Reconhecendo Padrões em Planilhas no domínio de uso da Biologia

Ivelize Rocha Bernardo, André Santanchè, Maria Cecília Calani Baranauskas
Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)
13.083-970 – Campinas, SP – Brazil
ivelize@lis.ic.unicamp.br, {santanche,cecilia}@ic.unicamp.br

Abstract. *Most of research data handled by biologists are in electronic spreadsheets, which are easy to implement as isolated entities, but are inappropriate for integration with other data sources or for enhanced queries. Several initiatives aim to interpret implicit schemas of spreadsheets, making them explicit, in order to drive their mapping process to open standards of interoperability. However, such process is detached of the spreadsheet creation context. In this paper we present a strategy for characterizing spreadsheets, centered in their creation context, and we investigate how this characterization can be used to improve an automated interpretation and mapping process of their respective schemas in the Biology usage domain.*

Resumo. *Grande parte dos dados de pesquisa tratados por biólogos estão em planilhas eletrônicas, que são fáceis de implementar como entidades isoladas, mas são inapropriadas para integração com outras fontes de dados ou para consultas avançadas. Várias iniciativas buscam a interpretação e explicitação de esquemas implícitos em planilhas para subsidiar seu mapeamento para padrões abertos de interoperabilidade. Entretanto, tal processo é dissociado do contexto de criação da planilha. Neste artigo apresentamos uma estratégia para a caracterização de planilhas, centrada no contexto de sua criação, e investigamos como ela pode ser usada para aprimorar a interpretação e explicitação automatizadas de esquemas no domínio de uso da Biologia.*

1. Introdução

Grande parte da informação digital disponível no mundo está representada em planilhas eletrônicas [Syed *et al.* 2010]. Apesar de sua flexibilidade em termos de representação da informação, as planilhas foram originalmente concebidas para utilização individual, sendo armazenadas em arquivos independentes, que não são facilmente interligados com dados de outras planilhas. Por esta razão, há uma crescente preocupação em encontrar formas de tornar seus dados mais flexíveis e compartilháveis [Zhao *et al.* 2010], de forma que outros aplicativos possam também interpretá-los.

Ao contrário das planilhas, abordagens mais sistematizadas para armazenamento de informações, por exemplo, envolvendo a criação de um banco de dados, predefinem os esquemas em dicionários de dados, a serem seguidos em seu registro. Tais esquemas podem ser considerados metadados, que conferem semântica aos dados armazenados. Planilhas eletrônicas, por outro lado, não possuem um esquema explícito. Os dados e metadados – que operam como um esquema implícito interpretável por pessoas – se misturam em um mesmo espaço tabular.

A planilha apresentada na Figura 1, por exemplo, tem o objetivo de catalogar espécies de um museu de biologia. Suas colunas que identificam espécie, filo e classe,

permitem que pessoas – principalmente especialistas no domínio – infiram o propósito da planilha e sua organização. Entretanto, tal esquema não está explícito para um programa de computador.

	A	B	C	D	E	F	G	H
1	**Registro-Catalogo**	**Especie**	**Filo**	**Classe**	**Pais**	**Estado**	**Municipio**	**Localidade**
2	xrb1358	Hirudo medicinalis	Annelida	Polychaeta	EUA	Carolina do Sul	Charleston	Folly Beach
3	akn9846	Achatina fulica	Mollusca	Bivalvia	Brasil	Pernambuco	Fernando de Noronha	Baía doa Golfinhos
4	lat5629	Crinoidea	Echinodermata	Ophiuroidea	Austrália	Austrália Ocidental	Perth	Silver Sands
5								

Figura 1. Planilha catálogo de espécies preenchida. [Instituto de Biologia da Unicamp]

A integração de dados da planilha da Figura 1 com uma segunda planilha que contenha as mesmas informações organizadas de uma forma diferente – *e.g.*, com campo “Filo” em um local diferente em cada planilha – usualmente é feita manualmente. Há diversos aspectos da planilha que dificultam o reconhecimento do seu esquema implícito para integração como, por exemplo, diferenças no local onde se encontra o esquema e como ele é disposto, na ordem das colunas, no rótulo usado para a identificação de campos e sua respectiva semântica, ou na estratégia para atribuir valores aos campos.

Uma abordagem para a integração destas planilhas consiste em reconhecer seus esquemas, distinguindo-os do restante dos dados, de modo a mapeá-los para padrões abertos de interoperabilidade – processo que passaremos a chamar explicitação do esquema. Tal explicitação permite que outros programas sejam capazes de interpretar os dados e executar automaticamente tarefas, como a da integração de dados.

Neste sentido, a interoperabilidade é um foco central das pesquisas envolvendo planilhas eletrônicas. As diversas abordagens encontradas variam desde processos de mapeamento manual para padrões abertos da Web Semântica [Han *et al.* 2008] [Langegger *and* Wob 2009] [O’Connor *et al.* 2010], até propostas para reconhecimento automático de estruturas [Syed *et al.* 2010], pela associação dos elementos da planilha a conceitos disponíveis em bases de conhecimento da Web – *e.g.*, DBpedia (http://dbpedia.org). Verificamos que em todos os casos tal explicitação é desvinculada de um reconhecimento prévio do contexto em que a planilha está inserida.

Há diversas maneiras de caracterizar um contexto e, neste trabalho, trataremos de contextos associados a domínios de uso da informação. Mais especificamente, ao considerarmos o contexto da Biologia como foco específico desta pesquisa, nos referimos àquelas planilhas cujo conteúdo está no domínio de uso de biólogos. Um processo de explicitação de esquemas guiado por um contexto, previamente caracterizado, permite o reconhecimento mais especializado de padrões de construção de planilhas, subsidiando a geração de resultados mais consistentes e com mais riqueza semântica. Retomando o exemplo anterior, a identificação do contexto possibilita reconhecer o padrão de construção da planilha – catalogação de espécies – que é usual no domínio da Biologia. Adicionalmente, a palavra “classe”, por exemplo, desassociada de qualquer domínio possui diversos significados, porém se associarmos esta palavra ao domínio de uso da Biologia é possível definir com mais precisão a sua semântica.

Este artigo apresenta os resultados alcançados envolvendo uma estratégia para o reconhecimento de padrões utilizados por biólogos na construção de planilhas eletrônicas. Os resultados obtidos irão validar a hipótese de que é possível categorizar tais padrões, bem como utilizá-los para aperfeiçoar a explicitação automatizada dos esquemas destas planilhas, produzindo resultados semanticamente mais ricos. Apesar de

termos um trabalho em andamento que implementa o processo proposto em uma ferramenta de reconhecimento automatizado de esquemas, o foco deste artigo está na descrição de tal processo e no detalhamento de sua concepção.

O presente documento está organizado da seguinte forma: a Seção 2 apresenta uma visão geral de trabalhos relacionados; a Seção 3 apresenta a nossa pesquisa envolvendo uma estratégia para o reconhecimento de padrões de construção de planilhas no domínio de uso da Biologia; a Seção 4 apresenta as conclusões deste trabalho e os próximos passos na pesquisa.

2. Trabalhos Relacionados

O primeiro desafio para integração de dados representados em formato tabular é a extração do esquema implícito que guiará a posterior interpretação destes dados. Syed *et al.* (2010) destacam que esta questão remete a um problema mais genérico de extrair esquemas implícitos de fontes de dados – sejam elas, bancos de dados, planilhas etc.

Uma abordagem para tornar a semântica das planilhas interoperável, promovendo a integração dos dados, consiste na associação manual de elementos destas planilhas a conceitos em bases que adotam padrões abertos da Web semântica. A Web semântica é uma iniciativa do W3C (http://www.w3.org) cujo objetivo é tornar a semântica dos dados da Web interpretável por máquinas, de tal modo que estas desempenhar tarefas que vão além da simples recuperação e apresentação de informações, tais como integração e reuso de dados. Especificamente, é usado o RDF (*Resource Description Framework*) [Manolla *and* Miller 2004], um modelo e linguagem para descrição de recursos, cuja função, dentro do conjunto de padrões da Web, é estabelecer a interoperabilidade semântica dos dados. Neste contexto, as ontologias cumprem um papel importante. Elas subsidiam a formalização e reutilização de conceitualizações compartilhadas por uma comunidade. O vocabulário OWL [W3C 2009], associado ao RDF, é usado para a construção de ontologias.

[Han *et al.* 2008] utilizam uma abordagem de mapeamento "*entity-per-row*" apta apenas para tabelas de estruturas simples. Nesta abordagem, cada linha da tabela deve descrever uma entidade diferente e cada coluna um atributo para essa entidade. A planilha da Figura 1, por exemplo, segue este tipo de construção. Cada coluna corresponde a um atributo – e.g., Espécie, Filo, Classe – e cada linha a um objeto depositado no museu (entidade). [Han *et al.* 2008] prevê o mapeamento manual dos atributos para torná-los interoperáveis semanticamente. Inicialmente o usuário deve eleger a célula que rotula a coluna contendo a identificação principal da entidade – o equivalente à chave primária do banco de dados –, que no exemplo da Figura 1 seria o campo "Registro-Catalogo". Em seguida, o sistema permite a associação manual de cada rótulo em células na mesma linha a um atributo da entidade, considerando que cada um deles encabeça uma coluna contendo os respectivos valores daquele atributo.

O resultado final, ilustrado na Figura 2, é representado na forma de um grafo RDF contendo o esquema. No centro do grafo está representado um nó que identifica o recurso (uma tupla da tabela). O símbolo $1 indica que o valor deste nó será obtido a partir do campo "Registro-Catalogo", que é o primeiro campo do esquema reconhecido. Cada aresta em um grafo RDF representa uma propriedade; ela liga o recurso ao valor da respectiva propriedade. Portanto, cada propriedade representa um atributo mapeado da planilha e seu valor será obtido da posição indicada, e.g., $2 indica que será obtido

da segunda posição e assim por diante.

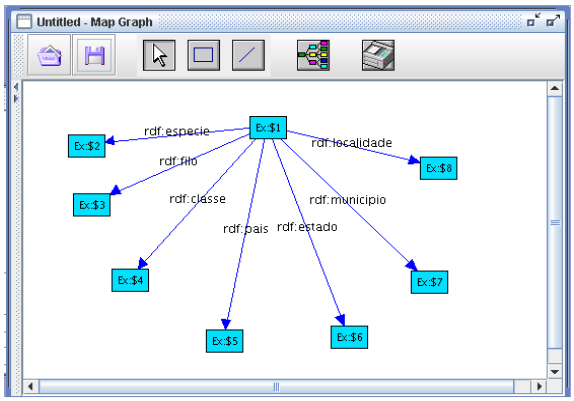

Figura 2. Grafo RDF planilha catálago de espécies. [RDF123 Application]

[Langegger *and* Wob 2009] possuem uma solução similar àquela de [Han *et al.* 2008] para planilhas com mapeamento "*entity-per-row*", que é mais flexível no mapeamento de esquemas. Dentre as possibilidades, está aquela de descrever hierarquias implícitas, por exemplo, uma coluna pode ser subdividida em sub-colunas. No exemplo da Figura 1, os campos País, Estado, Município e Localidade se referem ao local onde o espécie foi coletada. É usual que autores criem um rótulo que se estende por toda a faixa acima destas colunas – *e.g.*, "Local de Coleta" – para indicar que todos estes campos são subdivisões do campo maior. Além de representar o esquema da planilha em RDF, [Langegger *and* Wob 2009] também representam em RDF o mapeamento entre estruturas da planilha e elementos do grafo.

Por ser um padrão aberto que possibilita a interoperabilidade sintática e semântica de dados, o RDF permite a integração de dados de várias planilhas. [Langegger *and* Wob 2009] propõem o acesso a estes dados através do uso da linguagem SPARQL [Pérez *et. al* 2006] – uma linguagem de *query* para acesso a RDF. [O'Connor *et al.* 2010] propõem uma solução semelhante à de [Langegger *and* Wob 2009], mas utilizam OWL.

Uma segunda abordagem para o problema é desvincular os dados da planilha de sua estrutura tabular, pois segundo [Zhao *et al.* 2010] o motivo da baixa interoperabilidade semântica das planilhas é que a relação entre os elementos está associada à sua disposição na estrutura, ao invés de ser estabelecida a partir de sua caracterização semântica. Assim [Zhao *et al.* 2010] propõem transformar os dados das planilhas em objetos de dados semânticos – em que cada registro da planilha se tornará um objeto com atributos e valores – e criar um novo modelo de planilha que possa ser configurável e compatível com esses objetos de dados semânticos. Da mesma forma que os trabalhos anteriores, o processo de mapeamento do esquema é realizado de forma manual. Aplicando esta abordagem no exemplo da Figura 1, o esquema para catalogação de espécies se torna uma classe, cujos atributos são os campos (colunas) da planilha. Cada entrada de registro – linha contendo dados de um espécime coletado – se transforma em uma instância desta classe.

Analisando as soluções propostas anteriormente, verificamos que a base de todas elas é a extração do esquema implícito existente em dados tabulares [Syed *et al.* 2010], exigindo a construção manual do esquema de mapeamento.

Outra maneira de resolver o problema é automatizar o mapeamento semântico

dos dados utilizando *Linked Data*. Syed *et al.* [2010] argumentam que mapear os dados semanticamente de forma manual é inviável, portanto, sua proposta visa automatizar o mapeamento semântico através da ligação dos dados existentes nas planilhas a conceitos disponíveis em bases de conhecimentos, como DBpedia (http://dbpedia.org) e Yago (http://www.mpi-inf.mpg.de/yago-naga/yago/). Yago é uma grande base de dados semântica, cujo conteúdo é extraído, entre outros, da Wikipédia e do WordNet (http://wordnet.princeton.edu) – uma base léxica da língua inglesa que relaciona semanticamente as palavras. Isto possibilita, por exemplo, a realização de buscas a partir dos rótulos da planilha da Figura 1, com o objetivo de associá-los a conceitos do Yago. O termo Filo da planilha pode ser encontrado na base e relacionado com outros conceitos, como Classe e Espécie.

Dentre as vantagens desta última abordagem está o fato de que tais bases são constantemente mantidas e atualizadas por pessoas de várias partes do mundo. Por outro lado, a busca por rótulos destituídos de seus contextos pode gerar ligações ambíguas – *e.g.*, o rótulo "Classe" da Figura 1 pode ter diferentes interpretações, a depender do contexto em que é aplicado. Os dados destas bases também podem apresentar inconsistências, isto é, as pessoas que as alimentam podem ter opiniões divergentes entre si e/ou fornecer conceitos equivocados.

Dentre as soluções para o problema apresentadas, notamos que todas elegem os dados das planilhas individuais – destituídas de contexto – como estratégia central no reconhecimento do esquema da planilha e realização do mapeamento semântico. Neste trabalho partimos do pressuposto de que tal reconhecimento e mapeamento podem ser mais efetivos se considerarem o contexto em que a planilha está inserida.

Por esta razão projetamos um processo de reconhecimento e explicitação de esquemas dirigido pelo contexto, que será detalhado na próxima seção. Nosso processo também pode subsidiar programas que fazem o reconhecimento automático do esquema e associação entre campos/registros das planilhas a conceitos disponíveis em ontologias. Tal reconhecimento partirá de um conjunto de campos caracterizados de forma mais precisa dentro de seu domínio de uso. Observamos que nenhuma das abordagens analisadas é capaz de categorizar as planilhas conforme a natureza da informação que representam. Tal categorização é essencial para tarefas como:

- Definir a semântica e aplicabilidade dos dados extraídos. Por exemplo, os dados de uma planilha contendo eventos podem ser ordenados e apresentados em uma linha de tempo.
- Estabelecer o modo como dados de diferentes planilhas podem ser combinados conforme o seu tipo. Por exemplo, dados de espécies em um museu (objetos) podem ser associados a registros de suas coletas (eventos) de uma maneira específica.

3. Identificando Padrões

O diferencial do processo de explicitação de esquemas que propomos consiste em caracterizar a natureza da planilha, bem como o contexto no qual ela se insere e utilizá-los para guiar a sua interpretação. O diagrama da Figura 3 sintetiza o ciclo de execução do nosso processo para explicitação de esquemas. Os retângulos indicam tarefas e as setas indicam fluxos de dados entre tarefas.

Seguindo o fluxo dos rótulos numerados da figura, o processo inicia a partir do

reconhecimento do esquema da planilha e dos campos que o compõem ①. Na medida em que os campos são reconhecidos, eles são classificados em categorias abstratas ②, em que cada campo responde uma das seis questões exploratórias: quem, o quê, onde, quando, por quê e como (em inglês: *who, what, where, when, why, how*). Em paralelo, cada campo reconhecido subsidia o reconhecimento do domínio de uso da planilha ②. Por exemplo, o campo Espécie é um forte indicador de que a planilha deve pertencer ao domínio de uso da Biologia. Os campos abstratos e a ordem em que eles aparecem são usados caracterizar a natureza da planilha ③. Por exemplo, planilhas que registram eventos tendem a colocar a informação de tempo (*quando*) nas primeiras colunas. Como a natureza da planilha sempre se insere dentro de um domínio de uso, tal informação também subsidiará a caracterização da natureza ③. Por exemplo, uma planilha de registro de eventos típica no domínio da Biologia é o registro de coletas. Uma vez reconhecida a natureza da planilha, é possível prever e reconhecer um padrão de construção da mesma ④.

É importante ressaltar duas características deste processo: (i) ele funciona de forma incremental, ou seja, na medida em que cada tarefa obtém resultados eles são transferidos para a tarefa seguinte, que não espera a conclusão da tarefa anterior; (ii) ele é cíclico, pois dados obtidos em etapas posteriores retroalimentam e refinam a ação de etapas anteriores (setas tracejadas). O reconhecimento do padrão de construção da planilha, assim como a caracterização do seu domínio de uso, tornam mais efetivo o reconhecimento do esquema e dos campos; a caracterização da natureza de uma planilha reforça a caracterização de seu domínio.

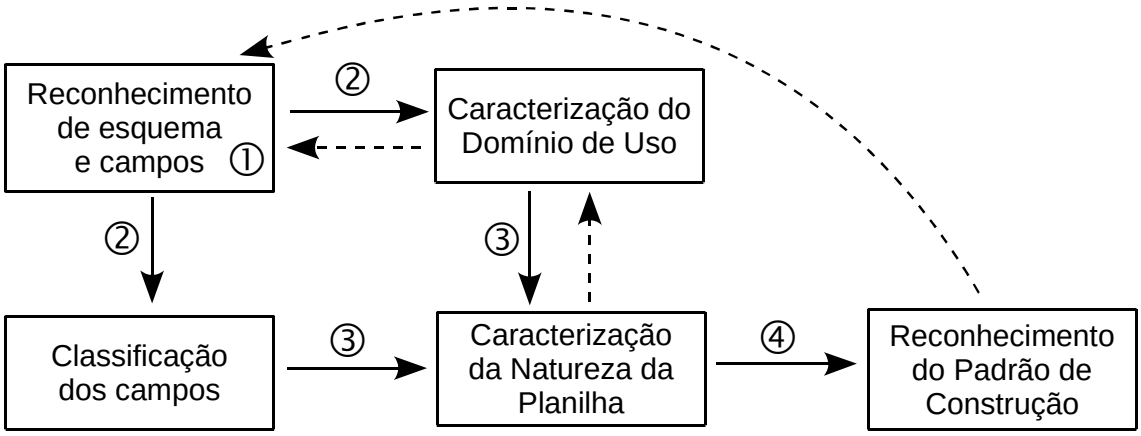

Figura 3. Processo de explicitação de esquemas.

A concepção deste processo partiu de uma análise sistemática de um conjunto de planilhas no domínio de uso da Biologia, que é apresentada neste artigo. Tal análise permitiu formular um conjunto de hipóteses, sobre as quais está fundamentada a nossa proposta de explicitação mais efetiva de esquemas a partir do reconhecimento do contexto e padrão de construção.

A análise sistemática envolveu um estudo de campo cuja metodologia compreendeu três atividades:

- Coleta e análise de planilhas preexistentes no domínio de uso da Biologia;
- Classificação das planilhas por natureza e caracterização de padrões de construção;

- Descrição de um processo para reconhecimento e caracterização de planilhas, conforme seu padrão de construção, passível de ser automatizado.

3.1. Escopos de Coleta e Análise das Planilhas

Selecionamos amostras de planilhas em dois escopos diferentes para análise: planilhas utilizadas pelo Instituto de Biologia (IB) da UNICAMP e planilhas compartilhadas publicamente na Web, alcançadas a partir de uma ferramenta de busca.

Observações feitas em projetos conjuntos com o IB motivaram o desenvolvimento deste estudo de campo. Durante o desenvolvimento de um sistema para catalogar coletas e espécies depositadas no Museu de Zoologia da UNICAMP – como parte do projeto BioCORE (http://www.lis.ic.unicamp.br/projects/biocore) – verificamos que os biólogos armazenam a maior parte de seus dados em planilhas. As planilhas dão aos biólogos autonomia para conceber e implementar modelos de registro e manipulação de dados, que lhes satisfazem até o ponto em que não precisam interligar seus dados com dados de outras planilhas, ou realizar operações mais complexas, típicas de bancos de dados.

No IB, foi coletada uma amostra em que foram identificados seis tipos de planilha. Sua análise desenvolveu-se, neste primeiro momento, sem interações com os biólogos. Esta estratégia nos permitiu buscar o reconhecimento das categorias e padrões partindo exclusivamente da observação formal da planilha.

A pesquisa realizada na Web utilizou a ferramenta Google para encontrar as planilhas que compuseram a amostra para análise. A estratégia de busca envolveu palavras-chave no domínio da Biologia em português e inglês, *e.g.*, biodiversidade, catálogo de espécies, chave de identificação etc. Dentre os muitos resultados foram filtrados apenas aqueles pertinentes ao escopo desta pesquisa. Foram analisadas de forma manual 42 planilhas pertencentes aos seguintes países: Tailândia, Buenos Aires, Canadá, Espanha, Brasil, Inglaterra, México e planilhas de cadastro de órgãos internacionais.

A divisão de escopos nos propiciou perspectivas diferentes sobre a construção de planilhas. No primeiro escopo é possível observar a cultura específica de um determinado grupo de usuários. As observações feitas neste estágio poderão ser confrontadas e refinadas, a partir de uma interação com os próprios autores, prevista em estágios subsequentes da pesquisa. O segundo escopo abarcou a diversidade de estratégias utilizadas em um contexto global de usuários.

Por envolver uma amostra menor, analisada no início da pesquisa, as observações do primeiro escopo foram qualitativas e guiaram a condução das etapas subsequentes. Por envolver um número maior de planilhas, as análises do segundo escopo foram tanto qualitativas quanto quantitativas. Os dados estatísticos apresentados a seguir se referem a tabulações realizadas no segundo escopo.

3.2. Planilhas de Objetos e Eventos

A análise das planilhas do primeiro escopo nos permitiu distinguir inicialmente duas categorias genéricas de planilhas, cuja observação foi verificada no segundo escopo. Percebemos que as planilhas se subdividiam em dois grupos distintos:

Grupo 1 – objetos: planilhas voltadas ao registro de informações sobre objetos.

Por exemplo, o recorte (linhas e colunas omitidas) de planilha ilustrado na Figura 4 registra espécies disponíveis em um museu; cada linha corresponde a o registro de uma espécie (objeto).

Grupo 2 – eventos: planilhas direcionadas a registros de eventos de coletas. Por exemplo, o recorte de planilha ilustrado na Figura 5 registra coletas de amostras feitas em campo; cada linha se refere a uma coleta (evento) realizada em uma data e local específicos.

	A	B	C	E	H	I
2	**Phylum**	**Class**	**Order**	**Family**	**Genus**	**Species**
3	Arthropoda	Insecta	Ephemeroptera	Baetidae	Acentrella	Acentrella insignificans
4	Arthropoda	Insecta	Ephemeroptera	Baetidae	Baetis	Baetis tricaudatus
5	Arthropoda	Insecta	Odonata	Coenagrionidae	Argia	Argia emma

Figura 4. Exemplo de planilha de objetos.
[http://id.water.usgs.gov/projects/spokane/macro/7mile99RQ.xls]

	A	C	D	E	F	G	H	J
3	**Data**	**N. da Estação**	**Latitude**		**Longitude**		**Prof.**	**Classificação**
4			**Graus**	**Minutos**	**Graus**	**Minutos**	**(metros)**	**Larsonneur *et al.*(1982)**
5	14/12/1997	6644	25	45.80´	45	11.77´	485	Litoclástico
6	14/12/1997	6645	25	44.09´	45	13.93´	256	Litobioclástico
7	14/12/1997	6646	25	43.78´	45	16.06´	198	Biolitoclástico

Figura 5. Exemplo de planilha de eventos. [Instituto de Biologia da Unicamp]

Confrontando a estrutura e as informações das duas categorias de planilhas, observamos que:

- As colunas representam os campos e as linhas são os registros.
- No Grupo 1 todas as planilhas possuem muitos campos que respondem às perguntas *o quê* e *quem*. Em 80% delas estes campos estão localizados nas colunas iniciais. Estes dois tipos de campo podem ser considerados chave de identificação e tendem a ser únicos. Eventualmente este grupo possui campos respondendo às perguntas *quando* e *onde* o objeto foi encontrado. Ao contrário do Grupo 2, dados referentes a *onde* têm a tendência de ser menos precisos e mais orientados à leitura humana, *e.g.*, nome da cidade ao invés da sua localização geográfica.
- O Grupo 2 todas as planilhas possuem muitos campos que respondem às perguntas *quando* e *onde*. Em 50% delas estes campos estão localizados nas colunas iniciais. Estes dois tipos de campo podem ser considerados chave de identificação e tendem a ser únicos. Dados relacionados a *onde* tendem a ser bastante precisos, *e.g.*, coordenadas geográficas.

O diagrama da Figura 6 sintetiza estas observações.

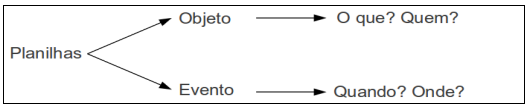

Figura 6. Síntese da primeira categorização.

3.3. Classificação de Planilhas e a Ontologia SUMO

A análise de um conjunto maior e mais diversificado de planilhas do segundo escopo possibilitou um refinamento da categorização inicial, a partir da observação de dois

novos grupos:

Grupo 3 - classificação: planilhas que sistematizam classificações taxonômicas. A Figura 7 apresenta à esquerda o recorte (linhas e colunas omitidas) de uma planilha de classificação de plantas e à direita uma planilha de classificação de animais. A planilha à direita chama a atenção para o fato de que existe uma minoria de planilhas que utilizam linhas para campos e colunas para registros. Um programa de reconhecimento deve estar preparado para esta possibilidade.

Grupo 4 - modelos: meta-planilhas cujos registros descrevem um esquema para a construção de outras planilhas. A Figura 8 apresenta um recorte de uma meta-planilha que descreve os campos do padrão para registro de dados de biodiversidade Darwin Core (http://rs.tdwg.org/dwc/).

	A	B	C	D	E
7		**Nome Científico**	**Nome(s) Comum(ns)**	**Nome(s) em Inglês**	**Classe (Classificação de Nice)**
8		**Scientific Name**	**Costumary Term(s) in Portuguese**	**Term(s) in English**	**Class (Nice Classification)**
9		**Nombre Científico**	**Término(s) Habitual(es) en Portugués**	**Término(s) en Inglés**	**Clase (Clasificación de Niza)**
10		**Nom Scientifique**	**Terme(s) Usuel(s) en Portugais**	**Terme(s) en Anglais**	**Classe (Classification de Nice)**
11	1	*Abelmoschus caillei (A. Chev.) Stevels*	quiabo; quiabeiro	West african okra	29; 31
12	2	*Abelmoschus esculentus (L.)*	quiabo; quiabeiro	Gumbo; Okra; Lady's fingers	05; 29; 31

	A	B	C	D
24	***Birds***			
25	**Kingdom :**		Animalia	
26	**Phylum :**		Chordata	
27	**Class :**		Archosauria	
28	Superorder :		Dinosauria	
29	**Order :**		Saurischia	
30	Suborder :		Theropoda	
31	(unranked)		Tetanurae	
32	(unranked)		Coelurosauria	
33	(unranked)		Maniraptora	
34	**Class :**		Aves	

Figura 7. Exemplos de planilhas de classificação.
[www.mdic.gov.br/arquivos/dwnl_1244660401.xls]
[http://www.reptileland.org/images/class/Book1.xls]

	A	B	C	D	E
11	Curatorial Extension Elements ข้อมูลเชิงพิพิธภัณฑ์	CatalogNumberNumeric	N	Numeric(Double)	เลขรหัสประจำตัวอย่าง
12		IdentifiedBy	N	Text	ชื่อผู้ระบุตัวอย่าง
13		DateIdentified	N	DateTime	วันที่ระบุตัวอย่าง
14		CollectorNumber	N	Text	หมายเลขผู้เก็บตัวอย่าง

Figura 8. Exemplo de meta-planilha.
[http://www.thaibiodiversity.org/Thai20DarwinCore.xls]

Confrontando as informações dos grupos 3 e 4, observamos que:

As taxonomias em Biologia (Grupo 3) têm características equivalentes às planilhas do Grupo 1, mas descrevem objetos abstratos. Alguns diferenciais em relação ao Grupo 1 são: os dados se concentram em descrever *o quê*, não havendo, nas planilhas observadas, dados de *quem*, *quando* e *onde*. Os registros refletem classificações hierárquicas, nas quais os campos vão aumentando a especialização da esquerda para a direita, *e.g.*, da esquerda para a direita: reino, filo, classe, ordem etc.

O Grupo 4 compreende a categoria de meta-planilhas, que possuem como registros o nome de campos, que serão usados para produzir esquemas de outras planilhas. Seu conteúdo responde à pergunta *como*. O conteúdo de uma das colunas iniciais contém o nome dos campos, a serem usados no esquema da planilha descrita.

Como está ilustrado na Figura 9, para classificarmos os grandes grupos das planilhas optamos por alinhar as nossas classes com aquelas definidas na ontologia SUMO [Niles and Pease 2001] – uma ontologia de nível superior (*upper ontology*) mantida pelo IEEE e amplamente adotada. Há duas razões importantes para esta decisão: (i) a associação com uma classificação formal pode ser usada como base no processo de automatização de reconhecimento, bem como pode guiar a geração de resultados em padrões da Web Semântica; (ii) a ontologia SUMO representa um modelo de classificação amplamente discutido e refinado pela comunidade.

As categorias foram reordenadas da seguinte maneira: os conteúdos descritos estão divididos em físicos e abstratos. Na categoria dos físicos estão o Grupo 1 (objetos) e Grupo 2 (eventos), pois descrevem objetos do mundo físico ou eventos que aconteceram no mundo físico. Na categoria dos abstratos estão o Grupo 3 (classificação) e Grupo 4 (modelo).

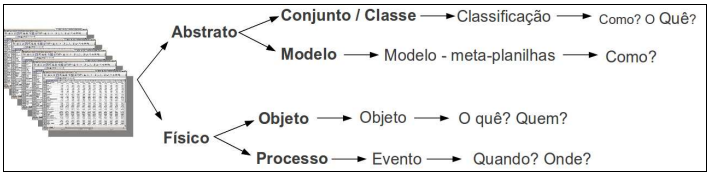

Figura 9. Síntese da categorização alinhada com o SUMO

3.4. Consolidação dos Resultados

Como resultados das observações e análises apresentadas nas subseções anteriores, elaboramos um conjunto de hipóteses preliminares para guiar o reconhecimento automático de padrões. Tais hipóteses estão sendo validadas em estágios subsequentes da pesquisa, a partir da interação com biólogos e da implementação de reconhecedores automáticos. A seguir apresentamos tais hipóteses:

H1: A organização da grande maioria das planilhas segue o padrão de colunas como campos e linhas como registros.

H2: Em sua maioria, os campos podem ser classificados como resposta a uma das seis questões exploratórias: quem, o quê, onde, quando, por quê e como.

H3: Os tipos de campo e a sua ordenação (e.g., campos que aparecem no início) normalmente expressam a categoria da planilha, conforme foi apresentado anteriormente.

Com base nas observações, um sistema de categorização automático deverá ser capaz de:

1. **Diferenciar o esquema dos registros** – Neste sentido, será adotada a hipótese da organização (H1). Visto que observamos que existem algumas planilhas com o esquema em outra direção, ou seja, linhas como campos e colunas como registros, será necessário desenvolver um mecanismo auxiliar para identificá-las. Uma possível direção consiste em verificar a repetição do mesmo tipo de dados, ou seja, se o mesmo tipo de dados se repete de uma linha para outra e varia de uma coluna para a outra, a tendência é que as colunas sejam os campos.
2. **Classificar os campos em quem, o quê, onde, quando, por quê e como** – Tal classificação pode ser alcançada a partir de um vocabulário, associando termos do

domínio de Biologia a cada uma destas perguntas.

3. **Categorizar as planilhas** – Seguindo os padrões de construção observados, o sistema deverá ser capaz de categorizar um conjunto significativo das planilhas conforme a sua natureza.

As conclusões alcançadas ao final desta pesquisa subsidiaram o projeto do processo ilustrado na Figura 3.

4. Conclusão e Trabalhos Futuros

É sabido que planilhas eletrônicas tornaram-se um veículo disseminado para registro e representação de informação no formato digital em muitos domínios do conhecimento, especialmente nas ciências naturais. A facilidade de uso de tais sistemas, ao mesmo tempo em que possibilitam aos usuários, profissionais de domínios diversos do conhecimento, autonomia para representação de seus dados, dificultam a eles a interligação de seus dados com dados de outras planilhas, em operações mais complexas. Este artigo buscou reconhecer e classificar padrões encontrados em planilhas utilizadas no domínio da Biologia, com vistas a subsidiar um processo de reconhecimento automático centrado no contexto de tais planilhas. Apesar do enfoque ter sido em Biologia, o processo apresentado na Figura 3 foi projetado em uma perspectiva genérica. Além disto, o método de caracterização de natureza da planilha, que abstrai o papel dos campos pelo uso das 6 perguntas exploratórias, é apto à generalização.

A partir da análise de 42 planilhas contendo dados no domínio de uso da Biologia, alcançamos os seguintes resultados:

- Um sistema de categorização de planilhas associado a um conjunto de características para subsidiar o reconhecimento automatizado de padrões de construção no domínio de uso da Biologia.
- Um conjunto de hipóteses que estão sendo validadas em etapas.
- O projeto de um processo de explicitação de esquemas, que está sendo implementado para reconhecer automaticamente padrões de construção de planilhas no domínio de uso da Biologia.

O programa de reconhecimento automático de planilha é um trabalho em andamento, mas seus testes preliminares têm apresentado resultados promissores. Tal programa implementa o processo da Figura 3, a fim de explicitar esquemas e mapear dados para padrões abertos da Web Semântica. Para a realização de testes ele inclui um módulo para buscar, recuperar e analisar automaticamente planilhas na Web a partir de palavras-chave. Ao aplicar este programa na amostra de 42 planilhas do segundo escopo, obtivemos um reconhecimento de 78,6% das planilhas. Um segundo teste envolveu a recuperação automática feita pelo programa de 1.914 planilhas, com as mesmas palavras-chave do segundo escopo. Neste caso, não houve seleção manual de planilhas pertinentes. O programa selecionou e reconheceu automaticamente o esquema de 137 das planilhas – 7% da entrada. É importante ressaltar que, neste segundo teste, por não ter havido uma seleção manual das planilhas pertinentes, a recuperação direta por palavras-chave feita pelo programa traz uma grande quantidade de planilhas que não estão relacionadas ao domínio de uso analisado. Apesar dos resultados de execução do programa serem preliminares, ele apresenta indicadores positivos da viabilidade de

automatização do processo proposto.

A partir dos resultados alcançados no domínio da Biologia, um dos trabalhos futuros envolverá investigação sobre sua adequação a outros domínios.

5. Agradecimentos

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

Referências

Han, L., Finin, T. W., Parr, C. S., Sachs, J. and Joshi, A. (2008) "RDF123: From Spreadsheets to RDF", In: International Semantic Web Conference (ISWC), Springer, vol. 5318, p. 451-466.

Langegger, A. and WoB, W. (2009) "XLWrap - Querying and Integrating Arbitrary Spreadsheets with SPARQL", In: International Semantic Web Conference (ISWC), Springer, vol. 5823, p. 359-374.

Manola, F. and Miller, E. (2004) "RDF Primer – W3C Recommendation" w3.org/TR/2004/REC-rdf-primer-20040210.

Niles, I. and Pease, A. (2001) "Towards a standard upper ontology". Formal Ontology in Information Systems.

O'Connor, M. J., Halaschek-Wiener, C. and Musen, M. A. (2010) "Mapping Master: A Flexible Approach for Mapping Spreadsheets to OWL", In: International Semantic Web Conference (ISWC), Springer, vol. 6497, p. 194-208.

Pérez, J., Arenas, M. and Gutierrez, C. (2006) "Semantics and Complexity of SPARQL". In The Semantic Web – ISWC, vol. 4273, p. 30-43.

RDF123 Application, http://rdf123.umbc.edu/, 15 January 2012.

Syed, Z., Finin, T., Mulwad, V. and Joshi, A. (2010) "Exploiting a Web of Semantic Data for Interpreting Tables", In: Proceedings of the Second Web Science Conference.

W3C OWL WG. (2009) "OWL 2 Web Ontology Language – Document Overview" – W3C Recommendation 27 October 2009. http://w3.org/TR/2009/REC-owl2-overview-20091027/.

Zhao, C., Zhao, L. and Wang, H. (2010) "A Spreadsheet System based on Data Semantic Object", In: IEEE International Conference on Information Management and Engineering (ICIME), Chengdu, China, p. 407-411.