

# Toward Distributed Development of Ontologies

Mark A. Musen  
Stanford University

# Manual, top-down development of ontologies is the norm

- Yahoo! taxonomy of Web pages
- Most clinical terminologies
- All the worlds' upper-level ontologies
  - SUMO
  - Open CYC
  - HL7 RIM



**Srinija Srinivasan**

ONTOLOGIST

She may be the best-kept secret at Yahoo!, the company that produces the wildly popular Web search engine. Trained in library science, Srinivasan is the one who decides how the thousands of Web pages submitted to Yahoo! should be categorized and classified, making it as intuitive, expandable and maintainable as possible.

42 NEWSWEEK DEC. 25, 1995/JAN. 1, 1996

# The Open Directory Project

- Started by Netscape in 1999
- “Editors” volunteer to develop categorizations of their domains of expertise
- There are explicit expectations of editors and review of their work
- At present, thousands of editors, yet only 0.5 FTE of paid management!



# Gene Ontology Consortium

- Outstanding example of community-driven ontology development
- Small group of volunteer developers work tirelessly to expand their work
- A new version of GO is released every 30 minutes!
- But there are lots of tensions in the GO community

© Carole Goble

# Prologue

— Carole Goble

Two households, both alike in dignity,  
In fair genomics, where we lay our scene,  
(One, comforted by its logic's rigour,  
Claims ontology for the realm of pure,  
The other, with blessed scientist's vigour,  
Acts hastily on models that endure),  
From ancient grudge break to new mutiny,  
When "being" drives a fly-man to blaspheme.  
From forth the fatal loins of these two foes  
Researchers to unlock the book of life;  
Whole misadventured piteous overthrows  
Can with their work bury their clans' strife.  
The fruitful passage of their GO-mark'd love,  
And the continuance of their studies sage,  
Which, united, yield ontologies undreamed-of,  
Is now the hours' traffic of our stage;  
The which if you with patient ears attend,  
What here shall miss, our toil shall strive to mend.

Based on an idea by Shakespeare

# A Portion of the OBO Library



# Moving from cottage industry to the industrial age

- There must be widely available **tools** that are open-source, that are easy to use, and that adhere to knowledge representation standards
- There must be a large user **user community** of developers who use the tools and who can provide feedback to one another and to the core team of tool builders
- There must be an **organizational structure** to provide a modicum of coordination and quality control
- If all these pieces are in place, people can do great work from their cottages!

# The NCI Thesaurus

Thesaurus Protégé 3.0 beta (file:\C:\projects\owl\Thesaurus.pprj, OWL Files)

File Edit Project OWL Code Window Help

owl:Thing

- Abnormal\_Cell\_Kind
- Anatomy\_Kind
- Biological\_Process\_Kind
- Chemicals\_and\_Drugs\_Kind
- Chemotherapy\_Regimen\_Kind
- Clinical\_or\_Research\_Activity\_Kind
- Diagnostic\_and\_Prognostic\_Factors\_Kind
- Drug\_Mechanism\_of\_Action\_Kind
- Drug\_Physiologic\_Effect\_Kind
- EO\_Anatomy\_Kind
- EO\_Findings\_and\_Disorders\_Kind
  - Experimental\_Organism\_Diagnoses
    - Experimental\_Allergic\_Encephalomyelitis
    - Mouse\_Pathologic\_Diagnoses
      - Mouse\_Cancer-Related\_Conditions
        - Benign\_Plasma\_Cell\_Proliferations\_of
          - Hyperplasia\_of\_the\_Mouse\_Intestinal
          - Hyperplasia\_of\_the\_Mouse\_Pulmonar
          - Melanocytic\_Proliferative\_Disorders\_c
          - Mouse\_Noncancerous\_Conditions
            - Benign\_Conditions\_of\_the\_Mouse
              - Congestion\_of\_the\_Mouse\_In

CLASS EDITOR

For Class: Benign\_Conditions\_of\_the\_Mouse\_Intestinal\_Tract (instance of owl:Class)

Name: Benign\_Conditions\_of\_the\_Mouse\_Intestinal\_Tract

rdfs:comment

Annotations

Property	Value	Lang
D code	C22102	
D DesignNote	Autonomous new grov...	
D Display_Name	Benign Conditions of th...	
D FULL_SYN	<term-name>Benign Co...	
D FULL_SYN	<term-name>Benign Co...	
D hasType	primitive	
D Preferred_Name	Benign Conditions of th...	

Properties and Restrictions

- rEO\_Disease\_Has\_Associated\_EO\_Anatomy (someValuesFrom Gastrointestinal\_Tract\_MMHCC, someValuesFrom
  - Gastrointestinal\_Tract\_MMHCC
  - Digestive\_System\_MMHCC [from Mouse\_Digestive\_System\_Disorder]
- rEO\_Disease\_Has\_Associated\_Cell\_Type
- rEO\_Disease\_Has\_Property\_or\_Attribute
- rEO\_Disease\_Maps\_to\_Human\_Disease

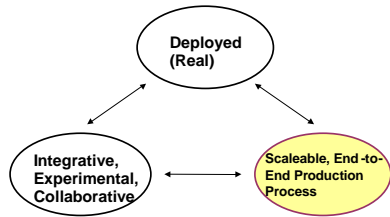
Superclasses

- Mouse\_Noncancerous\_Conditions
- Mouse\_Digestive\_System\_Disorder

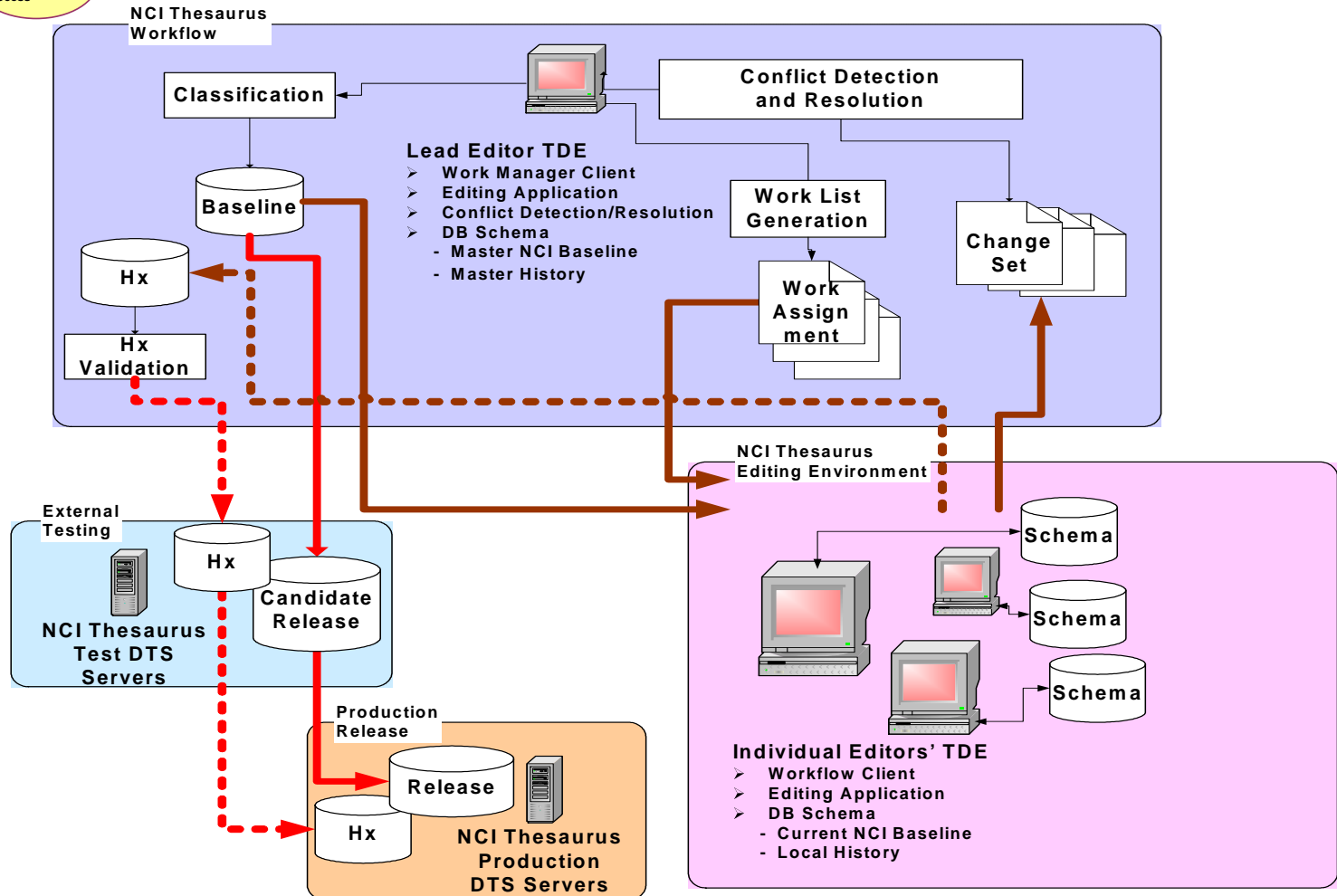
Disjoints

Logic View Properties View

NCI Thesaurus Today:  
Three Critical Pieces



# NCI Thesaurus Production Environment with History



# Egyptian Crop Pests

egypt\_crop\_pests4 Protégé 3.0 beta (file:K:\projects\owl\egypt\_crop\_pests4.pprj, OWL Files)

File Edit Project OWL Code Window Help

owlClasses Properties Forms Individuals Metadata

### SUBCLASS RELATIONSHIP

For Project: egypt\_crop\_pests4

#### Asserted Hierarchy

- owl:Thing
  - أجزاء النبات
  - أنواع التربة
  - الأفات المحصولية
    - الطيور
    - حشائش
    - حشرية
      - الرخويات
      - حشرات آكلة الجذور
      - حشرات آكلة السوق
      - حشرات آكلة للأوراق
      - حشرات العقص
      - حشرات ناقية الخشب
      - حشرات ضارة بالبذور
      - حشرات ماصة
      - حشرات متلفة للثمار
      - حشرات مخربة للأزهار
      - عناكب
      - قوارض
    - مرضية
      - نيماتودا
    - المبيدات الزراعية
    - المحاصيل
      - محاصيل الفلحة

### CLASS EDITOR

For Class: (instance of owl:Class) الأفات المحصولية

Name SameAs DifferentFrom

الأفات المحصولية

rdfs:comment

يقصد بها الحشرات الاقتصادية والعناكب والأمراض النباتية) الفطرية والبكتيرية والفيروسية) والحشائش والنيماتودا والقوارض والطيور والرخويات الضارة بالزراعة.

#### Annotations

Property	Value	Lang
rdfs:comment	والرخويات الضارة بالزراعة...	

#### Asserted

#### Asserted Conditions

NECESSARY & SUFFICIENT

NECESSARY

- owl:Thing
- الأجزاء المعرضة للإصابة أجزاء النبات
- String# الأسم
- String# الأسم العلمي
- String# الأهمية الاقتصادية
- String# الظروف الملائمة لانتشار الإصابة
- النباتات المعرضة المحاصيل
- String# طرق الوقاية والمكافحة
- مناطق الآفة المناطق الزراعية
- نوع الآفة = 1

#### Properties

- أسباب الإصابة (single String)
- أعراض الإصابة (single String)
- الأجزاء المعرضة للإصابة (multiple)
- الأسم (single String)
- الأسم العلمي (single String)
- الأهمية الاقتصادية (single String)
- الظروف الملائمة لانتشار الإصابة (single String)

#### Disjoints

# A thousand flowers are blooming!

- Ontologies are being developed by interested groups from every sector of academia, industry, and government—often “on the cheap”
- Many of these ontologies have been proven to be extraordinarily useful to wide communities
- Many of these same ontologies have been shown to have structural flaws
- We finally are at the stage where we have open-source tools and standard representation languages that can help us to create durable and maintainable ontologies with rich semantic content

# Research Questions

- How do we measure the “quality” of ontologies developed via industrial engineering and by grass-roots efforts?
- Can we perform controlled studies to compare the cost-effectiveness of top–down versus bottom–up approaches?
- Can we evaluate the cost-effectiveness of alternative management structures to facilitate distributed ontology development?
- Can we develop correlates of *peer review* to ensure open quality control of ontologies created in a distributed fashion?