Protégé-2000: An environment for knowledge-base systems

John Gennari
Biomedical & Health Informatics
University of Washington
Generic Protégé definition

Protégé is...
Generic Protégé definition

Protégé is...

- A methodology for building knowledge-based systems:
  - Ontology development
  - Knowledge base development (Frame-based)
- A tool for building knowledge-acquisition tools (a meta-tool)
  - Ontology editor
  - Knowledge-base editor
  - KA-tool generator and extensible UI
Protégé history

- Protégé (Musen 1986)
- Protégé-II (Eriksson 1991)
- Protégé/Win (1996)
  - Open source licensing
  - More than 6500 unique downloads world-wide
  - More than 1000 Protégé mailing list members
Example

(Live demo of Protégé, showing default UI and KA-tool generation)
Protégé-2000 architecture

User Interface

Protégé Default User Interface

Slot Widgets

Tab Widgets

Other KB Applications

Core Protégé-2000

Protégé Knowledge Model API

Protégé Knowledge Model

Mappings for read/write functions

Persistent Storage

CLIPS

RDF

Other files

Generic Schema

Other schema

Flat file storage

RDBMS storage
The Protégé knowledge model

- The Protégé knowledge model is frame-based:
  - Classes, arranged in a hierarchy
  - Instances
  - Slots (attributes of instances)
  - Facets (constraints on slots)
- Inheritance: slots are inherited from superclass to subclass (multiple parents are allowed)
- Meta-classes: classes may have other classes as instances
- A slot is a frame; it may be attached to more than one class
Frames, individuals, slots, facets

The universe is a set of Frames:

- Classes
- Individuals
- Facets
- Slots

(Picture adapted from Chaudhri et al, AAAI ‘98 OKBC paper)
Template slots vs. own slots

Has-subclass link: template slots are inherited

Meta-class “Animal”
Own slots:
S1: value, V1
Template slots:
T1: type integer
T2: allowed class C2

Class “Dog”
Own slots:
T1: 3
T2: instance1 of C2
Template slots:
Weight: type float

Individual “Fido”
Own slots:
Weight: 23.5

Class “Greyhound”
Own slots:
T1: 14
T2: instance2 of C2
Template slots:
Weight: type float
Speed: type float

Individual “Speedy”
Own slots:
Weight: 45.2
Speed: 30.54

Has-instance link: template slots become own slots

SW seminar, Feb 26, ’03
What about inference?

- Protégé doesn’t do inference
What about inference?

• Protégé doesn’t do inference
• Exceptions: Inheritance, inverse slots, defaults, facet constraints
• Declarative vs. Procedural knowledge
  – A KB of frames
  – A problem-solving method (PSM) to do inference or specialized querying
• A PSM for description logic on top of Protégé (?)
Feb 10, 2003:
“I would like to introduce a new member of the Protégé group . . . He will be concentrating on making Protégé a better tool for the Semantic web. He will initially be focused on integrating Protégé with description-logic systems and classifiers, and, in particular, with the OWL language.”
More demos

- Big ontologies
  - The Gene Ontology
  - The Foundation Model of Anatomy (FMA)
- Use of API: the FM Explorer
- Tab plug-ins
  - Eligibility screening
  - Prompt: Ontology merging
  - DataGenie: Importing from relational DBs
- Backend plug-ins
  - RDFS
  - XML Schema
Demo: The FMA

- The Digital Anatomist Project: Building a KB for teaching (and standardizing) human anatomy
- The Foundational Model of Anatomy (FMA)
- Uses a relational database system as a back-end
  - 66,766 anatomic concepts
  - More than 171,000 frames, 1.4 million records
  - Use of RDBMS is transparent to the user
- Extensive use of meta-classes as templates
Example

(Live demo of FMA)
Demo: The FM Explorer

- A stand-alone web application
- Simple viewing only
- Only uses core Protégé API
Anatomical structure, which consists of the maximal set of organ parts so connected to one another that together they constitute a unit of macroscopic anatomy, structurally distinct from other such units. Together with other organs, an organ constitutes an organ system or a body part. An organ is divisible into organ parts but not organs. Examples: femur, biceps, liver, heart, skin, tracheobronchial tree, sciatic nerve, ovary.
Demo: Prompt

- A tab plug-in developed by Natasha Noy (see AAAI ’99, AAAI ’00, IJCAI ’01)
- An instance of Bernstein’s *model management*
- Goal is to merge ontologies
  - Semi-automatic
  - A hybrid of linguistic matches and structure (graph) matching
Demo: Eligibility Screening

- Simple ontology: eligibility criteria and clinical-trial protocols
- Simple problem-solving method: screen patients
- Developed as a Protégé tab plug-in
- Installed at the UCI National Chao Cancer Center, February ‘00
  - Goal: Increase patient enrollment
  - Primary users: oncology nurses & staff from the clinical research office
Demo: The Gene Ontology

- A controlled vocabulary to facilitate queries across genetic DBs
- Three organizing principles:
  - molecular function
  - biological process
  - cellular component
- Geneontology.org provides several browsers
- Partial port to Protégé:
  - Fly and Yeast databases of Genes
  - > 7800 subclasses of function, process & components
Demo: RDFS support

- Back-end plug-in for Protégé-2000 file access
- Reads and writes
  - ontologies as RDF-schema documents
  - knowledge-base instances as RDF files
- Currently supports RDF semantics compatible with those of OKBC
Demo: XML Schema support

• Back-end plug-in for Protégé-2000
• Reads and writes XML Schema
• Uses a single schema that captures the Protégé knowledge model: Classes, slots, individuals
Summary & conclusions

• Protégé provides:
  – An extensible knowledge-base development environment
  – A methodology for building knowledge-based solutions
  – A clear specification of its representational semantics (the frame-based Protégé knowledge model)
  – A variety of persistent storage options

• In sum, a platform of un-paralleled flexibility for knowledge-base system development
Shameless plug

- Knowledge representation and applications, MEDED 598, 3 credits, MW 2-3:20
- See faculty.washington.edu/gennari/#teaching
- Issues:
  - What is a knowledge representation?
  - Why are issues in knowledge representation important for application builders?
  - What is the relationship between knowledge and data, between knowledge bases and data bases?
  - What is the tradeoff between expressivity of the KR and tractability of inference?
Acknowledgements

• The current Protégé team:
  Monica Crubezy       Mark Musen
  Ray Fergerson        Natasha Noy
  Jennifer Vendetti    Samson Tu

• FMA team: Cornelius Rosse, Onard Mejino, Augusto Agoncillo, Todd Detweiler, Kurt Rickard

• Gene Ontology: www.geneontology.org/

• Protégé: www.smi.stanford.edu/projects/protege/

• Me: faculty.washington.edu/gennari/