Extraindo e Integrando Semanticamente Dados de Múltiplas Planilhas Eletrônicas a Partir do Reconhecimento de Sua Natureza

Ivelize Rocha Bernardo, Matheus Silva Mota, André Santanchè

Instituto de Computação - Universidade Estadual de Campinas (UNICAMP)
13.083-970 - Campinas, SP - 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 ease of accessing associated with the creative freedom offered by spreadsheets 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 these 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 standards to be automatically recognized by our system, in a data extraction process and schema recognition. In this paper 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.

Resumo. Planilhas eletrônicas são populares entre usuários e organizações e se tornaram um instrumento fundamental de gerenciamento de dados. A facilidade de acesso associada à liberdade de criação oferecidas pelas planilhas resultaram no aumento do volume de dados disponíveis nesse formato. No entanto, planilhas não foram concebidas para a integração de dados de diferentes origens e desafios aparecem quando se trata da sistematização dos processos de reutilização e combinação de seus dados. Muitos trabalhos da literatura buscam a integração dos dados contidos em planilhas concentrando-se em aspectos léxicos e sintáticos, entretanto, a exploração da semântica associada a estes dados ainda é um desafio a ser superado. Alguns trabalhos propõem o mapeamento do conteúdo das planilhas para padrões abertos de interoperabilidade, principalmente aqueles da Web Semântica. A principal limitação destes trabalhos consiste no pressuposto de que é possível reconhecer e explicitar os esquemas e a semântica das planilhas automaticamente independentemente do seu domínio. Este trabalho se diferencia por considerar o papel essencial contexto – principalmente do domínio em que foi concebida a planilha – para se traçar o conjunto de práticas compartilhadas pela comunidade em questão, que estabelece padrões de construção a serem reconhecidos automaticamente pelo nosso sistema, em um processo de extração de dados e explicitação de esquemas. Neste artigo apresentamos o resultado de um experimento prático envolvendo tal sistema, no qual integramos os dados de centenas de planilhas eletrônicas disponíveis na Web. Tal integração foi possível pela capacidade única da nossa abordagem de reconhecer a natureza da planilha, analisada dentro de seu contexto de criação.

Categories and Subject Descriptors: H.Information Systems [**H.m. Miscellaneous**]: Interoperability

General Terms: Management

Keywords: interoperabilidade, planilhas eletrônicas, web semântica, biologia

1. INTRODUÇÃO

Planilhas eletrônicas têm assumido o caráter de "bases de dados populares". Através destas planilhas, autores não especialistas encontram autonomia para projetar tabelas em que registram e administram seus dados. Por um lado, tal facilidade de acesso combinada com o crescimento da capacidade computacional – acompanhado pelo avanço dos sistemas, que são capazes de manipular planilhas cada vez maiores – têm fomentado uma ampla multiplicação destas "bases populares" nos mais diversos contextos. Por outro lado, este fenômeno tem como efeito colateral a fragmentação dos dados, dispersos em diversos arquivos, contendo esquemas informais e implícitos, que foram projetados para atuar de forma isolada, dificultando a integração e articulação de dados de diferentes arquivos.

Há uma crescente preocupação no sentido de transformar dados tabulares em padrões abertos aptos ao reúso e integração [Han et al. 2008; Langegger and Wöß 2009; Oconnor and Halaschek-Wiener 2010; Syed et al. 2010; Venetis et al. 2011]. Tal como acontece em outras estratégias de extração de dados, abordagens para se obter interoperabilidade semântica podem ser divididas em três grupos: (i) mapeamento manual feito pelo usuário; (ii) reconhecimento automático de esquemas implícitos; (iii) reconhecimento semiautomático assistido pelo usuário. Em todos os casos, a meta da maioria dos trabalhos envolve mapear o esquema e seus dados para padrões abertos da Web Semântica. Nas abordagens (ii) e (iii) o reconhecimento pode ser incrementado pela associação dos elementos da planilha a conceitos disponíveis em bases de conhecimento da Web.

Dentre as abordagens que envolvem reconhecimento e explicitação do esquema implícito de uma planilha, as iniciativas analisadas se propõem a um reconhecimento genérico em qualquer contexto. Isto resulta num universo amplo de possibilidades de construção, em que não há restrição de um domínio específico, que possibilite o direcionamento do reconhecimento. Deste modo, ao invés de buscar a identificação de um padrão de construção, usualmente caracterizado pela natureza da planilha, as iniciativas se concentram no reconhecimento de rótulos individuais ou coocorrências. Por exemplo, planilhas que catalogam espécimes em um museu (natureza da planilha) usualmente compartilham um padrão de construção – não analisado por trabalhos relacionados – que pode guiar seu reconhecimento.

Neste trabalho partimos do pressuposto de que tal reconhecimento e mapeamento podem ser mais efetivos se considerarmos o contexto no qual a planilha foi criada. Usuários dentro de um contexto – por exemplo, um domínio de uso como o da biologia – compartilham práticas que resultam em padrões de construção. Em um trabalho anterior [Bernardo et al. 2012] demonstramos que muitos destes padrões são passíveis de serem reconhecidos por programas de computador e introduzimos nossa estratégia para reconhecimento automático de tais padrões.

Neste artigo apresentamos como tal processo de reconhecimento e explicitação foi usado na construção de um sistema capaz de transformar diversas planilhas eletrônicas em um repositório de dados unificado e integrado. Nosso processo inclui o reconhecimento automático do esquema e associação entre campos/registros das planilhas entre si e a conceitos disponíveis em ontologias. Este sistema demonstra o diferencial da nossa abordagem que, ao contrário dos trabalhos relacionados, é capaz de reconhecer a natureza de diversas planilhas analisadas e produzir dados com tal semântica, que direcionam operações consistentes de combinação destes dados.

Esta pesquisa foi motivada por um projeto maior em que está inserida, envolvendo a cooperação com biólogos para a construção de bases que integram dados de biodiversidade. Observamos que os biólogos mantêm uma parcela significativa de seus dados em planilhas eletrônicas e identificamos trabalhos voltados a tornar dados biológicos mais flexíveis [Yang et al. 2005; Ponder et al. 2001] e compartilháveis. Eles salientam que embora as informações sejam ricas em conteúdo semântico, não são exploradas o suficiente por estarem em um formato de difícil acesso e manipulação. Por esta razão, esta pesquisa adotou o contexto da biologia e planilhas eletrônicas voltadas ao gerenciamento de dados como seu foco específico. O restante do artigo está organizado da seguinte forma. A Seção 2 apresenta uma visão geral da literatura correlata. A Seção 3 introduz nosso processo de explicitação

esquema das planilhas. A Seção 4 apresenta nosso sistema que integra dados de planilhas a partir do reconhecimento de sua natureza. A Seção 5 apresenta a conclusão e trabalhos futuros.

2. REVISÃO DA LITERATURA

Há diversas pesquisas que se propõem a alcançar interoperabilidade semântica para dados tabulares, de modo que se possa realizadar a integração de dados de diversas fontes. O gerenciamento de dados em planilhas eletrônicas pode ser tratado como um subconjunto especializado deste universo. A seguir apresentaremos alguns trabalhos relevantes neste sentido. A Figura 1 apresenta uma planilha que registra informações sobre coleta de espécimes de abelhas, que será usada como exemplo para ilustrar a análise dos trabalhos relacionados.

	A	B	C	D	E	H	I	M	N	O
1	**Planilha Geral dos Dados da Biota**									
2	**Arquivo de origem: Abelhas 2o. Periodo**									
3										
4	ID *	Número da unidade de coleta **			Espécie	Grupo	Bioma ***	dia	mês	ano
5		Área	Coluna	Linha						
6	1	A1ME	1ª coluna	1º Linha	Eulaema cingulata	Abelhas	Amazônia	27	1	2008
7	2	A1ME	1ª coluna	1º Linha	Euglossa sp1	Abelhas	Amazônia	27	1	2008
8	3	A1ME	1ª coluna	1º Linha	Eulaema cingulata	Abelhas	Amazônia	27	1	2008

Fig. 1. Exemplo de planilha de registro de coleta [http://siscom.ibama.gov.br].

O principal fator para a transformação dos dados de uma planilha em um padrão aberto é o reconhecimento e explicitação de seu esquema. Conforme foi apresentado, este processo pode ser automático, manual ou semiautomático. No processo manual, o usuário deve localizar na planilha elementos que representam campos específicos de registros, associando-os a elementos de uma ontologia. Na maioria dos casos a ontologia estará representada nos padrões da Web Semântica – RDF (*Resource Description Framework*) e OWL (*Web Ontology Language*) – pautados sobre um modelo de grafos como aqueles apresentados nas Figuras 2 e 4. Trata-se, portanto, de transformar dados em um formato tabular para um grafo.

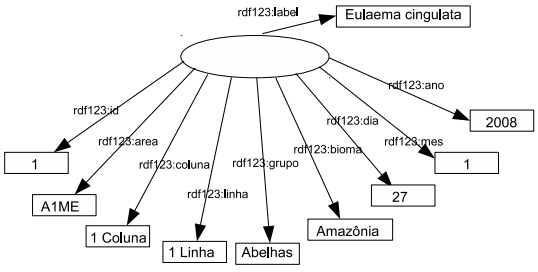

Fig. 2. Exemplo com resultado de mapeamento semântico realizado por [Han et al. 2008].

[Han et al. 2008] utilizam uma abordagem de mapeamento manual *entity-per-row* [Oconnor and Halaschek-Wiener 2010] apta apenas para tabelas de estruturas simples. Nesta abordagem, cada linha da tabela descreve uma entidade diferente, a ser mapeada para uma instância RDF. Cada coluna se refere a um atributo descritivo que se converte em uma propriedade RDF. A Figura 2 apresenta o grafo RDF resultante do mapeamento semântico de [Han et al. 2008] em uma das linhas da planilha apresentada na Figura 1. A elipse no centro se refere a uma instância RDF gerada a partir da primeira linha de dados da planilha. Os atributos se convertem em arestas (propriedades), cujos valores são vértices apontados pelas arestas. É importante ressaltar que a instância gerada não se refere a nenhuma classe específica. Isso acontece porque o foco desta abordagem, bem como o

de todas as que apresentaremos nesta seção, é a associação semântica de atributos individualmente. Entretanto, os atributos se combinam visando um propósito maior, que denominamos a natureza da planilha. Nossa abordagem vai além. Ela é capaz de reconhecer a natureza de diversas planilhas no domínio de uso da biologia. Isto se reflete em uma caracterização semanticamente mais rica das instâncias geradas.

[Langegger and Wöß 2009] vão além do *entity-per-row* e propõem esquemas de mapeamento de hierarquias implícitas encontradas em planilhas. Sua abordagem é capaz de interpretar o agrupamento de células "número da unidade de coleta" (Figura 1), transformando-o em uma hierarquia de objetos.

Tratando de dados tabulares em documentos Web, [Syed et al. 2010] consideram que realizar o mapeamento semântico de forma manual é inviável. Sua proposta é automatizar esse processo e sua abordagem se propõe a ser aplicada a qualquer contexto. Para mapear os atributos e valores encontrados na planilha para propriedades e valores RDF, é feita uma associação entre atributos da planilha e conceitos disponíveis em bases de conhecimentos, como DBpedia (http://dbpedia.org) e Yago (http://www.mpiinf.mpg.de/yago-naga/yago/). Dentre as vantagens desta abordagem está o fato de que tais bases são mantidas e atualizadas por pessoas de todas as partes do mundo. Uma limitação da mesma está no fato de que ela pode gerar ligações ambíguas e inconsistentes. Além disto, seu enfoque também está nos atributos.

Aplicando esta estratégia no exemplo da Figura 1, uma inconsistência seria gerada ao analisar a coluna `Grupo`, que terá diferentes interpretações em contextos distintos. [Venetis et al. 2011] abordam a problemática da ambiguidade tratando a correlação de dados de células em tabelas como se fosse a correlação entre fragmentos de texto. Assim, [Venetis et al. 2011] tentariam resolver a ambiguidade do atributo `Grupo` relacionando-o com `Espécie` ou `Bioma`. Apesar de aprimorar a associação entre atributos e termos em ontologias, a interpretação continua com o enfoque fragmentado nos atributos.

O enfoque dado nos atributos, pelos trabalhos relacionados, impede uma interpretação mais ampla da natureza e propósito da planilha. A planilha da Figura 1, por exemplo, registra eventos relativos a coletas feitas por biólogos em campo. Os trabalhos relacionados são capazes de reconhecer atributos individuais, mas não o fato de que cada registro se refere a um evento (coleta). Isto tem um impacto direto nas possibilidades de integração e articulação dos dados resultantes, por exemplo, se desejarmos articular uma instância da planilha na Figura 1 com as planilhas ilustradas na Figura 3. Tal como a planilha da Figura 1, a planilha da Figura 3(a) também registra eventos de coleta. Uma operação de combinação compatível com a natureza de ambas as planilhas é aquela de *merge*, em que os dados de uma podem complementar os da outra. A planilha da Figura 3(b) é de outra natureza, trata-se de um catálogo de espécimes. Apesar de não fazer sentido um *merge*, dados das planilhas da Figura 1 e Figura 3(a) podem se articular com aqueles da Figura 3(b). Por exemplo, as abelhas indicadas no registro de coleta podem ser associadas àquelas do catálogo. Como demonstraremos adiante, nossa proposta é capaz de reconhecer tais naturezas, que funcionam como uma "cola" inter-relacionando a semântica de cada campo com aquela da planilha como um todo. O reconhecimento destas naturezas direcionará a aplicação dos tipos de operação compatíveis com os dados das planilhas.

	A	B	C	D	E	F	G
1	**Espécie**	**Área**	**Data**	**Hora**	**Isca**	**Amostra**	**coleta**
22	*Euglossa variabilis*	Bom Jardim	8/7/2008	11:47	S.M.	1	2
23	*Eulaema cingulata*	Bom Jardim	8/7/2008	11:01	S.M.	1	2
24	*Eulaema meriana*	Bom Jardim	8/7/2008	13:01	S.M.	1	2

(a) Exemplo planilha registro de evento [http://siscom.ibama.gov.br]

	C	D	E	F	G	H	I	J	K	M	O
1	POLINIZADOR										
2	**Filo**	**Subfilo**	**Clase**	**Orden**	**Superfamilia**	**Familia**	**Subfamilia**	**Género**	**Especie**	**Etapa de Vida**	**Función / Act**
73	Arthropod	Hexapoda	Insecta	Hymenoptera	Apoidea	Apidae	Apinae	*Eulaema*	*Eulaema bombiformis*	Adulto	Polinizadores de varios
74	Arthropod	Hexapoda	Insecta	Hymenoptera	Apoidea	Apidae	Apinae	*Eulaema*	*Eulaema cingulata*	Adulto	Polinizadores de varios
75	Arthropod	Hexapoda	Insecta	Hymenoptera	Apoidea	Apidae	Apinae	*Eulaema*	*Eulaema meriana*	Adulto	Polinizadores de varios

(b) Exemplo planilha catálogo de espécimes [www.raaa.org.pe]

Fig. 3. Exemplo de planilhas.

Esse processo segue a mesma linha utilizada pela metodologia Semântica In Loco [Santanchè and Silva 2010], pois interpreta padrões de organização e comportamento do usuário, a fim de automatizar parte do processo envolvido na identificação da semântica e mapeamento. Tal metodologia está

pautada sobre os seguintes princípios: *Anotação In Loco*: ao criar o conteúdo, o autor segue alguns padrões de comportamento e organização que são interpretados como anotações; como resultado, este processo de anotação implícito acontece junto com a produção de conteúdo. *Integração de Metáforas*: as metáforas e modelos utilizados para anotação implícita do conteúdo estão alinhadas com aquelas utilizadas para a produção do mesmo. *Interoperabilidade*: as estratégias de anotação são projetadas a fim de possibilitar identificação automática dos esquemas implícitos e conversão para padrões abertos da Web Semântica. *Persistência Semântica*: elementos da anotação e esquemas explicitados são associados a ontologias unificadoras, que irão garantir interpretações equivalentes em diferentes contextos, subsidiando persistência semântica entre as transformações. Trabalhos anteriores da Semântica In Loco tiveram como foco o reconhecimento e extração de dados a partir de documentos textuais.

3. EXPLICITAÇÃO DE ESQUEMA DIRIGIDA PELA NATUREZA DA PLANILHA

Conforme mencionado anteriormente, esta proposta envolve a implantação de um processo em que os dados das planilhas são extraídos e transformados em RDF/OWL, para serem armazenados em um repositório. A questão central é que planilhas possuem esquemas implícitos, cujo processo de interpretação envolve a análise da organização dos dados, que é fortemente influenciada pela natureza da planilha e centrada no contexto. Ao contrário dos trabalhos relacionados, nossa abordagem não se propõe a ser genérica e interpretar qualquer tipo de planilha. Ela parte de um domínio específico e busca reconhecer, dentro do mesmo, padrões compartilhados de construção de planilhas. Por exemplo, num cenário de vendas de produtos, se a intenção é catalogar produtos, normalmente o registro "nome do produto" estará entre os primeiros registros da planilha, no entanto, se o objetivo é registrar as vendas desses produtos, a data da venda estará entre os primeiros registros.

Em [Bernardo et al. 2012] sistematizamos padrões de construção de planilhas no domínio de uso da biologia, que serviu como base para o projeto de um processo baseado no reconhecimento destes padrões. Esse processo abstrai campos específicos da planilha enquadrando-os nas seis perguntas exploratórias *(who, what, where, when, why, how)*. Ele funciona de forma cíclica e incremental [Bernardo et al. 2012], em que cada novo termo e a sua disposição na planilha contribui para o reconhecimento da natureza da mesma. Recursivamente, na medida em que se configura uma natureza é possível definir com mais precisão a semântica de novos termos.

A partir de observações em campo, verificamos que a maioria das planilhas em biologia podem ser divididas em quatro grupos principais: Grupo 1 - Objetos: planilhas voltadas ao registro de informações sobre objetos, e.g., espécies no museu; Grupo 2 - Eventos: planilhas direcionadas a registros de eventos, e.g., coletas de amostras; Grupo 3 - Classificação: planilhas que sistematizam classificações taxonômicas; Grupo 4 - Modelos: meta-planilhas cujos registros descrevem um esquema para a construção de outras planilhas.

Na medida em que ampliamos o universo de análise de planilhas, tem sido crescente a necessidade de se criar um formato de representação que expresse a forma como autores e usuários pensam e organizam as planilhas, explicitando os padrões compartilhados por comunidades. Por esta razão estamos trabalhando em tal representação, para torná-la aberta e independente do programa que realiza a interpretação. Tal representação é passível de interpretação por máquinas, de modo a guiar o processo de reconhecimento.

4. MAPEAMENTO SEMÂNTICO A PARTIR DO CONTEXTO EXPLICITADO

A partir do processo de reconhecimento implementado nas etapas anteriores, neste trabalho desenvolvemos um processo de mapeamento semântico dos dados para RDF/OWL. Tal mapeamento explora o reconhecimento da natureza da planilha para gerar dados semanticamente mais ricos. O caso prático aqui implementado visa demonstrar o potencial de integração e articulação dos dados extraídos de planilhas, uma vez que sua natureza é reconhecida e explicitada.

O grafo RDF da Figura 4 sintetiza o resultado obtido do nosso processo de extração. A área destacada em cinza identificada como lado (A) representa o mapeamento RDF da planilha da Figura 1 (evento) e o lado (B) representa o RDF da planilha da Figura 3(b). Diferentemente dos trabalhos relacionados, a instância foi reconhecida como um registro de coleta e materializado no grafo RDF na forma de uma instância da classe `bio:Collect` (vide aresta representando a propriedade `rdf:type`). Por outro lado, a instância no lado (B) foi reconhecida como sendo uma espécime no museu e materializada em RDF como instância da classe `gs:SpeciesConcept`.

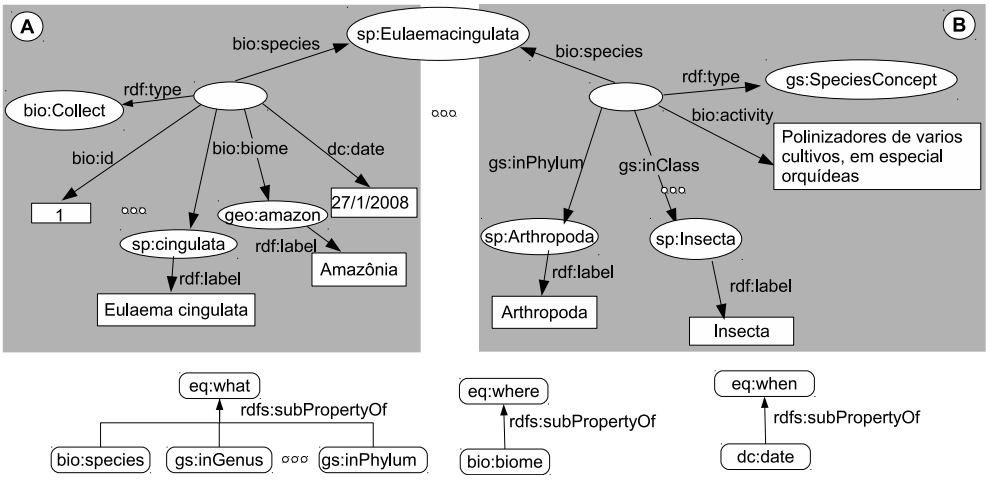

Fig. 4. Mapeamento semântico das planilhas Fig.2 e Fig.3.(b).

O experimento prático de validação deste sistema envolveu a coleta de 11.150 planilhas na Web. As planilhas foram localizadas a partir da ferramenta de busca Google, pelo uso de palavras chaves do domínio. Das 11.150 planilhas, foram reconhecidas e mapeadas automaticamente 1.151, em que 806 planilhas foram classificadas como objeto e 345 planilhas foram classificadas como evento. Ao todo foram reconhecidos 748.459 registros de espécimes dentro destas planilhas. Como resultado, foram produzidas 21.254 triplas RDF e foram reconhecidos 1.471 locais distintos, nos quais espécimes estão registrados. Dentre as razões para o não reconhecimento de muitas planilhas, está a estratégia usada para a sua captação, dado que a ferramenta de busca retorna muitas planilhas fora do contexto. Até o presente momento o sistema está preparado para o reconhecimento das planilhas do Grupo 1 e 2 (objetos e eventos) apresentados na seção anterior. Uma vez que o sistema é capaz de reconhecer a natureza da planilha e consequentemente das instâncias, os dados puderam ser combinados e refinados. Em especial, foi feito um merge de todos os registros de catálogo extraídos das planilhas.

A Figura 5 apresenta as etapas de execução do nosso sistema. Nas etapas 1 e 2 é realizada a extração de dados da planilha, o reconhecimento de sua natureza e de seu respectivo esquema. A extração de dados é realizada por meio de um módulo denominado *Document Data Extractor* desenvolvido em um trabalho relacionado [Mota et al. 2009; Santanchè et al. 2009]. Na etapa 3 o módulo de mapeamento transforma os dados em RDF. Na etapa 4 estes dados são armazenados em um banco de dados RDF chamado Virtuoso (http://virtuoso.openlinksw.com), que permite o acesso por uma interface Web.

Como ilustra a Figura 4, ao contrário dos trabalhos relacionados, em nossa abordagem o valor atribuído a cada propriedade não se limita a rótulos. Na instância da espécime, por exemplo, que está no lado (B) da Figura 4, é possível verificar que o valor da propriedade `gs:inPhylum` – que indica o filo do animal representado usando o vocabulário GeoSpecies (http://lod.geospecies.org/) – é por sua vez uma instância de um objeto. Neste caso, trata-se de uma instância que representa o filo Arthropoda (`sp:Arthorpoda`) do espécime. O sistema foi projetado para que todos os espécimes reconhecidos associados a este filo apontem para este mesmo objeto. Desta maneira, é possível congregar todos

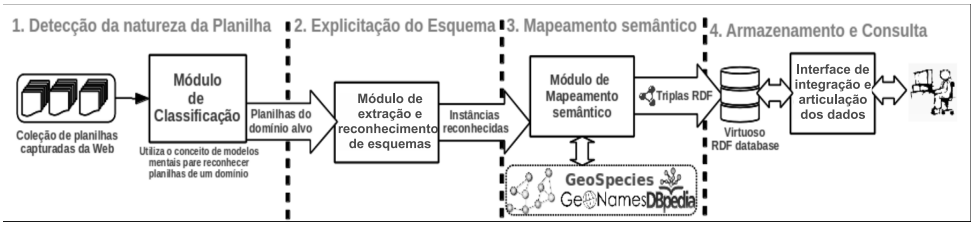

Fig. 5. Etapas de execução do sistema de reconhecimento e mapeamento de planilhas.

os dados extraídos da planilha em qualquer nível da caracterização de um ser vivo. Por exemplo, é possível compilar todos os dados de uma espécie específica, ou uma família inteira e assim por diante.

Para alcançar este mapeamento semântico, nossa abordagem combinou três estratégias: (i) ao reconhecer o padrão de construção da planilha e sua natureza – a partir dos campos usados e sua ordem e disposição – o sistema define a que classe pertencerão as instâncias produzidas; (ii) a natureza da planilha orienta o reconhecimento e mapeamento semântico das propriedades; (iii) uma vez identificada e mapeada a propriedade, seus valores poderão ser associados a bases de conhecimento – e.g., uma vez reconhecida a propriedade inPhylum, é possível associar seus valores a dados obtidos da base Geospecies.

Como ilustra a parte inferior da Figura 4, as propriedades mapeadas em RDF são categorizadas como sub-propriedades das seis perguntas exploratórias. Por exemplo, as propriedades de caracterização do espécime (`bio:species`, `gs:inGenus`, `gs:inPhylum` etc.) são sub-propriedades da propriedade `eq:what` e assim por diante. Esta classificação por propriedade possibilita usar as perguntas como chave de articulação. Instâncias de coletas podem ser articuladas com instâncias de espécimes em torno da propriedade *what*, já que a sua ocorrência em ambas indica uma informação em comum – a espécime coletada de um lado é a espécime da coleção de outro.

A Figura 6 apresenta uma cópia de tela do protótipo de consulta e visualização dos dados extraídos das planilhas. Esta interface está inserida na Etapa 4 da Figura 5. Ela mostra um exemplo prático de como exploramos o potencial de articulação de nossos dados em RDF. Neste protótipo agregamos os dados RDF dos 748.459 registros de espécimes obtidos. Os dados foram agregados por espécie e foram filtrados os registros georeferenciados ou que puderam ser relacionados automaticamente com a base de dados Geonames (http://www.geonames.org/) ou Geospecies. Foi desenvolvida uma interface interativa em JavaScript, sobre o framework de mapas OpenLayers (http://openlayers.org), na qual podem ser visualizados interativamente os registros.

5. CONCLUSÃO E TRABALHOS FUTUROS

As planilhas eletrônicas obtiveram grande aceitação entre usuários de vários segmentos, tornando-se "bases de dados populares", dispostas em arquivos de difícil integração. Como forma de solucionar este problema, muitos autores propuseram soluções utilizando diversas tecnologias que reconhecem esquemas implícitos e os mapeiam para padrões da Web Semântica. Este trabalho se diferencia por considerar o contexto em que foi concebida a planilha essencial para se traçar o conjunto de práticas compartilhadas pela comunidade, que estabelece padrões de construção a serem reconhecidos automaticamente por nosso sistema, em um processo de extração de dados e explicitação de esquemas.

Foi implementado o protótipo de um sistema, apresentado neste artigo, capaz de reconhecer esquemas e extrair dados de centenas de planilhas obtidas na Web. Por reconhecer a natureza das planilhas, cuja semântica se reflete nos dados produzidos, o sistema foi capaz de realizar combinações consistentes entre os dados. Este é um experimento preliminar de integração de dados. Estamos cientes das

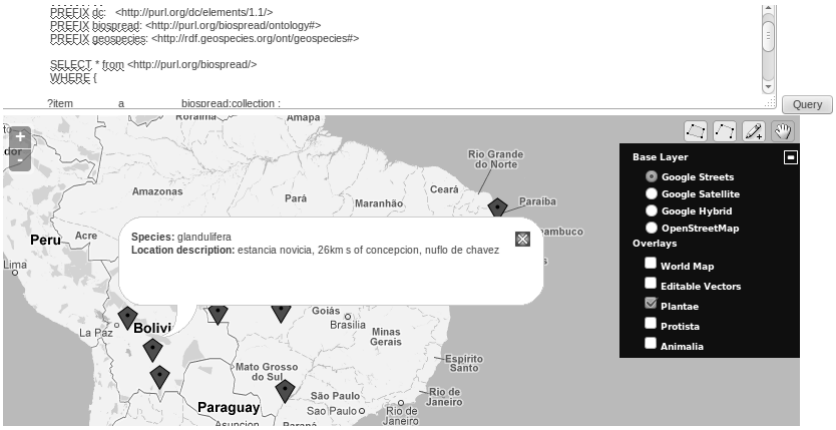

Fig. 6. Cópia de tela da interface de consulta do protótipo desenvolvido.

.

suas limitações, principalmente no que diz respeito à qualidade dos dados, provenientes de diversas fontes. Entretanto, ele serviu para validar nossa abordagem e demonstrar seu potencial de integração.

Esta pesquisa deu origem a novos desafios a serem investigados, como a descoberta automática de possibilidades de articulação dos dados de planilhas distintas – ainda que elas sejam de naturezas diferentes – e sua respectiva integração. Tal integração possibilitará inferências que emergirão da combinação desses dados e que não seriam obtidas a partir de uma análise dos documentos individuais.

6. AGRADECIMENTOS

Este trabalho foi parcialmente financiado por CAPES, CNPq, FAPESP, CAPESCOFECUB (projeto AMIB) e INCT em Web Science (CNPq 557.128/2009-9).

REFERENCES

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.

Han, L., Finin, T., Parr, C., Sachs, J., and Joshi, A. RDF123: from Spreadsheets to RDF. In *Seventh International Semantic Web Conference*. Springer, 2008.

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.

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.

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.

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.