Object Modeling for Semantic Meaning

Baris E. Suzek, Hongzhan Huang, Hsing-Kuo Hua, Peter McGarvey, Cathy H. Wu
Protein Information Resource, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC, 2007

Protein Information Resource (PIR)

Iterative Information Model Development

First Model
- Problems:
  - Record modeling: Limited semantics and scientific meaning
  - XML schema constraints: A lot of objects without attribute
  - Unidirectional associations, association names missing: Limited navigability

Second Model
- Problems:
  - Semantics needed improvement; better class naming, removal of type attributes
  - Attributes with complex data types such as List, Collection

Towards Final Model

Removing Complex Data types

Improving semantics
- Using "type" hides the semantic value such Protein annotation
- Can't meet use cases such as:
  - Find me all the proteins which "calcium Binding Region" or "Zinc Finger Region"

Towards Final Model

Improving semantics

- One object for each Protein Feature type

Final Model

Protein and Gene

Taxonomy

Conclusion
- Better advertisement and discovery of gridPIR data service
- gridPIR as a Gene/Protein resource
- gridPIR as a Protein Annotation resource
- gridPIR as a Taxonomy information resource
- Better understanding of scientific content of the service
- Better usability:
  - Retrieve proteins for a gene
  - Retrieve sequence for a protein
  - Retrieve proteins containing calcium binding sites
  - Retrieve proteins containing zinc finger regions
  - Retrieve proteins for a taxon in taxonomy tree