Automatic Interpretation Biodiversity Spreadsheets Based on Recognition of Construction Patterns

Ivelize Rocha Bernardo, André Santanchè and Maria Cecília Calani Baranauskas
Institute of Computing - Unicamp, Avenida Albert Einstein, 1251, Cidade Universitária, Campinas, Brazil

Keywords: Spreadsheet Interpretation, Semantic Mapping, Information Integration, Patterns Recognition, Biodiversity Data Integration.

Abstract: Spreadsheets are widely adopted as "popular databases", where authors shape their solutions interactively. Although spreadsheets have characteristics that facilitate their adaptation by the author, they are not designed to integrate data across independent spreadsheets. In biology, we observed a significant amount of biodiversity data in spreadsheets treated as isolated entities with different tabular organizations, but with high potential for data articulation. In order to promote interoperability among these spreadsheets, we propose in this paper a technique based on pattern recognition of spreadsheets belonging to the biodiversity domain. It can be exploited to identify the spreadsheet in a higher level of abstraction – e.g., it is possible to identify the nature a spreadsheet as catalog or collection of specimen – improving the interoperability process. The paper details evidences of construction patterns of spreadsheets as well as proposes a semantic representation to them.

1 INTRODUCTION

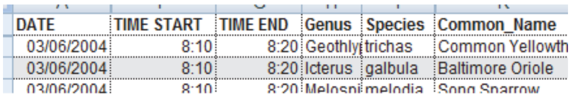
When producing spreadsheets, end-users have autonomy and freedom to create their own systematization structures, with few formal requirements. However, the product is driven to human reading, causing a side effect: programs provide poor assistance in performing tasks, since they are unable to recognize the spreadsheet structure and to discern the implicit schema – hidden in the tabular organization – from the instances and consequently the semantics of this schema. Therefore, it is difficult to combine and coordinate data among spreadsheets using conventional methods, because each new different schema may seem unknown.

But, how much different they are in fact? We present in this paper evidences that similarities in spreadsheets can indicate patters followed by groups.

Making a parallel, spreadsheets can be seen asclay, in which authors sculpt their elements according to their own experiences and/or conventions followed by the group to which they belong. For example, to carve a table, even though there is multitude of possibilities, there are patterns that have been consolidated in the author's community: tables have a top supported by one or more legs. These undocumented patterns are created and replicated according to users experience and their observations in the real world.

We consider that it is possible to map these patterns to a respective semantic description, through the recognition of structural reasons which leads a user to interpret a spreadsheet in one way and not another.

Thus, our strategy focuses on the detection of patterns to recognize similar spreadsheets. We argue that the specific way authors build their spreadsheets – i.e. the criterion to define elements, the approach to spatially organize them and the relationship between these elements – is directly related their daily experience in the community they belong.



DATE	TIME START	TIME END	Genus	Species	Common_Name
03/06/2004	8:10	8:20	Geothly	trichas	Common Yellowth
03/06/2004	8:10	8:20	Icterus	galbula	Baltimore Oriole
03/06/2004	8:10	8:20	Melospi	melodia	Song Sparrow

Figure 1: Example of a spreadsheet recording a collection [ecosystems.mbl.edu].

Figure 1 shows an example of a spreadsheet to register specimens collected in the field. A user can identify it due to the specific arrangement of the columns, registering event related fields in the leftmost columns, followed by genus and species.

Rocha Bernardo I., Santanchè A. and Cecília Calani Baranauskas M..
Automatic Interpretation Biodiversity Spreadsheets Based on Recognition of Construction Patterns.
DOI: 10.5220/0004898200570068
In *Proceedings of the 16th International Conference on Enterprise Information Systems* (ICEIS-2014), pages 57-68
ISBN: 978-989-758-029-1

However, these signs are not recognized by computer programs. Up to now, only humans are apt to infer the purpose of this spreadsheet and its organization.

The challenge of this research is to consider a computer system as a consumer of spreadsheets besides the user. Our approach involves to achieve a richer semantic interoperability for data from spreadsheets through pattern recognition.

Both (Tolk, 2006) and (Ouksel and Sheth, 1999) classify interoperability in progressive layers and consider that higher layers will subsidize more efficient operations.

Tolk (2006) proposes a more detailed classification and add the pragmatic, dynamic and conceptual layers, i.e. it is possible to consider aspects of context and user intentions in the interoperability process.

Beyond machines, (Haslhofer and Klas, 2010) define the highest interoperability level as the ability of humans and machines to share the same semantic.

In order to implement an interoperability technique, our proposal has two stages to map metadata, the first stage maps the terms to exploratory questions, in order to recognize the context – based on Jang et al., (2005) – and the second stage concerns the analysis of how these fields are arranged in the spreadsheet. This organization indicates construction patterns followed according to the conventions adopted by their community.

Most of the related work disregard this organization when implement strategies for seeking interoperability of tabular data. This paper argues that the structure, i.e. the organization of spreadsheet elements, must be considered, since it leads to the identification of construction patterns, which is related to the user intention/action. This technique allows us to go towards the pragmatic interoperability layer (Tolk, 2006).

We present in this paper evidences of spreadsheet construction patterns adopted by biologists and the application of these patterns in automatic recognition of their implicit schemas. To support our thesis we collected and analyzed approximately 11,000 spreadsheets belonging to the biodiversity domain.

This article is organized as follows: Section 2 gives an overview of some basic concepts and our research, Section 3 details the process of collecting and analyzing spreadsheets employed by biologists, as well as research hypotheses and their evaluation; Section 4 highlights evidences of construction patterns followed by biologists; Section 5 introduces our model to represent construction patterns; Section 6 compares our approach with related initiatives to recognize implicit schemas in spreadsheets; Section 7 presents our concluding remarks and the next steps of this research.

2 RESEARCH SCENARIO

According to Syed et al., (2010), a large amount of the information available in the world is represented in spreadsheets. Despite their flexibility, spreadsheets were designed for independent and isolated use, and are not easily articulated with data from other spreadsheets / files.

For this reason, there is a growing concern to make spreadsheet data more apt to be shared and integrated. The main strategies convert them into open standards to allow software to interpret, combine and link spreadsheet data(Connor et al., 2010); (Zhao et al., 2010); (Han et al., 2008); (Yang et al., 2005); (Ponder et al., 2010); (Doush and Pontelli, 2010); (Abraham and Erwig, 2006).

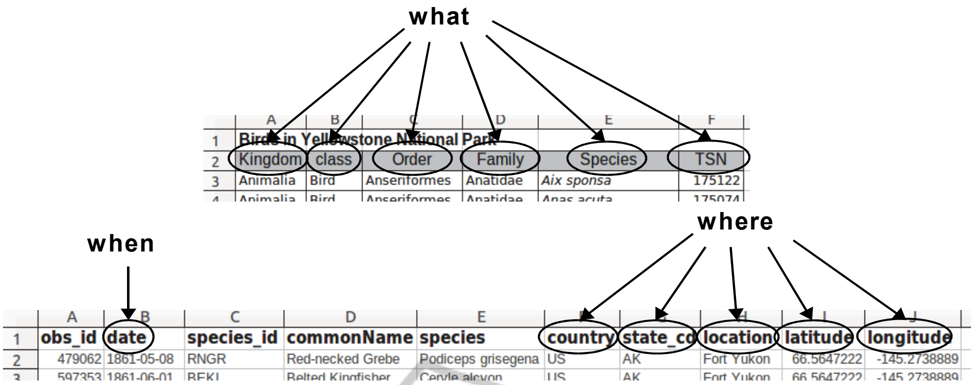
Related work address this problem mainly by manual mapping to Semantic Web open standards or by automatic recognition, relating spreadsheet elements to concepts available on Web knowledge bases such as DBpedia (http:/ / dbpedia.org).

Systematic approaches for data storage, such as databases, predefine explicit schemas to record data. These schemas can be considered as semantic metadata for the stored data. Spreadsheets, on the other hand, have implicit schemas, i.e. metadata and data merged in the same tabular space.

Many related work attempt to separate schema of their instances, as if one is more than the other. One of our hypotheses is that both are equally important and powerful in the search for a given semantically richer.

The central thesis behind our approach is that we can detect and interpret the spreadsheet's schema by looking for construction patterns shared by research groups. We propose in this paper a representation model able to capture such patterns, as well as to be processed by machines. Results of our analysis in thousands of spreadsheets indicate the existence of such recurrent patterns and that they can be exploited to recognize implicit schemas in spreadsheets.

There are several aspects that hinder the spreadsheet recognition and its implicit schema, such as differences between columns order, the label used to identify fields and their respective semantics etc.



	A	B	C	D	E	F
1	Birds in Yellowstone National Park					
2	Kingdom	class	Order	Family	Species	TSN
3	Animalia	Bird	Anseriformes	Anatidae	*Aix sponsa*	175122
4	Animalia	Bird	Anseriformes	Anatidae	*Anas acuta*	175074

	A	B	C	D	E	F	G	H	I	J
1	obs_id	date	species_id	commonName	species	country	state_co	location	latitude	longitude
2	479062	1861-05-08	RNGR	Red-necked Grebe	Podiceps grisegena	US	AK	Fort Yukon	66.5647222	-145.2738889
3	597353	1861-06-01	BEKI	Belted Kingfisher	Ceryle alcyon	US	AK	Fort Yukon	66.5647222	-145.2738889

Figure 2: Fields Characterization.

Although related work explore a subset of the common practices in tabular data – sometimes taking into account their context (Jannach et al., 2009); (Venetis et al., 2011); (Mulwad et al., 2010) – they do not define a mechanism or model to independently represent these patterns. Since the knowledge about how to recognize patterns is mixed with the programs, they cannot be decoupled from their code. We claim here that a representation to materialize the knowledge about these patterns as artefacts, independently of specific programs and platforms, enables to share, reuse, refine and expand such patterns among users and applications.

This research is driven by a larger project that involves cooperation with biologists to build biodiversity bases. We observed that biologists maintain a significant portion of their data in spreadsheets and, for this reason, this research adopted the context of biology as its specific focus.

We propose a model to represent construction patterns, departing from observations conducted through incremental steps, including spreadsheets collecting/catalog, formulating hypotheses/models and evaluation and they will be detailed in next sections.

3 METODOLOGY

As previously mentioned, our approach to represent construction patterns was based on a study of related work and field research in the biology domain.

Based on an initial analysis of how biologists of the Institute of Biology (IB) of Unicamp created their spreadsheets, we designed a process to automate the recognition of construction patterns, whose design involved (i) collecting and analyzing spreadsheet data;(ii) formulating hypotheses about construction patterns of spreadsheets; (iii) designing and implementing automatic recognizers for these spreadsheets.

3.1 Initial Data Collection and Analysis

Our analysis started with 9 spreadsheets belonging to the IB, in which we identified two main construction patterns, related to the nature of the spreadsheet: *catalogs* of objects – e.g., specimens in a museum – and *event* related spreadsheets, e.g., a log of samples collected in the field. We further will refer to these spreadsheet natures as *catalog* and *event*.

In order to address the significant differences among spreadsheet types we classified each field in six exploratory questions (*who, what, where, when, why, how*) (Jang et al., 2005). It enabled us to represent and recognize patterns in a higher level of abstraction, e.g., a *catalog* spreadsheet has as initial fields the taxonomic identification – classified as *what* question – on the other hand, a collection spreadsheet has as initial fields: date and locality – classified as *when* and *where* questions, as illustrated in Figure 2.

The next step involved collecting more 33 spreadsheets on the Web to compose our sample. To search spreadsheets belonging to the biology domain, we applied domain related keywords as criterion.

3.2 Hypotheses

According to the observation of these spreadsheets, we proposed the following pattern-related hypotheses:

H1: most of the spreadsheets organization follows the pattern of columns as fields and rows as records;

H2: in order to characterize the context (Jang et al., 2005)fields in the spreadsheets can be classified in one of the six exploratory questions;

H3: the first fields of a spreadsheet often define its nature, e.g., *catalog* or *event*, as well as its construction pattern.

We developed a system – SciSpread – to automatically recognize schemas based on these hypotheses. We found evidences, based on our hypothesis, that patterns can drive the recognition of the spreadsheet nature in a context, to make its schema explicit and to support its semantic annotation.

3.3 SciSpread

Figure 3 illustrates the overall SciSpread architecture. Figure 3 (A) represents the data input of the system, divided into two groups: spreadsheets collected on the Web and a configuration file, which guides the process of the spreadsheets recognition.

This configuration file contains data concerning construction patterns of spreadsheets, e.g., schema keywords, the relationship among these keywords and exploratory questions. It works as a dictionary of terms, mapping them to exploratory questions. Each term also receives a weight according to its relevance in a given pattern. For example, in order to recognize a specific category of spreadsheets, if the term species is 50% more relevant than latitude, its relevance-weight will be 10, while the latitude weight will be 5.

Figure 3 (B) details the processing of spreadsheets, in which the system extracts the data from these spreadsheets using a library and performs processing based on the model.

Thus, in the `Fields Recognition` stage, a search is performed looking for terms of the spreadsheet related to terms contained in the configuration file. Whenever a term is recognized, it follows to the `Fields Classification` step, in which it is linked to an ontology concept specified in the dictionary of the configuration file. The process continues until it finds a schema (`Schema Identification` step) that meets one of the expected patterns, otherwise, the corresponding spreadsheet is classified as unrecognized. The recognized schema/pattern conducts to the `Identification of the Spreadsheet Nature` step.

This search for terms/schema involves two aspects: the relevance-weight of the recognized spreadsheet terms and their spatial arrangement.

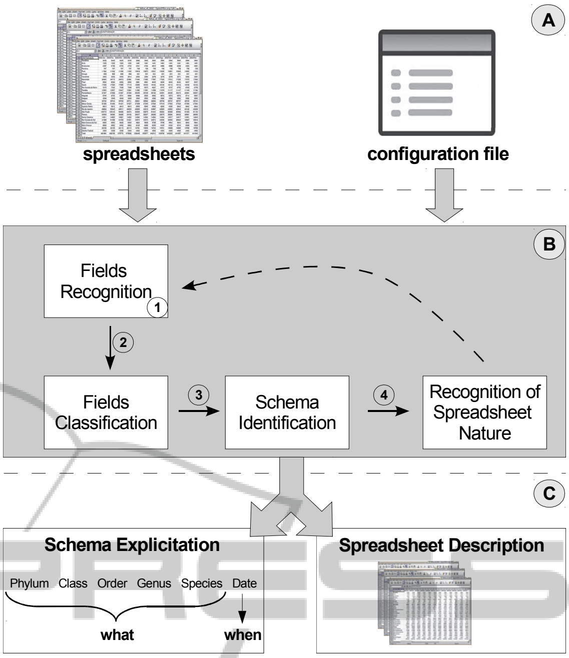


Figure 3: System Architecture.

Even though the recognition process operates in successive steps, the two last steps –to identify the spreadsheet schema and nature – are progressive, according to the recognition of terms and their arrangement as construction patterns.

Like related work – detailed in the previous section – this version of the system mixes with the program code most of the knowledge about how to recognize patterns and categorize spreadsheets, comprising the order and spatial relationships among fields. To address this limitation, we propose in this paper a model to represent construction patterns and their links with semantic representations based on ontologies. This model is independent of specific programs and platforms, enabling to share, reuse, refine and expand representations among users and applications.

Our model is founded in field observations detailed in the next section, which presents the results obtained by the SciSpread system after processing thousands of spreadsheets collected on the Web. A statistical analysis of these results indicates that spreadsheets follow building patterns shared by communities.

4 EVIDENCES OF CONSTRUCTION PATTERNS

The process of collecting and analyzing spreadsheets was developed as follows:

In the first step, we initially collected 9 spreadsheets of the Institute of Biology (IB) used for different purposes. The analysis of these spreadsheets revealed that common terms and patterns are followed by the community. Its results guided the next step: the collection of 33 additional spreadsheets on the Web.

Based in our preliminary assumptions, we looked for patterns concerning: schema layout (e.g., column labels), order and grouping of spreadsheet fields etc. A set of hypotheses – presented in the previous section – was defined and we developed an initial version of the automatic recognition system to validate these hypotheses.

The system was tuned to recognize all spreadsheets of this initial sample, whose nature fit in our context. We further randomly collected more 1,914 spreadsheets on the Web, finding them through the Google search engine, based on keywords extracted from previous spreadsheets: kingdom, phylum, order, biodiversity, species, identification key etc. The system recognized 137 spreadsheets (7%) of all 1,914 spreadsheets collected. The manual analysis of these spreadsheets showed that the system correctly recognized 116 spreadsheets and incorrectly recognized (false positives) 21 spreadsheets. Even though the latter spreadsheets have the expected construction pattern, they do not address the focus of our study, which are spreadsheets used for data management.

Increasing our sample size to 5,633 spreadsheets, the system recognized 7%; subsequently, increasing to about 11,000 spreadsheets, the system recognized 10.4%, which corresponds to 1,151 spreadsheets, in which 806 were classified in the *catalog* and 345 in the *event*.

We selected a random subset of 1,203 spreadsheets to evaluate the precision / recall of our system. The percentage of automatic recognition of the spreadsheets in the subset was approximately the same as the larger group. Our system achieved a precision of 0.84, i.e. 84% of retrieved spreadsheets were relevant; are call of 0.76, i.e. the system recognized 76% of all relevant spreadsheets; and an F-measure of 0.8. The accuracy was 93% and the specificity 95%, i.e. among all spreadsheets that the system does classified as not relevant, 95% were in fact not relevant.

The recognition rate of approximately 10.4% of the spreadsheets must consider that they were collected through a Web search tool. According Venetis et al., (2011), these search tools treat tabular structures like any piece of text, without considering the implicit semantics of their organization and thus causing imprecision in the search results. We further show an analysis of the data extracted from spreadsheets.

4.1 Pattern for Schema Location

The graph in Figure 4 shows that the spreadsheets schemas were concentrated on the initial lines and the percentage of matching per line of terms extracted from the spreadsheets against terms of our dictionary of terms. The quantity of spreadsheets that do not have schemas in the first lines decreases exponentially as we move away from the initial lines. We observed that most of the terms are located in the initial lines. Therefore, there is a tendency positioning schemas at the top followed by their respective instances.

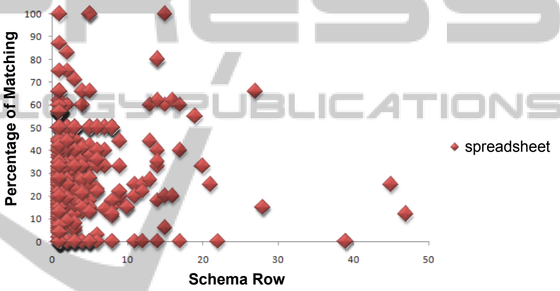


Figure 4: Terms by schema of initial lines.

4.2 Predominance of Terms and Spatial Distribution

In this stage of the analysis, we verified how much the predominant terms and their disposition in the schema can indicate of the spreadsheet nature: *catalog* or *event* (see explanation of these natures in Section 3.1).The schema fields were grouped in one of the six exploratory questions and they were weighted according to their position in the schema – a field will weigh less as it is far from the initial position.

Figure 5 shows the distribution of fields answering the *"what"* question in *catalog* spreadsheets as a gray map. Each block in the gray map represents a quadrant (set of cells) of the analyzed spreadsheets. Since spreadsheets have different sizes, the size (number of cells) of the quadrant to map each spreadsheet will vary proportionally to the size of the spreadsheet, in such a way that all spreadsheets are divided in the same number of quadrants/blocks.

The degree of gray indicates the percentage of spreadsheet fields answering the "*what*" question inside the quadrant. The results indicate that "*what*" fields are concentrated in the initial fields for *catalog* spreadsheets. Even though we found terms below the upper rows, the amount was incipient and thus we omitted of the figure.

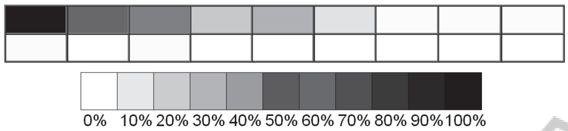


Figure 5: Distribution of what fields – catalog spreadsheets.

In *catalog* spreadsheets, the other five questions appear in smaller proportions. In order to perform a comparative analysis among proportions of the fields, we present a radar chart in Figure 6. Spreadsheets recognized as *catalog* tend to have many fields that answer the "*what*" question and some fields that answer the "*who*" question. The quantities of others questions were no significant. It delineates a pattern for *catalog* spreadsheets; it tends to have more fields to identify and detail specimens, with identification (what) fields in the beginning.

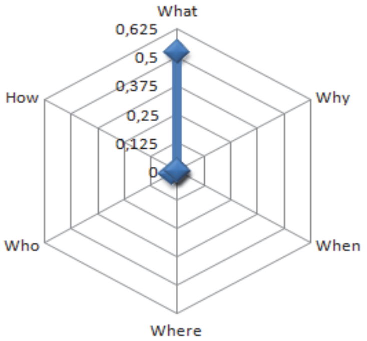


Figure 6: Proportions among fields – catalog spreadsheets.

Figure 7 and Figure 8 show the distribution of exploratory questions "*what*" and "*where*" in event spreadsheets. Compared to *catalog* spreadsheets, they have predominant fields answering the question "*where*" instead of "*what*".

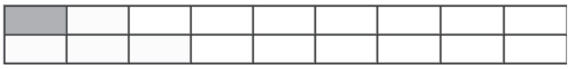


Figure 7: Distribution of what fields - event spreadsheets.

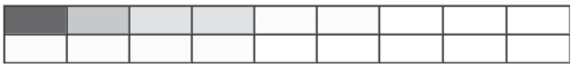


Figure 8: Distribution where fields - event spreadsheets.

Following the same approach of Figure 6, in Figure 9 we show the proportions of fields in *event* spreadsheets. In these spreadsheets there are lots of fields that answer the questions "*what*" and "*when*". Even though both have similar proportions in number, as we will show in the next chart, "*when*" fields predominate in the initial positions and hence by hypothesis H3, these field guides to identify the nature of the spreadsheet.

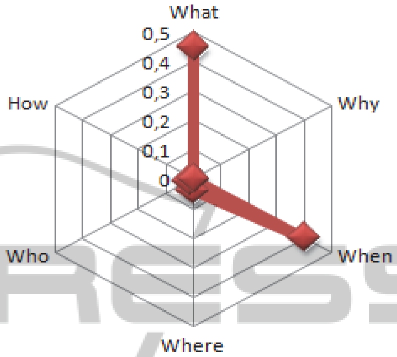


Figure 9: Relationship between the fields without positional weight - event spreadsheets.

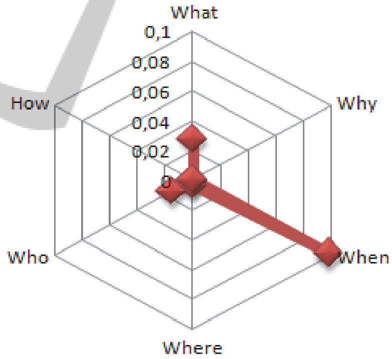


Figure 10: Relationship between the fields with positional weight - event spreadsheets.

Figure 10 emphasizes the importance of field positions in the schema. It is a variation of the chart in Figure 9, assigning weights to the fields that decrease exponentially as they move away from the initial columns. The chart in Figure 10 shows the sum of these weighted fields.

Analyzing Figure 9 singly, we tend to infer that "*what*" questions are as relevant as "*when*" questions, to characterize the pattern of *event* spreadsheets. However, Figure 10 indicates that fields answering "*when*" questions are mostly located in the initial positions. This positional differentiation can be exploited to drive the nature recognition process of spreadsheets. These observations motivated our proposition of a model to represent these patterns – presented in the next section – which takes into account relative positions of fields.

5 REPRESENTATION OF CONSTRUCTION PATTERNS

This section details the model, proposed in this paper, to capture and represent construction patterns in spreadsheets, which can be interpreted and used by machines. The characteristics of this model were based on field observations reported in the previous section. Therefore, even though we intend to conceive a generic model to represent patterns in spreadsheets for data management in general, in the present stage our analysis is focused in biology spreadsheets.

As detailed before, the schema recognition step involves analyzing patterns used by users to organize their data, which we argue to be strongly influenced by the spreadsheet nature inside a domain. Departing from our spreadsheet analysis, we produced a systematic categorization of construction patterns observed in biology spreadsheets, which supported the design of a process to recognize these patterns. Our process to recognize construction patterns and consequently the spreadsheet nature is focused on the schema recognition. The model presented in this section was designed to be used as part of this process, i.e. while a schema is recognized, the system tries to map it to a candidate model of a pattern, as illustrated in Figure 11.

Our representation approach considers that there is a latent conceptual model hidden in each spreadsheet, which authors express through patterns. How authors conceive models and transform them into spreadsheets is highly influenced by shared practices of the context in which the author is inserted, e.g., a biologist author cataloging specimens from of a museum. Her reference to build the catalog will be the specimens themselves, but also the usual strategy adopted by biologists of her community to tabulate data from specimens.

Thus, the construction patterns and the respective hidden conceptual models to be represented here reflect community or domain patterns and models. Figure 11 introduces two main patterns to be addressed by our representation approach, which follow the nature of the respective spreadsheet: a catalog and an event spreadsheet. Our analysis shows that a catalog spreadsheet contains taxonomic information of a specimen ("*what*" question) concentrated in the initial positions, defining their role as identifiers. On the other hand, an event spreadsheet contains temporal and location fields in the initial positions.

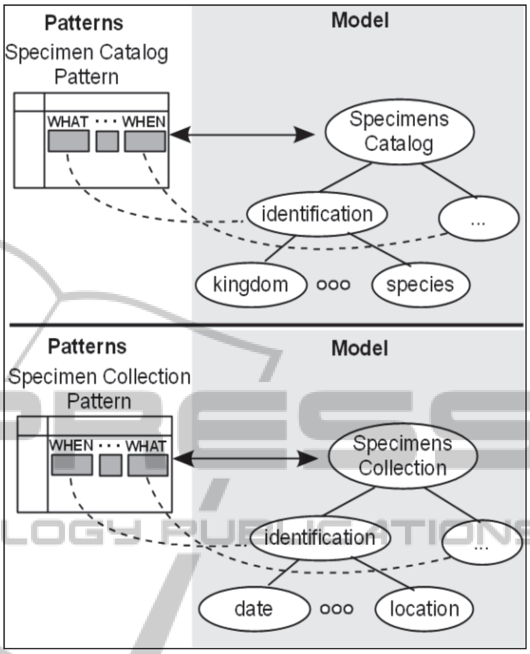


Figure 11: Matching between patterns and respective models.

Figure 12 shows a diagram that synthesizes a construction pattern for spreadsheets that *catalog* specimens. It was conceived from observations in the analysis described in the previous section. This diagram spatially delimits two main blocks of a spreadsheet: the implicit **Schema** and the data **Instances** which follow this schema. Delimited spaces identified by labels in the schema define the spatial organization of the hidden conceptual model.

This diagram visually introduces the rationale behind our model to represent a construction pattern and its relation with a conceptual model implicit in the organization. Elements of the conceptual model in the schema are represented by a hierarchy of labelled blocks. The innermost rectangles – e.g., kingdom and date – represent spreadsheet fields related to properties in an implicit schema. An inner block inside an outer block means a property which is part of a higher level property – e.g., kingdom is part of identifier.

A visual analysis of this diagram gives us directions of how the pattern is organized, e.g., schema up / instances down; identifier on the left, as a series of progressively specialized taxonomic references. To express these characteristics of the

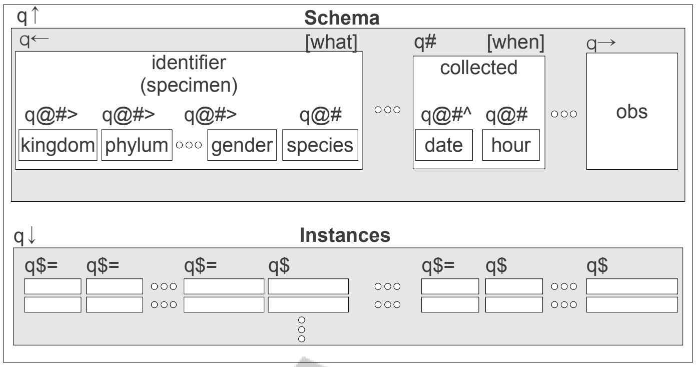


Figure 12: Construction pattern of catalog spreadsheet.

pattern in a computer interpretable representation, we represent them as qualifiers. We show qualifiers in the diagram of Figure 12 identified by the prefix "q", positioned on top of the upper left corner of the block they refer. They are categorized as follows:

Positional Qualifier – characterizes an element in a pattern according to its absolute position within a higher level element. There are four positional qualifiers: left (q←), right (q→), top (q↑) and bottom (q↓). In Figure 12 the positional qualifiers indicate that the schema is positioned on the top (above its instances); an identifier on the left of a schema and observations on the right.

Order Qualifier (q#) – characterizes an element in a pattern according to its relative order regarding its neighbouring elements. In Figure 12, each part of the identifier is recognized according its order (e.g., `kingdom` before `phylum` and `species` after `gender`); the collected property (date and time a specimen was `collected`) is positioned after the `identifier` and before `obs`.

Label Qualifier (q@) – indicates that the label characterizes the element. In the example, the label `species` identifies that this column refers to species.

Data Type Qualifier (q$) – characterizes the predominance of one data type in the instances of a given property. In the figure, the elements which are parts of the `identifier` are typed as strings.

Range Qualifier – specify if neighbour elements have generalization / specialization relations. The qualifier (q>) indicates that the left one is more general than the right one and (q<) the opposite.

Classified Qualifier – characterizes instances of a given property that are arranged in ascending order (q+) or descending order (q-). Figure 12 has no classified qualifiers. However, *event* spreadsheets further detailed have instances of date and time fields usually sorted in ascending order, receiving classified qualifiers in the pattern.

Redundancy Qualifier (q=) – characterizes redundancy of information in instances of a property. Such redundancy is typical, for example, in non-normalized relations among properties and composite properties, in which the values of a sub-property are broader or more generic of a related sub-property – usually the value of one sub-property embraces the value of the other. In the example, the `kingdom` sub-property embraces the `phylum` sub-property, which embraces the following sub-property and so on. Therefore, the `kingdom` is highly redundant, since several instances will have the same `kingdom`. The redundancy decreases while you move to more specialized sub-properties of identifier.

Besides the qualifiers in the diagram of Figure 12, we indicate between brackets the relation of elements with one of the six exploratory questions (*who*, *what*, *where*, *when*, *why*, *how*). This association will subsidize the characterization of construction patterns in a more abstract level. For example, looking at other kinds of *catalog* spreadsheets, outside the biology domain, we observed they define "*what*" fields as identifiers and they appear in the leftmost position (q←).

5.1 Formalizing the Model to Represent Patterns

In Figure 12 we informally introduced our model to represent patterns through a visual diagram. In this subsection, we will present a more formal representation, to be stored in digital format and to be read and interpreted by machines. This representation takes as a starting point the conceptual model implicitly expressed through the pattern. Figure 13 shows the representation of the construction pattern illustrated in Figure 12. The model is based on the OWL Semantic Web standard. The ovals represent classes (`owl:class`) and rectangles represent properties (`owl:ObjectPropertyorowl:DatatypePropert`).

The root class –`Specimen` in the figure –is related to the spreadsheet nature; in this case, instances of the `Specimen` class represent the instances of the spreadsheet, which catalogs specimens. A class will have a set of applicable properties, represented by a domain edge (`rdfs:domain`). Properties of this model are related to fields extracted from the spreadsheet. Range edges (`rdfs:range`)indicate that values of a given property are instances of the indicated class. For simplicity, the diagram omits details of the OWL representation.

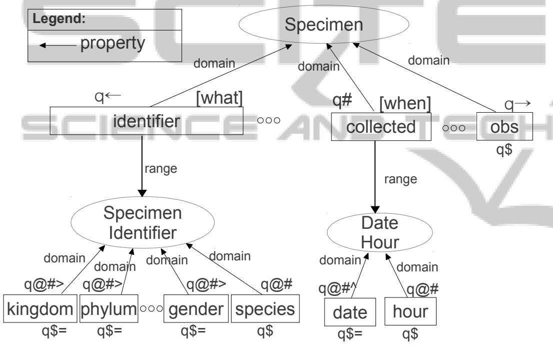


Figure 13: Conceptual model for catalog spreadsheets annotated with qualifiers.

Properties in our model are annotated by qualifiers presented in the previous section. Properties can be annotated in OWL through the `owl:AnnotationProperty`. In this case, annotations are objects that specify the qualifier and the pattern they are related. Qualifiers as annotations are depicted in Figure 13 above and/or below the properties they qualify. A qualifier above a property indicates that it is applied to the relationship between the property and the class to which it is applied by the domain relation. For example, the qualifier q← (left positional qualifier) is represented above the `identifier` property, indicating that when this property is used as a field in a spreadsheet describing a `Specimen`, we expect that it will appear in the left position.

A qualifier below the property means that it applies to property values – instances in the spreadsheet. For example, the qualifier q$= below the `kingdom` property indicates that a specific type (string) and redundancy are observed in the values of this property in the instances.

Properties are also annotated as answering one of the six exploratory questions. These annotations are depicted in Figure 13 inside brackets. There are additional concerns in the OWL model that are necessary to bridge it to the implicit spreadsheet schema, which are also represented as annotations: the order of properties and their relation with labels.

This OWL representation allows us to digitally materialize building patterns of spreadsheets, to be shared by users and applications. Figure 14 illustrates our OWL model applied to the characterization of an *event* spreadsheet used by biologists to record the log of specimens collected in the field – each specimen collected here is an event.

In this model we highlight that: in each instance (*event*), the time or the location is expected to be an identifier, positioned on the left (q←); values for time related properties (e.g., date and hour) will appear in ascending order (q+) in the instances.

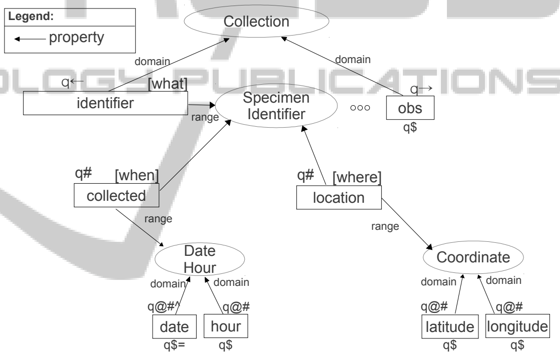


Figure 14: Conceptual model for collection spreadsheets enriched with qualifiers.

6 RELATED WORK

As discussed in Section 5, a fundamental characteristic of spreadsheets used for data management is the separation between schema and instances. The schema is presented above (q↑) or left (q←) and instances is below (q↓) or right (q→).

This observation appears in all the papers of related work, whose purpose is to recognize the implicit schema of spreadsheets. Syed et al., (2010) point out that this challenge leads to a more general problem of extracting implicit schemas of data sources – including databases, spreadsheets etc. One approach to make the semantics of spreadsheets interoperable, promoting the integration of data, is the manual association of spreadsheet fields to concepts in ontologies represented by open standards of the Semantic Web.

Han et al., (2008) adopt the simplest

approach to devise a schema and its respective instances, called entity-per-row(Connor et al., 2010). In this approach, besides the schema, each row of the table should describe a different entity and each column an attribute for that entity. The spreadsheet of Figure 1, for example, follows this kind of organization: each column corresponds to an attribute – e.g., `Date`, `Genus`, `Species` etc. – and each row to an event – a collection of a specimen. Han et al., (2008) and several related work assume the entity-per-row organization to support the process of manually mapping attributes, to make them semantically interoperable. Initially, the user must indicate a cell whose column contains a field which plays the role of identifier–equivalent to the primary key of a database. In the example of Figure1, it would be the field `date` and `time start`. Then, the system allows manual association between each cell of a field and an attribute of the semantic entity, considering that the respective column of the field will contain its values.

Langegger and Wob (2009) propose a similar, but more flexible, solution to map spreadsheets in an entity-per-row organization. They are able to treat hierarchies among fields, when a field is divided into sub-fields. In Figure 1, for example, the fields `Date`, `Time Start` and `Time End` refer to when the species was collected. It is usual that authors create a label spanning the entire range above these columns – e.g., labelled as "`CollectionPeriod`" – to indicate that all these fields are subdivisions of the larger field. This hierarchical perspective can be expressed in our model, since a property can be typed (`rdfs:range`) by a class, which in turn has properties related to it – e.g., the `identifier` property in Figure 13 is typed by the class `Specimen Identifier`, which affords the properties `kingdom`, `phylum` etc.

RDF has been widely adopted by related work as an output format to integrate data from multiple spreadsheets, since it is an open standard that supports syntactic and semantic interoperability. Langegger and WOB (Langegger and Wolfram, 2009) propose to access these data through SPARQL (Pérez et al., 2009) – a query language for RDF. Oconnor et al., (2010) propose a similar solution, but using OWL.

Abraham and Erwig (2006) observed spreadsheets are widely reused, but due to their flexibility and level of abstraction, the reuse of a spreadsheet by people outside its domain increases errors of interpretation and therefore inconsistency. Thus they propose a spreadsheet life cycle defined in two phases: development and use, in order to separate the schema of its respective instances. The schema is developed in the first cycle, to be used in the second cycle. Instances are inserted and manipulated in the second cycle guided by the schema, which cannot be changed in this cycle.

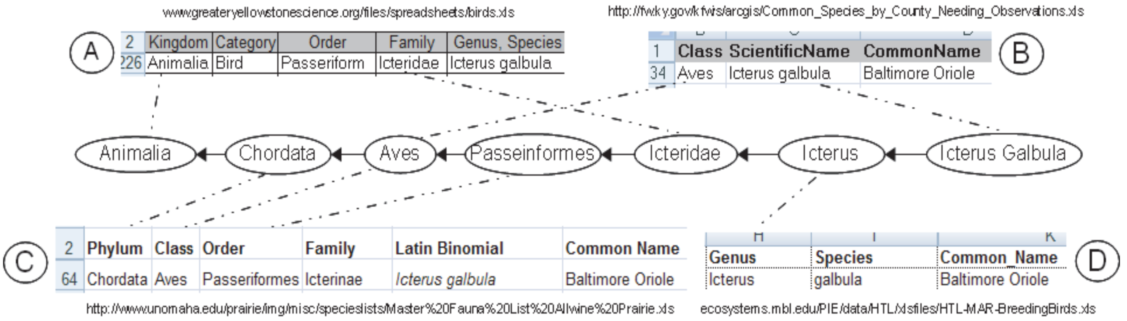
Another approach to address this problem is automating the semantic mapping using Linked Data. Syed et al., (2010) argue that a manual process to map spreadsheets is not feasible, so they propose to automate the semantic mapping by linking existing data in the spreadsheets to concepts available in knowledge bases, such as DBpedia (http://dbpedia. org) and Yago (http://www.mpi-inf.mpg.de/yago-naga/yago/).Yago is a large knowledge base, whose data are extracted, among others, from Wikipedia and WordNet (http://wordnet.princeton.edu). The latter is a digital lexicon of the English language, which semantically relates words.

Among the advantages of the last approach, there is the fact that such bases are constantly maintained and updated by people from various parts of the world. On the other hand, the search for labels without considering their contexts can generate ambiguous connections, producing inconsistencies. Thus, there are studies that stress the importance of delimiting a scope before attempting to find links.

Venetis et al., (2011) exploit the existing semantics in the tables to drive the consistent manipulation operations applicable to them. The proposal describes a system that analyzes pairs of terms heading columns and their relationship, in order to improve the semantic interpretation of them. Authors state that a main problem in the interpretation of tabular data is the analysis of terms independently. This paper tries to identify the scope by recognizing a construction pattern, which is related to a spreadsheet nature inside a context.

Jannach et al., (2009) state that the compact and precise way to present the data are primarily directed to human reading and not for machine interpretation and manipulation. They propose a system to extract information from web tables, associating them to ontologies. They organize the ontologies in three groups: **1. core:** concepts related to the model disassociated from a specific domain; **2. core + domain:** domain concepts of a schema related to the information to be retrieved; **3. instance of ontology:** domain concepts of instances. These ontologies aim at gradually linking the information to a semantic representation and directed by the user's goal.

Among these solutions, we note that some of them address individual pieces of information inside spreadsheets – devoid of context – and others



www.greateryellowstonescience.org/files/spreadsheets/birds.xls

2	Kingdom	Category	Order	Family	Genus, Species
226	Animalia	Bird	Passeriform	Icteridae	Icterus galbula

http://fw.ky.gov/kfwis/arcgis/Common_Species_by_County_Needing_Observations.xls

1	Class	ScientificName	CommonName
34	Aves	Icterus galbula	Baltimore Oriole

2	Phylum	Class	Order	Family	Latin Binomial	Common Name
64	Chordata	Aves	Passeriformes	Icterinae	*Icterus galbula*	Baltimore Oriole

http://www.unomaha.edu/prairie/img/misc/specieslists/Master%20Fauna%20List%20Allwine%20Prairie.xls

Genus	Species	Common_Name
Icterus	galbula	Baltimore Oriole

ecosystems.mbl.edu/PIE/data/HTL/xlsfiles/HTL-MAR-BreedingBirds.xls

Figure 15: Taxonomy chain create from spreadsheets.

consider the importance of identifying and characterizing the context. Even though all approaches rely on construction patterns of spreadsheets, none of them proposes a model to represent, exchange, reuse and refine these patterns, which is one of the main contributions of this work.

7 CONCLUSION AND FUTURE WORK

This paper presented our thesis that it is possible, from a spreadsheet structure, to recognize, map and represent how users establish construction patterns, which are reflected in the schema and data organization.

One of our main contributions here is a model to represent such patterns, as well as its association with a conceptual model to guide a process of automatic recognition.

Our process also involves the association between fields of spreadsheets and concepts available in ontologies. None of the related work departs from the characterization of the underlying conceptual models and their association with construction patterns, to categorize spreadsheets according to the nature of information they represent, and to recognize them.

Such categorization is essential to drive consistent operations according to the semantics and applicability of the extracted data, and to establish how data from different spreadsheets can be combined according to their type – e.g., (i) data from a spreadsheet containing *events* can be ordered and presented in a timeline; (ii) data from specimens in a museum (*catalog*) can be linked to records of their collections (*events*) in a specific manner.

To show the potential of this semantic characterization and the application of consistent operations over the extracted data, we implemented a prototype able to recognize schemas and extract data from several spreadsheets, mapping them to a semantic representation, which is combined with a taxonomic tree, see http://purl.org/biospread/?task=pages/txnavigator.

Figure 15 shows part of the interface of our prototype displaying a chain of the resulting taxonomic tree, populated combining partial fragments of spreadsheets “A”, “B”, “C” and “D”.

In spreadsheet “A”, the system finds the link between the Animalia Kingdom and the Icteridae Family. The instance in spreadsheet “B” links the Aves Class to the Icterus galbula Species –whose ScientificName term is related to the Species concept – and the Baltimore Oriole Common Name. In spreadsheet “C”, there are instances linking the Chordata Phylum, the Aves Class, the Passeriformes Order, the Icteridae Family and the Baltimore Oriole Common Name. Finally, in spreadsheet “D”, the Icterus Gender is linked to Baltimore Oriole Common Name.

Even though each spreadsheet has a partial fragment of the chain, their recognized schemas support instances integration and linking. Since spreadsheets were captured from several repositories on the Web, we observed that the instances showed greater diversity of format and data quality problems, whose proper integration is beyond the scope of this work and may be addressed in future work.

Our studies presented here have focused in the area of biodiversity. We intend to investigate its generalization to other domains of knowledge, extending this strategy to a semiotic representation.

In a starting approach, the proposal involves the creation and implementation of an interpretation model through the description of a system of signs. A sign is all that stands for something to someone and can be described as: sign = signifier + signified. The signifier is defined as the acoustic image/shape of the sign, the signified is the meaning/idea that the image has for each person or group (Saussure, 2011).

Extending this concept to spreadsheets, the signifier would be their structure and organization, which can represent of convention adopted by a community or user, reflected in a construction pattern. The significant would be the signifier associated with a particular meaning in the domain of the spreadsheet. As in linguistics, the relation between signifier/significant is indivisible; shape (signifier) depends on its usage (signified), as well as, idea (signified) depends on its representation (signifier) (Saussure, 2011).

Thus, our main concern in future work is to define and model spreadsheet elements in this sign model, besides interoperability issues, like name conflicts, domain conflicts etc.

ACKNOWLEDGEMENTS

Work partially financed by the Microsoft Research FAPESP Virtual Institute (NavScales project), the Center for Computational Engineering and Sciences - Fapesp/Cepid 2013/08293-7, CNPq (grant 143483/2011-0, MuZOO Project and PRONEX-FAPESP), INCT in Web Science (CNPq 557.128/2009-9), CAPES and FAEPEX- Unicamp, as well as individual grants from CNPq.

REFERENCES

Abraham, R. & Erwig, M., 2006. Inferring templates from spreadsheets. *Proceeding of the 28th international conference on Software engineering - ICSE '06*, 15, p.182.

Connor, M. J. O., Halaschek-wiener, C. & Musen, M. A., 2010. Mapping Master: a Flexible Approach for Mapping Spreadsheets to OWL. In *Proceedings of the International Semantic Web Conference*. pp. 194–208.

Doush, I. A. & Pontelli, E., 2010. Detecting and recognizing tables in spreadsheets. *Proceedings of the 8th IAPR International Workshop on Document Analysis Systems - DAS '10*, pp.471–478.

Han, L. et al., 2008. RDF123: from Spreadsheets to RDF. In *The Semantic Web*. Springer, pp. 451–466.

Haslhofer, B. & Klas, W., 2010. A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys*, 42(2), pp.1–37.

Jang, Seiie, Ko, Eun-Jung and Woo, W., 2005. Unified User-Centric Context: Who, Where, When, What, How and Why. In *Proceedings of the International Workshop on Personalized Context Modeling and Management for UbiComp Applications*. pp. 26–34.

Jannach, D., Shchekotykhin, K. & Friedrich, G., 2009. Automated ontology instantiation from tabular web sources—The AllRight system☆. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3), pp.136–153.

Langegger, A. & Wolfram, W., 2009. XLWrap – Querying and Integrating Arbitrary Spreadsheets with SPARQL. In *The Semantic Web*. pp. 359–374.

Mulwad, V. et al., 2010. Using linked data to interpret tables. In *Proceedings of the International Workshop on Consuming Linked Data*. pp. 1–12.

Ouksel, A. M. & Sheth, A., 1999. Semantic Interoperability in Global Information Systems A brief introduction to the research area and the special section. , 28(1), pp.5–12.

Pérez, J., Arenas, M. & Gutierrez, C., 2009. Semantics and complexity of SPARQL. *ACM Transactions on Database Systems*, 34(3), pp.1–45.

Ponder, W. F. et al., 2010. Evaluation of Museum Collection Data for Use in Biodiversity Assessment. , 15(3), pp.648–657.

Saussure, F. de, 2011. *Course in General Linguistics* R. Harris, ed.,

Syed, Z. et al., 2010. Exploiting a Web of Semantic Data for Interpreting Tables. , (April), pp.26–27.

Tolk, A., 2006. What comes after the Semantic Web - PADS Implications for the Dynamic Web. , pp.55–62.

Venetis, P. et al., 2011. Recovering Semantics of Tables on the Web. *Proceedings of the VLDB Endowment*, 4, pp.528–538.

Yang, S., Bhowmick, S.S. & Madria, S., 2005. Bio2X: a rule-based approach for semi-automatic transformation of semi-structured biological data to XML. *Data & Knowledge Engineering*, 52(2), pp.249–271.

Zhao, C., Zhao, L. & Wang, H., 2010. A spreadsheet system based on data semantic object. *2010 2nd IEEE International Conference on Information Management and Engineering*, pp.407–411.