Social-Ecological Systems Data Management

Beth Plale
Professor, School of Informatics and Computing
Director, Data To Insight Center
Indiana University
SocioEcoInfo Team at IU

Professors
• Beth Plale
• Xiaozhong Liu, School of Library and Information Science
• David Leake, School of Informatics and Computing

Postdocs
• Scott Jensen, Data To Insight Center, now at San Jose State
• Miao Chen, Data To Insight Center
• Inna Kouper, CLIR Fellow in Data Curation

Other
• Umashanthi Pavalanathan, student, now at Georgia Tech
• Julie England, Ostrom Workshop on Political Theory and Policy Analysis

The work presented here reflects the research of this group.
Motivation

“The ecological and social sciences have developed independently and do not combine easily” – Elinor Ostrom
What is the Problem?

- Social Ecological Systems are complex adaptive systems
  - Compartmentalized research
- Data collection requires extensive field work
  - Gaining trust of participants
  - Understanding formal and informal institutions involved
- Long-tail science data
  - Heterogeneous data sets
  - Local data collection, methods

“Socioeco Informatics”

Adapted from Krister Andersson
SES Framework

“a transdisciplinary language to simplify the analytical task confronting those trying to understand complex policy settings”

What is SES Framework?

e.g., Forests, fisheries, grazing, irrigation systems for agriculture

Unpacking SES Framework

Governance systems (GS)
- GS1 Government organizations
- GS2 Nongovernment organizations
- GS3 Network structure
- GS4 Property-rights systems
- GS5 Operational rules
- GS6 Collective-choice rules*
- GS7 Constitutional rules
- GS8 Monitoring and sanctioning processes

Social, economic, and political settings (S)

Resource units (RU)

Governance system (GS)

Interactions (I)

Users (U)
- U1 Number of users*
- U2 Socioeconomic attributes of users
- U3 History of use
- U4 Location
- U5 Leadership/entrepreneurship*
- U6 Norms/social capital*
- U7 Knowledge of SES/mental models*
- U8 Importance of resource*
- U9 Technology used

Related ecosystems (ECO)
SES (model) ↔ SocioEcoInfo Technology “Harness”

**Governance systems (GS)**
- GS1 Government organizations
- GS2 Nongovernment organizations
- GS3 Network structure
- GS4 Property-rights systems
- GS5 Operational rules
- GS6 Collective-choice rules*
- GS7 Constitutional rules
- GS8 Monitoring and sanctioning processes

**Users (U)**
- U1 Number of users*
- U2 Socioeconomic attributes of users
- U3 History of use
- U4 Location
- U5 Leadership/entrepreneurship*
- U6 Norms/social capital*
- U7 Knowledge of SES/mental models*
- U8 Importance of resource*
- U9 Technology used

**Semantic integration**
- Site XML objects
- Classification/Machine learning
- Ontologies

**Faceted search**
- Research Object

**Related ecosystems (ECO)**
- Interactions (I)
- Outcomes (O)
SES Coding categories

Social, Economic, and Political Settings (S)

Resource System (RS)
- RS1- Sector (e.g., water, forests, pasture, fish)
- RS2- Clarity of system boundaries
- RS3- Size of resource system
- RS4- Human-constructed facilities
- RS5- Productivity of system
- RS6- Equilibrium properties
- RS7- Predictability of system dynamics
- RS8- Storage characteristics
- RS9- Location

Resource Units (RU)
- RU1- Resource unit mobility
- RU2- Growth or replacement rate
- RU3- Interaction among resource units
- RU4- Economic value
- RU5- Size
- RU6- Distinctive markings
- RU7- Spatial and temporal distribution

Governance System (GS)
- GS1- Government organizations
- GS2- Non-government organizations
- GS3- Network structure
- GS4- Property-rights systems
- GS5- Operational rules
- GS6- Collective-choice rules
- GS7- Constitutional rules
- GS8- Monitoring and sanctioning processes

Users (U)
- U1- Number of users
- U2- Socioeconomic attributes of users
- U3- History of use
- U4- Location
- U5- Leadership/entrepreneurship
- U6- Norms/social capital
- U7- Knowledge of SES/mental models
- U8- Dependence on resource
- U9- Technology used

Action Situation

Interactions (I) → Outcomes (O)
- I1- Harvesting levels of diverse users
- I2- Information sharing among users
- I3- Deliberation processes
- I4- Conflicts among users
- I5- Investment activities
- I6- Lobbying activities
- O1- Social performance measures
- O2- Ecological performance measures
- O3- Externalities to other SESs
- O4- Social-ecological equilibrium

Related Ecosystems (ECO)
- ECO1- Climate patterns.
- ECO2- Pollution patterns.
- ECO3- Flows into and out of focal SES.
Int’l Forestry Resources and Institutions (IFRI) database: collection

- Data collected in 23 countries by 13 Collaborating Research Centers.
- Data collected using research instrument with 10 forms packages, totaling 180 pages, with some packages used up to 30 times or more per visit.
IFRI database: data

• Over 18 years longitudinal data on forest resources, use, and governance
• Consists of 346 separate site visits
• Relational database captures relations between data collection packages
• Responses to each question are a column in the database; 922 questions in the IFRI database
Number of site visits by country

Site visits / Revisits

Site elevation (meters)
Age of forests (years)

Who planted the forest

Policies impact on governance
Site Analysis 1996

Legend:
- Green: Forest
- Orange: Non-Forest
- Blue: May Creek Property
- White: Ownership in 1997

Legend for Stem count / hectare:
- High
- Low

Legend for Basal Area / hectare:
- Small
- Large

Created by: Wenjie Sun
Data source: GPEC Biocomplexity and Indiana Project
Site Analysis 2000
Site Analysis 2004

Legend
- Green: Forest
- Orange: Non-Forest
- Blue: May Creek Property
- White: Ownership in 1997

Stem count / hectare
- High
- Low

Basal Area / hectare
- Small
- Large

Created by: Wenjie Sun
Data source: GPEC Biocomplexity and Indiana Project
Site Analysis 2009

Legend
- Forest
- Non-Forest
- May Creek Property
- Ownership in 1997

Stem count / hectare
- high
- low

Basal Area / hectare
- small
- large
IFRI data densities seen through SES ordered heatmap

Data Density (Z-Score)

-2 0 2 4 6 8

IFRI map to SES for data discovery

SES Category

Study sites

Resource System

Resource Units

Governance System

Users

Interactions

Outcomes

Social Economic Political Settings
SocioEcoInfo Research Goals

• **Data Preservation**: stable, non-proprietary representation of IFRI (and more broadly social-ecological systems) data

• **Access and discovery**: SES leveraged semantic and metadata harness for data sharing, data access, semantic linkages of social ecological systems data

• **Constraints**: sensitivities of data: location, species, organizational restrictions, admissions of use that could compromise arrangements
IFRI Site Visits

- Data gathered by visit to a site
- Survey instrument used
- Questions asked in four SES areas: RU, RS, GS, U

Post visit analysis reveals interactions and outcomes. Power of data over time comes in its predictive ability.
Data Documentation Initiative (DDI) as solution to metadata description

- DDI codebook is most relevant to the IFRI survey instrument
- DDI can describe the structure of the instrument
  - Structure could be used to identify/categorize the questions in the IFRI database for machine learning
- DDI can be useful in access/discovery work to distinguish between instruments. But need something else to capture what is needed for access and discovery of data itself. Hence socioecoinfo semantic/metadata harness.

http://www.ddialliance.org
DDI applicability, cont.

DDI is targeted towards coded data.

Over 50% of IFRI data is nominal questions and 23% are numeric, so these could be represented as coded data

But some of richest information is in free text questions

16% of questions in research instrument
Preservation: archivable research object

- Goal: preservable database snapshot: Archivable Research Object
- Context:
Preservation: archivable research object

• Mental framework for thinking about IFRI archivable research object: Sustainable Environments Actionable Data (SEAD) project
  – $8,000,000 project, 2012-2016
  – UMichigan, Margaret Hedstrom, PI
  – IU, Beth Plale, co-PI, and Robert McDonald
  – UIUC, Praveen Kumar, co-PI, and Beth Sandore
Component Interactions: VA and IRs

Successful *automatic* ingest into UIUC IDEALS repository
Archivable IFRI Research Object

• Periodic snapshot of database
• Snapshot is single Research Object that contains:
  – One self-describing XML file per site visit
    • Unique ID is triple: siteID/countryID/researchorgID
  – Script that guides reconstitution of DB
    • Script able to harvest DB structure from XML site visit files
  – Set of relationships
    • Relate each site visit to snapshot
    • Relate script to snapshot
    • Relations across XML files within snapshot
  – Metadata: DDI for instrument, relationships, SES categorization
Access and Discovery of SES Data

• Increased emphasis on making social-ecological data reusable

• Research is on-going, involving longitudinal studies, data formats evolve and need to be migrated

• Researchers need an accessible, non-proprietary, stable data format that can be used without database training

• Can social-ecological data sources be represented in a way that generalizes across such cross-domain data sources, but still carries useful information?
Map of Research Instruments to Objects Based on the SES Framework

<table>
<thead>
<tr>
<th>Research Instrument</th>
<th>SES Subcategories</th>
<th>Logical Object Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site Visit</td>
<td></td>
<td>Logical Object: Site Visit</td>
</tr>
<tr>
<td>Forest A</td>
<td></td>
<td>Logical Object: Forest A</td>
</tr>
<tr>
<td>How old is forest</td>
<td></td>
<td>RS4 Was forest planted</td>
</tr>
<tr>
<td>Was forest planted</td>
<td></td>
<td>RS5 How old is forest</td>
</tr>
<tr>
<td>Describe forest history</td>
<td></td>
<td>RS9 Describe forest history</td>
</tr>
<tr>
<td>Reforestation</td>
<td></td>
<td>Topology</td>
</tr>
<tr>
<td>Topology</td>
<td></td>
<td>Topology</td>
</tr>
<tr>
<td>Change in tree density</td>
<td></td>
<td>O3 Change in tree density</td>
</tr>
</tbody>
</table>

| Forest B            |                   | Logical Object: Forest B |
| How old is forest   |                   | RS4 Was forest planted |
| Was forest planted  |                   | RS5 How old is forest     |
| Describe forest history |               | RS9 Describe forest history |
| Reforestation       |                   | Topology               |
| Topology            |                   | Topology               |
| Change in tree density |              | O3 Change in tree density |
From Database to Logical Objects

- XML representation must be usable without “reconstituting” the relational database
- Identifying a top-level object for generating the XML representation
  - Table without foreign key relations to it
  - In IFRI this is the site visit
Data Characteristics

<table>
<thead>
<tr>
<th>SES Framework 2nd-Tier Categories</th>
<th>Nominal</th>
<th>Text</th>
<th>Integer</th>
<th>Doubles</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS</td>
<td>142</td>
<td>24</td>
<td>14</td>
<td>1</td>
<td>181</td>
</tr>
<tr>
<td>I</td>
<td>92</td>
<td>9</td>
<td>8</td>
<td>5</td>
<td>114</td>
</tr>
<tr>
<td>O</td>
<td>38</td>
<td>20</td>
<td>2</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>RS</td>
<td>18</td>
<td>15</td>
<td>7</td>
<td>18</td>
<td>58</td>
</tr>
<tr>
<td>RU</td>
<td>29</td>
<td>10</td>
<td>13</td>
<td>7</td>
<td>59</td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>U</td>
<td>73</td>
<td>27</td>
<td>66</td>
<td>8</td>
<td>174</td>
</tr>
<tr>
<td>Total</td>
<td>393</td>
<td>105</td>
<td>111</td>
<td>39</td>
<td>648</td>
</tr>
</tbody>
</table>

- Of 922 questions in IFRI, 648 are mapped to the SES and approximately 30% capture metadata.
- Roughly ½ of the questions have enumerated results stored as an integer.
- The column names do not capture the semantics of the research questions, but represent a code for the corresponding question in the research instrument.
Automated Classification of Survey Questions to SES

• Categorization of each question in instrument to SES Framework in automated way

• Carry out word frequency calculation on instrument questions and SES categories. Used this to cluster questions.

• Apply machine learning to avoid need for learning data set
  – Examined decision tree classifiers, Naïve Bayes, support vector machine (SVM). Decision tree classifiers performed best*.

Graph-based Navigation: Country
Discovery through faceted search

This research is funded by the Indiana University, USA with the grant support of the Ford Foundation, USA.
Select Bolivia; facets change
Detail of study reflects data sensitivities
Ontological tools to integrate heterogeneous resources

• Six types of resources important in social-ecological research
  – Domain concepts (e.g. SES semantic wiki)
  – Theories & models (e.g. Ostrom’s SES framework)
  – Research data sets (e.g. IFRI, DDI data sets)
  – Publication (e.g. Chhatre & Agrawal (2007)*)
  – Instrument (e.g. IFRI survey questionnaire)
  – Research design (e.g. in

• The resources are actually related, and ontologies used to model the connections.

Resource types in ontology

Socio-Ecological Ontology

- Domain concepts
- Classes
- Instances

Related Data Instance

- SES framework
- Semantic Wiki
- Theories & Models
- IFRI Access Database
- IFRI Xml files
- Research Data Set
- Object-based DB (Mongo DB)
- Research Design
- Research Data Set
- Object Visit Objects
- Forest Object
- ...
An example of the linkage

Chhatre & Agrawal’s (2007) paper, addressing tradeoff and synergies of carbon storage and livelihood

Carbon Storage concept

Livelihood concept

Measured from

use

write

contains

Author: A. Chhatre

Author: A. Agrawal

IFRI
Summary

• Archival achieved through periodic snapshot of RDBMS. Each snapshot yields a Research Object
• Data in IFRI DBMS categorized against the SES Framework using manual categorization and machine learning
• SES Framework: provides a general framework around which other SES datasets (and their survey instruments) can be understood.
• Held two usability sessions with social ecological researchers in April 2013
• Ongoing work: assess what SEAD can provide for archiving and access
Funded in part by the National Science Foundation and Pervasive Technology Institute