Generalized Representation and Mapping for Social-Ecological Data: Freeing Data from the Database

Scott Jensen[1], Beth Plale[1], Xiaozhong Liu[2], Miao Chen[1], David Leake[1], Julie England[3]
[1]School of Informatics and Computing
[2]School of Library and Information Science
[3]Vincent and Elinor Ostrom Workshop in Political Theory and Policy Analysis
Indiana University
Bloomington, Indiana U.S.A
{scjensen, plale, liu237, miaochen, leake, england}@indiana.edu

***Abstract*— Scientific discovery increasingly requires collaboration between scientific sub-domains that often have different representations for their data. To bridge gaps between varying domain representations, researchers are developing metadata and semantic representations meaningful to broader communities. Through exploiting these representations we propose a logical model and architecture by which cross-domain researchers can more easily discover, use, and eventually archive, data. In this paper we present an architecture, intermediate data model, and methodology for mapping diverse social-ecological data sources stored in relational databases to a common representation, and for classifying textual data using machine learning. The results are visualized through client views that are built against the general logical model, and applied against a longitudinal database from social-ecological research.**

Keywords: scientific data; sustainability; data model; machine learning

I. INTRODUCTION

Domain science research communities are actively undertaking the development metadata and semantic representations that are meaningful to them. A computer science problem emerging as a result of these developments is the generalization and enhancement of these representations in new ways from which the domain researchers can benefit.

The representation of scientific data in a unifying framework through which heterogeneous scientific studies can be related is particular pressing in light of recent trends, including:

- An increased emphasis on reuse of scientific data both from scientific communities and from agencies that fund research,
- An increased need to integrate data across scientific disciplines; particularly in addressing global issues such as climate change,
- Recognition that the data deluge is not only from large data sets, but also from a plethora of smaller, heterogeneous data sets in the long tail of science that are irreplaceable and involve considerable effort to collect,
- Recognition of the need for stable representations of data that can capture data in a common format that is accessible to future researchers without requiring proprietary software.

A field whose data exemplify these issues is sustainability science. Sustainability science is inherently cross-domain, involving both the social and physical sciences. Social-ecological research, a field within sustainability science, studies complex human-natural resource systems - for instance common pool resources such as forests, fisheries, and water systems for irrigation. As noted by Ostrom in [16], *"the ecological and social sciences have developed independently and do not combine easily"*. Additionally, the data collected involves extensive field work to gain the trust of participants and knowledge of the institutions involved [18].

Sustainability scientists have identified the need for a consistent set of classifications to enable the cross-discipline examination of complex social-ecological systems. The SES Framework [15][16] pioneered by Lin Ostrom, 2011 Nobel Laureate in Economics, which is such a framework, has seen increasing uptake for describing resource systems and the relationships between their components. The SES Framework intuitively captures the relationship between a common pool resource, the governance mechanisms that impact it, and the social structures that interact with it. By capturing these key elements, characterizations in the framework enable reasoning about actions on the common pool resource (by governance and user groups) and their potential outcomes. Characterizations in the SES Framework facilitate comparing one common pool resource, such as a fishery, with another, such as a forest, and drawing general conclusions that would not be possible without the abstraction provided by the SES Framework.

The contribution of this paper is an architecture, intermediate data model, and methodology for mapping social-ecological data from a relational database, to the higher level model of social ecological system interaction embodied in the SES Framework. The approach, as outlined in Fig. 1, has the following strengths:

This work funded in part by a grant from the National Science Foundation OCI, #1058452.

978-1-4673-4466-1/12/$31.00 ©2012 IEEE

1. Stable representation for social-ecological data, because it is based on the accepted SES Framework and does not require proprietary database software, better positioning the data for archiving in community repositories,
2. Techniques for automated mapping of a database to a logical object representation,
3. A machine learning approach to mapping data in an existing database to the logical object form and the SES Framework, and
4. Using the stable representations generated, client views can easily be created. We discuss two here: a data discovery tool using heatmaps, and a browse tool.

This paper applies our methodology and approach to relational data from a long-term collaborative research network, the International Forestry Resources and Institutions (IFRI) research program, which studies the relationship between forests, humans, and global climate change. The IFRI network consists of centers in 11 countries with two additional countries being established [10,22]. The IFRI database captures over 18 years of longitudinal data on forest resources, their use, and governance for forests worldwide, including 346 visits to sites. The database contains a largely undertapped wealth of data on resource governance [6].

The remainder of this paper is organized as follows. We define terminology and a representational formalism in Section II. Section III introduces a representation of the logical object, and in Section IV a mapping methodology is given to map diverse social-ecological relational databases to logical objects using automated and semi-automated approaches. Section V applies machine learning to classifying the data making up logical objects, and does the classification in terms of the SES Framework. Section VI shows useful tools that can be deployed against the logical object representation. Section VII gives related work and Section VIII concludes the paper with future work.

II. Logical Object Data Model

At the core of our generalized representation is the logical object. The "*logical object*" is a unit of representation that is coherent, and represents the broadest concept in the researcher's model of their data. As is familiar from knowledge representation research, mappings of a database to logical objects must respect the task-relevant conceptualizations of users in order to be useful. In the IFRI domain, for example, arguments could be made for mapping different aggregations of the data as a logical object, to serve particular uses of the data. However, for the approach to be generalizable across databases to enable cross-domain access—our key goal—our focus is on how each data source identifies logical objects sufficiently broad to have cross-domain relevance, yet sufficiently narrow to carry useful information.

The study of social-ecological interactions of resource systems is done using a *research instrument*, which often in IFRI is a survey that involves asking villagers questions,

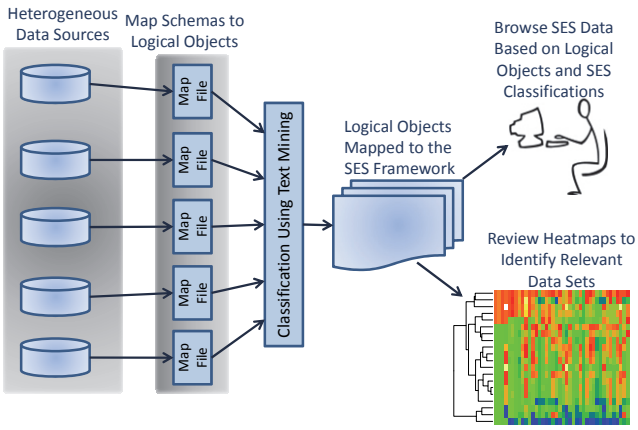

Figure 1: Architecture for mapping heterogeneous databases to logical objects based on a common framework for data discovery.

measuring the breadth of trees, examining the vegetation, etc. Responses to each question are then represented as separate columns in a database. A "*site visit*" is a visit to the field carried out by a team, and is the level at which the research instrument is applied. A site is often visited multiple times over a span of years. When thinking about organizing such a database into logical objects, we could consider numerous alternatives for what constitutes a logical object, such as treating the database itself as a single logical object or defining logical objects corresponding to a single site visit, to a single forest site visited through time, or to all visits within a fixed time period such as a single year.

We propose that the best conceptual representation for the logical object is the "site visit". As a rule of thumb, such objects can be identified in a database schema by being a table that has no foreign-key relationships to other tables. If there are multiple tables with no foreign key relations, we use the heuristic of identifying the table with the deepest hierarchy of foreign key relations back to it. In the IFRI database, the table with no foreign key relations and the deepest hierarchy of foreign key relations back to it is the site visit table.

A. Mapping Data within the Logical Object

The SES framework consists of 8 major categories, also referred to as first-tier variables or classes. Four categories represent the core components of the SES Framework (Resource System, Resource Unit, Governance System, and Users), two additional categories represent the interactions between these core categories and the resulting outcomes, and the final two categories identify interactions with external social and ecological systems [15]. Each of the first-tier categories is assigned a text label in the framework, analogous to namespaces assigned a prefix in XML (e.g., the first tier category "Resource System" is identified as "RS" in the Framework). Each category is decomposed into subcategories referred to as second-tier variables or subclasses. These subcategories identify more detailed characteristics of the first-tier categories. For example, RS3 (the size of the resource system) is a second-tier category within the first-tier category Resource System. In the SES Framework, there are 51 such second-tier categories.

To enable a generalizable approach based on the SES Framework (or similar taxonomies developed in other fields), we map a database's columns (which in IFRI represent questions in the research instrument) to second-tier subcategories of the SES Framework. However, the relations between tables as captured in a database, such as the 1:n relationship between site visits and forests within the IFRI database, are lost if columns are only mapped directly to SES subcategories. So, the relationships between the tables containing the columns are captured as a hierarchy of logical objects in our representation. For example, the IFRI database captures whether each forest has undergone change in tree density (mapped to the Outcomes category O3). Without a hierarchy of logical objects, it is not possible to determine which forest (represented as a tuple in the Forest table) has undergone such a change – only that some forest, somewhere, at some time underwent a change.

B. Hierarchy of Logical Objects

Logical objects can nest. In a normalized relational database, 1:n relationships are split into a separate tables based on foreign-key relationships. In our logical object representation, tables in a 1:n relationship represent separate child logical objects that form a hierarchy of logical objects based on foreign keys. In representing the objects captured in a database as a hierarchy, tables can have foreign key relationships with more than one table; such as the forest-user group relation captured in the IFRI database. In these cases, the relationship is represented as a child object of one of the tables and the foreign key for the other table is captured as metadata. For example, for IFRI, the forest-user group relationship is captured as a child object of the forest object, and, based on the foreign key relationship to the user group table, the name of the user group is included as metadata of the forest-user group object.

In representing the relationship between logical objects and the database, we use T_i to denote table i and T_{ij} denotes the values captured in column j of table T_i. Each column is mapped to a subcategory in the SES Framework such that: $T_{ij} \rightarrow C_{mn}$ where C_{mn} represents the SES second-tier subcategory n within the first-tier category m. This mapping is applied at the column-to-subcategory level where the values are atomic, but not at the aggregate level of tables and SES categories, so $T_{ij} \rightarrow C_{mn}$ does not imply that $T_i \rightarrow C_m$. Instead each T_i represents either the overall logical object or a child logical object as shown in Fig. 2 where Site Visit is a top-level logical object and Forests A and B are second-level logical objects. The reason for this difference is that the tables and SES categories can represent two different views of the data.

The IFRI research instrument is broken out into forms, where each form contains a set of questions. In contrast, the categories in the SES Framework form a taxonomy of characteristics of social-ecological systems. During data collection, the interview form for villagers (resource users) may include questions about interactions, outcomes, resources owned by the users, or how they interact with government agencies. These correspond to first-tier categories in the SES Framework. Logical objects can be formed based on the data collection instrument. They can also be formed based on the database table structure and categorization of table columns, allowing alternative means for identifying the logical object. The generic architecture of the "logical object" positions it for application to any social-ecological data source that is captured and stored in a relational database or tabular format.

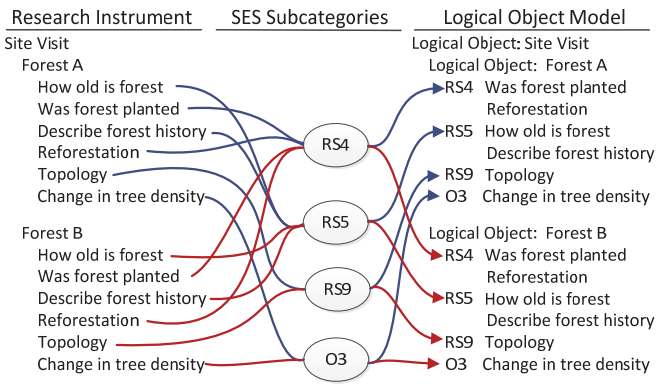

Figure 2: Using a hierarchy of lightweight logical objects to preserve 1:n relationships when mapping database columns to the SES Framework. Each forest is represented as a separate tuple in the database.

III. GENERALIZED DATA REPRESENTATION

A generalized representation serves several purposes. It provides a:

1. Stable representation of the data that can be preserved and reused by future researchers without requiring proprietary formats or requiring a database to be fully reconstituted to access the data,
2. Unifying generalized structure to which individual sustainability databases can be mapped,
3. Common representation for applying machine learning to map diverse data sources to the SES Framework, and
4. Consistent format for building different visualizations of sustainability data that integrate independent logical object views using the SES Framework.

We propose a generalized representation where a separate XML document is generated for each top-level logical object in the database. This representation conforms to a schema that defines a top-level *archive* root element for a hierarchy of logical objects represented by *archiveObject* elements for the tables in the database. Those columns in the database that are mapped to the SES Framework are *archiveElements* within an *archiveObject*. While all columns within the database are captured in this general representation, as discussed further in Subsection A below, not all columns in the database are mapped to the SES Framework. This provides a flexible architecture that we believe can be applied across other sustainability data sources that originate in a relational database form. The form consists of two main components, *archiveObjects* and *archiveElements* as illustrated in the following Extended Backus-Naur Form :

```
archive   :==   archiveId archiveDataSource
                archiveObject
```

```
      archiveId   :==   string
  archiveObject   :==   {metadata}* {archiveObject}*
                        {archiveElement}* objectName
                        uri
     objectName   :==   string
            uri   :==   anyURI
       metadata   :==   name value tableName
                        fieldName
           name   :==   string
          value   :==   string
      tableName   :==   string
      fieldName   :==   string
 archiveElement   :==   [sesSubcategory]
                        (archiveEnum|archiveString|
                        archiveInt|archiveLong |
                        archiveDouble|archiveDecimal
                        |archiveDateTime
                        |archiveBoolean)
                        {metadata}*
                        archiveFldAttributes
 sesSubcategory   :==   RS1|RS2|RS3|RS4|RS5|RS6|RS7|
                        RS8|RS9|RU1|RU2|RU3|RU4|RU5|
                        RU6|RU7|GS1|GS2|GS3|GS4|GS5|
                        GS6|GS7|GS8|U1|U2|U3|U4|U5|
                        U6|U7|U8|U9|I1|I2|I3|I4|I5|
                        I6|O1|O2|O3|S1|S2|S3|S4|S5|
                             S6|ECO1|ECO2|ECO3
```

In the above notation, the details for *archiveDataSource* and *archiveFldAttributes* are omitted. These capture metadata that is not captured in the research instrument about the database (e.g., version) and database columns (e.g., table and field names) respectively. The details for each of the data types within the *archiveElement* have been omitted due to space, but these capture the element's value, applicable database metadata based on the data type (e.g., maximum field size, default value, etc.), and optional characteristics such as the units-of-measure. Details of the definition for the XML schema *string* and *anyURI* types as defined in [1] are also omitted and the production rules based on these types are shown only as "string" or "anyURI".

In our representation, the *site visit* is the top level *archiveObject*. The hierarchy of logical objects is specific to a research instrument, and will likely differ depending on the type of system for which data is being captured, e.g., the forestry resource systems of the IFRI database versus the Nepal Institutions and Irrigation Systems (NIIS) database [17] for sustainability data about irrigation systems.

A. Metadata

In mapping the IFRI database to subcategories in the SES Framework, we identified cases where columns did not map to the Framework but instead represent metadata either about a logical object or another column in the database. Examples include questions in the header of each survey document such as who performed the data collection or when the research was undertaken. In addition, questions in the body of the forms may represent metadata such as sampling methods used or the authority used to identify the species of plants. Similarly, some columns represent metadata related to another column, such as the name of a governing institution. In all of these cases, the database mapping identifies the columns as metadata. The values captured in these columns are associated with the object or column to which they relate as metadata recorded as name/value pairs. This approach is reflected in the representation of the logical object format shown above. It allows the XML schema to remain generic, while enabling detailed metadata to be captured at any level where it is available.

B. Generalizability

The logical object model uses a set of constructs that can be applied to other social-ecological databases through the SES Framework, or possibly other domains contingent on two criteria:

1) The database has a framework or ontology such as the SES Framework for cross-domain linking, and
2) Data is captured in a way that individual columns can be classified based on the framework. The IFRI instrument is a survey, so each question can be classified using the SES Framework (unless the question represents metadata) since there is a 1:1 relationship between questions and columns. This approach is not well suited to a database that represents non-survey data if a 1:1 mapping is not feasible.

IV. Ingesting Social-Ecological Data

To extract relational data into a logical object representation from diverse databases, our methodology (Fig. 3) employs an XML file that maps tables and columns into the generalized model of *archiveObjects*, *archiveElements*, and *metadata* as discussed in Sections II and III. The main components of the mapping schema are *objectMap* and *elementMap* that define the mapping to *archiveObjects* and *archiveElements* respectively. Additionally there is a single *dbMap* element at the root of the schema that contains a query to select the IDs that identify the archive objects generated as separate XML documents. The *dbMap* element contains a single *objectMap* element that represents the top-most logical object (site visit in IFRI). Details are omitted due to space, but the major components of the mapping schema are outlined in Fig. 3.

Drawing from D2R [3] for mapping relational databases to RDF, the mapping to logical objects is based on the relational database's table and column structure. Each *objectMap* contains a *query* element that defines a query for selecting object instances from the database. Fig. 4 shows a snippet of the mapping file for the IFRI database that includes the start of the *objectMap* for "Forest" objects within the top-level "Site Visit" logical object. Since the hierarchy of logical objects is based on database table definitions, the queries within each object are generally straightforward; the query in Fig. 4 selects all rows from the FOREST table where the foreign key relation to the oversight table matches the *queryParameter* specified. A query can have multiple parameters, and each parameter contains a field name, position, and data type. Position refers to the parameter's position in the prepared statement for the query – in this case there is only one parameter. The field name refers to a field in the query for the parent object. Since "Forest" is a child of "Site Visit" in IFRI, the *fieldName* refers the ID column in a row from the query for the site visits:

```
select * from OVERSITE where ID = ?.
```

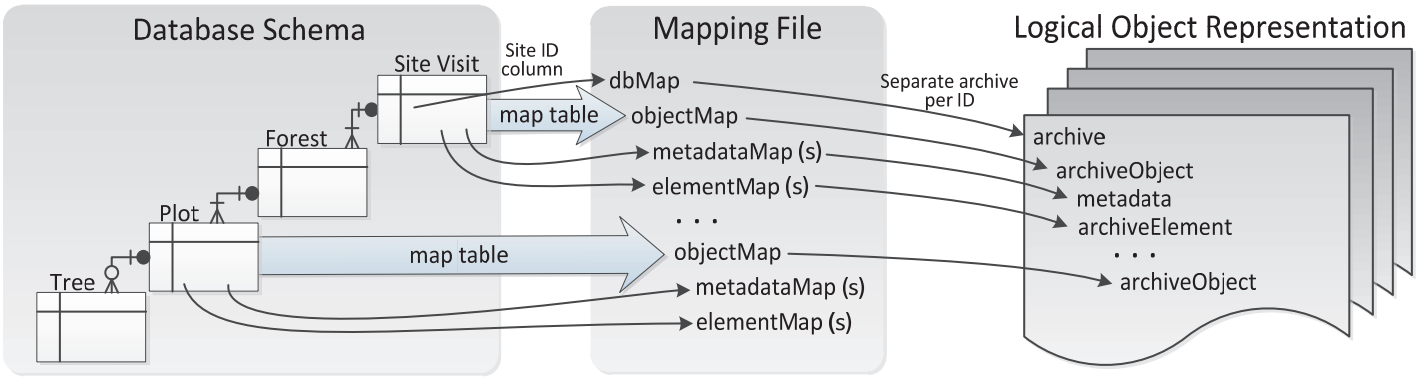

Figure 3: Mapping a social-ecological database to an XML-based generalized logical object representation that enables classification using the SES Framework.

For an *elementMap* within the "Forest" objects as shown in Fig. 4, the *queryField* element identifies the column within the object's query that is used to populate the corresponding archive element. Additionally, the *elementMap* contains the database table and field name which are used when processing the map to determine the element's data type and additional metadata that can be extracted from the database.

Although the mapping process is straightforward, manually creating the mapping file is tedious for a database such as IFRI which contains 922 columns, of which 70% (648) map as elements and the remaining 30% represent metadata. However, as for mapping databases in D2R [3] (where tables are mapped to RDFS classes based on the table's name, and columns within a table are mapped as properties of that RDFS class), much of the semantics can be lost in an automated default mapping based solely on table and column names.

The loss of semantics is greatest in an e*lementMap*. This is illustrated in Fig. 4, where the *label* represents the survey question: *"Was most of the vegetation in the forest planted, or is it a result of natural growth?"*. In the research instrument, the scientist selects from an enumerated list of four options. If the second of the four options is selected, *"People have planted woody vegetation, but their efforts have been aided by natural regeneration and seeding"*, a default mapping based on column names and their corresponding values would generate an archiveElement with the name "FVEGORIGIN" that has the value "2". When processing a database using the mapping file, an additional csv file is loaded that maps an index for each enumeration to the full text for its options. The *enumClass* element in Fig. 4 maps the responses contained in the database field FVEGORIGIN to the human-readable responses in the enumeration with an ID of "18" in this separate file. This additional file allows the full text of the enumerated responses to be associated with each question instead of only the numerical index.

```
<soc:objectMap>
 <soc:query>
  <soc:sql>
     SELECT * FROM FOREST WHERE FK_OVERSITE = ?
  </soc:sql>
  <soc:queryParameter>
    <soc:fieldName>ID</soc:fieldName>
    <soc:parameterPosition>1</soc:parameterPosition>
    <soc:parameterDataType>LONG</soc:parameterDataType>
  </soc:queryParameter>
 </soc:query>
 <soc:description>Forest</soc:description>
 <soc:mapComponents>
       • • •
  soc:elementMap>
    <soc:queryField>FVEGORIGIN</soc:queryField>
    <soc:label> Was most of the vegetation in the forest
                planted, or is it a result of natural growth
    </soc:label>
    <soc:tableName>FOREST</soc:tableName>
    <soc:fieldName>FVEGORIGIN</soc:fieldName>
    <soc:enumClass>18</soc:enumClass>
    <soc:sesSubClass>RU2</soc:sesSubClass>
  </soc:elementMap>
```

Figure 4: Section of the map for the IFRI database that maps the "Forest" child logical object.

To address the loss of semantics inherent in this fully automated approach, while avoiding the tedium of completing a manual XML mapping file, we employ a spreadsheet-based semi-automated approach in which each row represents a column in the database. Much of the information in the spreadsheet can be prepopulated in an automated fashion, and for completing fields which cannot automatically be populated from the database, (such as the *enumClass* or *label*) it is much easier for non-programmers to populate and integrate data sources into a spreadsheet than an XML file. Using IFRI as an example, additional tables in the database that are used to populate the user interface contain the text for each question. Similarly, other systems may either use property files to manage questions or the definition for each question may be captured in a metadata format such as the Data Documentation Initiative (DDI) [7] which is used to define questionnaire codebooks in the social sciences. This semi-automated approach reduces the effort required to create the database mapping while preserving the semantics that are necessary for automating the mapping and generating the client views.

V. Automated Classification through Machine Learning

In mapping databases to the logical object representation, one of the time-consuming tasks is the mapping of database columns to the 51 subcategories of the SES Framework. In Fig. 4, the *sesSubClass* element maps the e*lementMap* definition to the SES subcategory "RU2" (growth or replacement rate). One of the goals of this research is to enable data discovery – allowing scientists to find data sets relevant to the particular research question they are trying to address. As

noted in [18], it is not expected that a researcher investigating a well-designed theoretical question would capture data for all 51 of the SES subcategories, however, as discussed in [15], data must exist in a dataset for the categories relevant to the research question being investigated. Automating the classification of datasets based on the SES Framework would enable researchers to quickly identify potentially relevant datasets from a large heterogeneous population of datasets. Our research described in this section examines whether machine learning can be applied to social-ecological datasets to automatically classify the individual questions (database columns) based on the 51 subcategories of the SES Framework – allowing other researchers to determine if a dataset potentially captures data relevant to their research question.

In this context, the focus is on classifying the questions in a research instrument and not classifying each instance of a resource system captured with that instrument. This allows the approach to be generalized to other similar data sources since we are mapping columns ($T_{ij} \rightarrow C_{mn}$) instead of mapping resources described by different schemas. For the IFRI database, the goal is not to classify each of the 346 site visits, but instead to classify the 648 survey questions captured in the database. To evaluate different classification methods, we first manually classify all of the IFRI questions based on the 51 subcategories of the SES Framework; this is based in part on a prior partial classification of IFRI by Kashwan and Kreitmair [14]. The XML representations generated from the mapping files as discussed in Section 4 are then used to generate data files that could be processed using the Weka machine learning API [11], which implements numerous machine learning algorithms.

Since the algorithms used in our classifiers cannot handle text data directly, all of the text fields are converted to numeric values using the bag-of-words approach in which each word is assigned a value based on its importance in the text (with tfidf weighting). To evaluate different approaches to classification based on the SES Framework, we initially tested Naïve Bayes, support vector machine (SVM), and decision tree (J48) classifiers. However, because the decision-tree classifiers significantly outperformed both Naïve Bayes and SVM on the IFRI data, all further measurements are based on the J48 decision tree classifier. Although all of the text fields related to a column could be converted to a single word vector, as shown in Table 1, this significantly degrades the performance of the classifier, so each field is converted to a separate word vector. For columns capturing free-text responses, the values for all 346 site visits are concatenated. For the questions where responses are based on an enumerated list, all of the distinct options actually selected for at least one site are concatenated. If an option was not used in any of the 346 site visits, it is not included. For numeric data, we used the min, max, mean and standard deviation across all site visits as machine learning features. Although the IFRI data is a sizable research instrument from the viewpoint of social-ecological scientists, with 648 questions and 51 second-tier subcategories in the SES Framework, due to the relatively small size of the dataset we applied a leave one out cross validation (LOOCV) approach both when assigning the 8 first-tier categories of the SES Framework and the 51 second-tier subcategories. As reflected in Fig. 1, the machine learning classification is based on the logical object representation generated using the mapping files.

Table 1: A semi-automated approach exhibits good performance at both the first tier categories and second-tier subcategories when mapping databases to the SES Framework.

	Mapping Columns to SES Categories		
	Single Word Vector	*Default Mapping*	*Semi-Automated Mapping*
Precision	0.614	0.520	0.694
Recall	0.617	0.525	0.677
F-Measure	0.615	0.517	0.675
Mapping Columns to SES Subcategories			
Precision	0.471	0.329	0.62
Recall	0.471	0.34	0.597
F-Measure	0.466	0.323	0.597

Table 1 shows the performance of the classifier when (a) all text fields are combined as a single word vector, (b) When the mapping is based solely on an automated default mapping, and (c) when mapping is based on a semi-automated approach. Both the default and semi-automated approaches are based on separate word vectors for each text field in the data because that approach outperforms using a single vector. The lower performance of classification based on a fully automated mapping is to be expected. Although the text from questions with a free-text response is still available for classification, only 16% of the fields in the IFRI instrument use a free-text response, whereas 60% of the fields are nominal questions based on an enumerated set of responses. In a fully automated approach based solely on the database, these become only integer index values, so the set of responses to one question in the survey cannot be differentiated from the set of responses to any other nominal question. Additionally, similar to other database mapping approaches such as D2R's mapping to RDF, an automated default mapping has only the column name from the database metadata to represent each question in a research instrument instead of the full text of the question.

The results in Table 1 for classification based on the SES Framework's 51 subcategories is based on a "flat" approach in that all 51 subcategories are evaluated as independent classes. This means that although RS5 (productivity of the resource system) and RS7 (predictability of system dynamics) are both within the first-tier RS category for Resource Systems, this membership in the same first-tier category of the framework is not considered in the classification. An alternate approach is to use a hierarchical classifier in which instances are first classified based on their first-tier category (e.g., RS vs. GS), and then the instances within a first-tier category are classified for the second tier using a model trained using only that category's data. To evaluate a hierarchical approach using the semi-automated mapping, each instance is classified based on the first tier SES categories using a LOOCV approach and then classified using the second tier of the Framework for each category that was represented in the first-tier distribution. In initial calculations using a multiplicative approach to hierarchical classification as in [8], performance is lower than

when the hierarchy is flattened to consider only the second tier as reported in Table 1. To evaluate the performance impact of the first and second tier subcategories of the SES on hierarchical classification, we applied the following weighting formula (1):

$$\alpha * \log(\theta_{tier1}) + (1- \alpha) * \log(\theta_{tier2}) \quad (1)$$

Where "θ_{tier1}" is the distribution value for each class based on the 8 first tier SES categories and "θ_{tier2}" is the distribution from classification based on classifying within the first tier category. The weighting factor (α) is varied from 0.1 to 0.9 in increments of 0.1. We found weighting the first and second tiers of the SES hierarchy has a marginal impact on the classification performance. The F-measure across all of the weights varied from 0.523 to 0.526, which is lower than the result of 0.597 for the flat classification using semi-automated mapping shown in Table 1.

Statistically, the performance of hierarchical classification for the IFRI data is lower than flat classification; this is most likely due to the insufficiency of tier 2 training instances. The performance of the flat classification based on the 648 IFRI survey questions used to train the model is good for the size of the training set, but the tier 2 model only employs the instances that belong to a specific tier 1 class - resulting in a much smaller training corpus for each tier 2 learning model. In the future, in order to enhance the performance of the hierarchical classification, additional social-ecological data sources need to be identified and classified based on the SES Framework to provide a larger corpus of training instances.

VI. USES

Shown on the right hand side of Fig. 1 are two alternative uses of the resulting representation: browsing and heat maps. The goal in both of these client views is data discovery. The logical object representation is XML-based. While native XML databases can significantly underperform relative to relational databases when querying data [12], neither of these client views actively queries the XML.

The browse capability is made possible because of an XSLT transformation on the logical object. The XSLT transformation formats the XML and presents it to the researcher with questions, responses, and associated metadata ordered as logical objects based on the SES Framework. When classification is automated as discussed in Section V, the XML schema contains an additional attribute that flags whether each *archiveElement* has been manually classified or automated based on machine learning. The XML-based representation has the advantage of simplifying the generation of alternate web-based presentations of the data, such as suppressing or redacting selected questions (or just the responses) based on the class of user accessing the data. Some social-ecological research programs capture data regarding sensitive or endangered species, and opening a database such as IFRI to public access without limitation could provide an unintended roadmap for users without a legitimate research purpose.

Another client view that we explored is the heat map, used to reflect the volume of data captured in a site visit for each of the SES subcategories. We generate the heat maps using the R statistics package [19]. Site visits are displayed as rows and clustered based on their similarity. The SES subcategories are displayed on the x-axis, but not clustered, so that the ordering within first tier categories is preserved as illustrated in Figure 5.

With both client views, the goal is to enable scientists to identify data sets that are potentially relevant to the research question they are investigating. These two views can also be used in tandem, with a scientist first using heat maps of the entire set of datasets to identify datasets potentially of interest based on the volume of data for the SES subcategories relevant to their research question and then zooming in to browse the details of those datasets.

VII. RELATED WORK

One of the goals of the logical object model is to provide a generalized format for capturing a stable and non-proprietary representation of sustainability data captured using diverse relational schemas. Commercial products [4], as well as other researchers [5], have addressed capturing relational data in a stable XML representation for the purpose of archiving databases. The SIARD Suite [21] developed by the Swiss Federal Archives (SFA) is a tool for archiving relational databases (Oracle, Microsoft SQL Server, and Microsoft Access) that creates a zipped archive containing an XML file for each table in the database and a single XML file for the metadata (structure and constraints) of the database. However, the SIARD format cannot be accessed directly; it is instead intended to enable reconstituting a database at some future date (assuming future database platforms support SQL:1999).

The mapping of complex relational databases has similarities with other database conversion approaches such as the D2RQ mapping language [9] and the earlier D2R Map Language [2]. The D2RQ mapping language is used with D2R Server to present relational data as semantic web data by mapping a relational database to RDF. The D2R Server can be used with D2RQ to: (a) translate SPARQL queries over RDF to SQL queries against a relational database store on-the-fly, (b) map relational data as RDF linked data on-the-fly, or (c) create an RDF dump of a relational database that can be accessed directly as RDF when loaded into a triple store such as Jena or Sesame. The goal of the on-the-fly mapping differs from the goal of our representation in that as Bizer and Cyganiak note in [3], a large volume of data is still in relational databases, so D2R server can provide a means to populate the semantic web with relational data without replicating the data as RDF. Mapping to a generic format shares similarities with our approach, but D2RQ differs in that one of the goals of our research is to provide a stable non-proprietary and non-relational data representation. Additionally, by generating a snapshot of the database, researchers can reference a version of the database - an open research issue in production databases.

Because a substantial portion of our data is text, in the form of responses to survey questions in the research instrument, or detailed text in enumerated responses, our categorization focus is on text categorization. Decision tree, SVM, and Naïve Bayes are three important text classification algorithms out of the many that the machine learning research has identified [20] and are thus applied in our experiments.

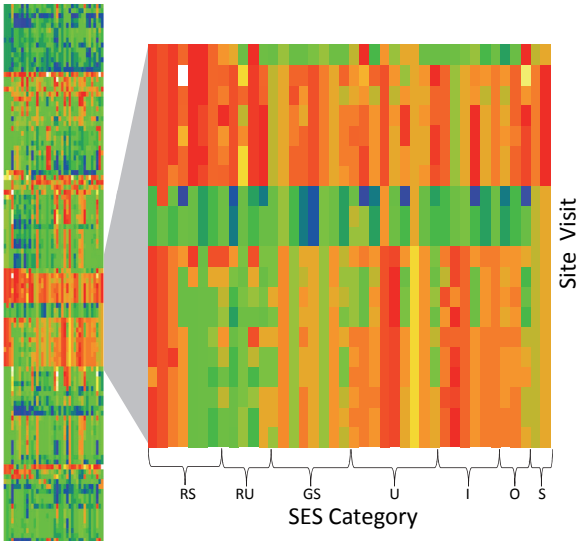

Figure 5: Heat maps allow scientists to visualize the coverage of a large volume of data based on the SES Framework and selected datasets relevant to their research.

VIII. Conclusions and Future Work

Our results show that a semi-automated approach to mapping social-ecological data sources to the generalized logical object model can enable a common stable representation against which useful and generalized tools can be built. We demonstrate this through two client views, a SES Framework aware browse tool and heat maps. Future work involves assessing the generality of the approach on other social ecological data sources.

Additionally, from this representation automated classification based on the SES Framework can be done. In our current work on classifying the IFRI data using machine learning, each question captured in the database is mapped to a single SES subcategory. However, some questions could arguably be mapped to multiple subcategories. Because the goal is to identify datasets potentially applicable to a research question, allowing for multiple classifications as in [8], may be preferable.

A key avenue of future research is establishing the utility of the logical object model in serving as a mechanism for the long term archiving of social-ecological data. Long term archiving of social ecological data from relational databases is a known problem for the reasons mentioned in this paper. Social ecological research is often described using the DDI schema. In our work thus far we have successfully developed and employed a semi-automated approach to mapping the relational database to the SES Framework. In application of the existing solution to archival, each logical object could be an OAIS-conforming archive Submission Information Package (SIP). This approach would archive each site visit as an independent entity. While this may be adequate, information about the collection of site visits would be lost. The open questions are first to assess the impact of the lost information, and second to make ties from the logical data object back to DDI.

Acknowledgment

The authors thank the late Dr. Elinor Ostrom and the Vincent and Elinor Ostrom Workshop in Political Theory and Policy Analysis. Dr. Ostrom pioneered the SES Framework and will be dearly missed by all who knew her.

References

[1] P.V. Biron and A. Malhotra (ed.), XML Schema Part 2: Datatypes Second Edition, W3C, http://www.w3.org/TR/xmlschema-2/, accessed Jul 2012

[2] C. Bizer, D2R MAP – A database to RDF Mapping Language, *12th Int'l World Wide Web Conf (WWW)*, Budapest, Hungary, May 2003.

[3] C. Bizer and R. Cyganiak, D2R Server: Publishing Relational Databases on the Semantic Web, *5th Int'l Semantic Web Conf*, Athens, GA, Nov 2006.

[4] S. Brandl and P. Keller-Marxer, Long-term Archiving of Relational Databases with Chronos, *1st Int'l Workshop on Database Preservation (PresDB07)*, Edinburgh, Scotland, Mar 23, 2007.

[5] P. Buneman, S. Khanna, K. Tajima, and W. Tan, Archiving Scientific Data, *ACM Trans on Database Systems*, 29(1), Mar 2004, pp. 2-42.

[6] A. Chhatre and A. Agrawal, Forest commons and local enforcement. *National Academy of Sciences*, 105(36), pp. 13286-13291, 2008.

[7] DDI-Codebook Version 2.5, Data Documentation Initiative. http://www.ddialliance.org/Specification/DDI-Codebook/2.5/, accessed Jul 2012

[8] S. Dumais and H. Chen, Hierarchical classification of Web Content, 23rd ACM Int'l Conf on Research and Development in Information Retrieval (SIGIR-00), Athens, GR, 2000.

[9] The D2RQ Mapping Language, Version 0.8 – 2012-03-12, http://d2rq.org/d2rq-language, accessed Jul 2012

[10] C. Gibson, M. A. McKean, and E. Ostrom, People and Forests, Cambridge: MIT Press, 2000.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations 11(1)

[12] Scott Jensen, Devarshi Ghoshal, and Beth Plale, Evaluation of Two XML Storage Approaches for Scientific Metadata, Indiana University Dept of Computer Science Tech Report 698, Oct. 2011.

[13] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, LS-8 Report 23, Univ of Dortmund, 1998.

[14] P. Kashwan and U. Kreitmair, Forests as Complex SES: Application of PNAS Framework to Forestry, unpublished, Feb 2010.

[15] E. Ostrom, A diagnostic approach for going beyond panaceas, *National Academy of Sciences*, 104(9), pp. 15181-15187, 2007.

[16] E. Ostrom, A General Framework for Analyzing Sustainability of Social-Ecological Systems, *Science*, 325(5939), pp. 419-422, 2009.

[17] E. Ostrom and R. Gardner, Coping with Asymmetries in the Commons: Self-Governing Irrigation Systems Can Work, *The Journal of Economic Perspectives*, 7(4), pp. 93-112, 1993.

[18] A.R. Poteete, M.A. Janssen, and E. Ostrom, Working together: collective action, the commons, and multiple methods in practice, Princeton University Press, Princeton, NJ, 2009.

[19] R Development Core Team, R: *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011. Available at: http://www.R-project.org

[20] F. Sebastiani, Machine learning in automated text categorization, *ACM Computing Surveys*, 34(1), pp. 1-47, 2002.

[21] Archiving Databases: SIARD Suite http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en, accessed Jul 2012

[22] E. Wollenberg, L. Merino, A. Agrawal, and E. Ostrom, Fourteen years of monitoring community managed forests: Leaning from IFRI's experience, *Int'l Forestry Review*, 9(2), pp. 670-684, 2007.