1">
 <xsd:annotation>
   <xsd:documentation>
     <ccts:Component>
       <ccts:ComponentType>BBIE</ccts:ComponentType>
       <ccts:DictionaryEntryName>Order. Issue Date. Date</ccts:DictionaryEntryName>
       <ccts:Definition> The date assigned by the Buyer on which the Order was
issued.</ccts:Definition>
       <ccts:Cardinality>1</ccts:Cardinality>
       <ccts:ObjectClass>Order</ccts:ObjectClass>
       <ccts:PropertyTerm>Issue Date</ccts:PropertyTerm>
       <ccts:RepresentationTerm>Date</ccts:RepresentationTerm>
       <ccts:DataType>Date. Type</ccts:DataType>
       <ccts:AlternativeBusinessTerms>OrderDate</ccts:AlternativeBusinessTerms>
     </ccts:Component>
   </xsd:documentation>
 </xsd:annotation>
</xsd:element>
```

OASIS UBL 2.0 order fragment

- The transaction/event verb is expressed as the transaction type, and the subject and rest of the predicate is contained in the transaction payload.



The MIT 2008 Information Quality Industry Symposium

B2B Transaction Packaging and Unpackaging – OASIS UML 2.0

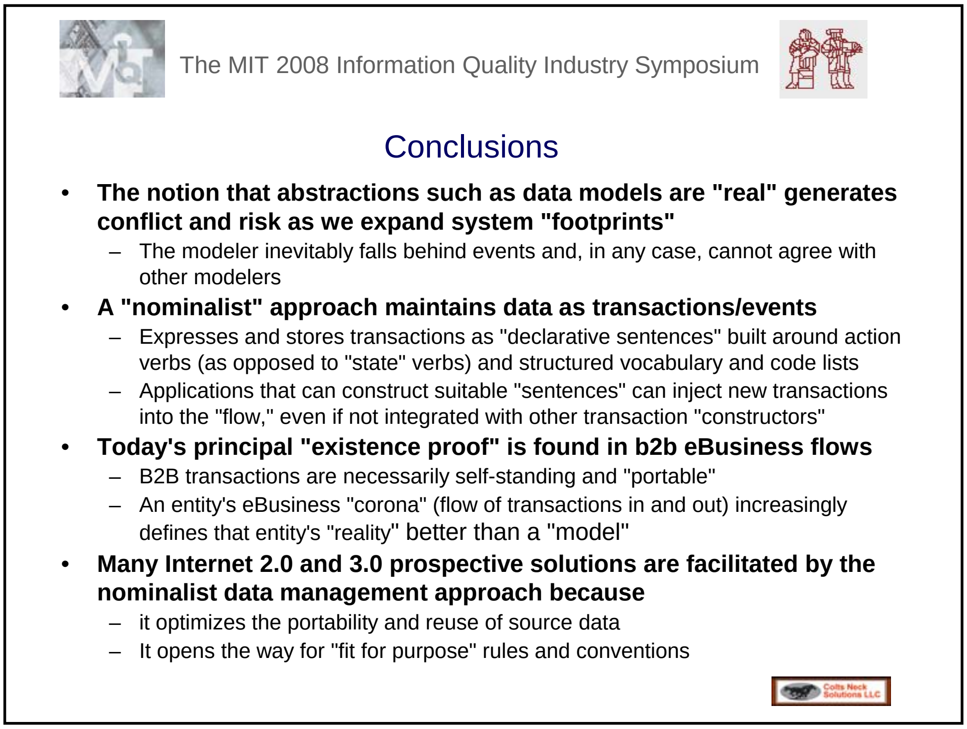
transaction constructor

Document Creation
Outgoing instance
Data Entry
Directed Authoring
XML
Run-time Processing
One-time Preparation
CVA
XML Vocabulary specified by this OASIS standard
Context/Value Associations
Values
GC
XML Vocabulary specified by the genericode OASIS standard

transaction destructor

Incoming instance
Structural and Lexical Constraints
Value Constraints
Application Definition
XML
W3C Schema
CVA Validation
Application Code
Run-time Processing
One-time Preparation
XSD
CVA
Document Schema
Context/Value Associations
XML Vocabulary specified by this OASIS standard
Values
GC
XML Vocabulary specified by the genericode OASIS standard

- In OASIS UBL 2.0, metadata/tags are built up from xBIEs (Business Information Entity) and Components



The MIT 2008 Information Quality Industry Symposium

Conclusions

- **The notion that abstractions such as data models are "real" generates conflict and risk as we expand system "footprints"**
  - The modeler inevitably falls behind events and, in any case, cannot agree with other modelers
- **A "nominalist" approach maintains data as transactions/events**
  - Expresses and stores transactions as "declarative sentences" built around action verbs (as opposed to "state" verbs) and structured vocabulary and code lists
  - Applications that can construct suitable "sentences" can inject new transactions into the "flow," even if not integrated with other transaction "constructors"
- **Today's principal "existence proof" is found in b2b eBusiness flows**
  - B2B transactions are necessarily self-standing and "portable"
  - An entity's eBusiness "corona" (flow of transactions in and out) increasingly defines that entity's "reality" better than a "model"
- **Many Internet 2.0 and 3.0 prospective solutions are facilitated by the nominalist data management approach because**
  - it optimizes the portability and reuse of source data
  - It opens the way for "fit for purpose" rules and conventions

Colts Neck Solutions LLC