Community Data Annotation/Curation
Community Annotation/Curation

Demo Project
- Open atlas
  - Individuals
  - Populations (??)

Success criteria
- Acceptance and participation by anatomy community
- Portability of tools to other projects
- At least one “good” atlas

Project cycles
- Identify customers (anatomists) and customer’s customers (radiology, surgery, algorithm developers, educators)
- “Extreme” approach, “release early, release often"

Feasibility studies
- Pick two anatomical areas (thorax, brain)

Deliverables
- Infrastructure/process
- Distributed atlas

Integration needs
- Visualization
- Federated database
- Ontologies

Issues
- Intellectual property
- Business model
Open Atlas: Requirements

Open data and open process
Customer GUI application
Software Toolkit
Methods for curation
Mechanism for consensus building
Mechanisms for quality control
Continuous process feedback
Provenance
Soup to nuts software
  • Reference implementation
    ▪ Visualization
    ▪ Editor
    ▪ Registration, model extraction, etc.
  • Query application
Outreach to customer’s customer
Local and web based
Open Atlas: Components

User interface
Segmentation tools + manual correction
Interface to multiple ontologies
Revision control
Automated quality assurance
Dashboards
Packaging/delivery
Data repository
API for programmatic access to data/annotations/tools
Core team
- Anatomists/Radiologists (Domain expert)
- Database design
- Ontology support
- Image analysis
- Image/Geometry editor
- Process support tools
Starting Points

U Wash FMA

NLM Visible Human Thorax
  • Original from EAI
  • Enhanced by Virtual Soldier Project

Brigham and Women’s Brain Atlas/Slicer
Community Data Annotation/Curation
Background Slides:

Open, Distributed and Collaborative Data Annotation

Bill Lorensen
Insight Software Consortium
Motivation

Many imaging communities are data starved

• Algorithm developers
• End users

Lots of raw data, but very little annotated data

• LIDC
• Notre Dame Biometrics Data Distribution
Forms of Annotation

Anatomy labels
Contours
Statistical
Anatomical landmarks
Templates
Ground truth
Problem Statement

Sensors are producing large amounts of data
Annotation adds value
Annotation of large data collections is expensive and error prone
Customers

Algorithm developers
Anatomists
Teachers
Sensor manufacturers
Solution

A distributed, coordinated community can efficiently and economically annotate large sets of data

- wikipedia
- wikimapia

Extreme programming techniques can be applied to the data annotation process
Examples

Anatomical atlases

Face recognition
- 2D photos
- 3D range data
Example – FBI Facial Reconstruction

Two data collections

• 300 CT datasets of heads
• 1000 photo and range data of faces

Challenge

• Extract models of eyes, noses and mouths from range data
• Replace eyes, noses and mouths in CT data with range data models
Face Template

Photo

Range Data
Mouth
Multidisciplinary Project

Image Analysis
Anatomy
Databases
Ontologies
Software Engineering
Quality Assurance
Visualization
Menu for Success

A Community with a common vision
A pool of talented and motivated developers/scientists
A mix of academic and commercial
An organized, light weight approach to product development
A leadership structure
Communication
A business model

Adopted from “Open Source Menu for Success”
Leadership Structure

Follow NCBC model

Algorithms
• Ontology creation
• Image analysis

Engineering

Driving Projects
• Open Atlas
• Radiology ground truth
Business Model

All core technology is open, without restriction
All NLM supported annotation is open, without restriction
Proprietary enhancement of annotated data is allowed
Annotated data can be used in commercial products without restriction
Guiding Principles
Extreme Data Annotation

The community owns the data

Although the origin of the data is retained, others are free to correct defects and enhance each other's data.

In the end, all of the data should appear as though one person annotated it.
Extreme Data Annotation

Release early, release often

Although people are tempted to keep their data under wraps until it is perfect, the process encourages them to release their data as soon as it passes some minimum quality control tests. The longer the data is visible to the community, the better integrated it will be.
Continuous integration

There is no scheduled porting to databases or model formats.

All new data is integrated into supported databases and data formats continuously.
Everyone agrees to keep the data free of defects

Although everyone is encouraged to submit their data early, the data must pass quality tests and integration tests nightly. A continuous QA process sends e-mails to people who check in data that does not meet quality control tests. More effectively, the community enforces the commitment though peer pressure.
## Software/Data Analogies

<table>
<thead>
<tr>
<th>Software</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>Annotated data</td>
</tr>
<tr>
<td>Text editor</td>
<td>Image editor</td>
</tr>
<tr>
<td>Compilation error</td>
<td>Collisions</td>
</tr>
<tr>
<td>Compilation</td>
<td>Model creation</td>
</tr>
<tr>
<td>Style</td>
<td>Ontology</td>
</tr>
</tbody>
</table>
Why NLM?

NLM produces, collects, annotates, stores and distributes data

• Medline
• Visible Human Project
• Mayo Data Collection

NLM has managed distributed, collaborative, multidisciplinary projects

• Insight Toolkit
• HPCC Internet 2
What is needed?

Select a pilot project
  • Open Atlas Project

Select customers

Select core team
  • Anatomists
  • Database design
  • Ontology support
  • Image analysis
  • Image/Geometry editor
  • Process support tools
Open Atlas Project

Create anatomical atlases from cross-sectional image data
Semi-automatic and manual labeling of structures
Engage the anatomy community