

# Extracting and Semantically Integrating Data from Multiple Spreadsheets based on Recognition of their Nature

Ivelize R. Bernardo, Matheus S. Mota e André Santanchè

University of Campinas, UNICAMP, Brazil

`ivelize@lis.ic.unicamp.br`, `mota@ic.unicamp.br`, `santanche@ic.unicamp.br`

**Abstract.** Spreadsheets are popular among users and organizations, becoming an essential data management tool. The easiness to handle spreadsheets associated with the creative freedom offered by them resulted in the increase of the data volume available in this format. However, spreadsheets are not conceived for integration of data from distinct sources and challenges arise involving systematization of processes to reuse and combine their data. Many related initiatives address integration of data inside spreadsheets focusing on lexical and syntactical aspects, however, the exploration of the semantics related to this data is still an open challenge. In this sense, some related work propose mapping spreadsheets contents to open interoperability standards, mainly Semantic Web standards. The main limitation of such proposals is the assumption that it is possible to recognize and make explicit the schema and the semantics of spreadsheets automatically regardless of their domain. This work differs from related work by assuming the essential role of the context – mainly the domain in which the spreadsheet was conceived – to delineate shared practices of the community, which establishes building patterns to be automatically recognized by our system, in a data extraction process and schema recognition. In this article we present a result of a practical experiment involving such a system, in which we integrated data from hundreds of spreadsheets available on the Web. This integration was possible due to a unique ability of our approach of recognizing the spreadsheet nature, analyzed inside its creation context.

Categories and Subject Descriptors: H.2 [Database Management]: Miscellaneous; H.3 [Information Storage and Retrieval]: Miscellaneous; I.7 [Document and Text Processing]: Miscellaneous

Keywords: biology, interoperability, semantic web, spreadsheets

## 1. INTRODUCTION

Spreadsheets have been giving autonomy to end users to design their own tables, used to register and manage data [Chambers and Scaffidi 2010; Scaffidi et al. 2005].

Among the roles played by spreadsheets, we are interested in a relevant subset in which they are applied as “popular databases”. [Chambers and Scaffidi 2010] noted that, among spreadsheets produced by end users, 25% are used as databases

The growth of computing power associated with the advance of systems – which are able to handle increasingly larger spreadsheets – fostered a proliferation of these “popular databases” in different contexts. This phenomenon has as side effect the fragmentation of data, scattered in various files, containing informal and implicit schemas, which are designed to operate as isolated entities. These factors hamper data integration and the combination of data from distinct files.

There is a growing concern in transforming tabular data to open standards suitable for reuse and integration [Han et al. 2008; Langegger and Wöß 2009; Oconnor and Halaschek-Wiener 2010; Syed et al. 2010; Venetis et al. 2011]. This process can be enhanced by associating elements of spreadsheets

---

This work was partially supported by the Microsoft Research FAPESP, Virtual Institute (NavScales project), the Brazilian Institute for Web Sciences Research, CNPq (MuZOO project), PRONEX-FAPESP<sup>1</sup>, CAPES (AMIB project), as well as individual grants from CNPq.

Copyright©2012 Permission to copy without fee all or part of the material printed in JIDM is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

	A	F	G	H	I	K
1	DATE	TIME START	TIME END	Genus	Species	Common Name
6	03/06/2004	8:10	8:20	Geothlypis	trichas	Common Yellowthroat
7	03/06/2004	8:10	8:20	Icterus	galbula	Baltimore Oriole
8	03/06/2004	8:10	8:20	Melospiza	melodia	Song Sparrow

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

to concepts in knowledge bases available on the Web.

Among the approaches involving recognition of implicit spreadsheet schemas to make them explicit, there are initiatives meant to be generic to any context, resulting in a too wide spectrum of possibilities. Therefore, they do not explore context specificities to drive their recognition process. Moreover, instead of identifying a construction pattern, usually characterized by the nature of the spreadsheet, these initiatives focus on the recognition of individual labels. For example, spreadsheets to catalog specimens in a museum (nature of the spreadsheet) usually share a construction pattern – not analyzed by related work – which can guide their recognition.

In this article we assume that such recognition and mapping process will be more effective if we consider the context in which the spreadsheet was created. Users in a context – for example, a usage domain of biology – share practices which result in construction patterns. In a previous paper [Bernardo et al. 2012] we demonstrated that many of these patterns are likely to be recognized by computer programs and we have introduced our strategy for automatic recognition of such patterns. This article presents how our process to recognize schemas making them explicit was applied in the construction of a system able to transform several spreadsheets into a unified and integrated data repository. Our process includes automatic schema recognition and association between fields/records of spreadsheets with concepts available in ontologies. This system demonstrates a distinctive characteristic of our approach. Unlike related work, it is able to recognize the nature of many of the analyzed spreadsheets, producing data with such semantics, which can drive consistent combination operations over them.

This research is part of a larger project involving cooperation with biologists to build databases that integrate biodiversity data. We observed that biologists maintain a significant portion of their data in spreadsheets. In parallel, there are initiatives aimed at making biological data more flexible [Yang et al. 2005; Ponder et al. 2001] and shareable. They point out that although the information is rich in semantics, it is not properly explored, as formats adopted to its representation hamper its access and manipulation. For this reason, this research adopted the context of biology and spreadsheets oriented to data management as its specific focus. The remaining article is organized as follows. Section 2 presents an overview of related work. Section 3 introduces our process to make spreadsheet schemas explicit. Section 4 presents our system that integrates data from several spreadsheets based on the recognition of their nature. Section 5 presents a comparative analysis between our propose and related work. Section 6 presents the conclusion and future work.

## 2. RELATED WORK

There are several initiatives aimed to provide semantic interoperability for tabular data, in order to subsidize integration of data from different sources. Data management through spreadsheets can be treated as a specialized subset of this universe. In this section we present some relevant works in this direction. Figure 1 presents a spreadsheet containing data about birds specimens collected in the field, which will be used as an example to illustrate the analysis of related work.

The main factor to transform data from a spreadsheet into an open standard representation is the recognition of its schema in order to make it explicit. This process can be automatic, manual or semiautomatic. In the manual process, the user must locate spreadsheet elements that represent specific record fields and associate them to elements of an ontology. In most cases, the ontology will

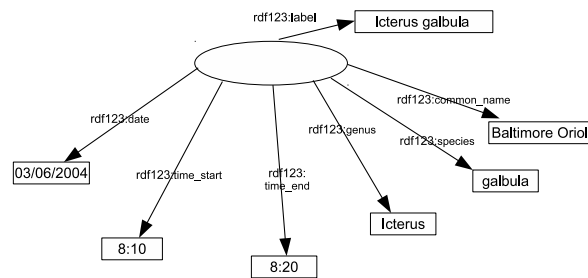


Fig. 2. Example with the result of a semantic mapping performed by [Han et al. 2008].

be represented in Semantic Web standards – RDF (*Resource Description Framework*) [Manola and Miller 2004] and OWL (*Web Ontology Language*) [WG 2009] – which are founded on a graph model, as those shown in Figures 2 and 4. Therefore, it will be a transformation of a tabular data into a graph.

[Han et al. 2008] applies the *entity-per-row* [Oconnor and Halaschek-Wiener 2010] manual mapping approach, apt only for tables with simple schemas. In this approach, each table row describes a different entity, to be mapped to a RDF instance. Each column refers to a descriptive attribute that becomes a RDF property. Figure 2 shows the resulting RDF graph of a semantic mapping, performed by using the [Han et al. 2008] technique, in one of the rows of the spreadsheet shown in Figure 1. The ellipse in the center refers to a RDF instance generated from the first row containing spreadsheet data. The attributes become edges (properties), whose values are the vertexes pointed by the edges. It is important to stress that the generated instance does not refer to any specific class. It reflects the focus of this approach – as well as all approaches which we will present in this section – in the recognition and mapping of attributes individually. However, in spreadsheets – as in other kinds of data management artefacts – attributes are combined to achieve a higher purpose, we call the *nature of the spreadsheet*. Our approach goes beyond. It is able to recognize the nature of several spreadsheets belonging the biology usage domain. It results in a semantically richer characterization of the generated instances. [Langegger and Wölk 2009] are not limited to the *entity-per-row* perspective and propose mapping implicit hierarchies found in spreadsheet schemas.

[Abraham and Erwig 2006] identified a specific subset of spreadsheets, which are adopted by users as templates to produce new ones, in a copy and adapt approach. However, people outside the creation context can produce errors and inconsistencies in the reuse process. Therefore, [Abraham and Erwig 2006] propose a life cycle for spreadsheets in two stages: development and use. The stages clearly devise the schema creation (development stage) from the data entry process (use stage). A schema created in the first stage cannot be changed in the second stage, reducing errors and inconsistencies.

In most cases, the manual semantic mapping is not feasible [Syed et al. 2010]. For this reason, some related work propose an automatic semantic mapping supported by external knowledge bases, as those provided by Semantic Web. [Syed et al. 2010] propose a generic mapping approach, which can be applied to any context. In order to map the attributes and values found in the spreadsheet to RDF properties and values, they associate spreadsheet attributes to concepts available in knowledge bases, as DBpedia (<http://dbpedia.org>) and Yago (<http://www.mpiinf.mpg.de/yago-naga/yago/>). One of the advantages of this approach is the fact that these bases are maintained and updated by people from all parts of the world. On the other hand, it can generate ambiguous and inconsistent links.

Applying this strategy to the case of Figure 1, an inconsistency could be generated by analyzing the **Genus** (Genus) column, which has different interpretations in different contexts. [Venetis et al. 2011] address the ambiguity problem making a correlation of table cells like a correlation between text fragments. Therefore, [Venetis et al. 2011] will address the ambiguity of **Genus** by relating it with **Species** (Species).

	A	B	C	D	E	F	G	H	I	J
1	observation_id	first_arrived_dt	species_id	common_name	scientific_name	country_cd	state_cd	town	latitude	longitude
10	507241	1873-04-08	BARS	Barn Swallow	Hirundo rustica	US	NJ	Caldwell	40,8398218	-74,2765366
11	302452	1873-05-22	INBU	Indigo Bunting	Passerina cyanea	US	NJ	Caldwell	40,8398218	-74,2765366
12	314326	1876-05-09	BAOR	Baltimore Oriole	Icterus galbula	US	MA	Worcester	42,2625932	-71,8022934
13	188815	1877-04-25	CHSW	Chimney Swift	Chaetura pelagica	US	NI	Caldwell	40,8398218	-74,2765366

(a) Spreadsheet sample recording events [<https://www.pwrc.usgs.gov>]

	A	B	C	D	E
1	Birds in Yellowstone National Park				
2	Kingdom	Category	Order	Family	Genus, Species
224	Animalia	Bird	Passeriformes	Icteridae	<i>Euphagus cyanocephalus</i>
225	Animalia	Bird	Passeriformes	Icteridae	<i>Icterus bullockii</i>
226	Animalia	Bird	Passeriformes	Icteridae	<i>Icterus galbula</i>
227	Animalia	Bird	Passeriformes	Icteridae	<i>Melospiza cinerea</i>

(b) Spreadsheet sample containing a catalog of species [[www.greateryellowstonescience.org](http://www.greateryellowstonescience.org)]

Fig. 3. Example spreadsheets

[Jannach et al. 2009] also apply a semantic mapping of terms during the Web tables extraction process. It involves three types of ontology: 1 - core: works as a meta schema describing a generic structure to be recognized; 2 - domain: elements of a schema in a specific domain to be recognized – it instantiates the core ontology; 3 - ontology instance: elements extracted from tables mapped to instances of the domain ontology. This process clusters related elements to put them in a context, improving the proper association with ontologies.

[Hermans et al. 2010] are able to automatically recognize the structure and content of a spreadsheet, transforming it in an UML representation. They adopt a three step approach: parse, prune and enrichment. In the first step, a parse tree is produced representing the internal structure of the spreadsheet, which is further pruned in the second step, to maintain just the relevant elements. In the last step, the pruned parsing tree is transformed in an UML class diagram through the recognition of patterns, represented as grammars.

[Limaye et al. 2010] adopt machine learning techniques to recognize the implicit schema. They start associating a type to each attribute and follow looking for binary relations between attributes. The recognized attributes are associated to concepts in the Yago knowledge base.

As we will further analyze in Section 5, most of related work focus their recognizing process on individual attributes, do not capturing the wider purpose of the spreadsheet perceived by the whole scenario, involving the combination of attributes and their disposition. The spreadsheet of Figure 1, for example, records events related to collections, created by biologists in the field. Most of Related work are able to recognize individual attributes, but not the wider scenario, i.e., that each record refers to an event (collection). It has a direct impact on the possibilities of integration and articulation of the resulting set, for example, if we wish to articulate an instance of the spreadsheet of Figure 1 with spreadsheets illustrated in Figure 3. The spreadsheet of Figure 1 records collection events as the spreadsheet of Figure 3(a). An operation to combine both spreadsheets, compatible with their nature, will be, for example, a merge operation, in which data from one spreadsheet can complement the other.

A third spreadsheet – illustrated in Figure 3b – has a different nature, as it contains a specimens catalog. Although it makes no sense merging this spreadsheet with data of Figures 1 and 3(a), their data can be articulated. For example, specific bee species indicated in the record collection can be linked to those of the catalog. As we will show below, our proposal is able to recognize such nature of each spreadsheet, which works as a “glue”, interrelating the semantics of each field with the semantics of the spreadsheet as a whole. The recognition of each nature will drive applications to apply consistent operations to data from spreadsheets.

This process follows the same methodology of the In Loco Semantics [Santanchè and Silva 2010], it interprets organization patterns and the user behaviour in order to automate part of the process involving in the identification and semantic mapping. This methodology is guided by the following

principles: *In Loco Annotation*: the annotation process occurs concomitant to the content production (in loco). *Metaphor Integration*: the metaphors and models adopted in content annotation are aligned with those adopted for content production. *Interoperability*: in loco annotation strategies are designed to enable automatic information extraction and conversion to Semantic Web open standards. *Semantic Persistence*: in loco annotation elements are connected to unification ontologies, which will guarantee their equivalent interpretations in different contexts, subsidizing the semantic persistence among transformations. In previous works of the In Loco Semantics focused on the recognition and data extraction of textual documents.

### 3. MAKING THE SCHEMA EXPLICIT DIRECTED BY THE SPREADSHEET NATURE

As mentioned earlier, this proposal involves the implementation of a process in which data from spreadsheets are extracted and transformed into RDF/OWL, to be stored in a repository. The central problem concerns the implicit schemas of spreadsheets, whose interpretation involves analyzing their data organization, which is strongly influenced by the nature of the spreadsheet and centered on its context. In contrast to related work, our approach is not intended to be generic, interpreting any kind of spreadsheet. It departs from a specific domain in order to recognize in it the shared patterns to build spreadsheets. For example, in a scenario of products sale, if the intention is to catalog products, the “product name” field will typically be among the first columns of the spreadsheet, however, if the purpose is to record each sale of these products, the “date of sale” will be among the first columns.

In [Bernardo et al. 2012] we have systematized building patterns observed in spreadsheets in the biology usage domain, which served as basis for the design of a process based on the recognition of these patterns. This process starts by mapping each field of the spreadsheet in the six exploratory questions (*who, what, where, when, why, how*). It works in a cyclical and incremental form [Bernardo et al. 2012], in which each new term and its disposition contributes to the recognition of the nature of a given spreadsheet. Recursively, new terms are more precisely semantically defined as far as the nature recognition evolves.

Departing from field observations, we noted that most of the spreadsheets in biology can be divided into four main groups: Group 1 - Objects: spreadsheets aimed at recording information about real world objects, e.g., specimens in the museum; Group 2 - Events: spreadsheets directed to recording events, e.g., sample collections; Group 3 - Classification: spreadsheets that systematize taxonomic classifications; Group 4 - Models: meta-spreadsheets whose records describe a schema for the construction of other spreadsheets.

As far as we expanded the universe for analysis of spreadsheets, increased the need to create a representation format that expresses how authors and users think and organize spreadsheets, explaining patterns shared by communities. For this reason, we are working on such a representation, making it open and independent of the program that performs the interpretation. This representation is apt to be interpreted by machines, in order to guide the recognition process.

### 4. SEMANTIC MAPPING, DATA INTEGRATION AND QUERYING

The recognition process, previous described, enabled the developed a semantic mapping process to RDF/OWL data. This mapping process exploits the recognition of the spreadsheet nature to generate semantically richer data. The prototype implemented here aims to demonstrate the potential for integration and linkage of data, extracted from spreadsheets, when its nature is recognized and made explicit.

Figure 4 presents an overview of the entire process behind our proposed and implemented strategy to extract and integrate data from spreadsheets. Steps 1 and 2 extract data from spreadsheets and then recognize their nature and respective schema. The extraction is performed by a module named

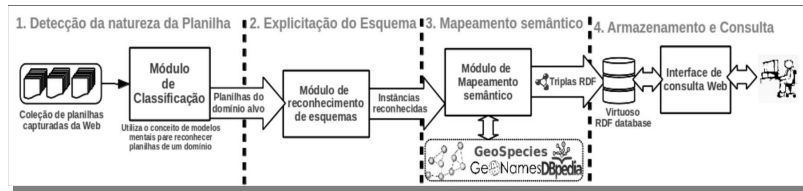


Fig. 4. Steps execution of recognition system and mapping spreadsheets

*Document Data Extractor*, developed in an associated work [Mota et al. 2009; Santanchè et al. 2009] and available at <http://code.google.com/p/ddex/>. In a previous work [Bernardo et al. 2012], we present details related to steps 1 and 2 of Figure 4.

In [Bernardo et al. 2012] we introduced how we took advantage of making the schema explicit and detecting the nature of the spreadsheet to enrich semantically the data extracted from spreadsheets (step 3 of Figure 4) and store them as RDF. The data are stored in a Virtuoso RDF database (<http://virtuoso.openlinksw.com>), which allows access through a WebService (available at <http://sparql.lis.ic.unicamp>).

In this work, we have evolved in the semantic mapping process (step 3) and have produced an enhanced system to explore these integrated data (step 4). The system takes advantage of the semantically richer data to link them to external datasets on the Web. It also includes a Web query interface that allows users to navigate and articulate the data extracted from the spreadsheets.

#### 4.1 Nature-driven Ontology Production: Running Example

As we presented in the previous section, the recognition of the spreadsheet nature plays an important role on our semantic mapping process and in to determine consistent operations over data. Returning to our examples in Figures 1 and 3, showing excerpts of two different spreadsheet files, they will be classified by our system (steps 1 and 2) in different natures – i.e., events and catalog – as presented before. It will reflect in distinct but connected RDF graphs.

The RDF graph of Figure 5 summarizes the result of our extraction process for both spreadsheets. The area highlighted in gray – identified as side (A) – represents the RDF mapping of the spreadsheet of Figure 3(a) (event) and the side (B) represents the RDF mapping of the spreadsheet of Figure 3(b). Unlike related work, the instance was recognized as a collection record and materialized in the RDF graph as an instance of the class `bio:Collect` (see an edge representing the property `rdf:type`). Moreover, the instance on the side (B) was recognized as a specimen in the museum and materialized as a RDF instance of the class `gs:SpeciesConcept`.

As illustrated in Figure 5, unlike the related work, in our approach the value assigned to each property is not limited to labels. In the specimen instance, for example, illustrated in the side (B) of Figure 5, it is possible verify that the property value for `gs:inFamily` – which indicates the animal’s family, represented using the GeoSpecies vocabulary (<http://lod.geospecies.org/>) – is an object instance. In this case, it is an instance of a specimen that represents the family Icteridae (`biospread:Icteridae`). Once a property is identified and mapped, the system tries to relate its values to knowledge bases – e.g., once the property `gs:inFamily` is recognized, the system tries to relate its values to the Geospecies base. The system was designed to link all specimens – recognized as members of this Icteridae family – to the same (`biospread:Icteridae`) object. Thus, it is possible congregate all the data from the spreadsheets at any level of characterization of a living being. For example, it is possible to compile all the data from a particular species or from an entire family and so on. As illustrated in the lower part of Figure 5, properties mapped into RDF are categorized as sub-properties of properties representing the six exploratory questions. For example, the properties to characterize a specimen (`bio:species`, `gs:inFamily`, `gs:inOrder` etc.) are sub-properties of the

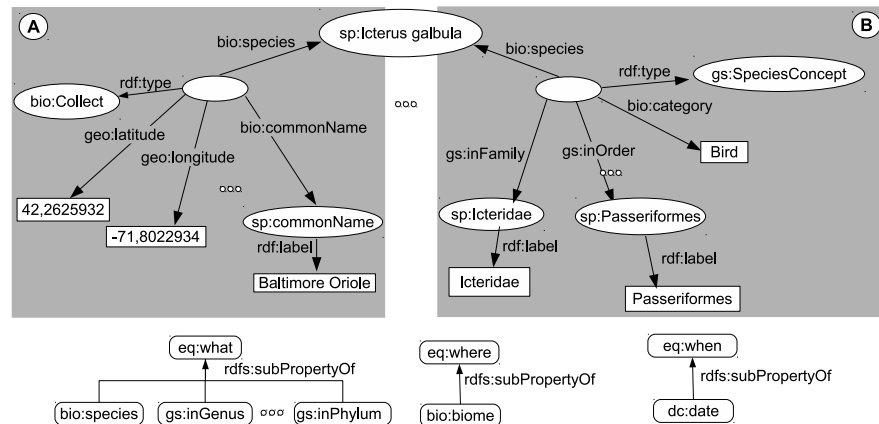


Fig. 5. Semantic mapping Fig.3.(a) and Fig.3.(b) spreadsheets

**eq:what** property and so on. This property classification allows to use the questions as a key for articulation. Collection instances can be articulated with specimen characterization instances around *what* properties, because their occurrence in both sides indicates common information – a specimen collected in one side is the specimen characterized in the other side.

From the implementation point of view, once the nature of the spreadsheet was recognized, the Extractor Module (step 2 of Figure 4) dispatches its RDF representation – already classified in its respective class – to the Semantic Mapping Module. The system tries to find references to a same object and unify them by pointing to an unified URI. For example, the mapper verify if there is already a URI for the species labeled as “Icterus Galbuga”. If already exists a URI, the mapper will link the data to it, if not, the module generate a new one. Additionally, the mapper use the labels forwarded by the extractor to relate terms to external knowledge bases. For instance, our module poses each label of species to the geospecies dataset, in order to find URIs that represent the same species. In the case of the species *Icterus Galbula*, our system was able to associate the produced URI (<http://purl.org/biospread/resource/species/icterusgalbula>) with the URI <http://lod.geospecies.org/ses/ePoGq>, from the geospecies database.

## 4.2 Practical Case and Query Interface

The practical case presented here aims to validate our prototype and demonstrate the potential for integration and linkage of data extracted from spreadsheets when its nature is recognized and made explicit. The practical experiment to validate our system involved gathering approximately 11,000 spreadsheets from the Web. They were located through the Google search engine, using keywords in Biology domain. The system automatically recognized and mapped 1,151 spreadsheets from which 806 were classified as Object spreadsheets and 345 were classified as Event spreadsheets. The graph shown in Figure 6 presents more information about the process. As can be seen, the process recognized 3 different kingdoms, 51 phyla and 33808 species. Also, 55248 different collection items were recognized, with 48034 georeferences (latitude, longitude).

Many spreadsheets were not recognized due to the strategy adopted to locate them through a search engine, which returns many spreadsheets out of the context. Up to now, the implemented system is able to recognize spreadsheets belonging to Group 1 and 2 (objects and events), as presented in the previous section. Once the system is able to recognize the nature of each spreadsheet and consequently of its instances, data could be combined and refined. The recognized nature of each record guided the application of consistent operations over it. In particular, it was possible to merge all catalog-typed records, extracted from the spreadsheets.

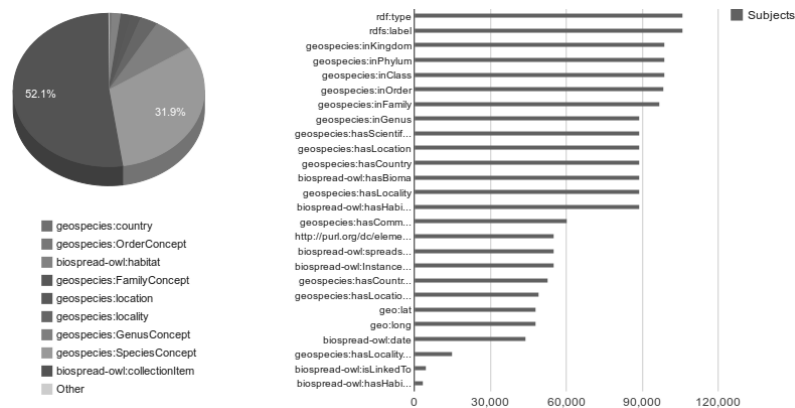


Fig. 6. Graph about Recognition Process [<http://purl.org/biospread/?task=pages/statistics>]

### A Taxonomy Navigator



Fig. 7. Screen print the query interface of the prototype [<http://purl.org/biospread/?task=pages/txnavigator>]

The Figure 7 presents a screenshot of our query and visualization prototype for data extracted from spreadsheets. This interface appears in step 4 of Figure 4.

It shows a practical example of how to explore the potential of articulation of our data in RDF. In this prototype, we aggregated 748,459 RDF specimen records obtained. By recognizing the nature of spreadsheets, it was possible to articulate data collected in the field with data describing species. Moreover, data of the same species were merged and aggregated in different levels taxonomic classification: kingdom, phylum, class, order, family, genus and species. Each aggregation level is filtered by a respective drop down box of the system's interface, illustrated in Figure 7. Every time the user characterizes a taxonomic level – e.g., by selecting a specific kingdom – the system will filter the records of the respective level. The georeferenced records are plotted over an interactive map – see bottom of Figure 7 – and might be automatically related to the Geospecies database. In this prototype version the data will be plotted over the map only when the user characterizes all taxonomic levels. The prototype has an interactive exploration interface in JavaScript, using the OpenLayers framework for maps (<http://openlayers.org>). This prototype is available in <http://purl.org/biospread/>



and both ontology and resources (instances) resulted from the practical case can be accessed through <http://purl.org/biospread/resource/> and <http://purl.org/biospread/ontology/> respectively.

Each flag in the map of Figure 7 represents a data collected in the field of a specimen. When the user clicks on the flag, it shows the data of the event (data collected), but enables also accessing the summary of data available of the respective species, by articulating this specimen with data collected of the same and other spreadsheets concerning the same species.

## 5. COMPARISON WITH RELATED WORK

Section 2 introduced our comparison with related work, evidencing that most proposals focus their recognition process in isolated attributes, do not exploring the nature of the spreadsheet, which can be captured by analyzing the combination of attributes and their disposition. Moreover, we further compare some specific issues.

Two of the related work [Zhao et al. 2010; Abraham and Erwig 2006] propose a change in the way that users produce spreadsheets, contrasting with our approach, which explores the latent semantics in the spreadsheet in their “natural habitat”. While [Zhao et al. 2010] tries to dissociate the semantic structure from the tabular structure of the spreadsheet, our proposal analyzes the pattern used by the user to organize the structure in order to infer the semantics from it. [Abraham and Erwig 2006] address an important phenomenon of spreadsheet reuse as templates, which contributes to establish building patterns. In our work we also explore such building patterns, but for an automatic recognition of the spreadsheet schema, since a systematized production process, as proposed by them, requires controlled environments.

Although the correlation between attributes, proposed by [Venetis et al. 2011], enhances the association between attributes and concepts in ontologies, their focus stay fragmented in isolated attribute interpretation.

Even though the grammar-driven mechanism proposed by [Hermans et al. 2010] is a powerful approach to capture patterns, our approach goes beyond a grammar by affording weight-based approximate patterns and several other kinds of pattern characterization, e.g., spatial relations. Moreover, our main target is semantically richer, as it addresses ontologies instead of UML. Besides the advantages provided by Semantic Web standards, RDF/OWL define properties as first class citizens. It was particularly relevant in our work, due to the importance of uniquely characterize each property to support merging and articulation of data, coming from distinct spreadsheets.

The [Limaye et al. 2010] achieve relevant results but, compared to our approach, recognize only binary relations instead of the nature of the whole spreadsheet. However, their approach have relevant contributions, which can complementary to our approach. Thus, we will explore them in a future work.

## 6. CONCLUSION AND FUTURE WORK

Spreadsheets have been having great acceptance among users of various segments, becoming “popular databases” arranged in files, which are difficult to integrate. To tackle this problem, many authors proposed solutions that recognize implicit schemas and map them to patterns of the Semantic Web. The differential of our work is to consider the context in which the spreadsheet was conceived essential to delineate the set of practices shared by the respective community, establishing building patterns to be recognized automatically by our system, in a process of data extraction, making schemas explicit.

We have implemented a prototype system, presented in this article, which can recognize schemas and extract data from hundreds of spreadsheets obtained from the Web. By recognizing the nature of spreadsheets, reflecting their semantics in the produced data, the system is able to perform consistent combinations with these data. This is a preliminary experiment of data integration. We are aware

of its limitations, especially regarding the quality of data coming from various sources. However, it validates our approach and demonstrates its potential for data integration.

This research inaugurated new challenges to be investigated, such as automatically discovering of articulation possibilities for data coming from different spreadsheets – even for data of different natures – and their respective integration. Such an integration will enable inferences that emerge from the combination of these data and which could not be obtained from an analysis of documents individually.

## REFERENCES

- ABRAHAM, R. AND ERWIG, M. Inferring templates from spreadsheets. In *Proceedings of the 28th international conference on Software engineering*. ICSE '06. ACM, New York, NY, USA, pp. 182–191, 2006.
- BERNARDO, I. R., MOTA, M. S., AND SANTANCHÈ, A. Extraíndo e integrando semanticamente dados de múltiplas planilhas eletrônicas a partir do reconhecimento de sua natureza. In *Simpósio Brasileiro de Banco de Dados (SBBD)*. pp. 256–263, 2012.
- BERNARDO, I. R., SANTANCHÈ, A., AND BARANAUSKAS, M. C. C. Reconhecendo padrões em planilhas no domínio de uso da biologia. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*. pp. 360–371, 2012.
- CHAMBERS, C. AND SCAFFIDI, C. Struggling to excel: A field study of challenges faced by spreadsheet users. *Visual Languages and Human-Centric Computing, IEEE Symposium on* vol. 0, pp. 187–194, 2010.
- HAN, L., FININ, T., PARR, C., SACHS, J., AND JOSHI, A. RDF123: from Spreadsheets to RDF. In *Seventh International Semantic Web Conference*. Springer, 2008.
- HERMANS, F., PINZGER, M., AND VAN DEURSEN, A. Automatically extracting class diagrams from spreadsheets. In *Proceedings of the 24th European conference on Object-oriented programming*. ECOOP'10. Springer-Verlag, Berlin, Heidelberg, pp. 52–75, 2010.
- JANNACH, D., SHCHEKOTYKHIN, K., AND FRIEDRICH, G. Automated ontology instantiation from tabular web sources—the allright system. *Web Semant.* 7 (3): 136–153, Sept., 2009.
- LANGEGGER, A. AND WÖSS, W. Xlwrap — querying and integrating arbitrary spreadsheets with sparql. In *Proceedings of the 8th International Semantic Web Conference*. ISWC '09. Springer-Verlag, Berlin, Heidelberg, pp. 359–374, 2009.
- LIMAYE, G., SARAWAGI, S., AND CHAKRABARTI, S. Annotating and searching web tables using entities, types and relationships. *Proc. VLDB Endow.* 3 (1-2): 1338–1347, Sept., 2010.
- MANOLA, F. AND MILLER, E. RDF Primer Ū W3C Recommendation. w3.org/TR/2004/REC-rdf-primer-20040210, 2004.
- MOTA, M. S., OLIVEIRA, N., COSTA, D. P., SANTANCHÈ, A., AND DALFORNO, C. Geração semanticamente dirigida e apresentação dinâmica de objetos digitais complexos na web. *VI Workshop de Trabalhos de Iniciação Científica – XV Simpósio Brasileiro de Sistemas Multimídia e Web (Webmedia)*., 2009.
- OCONNOR, M. J. AND HALASCHEK-WIENER, C. Mapping master: a flexible approach for mapping spreadsheets to owl. In *9th International Semantic Web Conference (ISWC2010)*, 2010.
- PONDER, W. F., CARTER, G. A., FLEMONS, P., AND CHAPMAN, R. R. Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology* 15 (3): 648–657, 2001.
- SANTANCHÈ, A., MOTA, M., COSTA, D., OLIVEIRA, N., AND DALFORNO, C. O. Componere: component-based in web authoring. In *Proceedings of the XV Brazilian Symposium on Multimedia and the Web*. WebMedia '09. ACM, New York, NY, USA, pp. 12:1–12:8, 2009.
- SANTANCHÈ, A. AND SILVA, L. A. M. Document-centered learning object authoring. In *IEEE Learning Technology Newsletter*. Vol. 12. pp. 58–61, 2010.
- SCAFFIDI, C., SHAW, M., AND MYERS, B. Estimating the numbers of end users and end user programmers. In *Visual Languages and Human-Centric Computing, 2005 IEEE Symposium on*. pp. 207 – 214, 2005.
- SYED, Z., FININ, T., MULWAD, V., AND JOSHI, A. Exploiting a Web of Semantic Data for Interpreting Tables. In *Proceedings of the Second Web Science Conference*, 2010.
- VENETIS, P., HALEVY, A., MADHAVAN, J., PAŞCA, M., SHEN, W., WU, F., MIAO, G., AND WU, C. Recovering semantics of tables on the web. *Proc. VLDB Endow.* 4 (9): 528–538, June, 2011.
- WG, W. O. OWL 2 Web Ontology Language Ū Document Overview. <http://w3.org/TR/2009/REC-owl2-overview-20091027>, 2009.
- YANG, S., BHOWMICK, S. S., AND MADRIA, S. Bio2x: a rule-based approach for semi-automatic transformation of semi-structured biological data to xml. *Data Knowl. Eng.* 52 (2): 249–271, Feb., 2005.
- ZHAO, C.-C., YONG ZHAO, L., AND LING WANG, H. A spreadsheet system based on data semantic object. In *Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on*. pp. 407 –411, 2010.