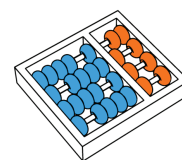


2012

Ivelize Rocha Bernardo

“Explicitação de Esquemas Orientada a Contexto para Promover Interoperabilidade Semântica”

CAMPINAS



Universidade Estadual de Campinas
Instituto de Computação

Ivelize Rocha Bernardo

“Explicitação de Esquemas Orientada a Contexto para Promover Interoperabilidade Semântica”

Orientador(a): **Prof. Dr. André Santanchè**

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para obtenção do título de Mestre em Ciência da Computação.

ESTE EXEMPLAR CORRESPONDE À
VERSÃO FINAL DA DISSERTAÇÃO DEFEN-
DIDA POR IVELIZE ROCHA BERNARDO,
SOB ORIENTAÇÃO DE PROF. DR. ANDRÉ
SANTANCHÈ.

Assinatura do Orientador(a)

CAMPINAS

FICHA CATALOGRÁFICA ELABORADA POR
ANA REGINA MACHADO - CRB8/5467
BIBLIOTECA DO INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E
COMPUTAÇÃO CIENTÍFICA - UNICAMP

Bernardo, Ivelize Rocha, 1982-
B456e Explicitação de esquema orientada a contexto para promover
interoperabilidade semântica / Ivelize Rocha Bernardo. – Campinas,
SP : [s.n.], 2012.

Orientador: André Santanchè.
Dissertação (mestrado) – Universidade Estadual de Campinas,
Instituto de Computação.

1. Planilhas eletrônicas. 2. Web semântica. 3. Recuperação da
informação. 4. Biologia - Processamento de dados. I. Santanchè,
André. II. Universidade Estadual de Campinas. Instituto de
Computação. III. Título.

Informações para Biblioteca Digital

Título em inglês: Promoting semantic interoperability by a context oriented
approach to make schemas explicit

Palavras-chave em inglês:

Electronic spreadsheets

Semantic Web

Information retrieval

Biology - Data processing

Área de concentração: Ciência da Computação

Titulação: Mestra em Ciência da Computação

Banca examinadora:

André Santanchè [Orientador]

Maria Cecília Calani Baranauskas

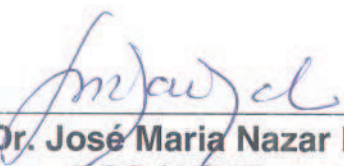
José Maria Nazar David

Data de defesa: 04-09-2012


Programa de Pós-Graduação: Ciência da Computação

TERMO DE APROVAÇÃO

Dissertação Defendida e Aprovada em 04 de Setembro de 2012, pela
Banca examinadora composta pelos Professores Doutores:



Prof. Dr. José Maria Nazar David
DCC / UFJF



Profª. Drª. Maria Cecília Calani Baranauskas
IC / UNICAMP



Prof. Dr. André Santanchè
IC / UNICAMP

Explicitação de Esquemas Orientada a Contexto para Promover Interoperabilidade Semântica

Ivelize Rocha Bernardo¹

04 de Setembro de 2012

Banca Examinadora:

- Prof. Dr. André Santanchè (Supervisor/*Orientador*)
- Prof. Dra. Maria Cecília Calani Baranauskas
Instituto de Computação – UNICAMP
- Prof. Dr. José Maria Nazar David
Departamento de Ciência da Computação – UFJF
- Profa. Dra. Claudia Bauzer Medeiros
Instituto de Computação – UNICAMP (Suplente)
- Prof. Dr. Luciano Antonio Digiampietri
Escola de Artes, Ciências e Humanidades – EACH – USP (Suplente)

¹Suporte financeiro de: Bolsa do CAPES (processo 01-P-04388-2010) 01/08/2010–31/07/2011, Bolsa da Fapesp (processo 2011/05088-8) 01/08/2011–31/07/2012

Abstract

The flexibility provided by spreadsheets allows their customization following mental models of their authors and makes them popular data management systems. Gradually there is a growing need of integrating and join data from different spreadsheets and, to enable machines assistance in this process, the challenge is how to automatically interpret their implicit schema, which is addressed to human interpretation. 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 apart from their domain. This work differs by assuming the essential role of the context and 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. Our proposal involves a strategy to characterize building patterns related to conceptual models of authors in spreadsheets building process, which results from an extensive research of practices shared among authors of spreadsheets in the Biology usage domain. In this document 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

A flexibilidade proporcionada por planilhas eletrônicas possibilita sua customização seguindo modelos mentais de seus autores e as tornam sistemas populares de gerenciamento de dados. Gradativamente tem crescido a necessidade de se integrar e articular dados de diferentes planilhas e, para que máquinas possam auxiliar neste processo, o desafio é como interpretar automaticamente o seu esquema implícito, que é dirigido à interpretação humana. 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 contexto e o domínio em que foi concebida a planilha essenciais para se traçar o conjunto de práticas compartilhadas pela comunidade em questão, que estabelece padrões de construção a serem reconhecidos automaticamente por nosso sistema, em um processo de extração de dados e explicitação de esquemas. Nossa proposta envolve uma estratégia para caracterização de padrões de construção associados a modelos conceituais de autores na construção de planilhas, que é resultado de uma ampla pesquisa de práticas compartilhadas por autores de planilhas no domínio de uso da Biologia. Neste documento 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 de nossa abordagem de reconhecer a natureza da planilha analisada dentro de seu contexto de criação.

Acknowledgements

Agradeço ao meu orientador, Doutor André Santanchè, pela orientação, dedicação e incentivo ao longo de todo o projeto.

Ao meu pai Santo Edson, por me inspirar com sua força de vontade e olhar otimista diante dos obstáculos e à minha mãe Walkíria, pela sua eterna dedicação, amor e paciência. À minha irmã Anelize, por ser a minha melhor amiga e sempre estar ao meu lado nessa caminhada e à minha avó Neyde, por me ensinar o amor pelo conhecimento. Aos meus amigos, sem os quais a vida não seria tão colorida.

Agradeço a todos os professores, funcionários e colegas da UNICAMP. Aos membros da banca, as agências de fomento CAPES (processo 01–P–04388-2010) e FAPESP (processo 2011/05088–8), e também ao projeto INCT in Web Science (CNPq 557.128/2009–9) pelo apoio.

Por fim, aos familiares e todos aqueles que de alguma maneira contribuíram para o meu crescimento, ajudando-me a chegar até aqui.

Sumário

Abstract	vii
Resumo	viii
Acknowledgements	ix
1 Introdução	1
1.1 Motivação	1
1.2 Objetivos e Contribuições	2
1.3 Estrutura da Dissertação	3
1.3.1 Capítulo 2	3
1.3.2 Capítulo 3	3
1.3.3 Capítulo 4	4
1.3.4 Capítulo 5	4
2 Reconhecendo Padrões em Planilhas no domínio de uso da Biologia	5
2.1 Introdução	5
2.2 Trabalhos Relacionados	7
2.3 Identificando Padrões	10
2.3.1 Escopos de Coleta e Análise das Planilhas	11
2.3.2 Planilhas de Objetos e Eventos	12
2.3.3 Classificação de Planilhas e a Ontologia SUMO	14
2.3.4 Consolidação dos Resultados	15
2.4 Conclusão e Trabalhos Futuros	16
3 Interpretação Automática de Planilhas Baseada no Reconhecimento de Padrões de Construção	18
3.1 Introdução	18
3.2 Cenário da Pesquisa	20
3.2.1 Tipos de Planilhas	22

3.3	Padrões de Construção	23
3.4	Metodologia	27
3.4.1	Coleta e análise inicial de dados	28
3.4.2	Hipóteses	28
3.4.3	Modelo de Reconhecimento	29
3.4.4	Sistema	29
3.5	Evidências de Padrões de Construção	31
3.6	Trabalhos Relacionados	35
3.7	Conclusão e Trabalhos Futuros	38
4	Extraíndo e Integrando Semanticamente Dados de Múltiplas Planilhas Eletrônicas a Partir do Reconhecimento de Sua Natureza	40
4.1	Introdução	40
4.2	Revisão da Literatura	41
4.3	Explicitação de Esquema dirigida pela Natureza da Planilha	45
4.4	Mapeamento semântico a partir do contexto explicitado	46
4.5	Conclusão e Trabalhos Futuros	49
5	Conclusões e Extensões	50
	Bibliografia	52

Lista de Figuras

2.1	Planilha catálogo de espécies preenchida [Instituto de Biologia da Unicamp].	6
2.2	Grafo RDF planilha catálogo de espécies [RDF123 Application].	8
2.3	Processo de explicitação de esquemas.	11
2.4	Exemplo de planilha de objetos [id.water.usgs.gov].	13
2.5	Exemplo de planilha de eventos [Instituto de Biologia da Unicamp]. . . .	13
2.6	Síntese da primeira categorização.	14
2.7	Exemplos de planilhas de classificação [mdic.gov.br] [reptileland.org]. . . .	14
2.8	Exemplo de meta-planilha [thaibiodiversity.org].	15
2.9	Síntese da categorização alinhada com o SUMO.	16
3.1	Planilha catálogo de espécies [Instituto de Biologia – Unicamp].	21
3.2	Síntese da categorização alinhada com o SUMO.	23
3.3	Modelo Conceitual Subjacente na Planilha do Tipo Catálogo.	24
3.4	Padrão de construção de imagem da planilha em uma planilha tipo catálogo. 25	
3.5	Modelo conceitual de Planilha de Objetos (Espécimes) enriquecido com qualificadores.	27
3.6	Modelo conceitual de Planilha de Eventos (Coletas) enriquecido com qua- lificadores.	28
3.7	Arquitetura do Sistema.	30
3.8	Localização do esquema.	32
3.9	Termos por esquema nas linhas iniciais.	33
3.10	Distribuição do <i>what</i> - planilhas de objetos.	33
3.11	Relação entre os campos - planilhas objeto.	34
3.12	Distribuição do <i>what</i> - planilhas de eventos.	34
3.13	Distribuição do <i>where</i> - planilhas de eventos.	35
3.14	Relação entre os campos sem peso posicional - planilhas tipo processo . .	35
3.15	Relação entre os campos com peso posicional - planilhas tipo processo . .	36

4.1	Exemplo de planilha de registro de coleta [siscom.ibama.gov.br].	42
4.2	Exemplo mapeamento semântico realizado por [Han et al. 2008].	43
4.3	Exemplo de planilhas.	44
4.4	Mapeamento semântico das planilhas Fig.2 e Fig.3.(b).	46
4.5	Etapas de execução do sistema de reconhecimento e mapeamento de planilhas.	47
4.6	Cópia de tela da interface de consulta do protótipo desenvolvido.	48

Capítulo 1

Introdução

1.1 Motivação

O ponto de partida deste trabalho foi o desenvolvimento de um processo capaz de reconhecer esquemas implícitos em planilhas eletrônicas, de modo que eles fossem explicitados e seus respectivos dados pudessem ser convertidos em padrões abertos da web semântica. Uma vez convertidos, se torna possível integrar e combinar dados de diferentes planilhas.

Ele foi motivado por dois fatores:

(i) Por um lado, um projeto em conjunto com o Instituto de Biologia da Unicamp nos chamou a atenção para o que denominamos “bases de dados populares”. Os biólogos mantêm em planilhas eletrônicas uma parcela significativa de seus dados. Eles são cativados pela facilidade de acesso destas planilhas eletrônicas e pela autonomia proporcionada pelos sistemas de edição das mesmas. Entretanto, a proliferação de planilhas, projetadas originalmente para uso na forma de arquivos isolados, causa problemas de integração e articulação de dados. Em muitos casos, observamos que os biólogos consomem uma parcela significativa de seu tempo “copiando e colando”, transferindo e adaptando dados entre planilhas.

(ii) Em paralelo, esta pesquisa tomou como referência uma pesquisa anterior de uma metodologia denominada Semântica In Loco [16, 17], na qual dados são extraídos automaticamente de documentos, a partir do reconhecimento de padrões de anotação feitos pelos usuários durante o processo de criação dos mesmos.

O principal desafio desta pesquisa residiu no fato de que os esquemas das planilhas são implícitos e baseados em práticas compartilhadas por domínios de uso, o que dificulta a identificação automatizada por computadores. Neste sentido, esta investigação buscou um diferencial em relação às abordagens existentes, pautadas no princípio de que é possível se desenvolver uma estratégia de reconhecimento automático de planilhas independente de domínio ou contexto.

Trabalhos relacionados se subdividem, principalmente, em dois grupos: os que adotam como estratégia de reconhecimento um processo manual realizado pelo usuário autor da planilha e os que adotam um reconhecimento automático.

Alguns trabalhos consideram o reconhecimento manual inviável e propõem o reconhecimento automático, entretanto, há necessidade de se adotar uma técnica de desambiguação dos termos, pois um reconhecedor semântico automático pode encontrar para um mesmo termo sintático diferentes significados, e.g. “classe”. Assim esses trabalhos seguem com a identificação de colunas individualmente, mas não se preocupam em identificar como estas colunas se combinam para formar um tipo de planilha e a qual comunidade ela pertence.

Nossa abordagem segue em uma direção distinta, adotando o contexto em que a planilha está inserida como referencial para a interpretação automática da planilha, dessa forma conseguimos caracterizar a natureza da planilha e consequentemente associá-la a uma semântica mais rica. Neste sentido a pesquisa se desenvolveu em etapas progressivas a saber:

1. Uma análise comparativa de amostras de planilhas nos levou à elaboração de hipóteses sobre padrões de construção destas planilhas – associados a um domínio – bem como a modelagem de um processo automático de reconhecimento [4].
2. Foi implementado um protótipo de um sistema que nos permitiu avaliar algumas das hipóteses e analisar um conjunto mais amplo de planilhas.
3. O resultado nesta análise nos levou à formulação de uma estratégia para a representação de padrões seguidos por autores ao criar planilhas eletrônicas. Tal representação subsidia a ação dos programas de reconhecimento de planilhas [3].
4. Construção de um protótipo capaz de reconhecer esquemas, extrair e integrar dados de múltiplas planilhas. Tal sistema se diferencia dos trabalhos relacionados por ser capaz de reconhecer a natureza das planilhas em questão [2].

1.2 Objetivos e Contribuições

O objetivo dessa pesquisa é desenvolver uma estratégia de integração semântica entre dados presentes em planilhas eletrônicas, através da explicitação de esquemas baseada na natureza das planilhas e orientada pelo contexto da comunidade.

A seguir são apresentadas as principais contribuições deste trabalho:

Elaboração de um processo de explicitação de esquemas baseado no reconhecimento da natureza da planilha a partir de seu padrão de construção: Planilhas apresentam um padrão de construção definido por práticas compartilhadas da

comunidade, seu propósito e tipo de informação que armazenam. Foi desenvolvido um processo capaz de reconhecer tais padrões de construção, através de uma análise incremental e iterativa dos termos das planilhas.

Estratégia para a representação de padrões de construção associados a modelos conceituais de autores na criação de planilhas: Tal estratégia conta com uma representação na forma de diagrama, passível de ser representada digitalmente e interpretada por computadores. Esses modelos permitirão a representação e intercâmbio de padrões para a construção de planilhas adotados em diversos contextos.

Protótipo de sistema para reconhecimento e explicitação de esquema em planilhas eletrônicas, bem como o mapeamento para RDF/OWL: o sistema engloba todas as fases da pesquisa, desde a caracterização e reconhecimento da natureza da planilha, até a extração dos dados e mapeamento para RDF/OWL.

1.3 Estrutura da Dissertação

1.3.1 Capítulo 2

O Capítulo 2 contém o artigo Reconhecendo Padrões em Planilhas no domínio de uso da Biologia, que foi publicado no VIII Simpósio Brasileiro de Sistemas de Informação - SBSI 2012. Este capítulo engloba a construção do conjunto de hipóteses sobre padrões de construção de planilhas seguidos pela comunidade de biólogos, bem como detalha nosso processo para reconhecimento do esquema implícito de planilhas a partir da caracterização de sua natureza.

1.3.2 Capítulo 3

O Capítulo 3 é formado pelo artigo Interpretação Automática de Planilhas Baseada no Reconhecimento de Padrões de Construção, que se tonará um relatório técnico. Neste capítulo apresentamos nossa proposta para caracterização de padrões de construção seguidos pelos biólogos no desenvolvimento de planilhas eletrônicas de gerenciamento de dados. Os fundamentos do modelo para tal caracterização são evidenciados pela análise de trabalhos relacionados e dados estatísticos resultantes da análise de centenas de planilhas. Esta modelagem permitiu a representação e compartilhamento de padrões seguidos por uma comunidade na construção de planilhas. Deste modo, este artigo expande o processo de reconhecimento de padrões para acomodar tal modelo.

1.3.3 Capítulo 4

O Capítulo 4 contém o artigo *Extraindo e Integrando Semanticamente Dados de Múltiplas Planilhas Eletrônicas a Partir do Reconhecimento de Sua Natureza*, que foi publicado no XXVI Simpósio Brasileiro de Banco de Dados - SBBD 2012. Neste capítulo fecha-se o ciclo e é apresentada uma aplicação prática que extrai e integra dados de múltiplas planilhas, conforme proposto anteriormente. Este artigo enfatiza que o método de reconhecimento proposto é capaz de caracterizar a natureza das planilhas analisadas, ampliando as possibilidades de combinação dos dados.

1.3.4 Capítulo 5

O capítulo 5 contém as conclusões desta dissertação. Nele apresentamos as contribuições da pesquisa, assim como as principais possíveis extensões e trabalhos futuros.

Capítulo 2

Reconhecendo Padrões em Planilhas no domínio de uso da Biologia

2.1 Introdução

Grande parte da informação digital disponível no mundo está representada em planilhas eletrônicas [18]. 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 [22], 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 2.1, 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 integração de dados da planilha da Figura 2.1 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

	A	B	C	D	E	F	G	H
1	Registro-Catálogo	Especie	Filo	Classe	País	Estado	Município	Localidade
2	xrb1358	Hirudo medicinalis	Annelida	Polychaeta	EUA	Carolina do Sul	Charleston	Folly Beach
3	akn9856	Achatina fulica	Mollusca	Bivalvia	Brasil	Pernambuco	Fernando de Noronha	Baía do Golphinhos
4	lat5629	Amphiodia atra	Echinodermata	Ophiuroidea	Austrália	Austrália Ocidental	Perth	Silver Sands

Figura 2.1: Planilha catálogo de espécies preenchida [Instituto de Biologia da Unicamp].

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 [5, 7, 12], até propostas para reconhecimento automático de estruturas [18], 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 plani-

lhas, 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 2.2 apresenta uma visão geral de trabalhos relacionados; a Seção 2.3 apresenta a nossa pesquisa envolvendo uma estratégia para o reconhecimento de padrões de construção de planilhas no domínio de uso da Biologia; a Seção 2.4 apresenta as conclusões deste trabalho e os próximos passos na pesquisa.

2.2 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. [18] 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*) [8], 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 [20], associado ao RDF, é usado para a construção de ontologias.

Han et al. [5] 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 2.1, 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. [5] 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 2.1 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 2.2, 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.

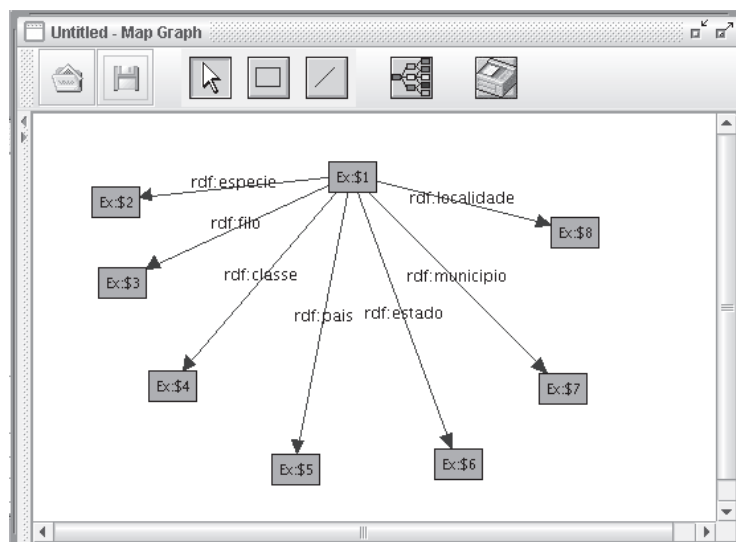


Figura 2.2: Grafo RDF planilha catálogo de espécies [RDF123 Application].

Langegger and Wöb [7] possuem uma solução similar àquela de Han et al. [5] 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 2.1, 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 Wöb [7] 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

Wöb [7] propõem o acesso a estes dados através do uso da linguagem SPARQL [13] – uma linguagem de query para acesso a RDF. Oconnor and Halaschek-Wiener [12] propõem uma solução semelhante à de Langegger and Wöb [7], mas utilizam OWL.

Uma segunda abordagem para o problema é desvincular os dados da planilha de sua estrutura tabular, pois segundo Zhao et al. [22] 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. [22] 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 2.1, 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 [18], 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. [18] 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 2.1, 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 2.1 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 reconhe-

cimento 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.

2.3 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 2.3 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.

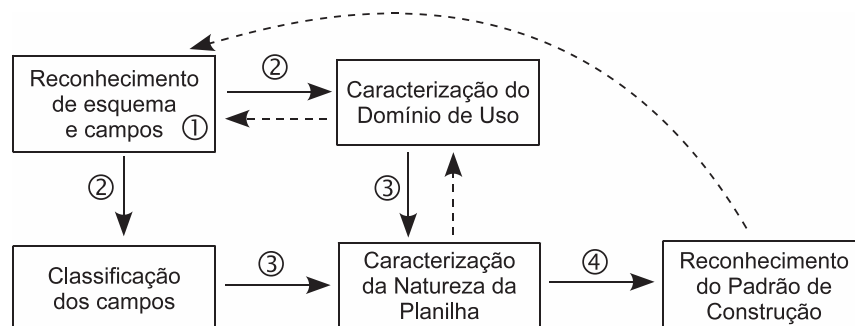


Figura 2.3: Processo de explicitação de esquemas.

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.

2.3.1 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 pu-

blicamente 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.

2.3.2 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 2.4 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 2.5 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 2.4: Exemplo de planilha de objetos [id.water.usgs.gov].

	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 2.5: 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 2.6 sintetiza estas observações.



Figura 2.6: Síntese da primeira categorização.

2.3.3 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 2.7 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 2.8 apresenta um recorte de uma meta-planilha que descreve os campos do padrão para registro de dados de biodiversidade Darwin Core (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	<i>Abelmoschus caillei</i> (A. Chev.) Stevels	quiabo; quiabeiro	West african okra	29; 31
12	2	<i>Abelmoschus esculentus</i> (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 2.7: Exemplos de planilhas de classificação [mdic.gov.br] [reptileland.org].

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

Figura 2.8: Exemplo de meta-planilha [thaibiodiversity.org].

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 2.9, para classificarmos os grandes grupos das planilhas optamos por alinhar as nossas classes com aquelas definidas na ontologia SUMO [10] – 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).

2.3.4 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:



Figura 2.9: Síntese da categorização alinhada com o SUMO.

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 2.3.

2.4 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, es-

pecialmente 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 2.3 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 2.3, 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.

Capítulo 3

Interpretação Automática de Planilhas Baseada no Reconhecimento de Padrões de Construção

3.1 Introdução

Artefatos de software – e.g., programas de computador, bancos de dados e planilhas eletrônicas – são produzidos a partir de modelos conceituais. A depender do método adotado no processo de produção, tais modelos podem ser projetados, registrados e debatidos previamente – como acontece no projeto de sistemas de software e bancos de dados – ou podem estar apenas na mente do autor no momento da criação do artefato, como é usual em planilhas eletrônicas.

Segundo Norman [11], na interação com estes artefatos o usuário contrói em sua mente um modelo mental. Ao contrário do modelo conceitual, um modelo mental usualmente não é preciso no que diz respeito aos aspectos técnicos do que ele representa. Modelos mentais não são estáticos, eles são construídos a partir da experiência do usuário, em geral e na interação com sistemas similares. Por isso, os modelos evoluem no processo de interação do usuário com o sistema e ao longo do tempo [11].

Há, portanto, uma relação entre o modelo conceitual, que subsidiou a criação do artefato de software, e o modelo mental que é construído na mente do usuário ao interagir com o mesmo. Como o modelo mental depende da experiência do usuário, pode-se afirmar que ao longo do tempo, na medida em que interage com artefatos similares, o usuário vai desenvolvendo uma expectativa, que será a base para a formação deste modelo mental [15]. Uma das bases para o reconhecimento de artefatos similares é a adoção de *padrões de*

construção.

Como planilhas eletrônicas – foco desta pesquisa – são artefatos cuja produção é acessível a usuários não especialistas em computação, com grande frequência os papéis de autor e consumidor de planilhas se mesclam. Deste modo, o mesmo autor que consome planilhas, e que desenvolveu uma expectativa sobre elas na formação de seu modelo mental, será o autor de planilhas e tenderá a reproduzir alguns dos padrões de construção. Isto cria uma sinergia entre modelos mentais, expectativas e padrões de construção, que concorrem para a solidificação dos padrões, especialmente dentro de domínios e comunidades.

Planilhas eletrônicas têm esquemas implícitos, ao contrário de abordagens mais rigorosas na estruturação dos dados – tal como um banco de dados – que especificam esquemas de dados, baseados nos modelos conceituais que lhe deram origem. O modelo conceitual percebido por um usuário ao interpretar a planilha está intimamente ligado ao modelo mental construído por ele na interação com a mesma, que por sua vez está associado aos padrões usados na sua construção.

O desafio desta pesquisa é considerar um sistema computacional como consumidor da planilha, ao invés de um usuário. Isto implica em tornar tal sistema capaz de reconhecer os padrões de construção e inferir um modelo conceitual.

O cenário de construção e uso de planilhas eletrônicas foi escolhido para a realização de nosso estudo, dada a liberdade de modelagem oferecida pela planilha, que a torna propícia a ganhar diferentes formas, facilitando a formação de padrões de construção emergentes. Fazendo-se um paralelo, a planilha funciona como uma argila em que o autor esculpe seu modelo conceitual. Se este autor quisesse esculpir uma mesa, ainda que haja uma infinidade de possibilidades, há padrões que se consolidaram na comunidade a que o autor pertence que estabelecem, por exemplo, que uma mesa será apoiada sobre um ou mais pés, em que se apoiará um tampo horizontal. Observamos que parte do reconhecimento dos padrões de construção da mesa consiste no reconhecimento de um conjunto de seus objetos componentes e certa relação espacial entre eles.

No âmbito das planilhas eletrônicas, os usuários finais têm autonomia e liberdade para produzir as suas próprias estruturas de sistematização, caracterizadas por esquemas implícitos dirigidas aos humanos. Esta autonomia e liberdade têm um efeito colateral: programas proporcionam uma assistência muito limitada para executar tarefas, uma vez que eles são incapazes de reconhecer os esquemas implícitos e conseqüentemente sua respectiva semântica, e.g., é difícil combinar e articular os dados a partir de duas planilhas que possuem esquemas distintos.

Há diversas iniciativas voltadas ao desenvolvimento de programas capazes de reconhecer e fazer uso deste esquema implícito em tarefas de integração de dados. Apesar de partirem de premissas derivadas de algumas práticas populares de construção de pla-

nilhas, nenhuma destas iniciativas desenvolveu um modelo para caracterizar padrões de construção de domínios e comunidades, de modo a explorá-los no processo de identificação.

O diferencial da nossa proposta consiste no desenvolvimento de um modelo para caracterizar padrões de construção que guie um processo automático de reconhecimento de esquemas implícitos em planilhas. Este artigo apresenta nossa abordagem para a representação de tal modelo. Ela foi resultado de uma extensa pesquisa envolvendo trabalhos relacionados, bem como coleta e análise de mais de 11.000 planilhas no domínio de uso da Biologia. Ainda que esta pesquisa tenha como ponto de partida o domínio da Biologia, a abordagem foi construída para que possa ser generalizada e aplicada igualmente a outros domínios.

O restante deste artigo está organizado da seguinte maneira: a Seção 3.2 apresenta o contexto a partir do qual se desenvolveu a proposta desta pesquisa; a Seção 3.3 introduz nosso modelo para caracterização e representação dos padrões de construção; a Seção 3.4 detalha o processo de coleta de dados em campo e análise das planilhas utilizadas pelos biólogos, bem como levantamento de hipóteses em relação à sua organização e o método de avaliação dessas hipóteses; a Seção 3.5 fundamenta nossa abordagem para a caracterização de padrões a partir da consolidação de evidências levantadas em trabalhos relacionados e dados em campo; a Seção 3.6 compara nossa abordagem com iniciativas relacionadas de reconhecimento de esquemas implícitos em planilhas; a Seção 3.7 apresenta nossas considerações finais e os próximos passos.

3.2 Cenário da Pesquisa

O ponto de partida para o desenvolvimento desta pesquisa foi a construção de uma estratégia que permita o reconhecimento e explicitação de esquemas implícitos, usualmente adotados naquelas planilhas eletrônicas voltadas ao gerenciamento de dados.

De acordo com Syed et al. [18], uma grande quantidade de informação disponível no mundo está em planilhas. Apesar de sua flexibilidade, as planilhas foram originalmente concebidas para uso independente e isolado, sendo tratadas como arquivos separados, que não são facilmente articulados com dados de outras planilhas/arquivos.

Por esta razão, há uma preocupação crescente em encontrar maneiras de tornar seus dados mais aptos ao compartilhamento e integração [5, 7, 22, 12, 18, 19], convertendo-os em padrões abertos e permitindo que os aplicativos possam interpretar, combinar e articular esses dados.

Há vários trabalhos relacionados que tentam resolver este problema, propondo desde o mapeamento manual para padrões abertos da Web semântica, até o reconhecimento automático de estruturas, pela associação de elementos em planilhas a conceitos disponíveis nas bases de conhecimento Web – por exemplo, DBpedia (<http://dbpedia.org>).

Abordagens mais sistemáticas para armazenamento de dados, tais como bancos de dados, predefinem esquemas explícitos para a gravação dos dados. Estes esquemas podem ser considerados como metadados que dão semântica aos dados armazenados. Planilhas, por outro lado, não têm um esquema explícito. O esquema (implícito) e seus respectivos dados – metadados e dados – fundem-se no mesmo espaço tabular.

A planilha apresentada na Figura 3.1, 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.

Registro-Catalogo	Espécie	Filo	Classe	Pais	Município	Localidade
xrb1358	Hirudo med	Annelida	Polychaeta	EUA	Charleston	Folly Beach
akn9846	Achatina ful	Mollusca	Bivalvia	Brasil	Fernando de N	Baía do Golfinhos
lat5629	O. fragilis	Echinoderm	Ophiuroid	Austrália	Perth	Silver Sands

Figura 3.1: Planilha catálogo de espécies [Instituto de Biologia – Unicamp].

Há diversos aspectos da planilha que dificultam o reconhecimento do seu esquema implícito para integração, como: diferenças no local onde se encontra o esquema e como ele está disposto; a ordem das colunas; o rótulo usado para a identificação de campos e sua respectiva semântica.

Para identificar um esquema implícito, algumas abordagens tentam capturar práticas comuns para a construção de planilhas. Considere-se o comportamento humano ocidental, por exemplo, na criação de uma lista de nomes. Os usuários tendem a colocar os nomes em células adjacentes da mesma coluna (lista vertical) ou a mesma linha (lista horizontal), mas não é habitual colocar nomes em diagonal. Em listas verticais, o rótulo que indica a que se refere a lista – parte do esquema – tende a aparecer no topo da respectiva coluna, geralmente em um estilo diferente.

Ainda que trabalhos relacionados explorem um subconjunto das práticas comuns na construção de planilhas e até mesmo o contexto em que se inserem [9, 19], até onde sabemos, em todos eles tais práticas aparecem implícitas e embutidas de forma estática no código que processa as planilhas. Nenhum deles considerou desenvolver um modelo para caracterizar os padrões de construção, que possa ser representado de forma independente do sistema.

A tese central de nossa abordagem consiste no fato de que a representação explícita de padrões de construção compartilhados por comunidades é a chave para uma interpretação automática mais eficaz de planilhas. A partir da caracterização e reconhecimento destes padrões, alguns resultados preliminares da nossa abordagem têm indicado que é possível um reconhecimento mais efetivo de esquemas implícitos em planilhas e do seu modelo

conceitual subjacente, bem como a produção de resultados semanticamente mais ricos.

Esta pesquisa é motivada por um projeto maior – no qual está inserida – que envolve 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. Por esta razão, esta pesquisa adotou o contexto da biologia como seu foco específico.

A estratégia para modelagem e representação de padrões de construção foi resultado de uma pesquisa que envolveu etapas incrementais, incluindo coleta de dados em campo, formulação de hipóteses/modelos e experimentação. Este processo será detalhado na Seção 3.4. Na subseção a seguir resumiremos observações e hipóteses preliminares feitas na primeira parte desta pesquisa, conforme apresentado em [4].

3.2.1 Tipos de Planilhas

Dentre os diversos tipos de planilhas em um mesmo domínio de uso, notamos que os esquemas seguem padrões que variam de acordo com a intenção do usuário. 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.

Verificamos que há uma tendência das planilhas se agruparem de acordo com as seis perguntas exploratórias: quem, o quê, onde, quando, por quê e como (em inglês: *who, what, where, when, why, how*). De acordo com as quantidades de perguntas respondidas e a ordem em que elas aparecem no esquema das planilhas, as organizamos em quatro grupos maiores:

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 de coletas.

Grupo 3 – Classificação: planilhas que sistematizam classificações taxonômicas.

Grupo 4 – Modelos: meta-planilhas cujos registros descrevem um esquema para a construção de outras planilhas.

Como está ilustrado na Figura 3.2, para classificarmos os grandes grupos das planilhas optamos por alinhar as nossas classes com aquelas definidas na ontologia SUMO [10] – 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; (ii) a ontologia SUMO representa um modelo de classificação amplamente discutido e refinado pela comunidade.

No alinhamento com o SUMO, 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 3.2: Síntese da categorização alinhada com o SUMO.

3.3 Padrões de Construção

Cada vez que um autor inicia o processo de construção de uma planilha eletrônica, ele parte de um modelo conceitual. Por exemplo, um modelo conceitual usual para planilhas do tipo catálogo de espécimes em um museu está ilustrado na Figura 3.4. Os retângulos na figura representam classe de dados e as arestas representam propriedades que relacionam objetos das classes. Este modelo materializará em uma disposição específica de dados sobre a estrutura tabular da planilha. Por exemplo, ao produzir um catálogo de objetos, o autor parte de um modelo conceitual de um conjunto, em que cada elemento corresponde a um objeto com uma identidade própria. Norman [11] chama a materialização de um modelo conceitual do autor em um artefato de software de imagem do sistema. Nesta pesquisa, o artefato de software ou sistema é uma planilha eletrônica, portanto, utilizaremos o termo imagem da planilha para nos referirmos à imagem do sistema em planilhas.

O modo como o autor concebe o modelo conceitual e o transforma em uma imagem da planilha é influenciado por práticas compartilhadas pelo contexto em que ele está inserido. Considere um autor biólogo que está catalogando espécimes de um museu. Seu referencial para a construção do catálogo serão os próprios espécimes, mas também a estratégia usual adotada por biólogos para tabular dados de espécimes. Assim, uma imagem da planilha de catálogo ilustrada na Figura 3.3, usualmente terá as informações taxonomicas da espécime concentradas nos campos iniciais dado que, para este tipo de planilha, eles exercem o papel de campos identificadores. Esse é um exemplo de padrão de construção da imagem da planilha seguido pelo biólogo para este tipo de planilha.

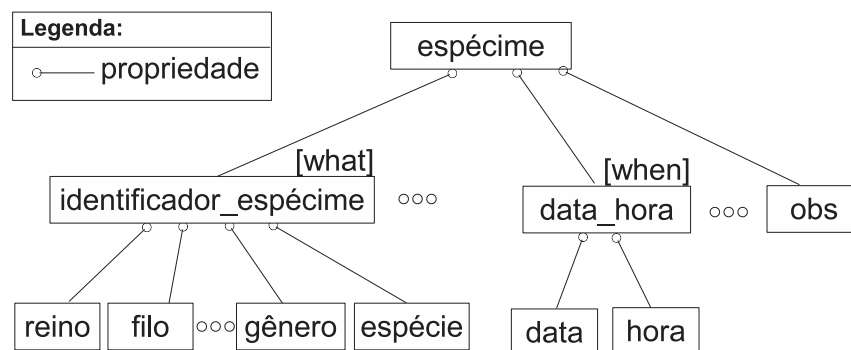


Figura 3.3: Modelo Conceitual Subjacente na Planilha do Tipo Catálogo.

A estratégia de reconhecimento de planilhas, proposta neste trabalho, considera que a identificação destes padrões de construção do seu modelo conceitual subjacente e consequentemente seu esquema implícito.

Por ser uma ferramenta flexível, uma planilha se presta a vários propósitos, e.g., cálculos em geral, orçamentos, gerenciamento de dados. Neste trabalho estamos interessados na aplicação de planilhas para o gerenciamento de dados.

Há aspectos nos padrões de construção associados a planilhas de gerenciamento de dados que são compartilhados pela maioria dos autores independentemente do seu domínio, por outro lado, há práticas especializadas, associadas a um domínio específico. Criar um modelo para representar estes padrões de construção de imagens das planilhas, de modo que possa ser interpretado e usado por máquinas – e especificamente pelo nosso programa de reconhecimento – é um problema fundamental tratado nesta pesquisa.

Todas as planilhas para gerenciamento de dados observadas definem implicitamente um esquema, separado das instâncias de dados que seguem este esquema. Os esquemas são definidos na forma de listas ordenadas de propriedades; cada instância dá valores às propriedades. Estes são pressupostos de base para a construção do modelo que apresentamos a seguir.

A Figura 3.4 apresenta um diagrama que representa um padrão que sintetizamos a partir da análise de planilhas para a catalogação de espécimes. Este diagrama combina os elementos de dados da planilha (modelo de dados) com sua relação espacial.

Em nossa abordagem, a codificação de um padrão de construção é representada como uma construção hierárquica de elementos. No exemplo da figura, os retângulos mais internos representam as propriedades do esquema (no bloco esquema) ou instâncias das propriedades (no bloco instâncias). Propriedades em retângulos internos estão subordinados a elementos em retângulos externos, que são parte do mapa conceitual a ser apresentado, e.g., reino e filo estão subordinados a identificador. Para representar o modo

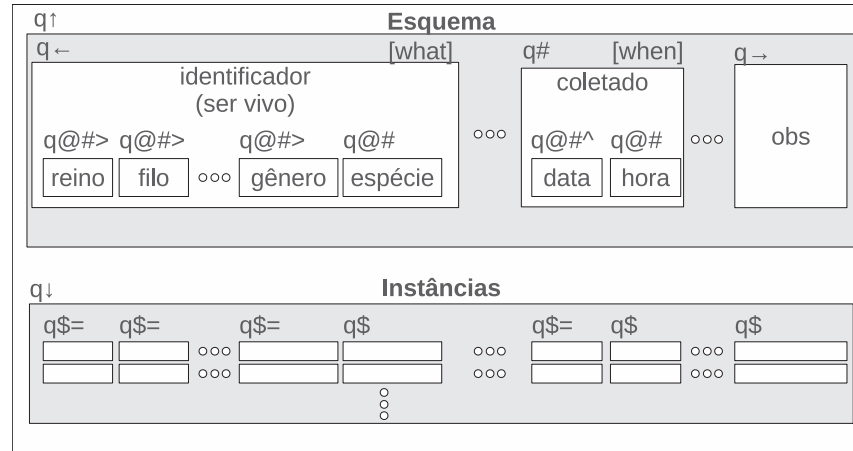


Figura 3.4: Padrão de construção de imagem da planilha em uma planilha tipo catálogo.

como os elementos se relacionam na construção de um padrão de modelo mental, definimos qualificadores que são identificados no diagrama pelo prefixo “q” e são posicionados no canto esquerdo superior do elemento a que se referem. Um diagrama de codificação de um modelo mental pode utilizar os seguintes qualificadores:

Qualificador posicional – Caracteriza um elemento a partir de sua posição dentro do elemento de nível superior. São quatro elementos posicionais: à esquerda ($q\leftarrow$), à direita ($q\rightarrow$), acima ($q\uparrow$) e abaixo ($q\downarrow$). Na Figura 3.4 os qualificadores posicionais indicam que um esquema se posiciona acima de suas instâncias; um identificador à esquerda e observações à direita.

Qualificador de ordem ($q\#$) – Caracteriza um elemento a partir de sua ordem em relação a elementos vizinhos. Na figura, cada parte do identificador é reconhecido a partir de sua ordem; um elemento de caracterização temporal (data_hora) se posiciona depois de identificadores.

Qualificador por rótulo ($q@$) – Indica que o rótulo caracteriza o elemento. No exemplo, o rótulo “espécie” permite identificar que aquela coluna se refere à espécie.

Qualificador de tipo de dados ($q\$$) – Caracteriza a predominância de um tipo de dados nas instâncias de uma propriedade. Na figura, as partes do identificador são qualificadas como strings.

Qualificador de abrangência – Caracteriza se elementos vizinhos têm relação de generalização/especialização. O qualificador $q\text{indica que o da esquerda é mais genérico que o da esquerda}$

Qualificador de ordenação – Caracteriza instâncias que se ordenam de forma crescente ($q+$) ou decrescente ($q-$). O exemplo da Figura 3.4 não tem qualificadores de ordenação. Em planilhas de eventos, as instâncias de data e hora estão usualmente orde-

nadas de forma crescente e recebem qualificadores de ordenação.

Qualificador de Redundância (q=) – Caracteriza a redundância de informação em instâncias de uma propriedade. Tal redundância é típica em propriedades compostas nas quais a propriedade da esquerda é mais genérica ou abrangente que a da direita. No exemplo, a propriedade reino é altamente redundante, dado que várias instâncias da planilha compartilharão o mesmo reino.

Além dos qualificadores, elementos são associados quando possível a uma das seis perguntas exploratórias (*who, what, where, when, why, how*), representadas no diagrama entre colchetes. Esta associação subsidia a generalização da codificação do padrão de construção. Por exemplo, observamos que usualmente planilhas de objetos – mesmo de outras naturezas – definem um identificador *what* mais à esquerda (q←).

Para a representação de tais diagramas em formato interpretável por máquinas, optamos por partir de um modelo conceitual, tal como ilustra a Figura 3.5, e enriquecê-lo com qualificadores e caracterizações das perguntas exploratórias. O diagrama da Figura 3.5 reflete a representação de codificação do padrão de construção da Figura 3.4. O modelo conceitual é baseado na linguagem RDF/OWL da Web Semântica. Na figura, as ovas representam classes (rdfs:class ou owl:class) e os retângulos propriedades (rdfs:property). As arestas identificadas como *domain* (rdfs:domain) indicam que uma dada propriedade é usada na caracterização daquela classe. As arestas *range* indicam que os valores da propriedade atendem à classe indicada. Para fins de simplificação, o diagrama omite detalhes de representação do RDF/OWL, por exemplo, a diferenciação de propriedades objeto – cujos valores são instâncias de classes – daquelas de dados.

Os qualificadores estão registrados no diagrama acima ou abaixo das propriedades que eles qualificam, indicando que aquele qualificador se aplica à relação entre a respectiva propriedade e à classe a que ela está associada por *domain* no modelo. Por exemplo, o qualificador q← (qualificador posicional à esquerda) é representado sobre a propriedade IDENTIFICADOR, indicando que quando esta propriedade é usada na descrição de espécime e o padrão é que ele pareça posicionada à esquerda. Quando o qualificador é posicionado acima significa que ele se aplica à propriedade no esquema. Por exemplo, q← está acima de IDENTIFICADOR indicando que o posicionamento à esquerda é verificado na propriedade dentro do esquema. Quando o qualificador é posicionado abaixo significa que ele se aplica aos valores da propriedade no bloco de instâncias. Por exemplo, o qualificador q\$= abaixo de reino indica que um tipo específico e redundância são observados nos valores desta propriedade.

Os qualificadores são representados no modelo de dados RDF/OWL como anotações sobre a relação entre a propriedade e a classe. Assim também acontece com as seis perguntas exploratórias, representadas entre colchetes.

Tal representação nos permite caracterizar padrões de construção de planilhas com-

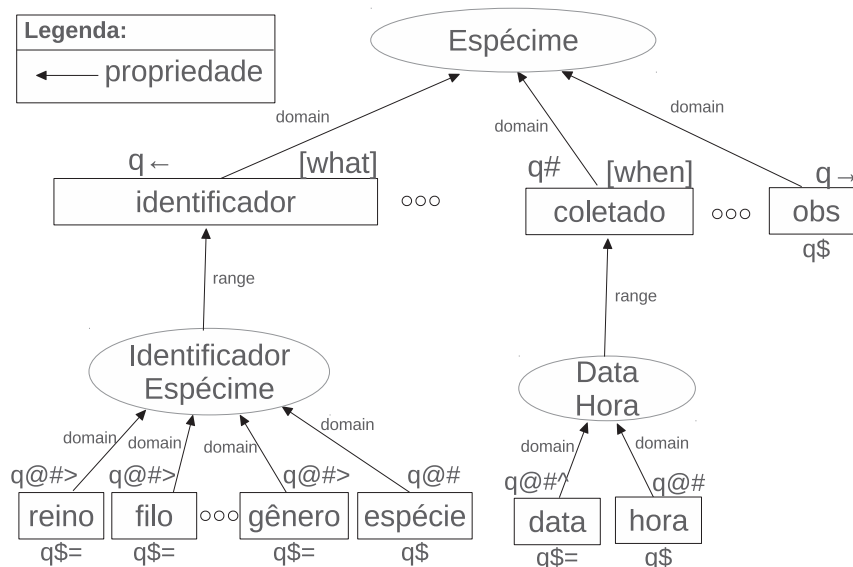


Figura 3.5: Modelo conceitual de Planilha de Objetos (Espécimes) enriquecido com qualificadores.

partilhados por usuários e sua relação com o modelo conceitual subjacente. A Figura 3.6 ilustra esta técnica aplicada na caracterização de uma planilha de coleta de espécimes, utilizada por biólogos para registro de coletas em campo. Trata-se de uma planilha em que cada instância corresponde a um evento – uma coleta.

Neste modelo destacamos que:

- O identificador de cada instância(evento)é o momento em que ele ocorreu no espaço e no tempo, por ser um identificador, sua posição é à esquerda ($q \leftarrow$).
- As propriedades referentes ao tempo aparecem em ordem crescente ($q +$) nas instâncias.

3.4 Metodologia

Conforme mencionado anteriormente, nossa abordagem para a representação de padrões de construção partiu de um estudo de trabalhos relacionados e uma extensa pesquisa em campo.

Com base na análise inicial de como os biólogos do Instituto de Biologia (IB) da Unicamp constroem suas planilhas, idealizamos um processo de reconhecimento automatizado baseado em padrões de construção [4], cuja concepção se deu em ciclos iterativos crescentes, que envolveram (i) coleta e análise de dados em campo; (ii) formulação de hipóteses sobre os padrões de construção de planilhas; (iii) projeto e implementação de

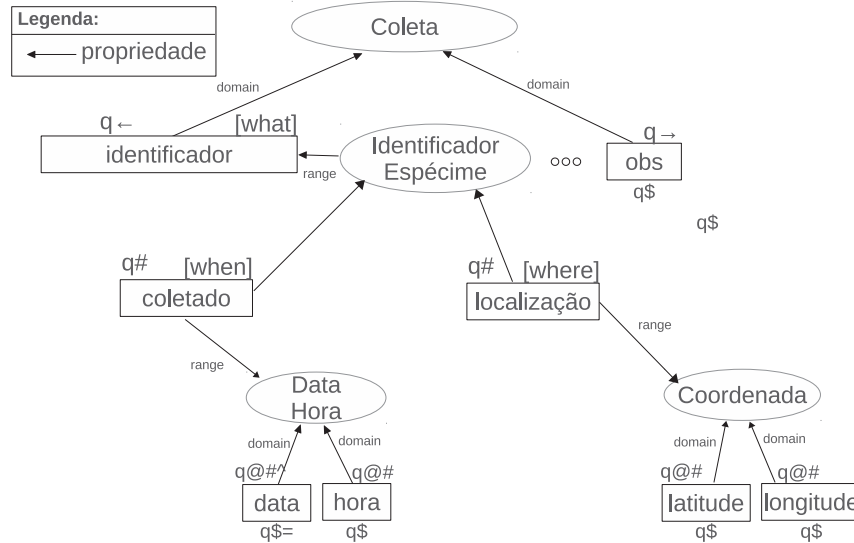


Figura 3.6: Modelo conceitual de Planilha de Eventos (Coletas) enriquecido com qualificadores.

reconhecedores automáticos destas planilhas. Os resultados obtidos pelos reconhecedores realimentaram a etapa (i) e fecharam o ciclo.

3.4.1 Coleta e análise inicial de dados

A observação iniciou com 9 planilhas pertencentes ao IB e, a partir dessas planilhas, traçamos dois perfis de construção para as mesmas - planilhas que relacionam: objetos e eventos. Verificamos que a abstração dos campos de uma planilha a partir das seis perguntas exploratórias facilita a caracterização de padrões de construção. O próximo passo foi coletar na Web mais planilhas para compor nossa amostragem. O critério empregado na busca foi utilizar palavras chave pertencentes ao domínio de uso da biologia. A análise da nova amostra de planilhas nos fez identificar mais dois perfis de construção: classe e modelo.

3.4.2 Hipóteses

A partir da observação das planilhas, verificamos um padrão de construção entre elas e levantamos as seguintes hipóteses, detalhadas em Bernardo et al. [4]:

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 e seu respectivo padrão de construção, conforme foi apresentado anteriormente.

3.4.3 Modelo de Reconhecimento

A partir destas hipóteses, foi proposto um modelo de reconhecimento automático, detalhado em Bernardo et al. [4]. Nesse modelo, propomos uma explicitação de esquemas diferenciada dos trabalhos relacionados, pois consideramos a natureza da planilha e o contexto no qual ela se insere, subsídios para alcançarmos uma interpretação semântica mais rica. Como resultado desse modelo, foi possível categorizar planilhas de acordo com a Figura 3.2.

3.4.4 Sistema

Baseados no modelo proposto, construímos um sistema com o objetivo de validar as hipóteses. A Figura 3.7 ilustra a arquitetura geral do sistema. Em (1) está representada a entrada de dados do sistema; estes dados se dividem em dois grupos: planilhas eletrônicas coletadas na Web e um conjunto de dados, responsáveis por guiar o processo de reconhecimento das planilhas, que chamaremos de Perfil.

No Perfil estão registrados os dados que caracterizam os padrões de construção de planilhas. Atualmente ele possui um dicionário de termos reconhecidos, em que é possível mapear estes termos às perguntas exploratórias. Cada termo recebe também um peso de acordo com a sua relevância, ou seja, se o termo espécie for 50% mais relevante que o termo longitude no reconhecimento de uma categoria específica de planilhas, então o peso-relevância de espécie será 10 e o de longitude será 5, por exemplo.

Outro tipo de configuração possível é o mapeamento de cada termo e pergunta exploratória à sua categoria da ontologia SUMO correspondente.

Na Figura 3.7 (2) está representado o processamento das planilhas, em que o sistema extrai dados destas planilhas utilizando uma biblioteca chamada DDEX e realiza o processamento baseado no modelo proposto. Vide detalhes em Bernardo et al. [4].

Assim, na fase de explicitação de esquemas, é realizada uma busca pelos termos relevantes da planilha coletada, em relação aos termos contidos no dicionário. Esta busca envolve dois aspectos: seu peso-relevância conforme o padrão de planilha buscado e sua posição espacial em relação a outros termos. O processo de reconhecimento funciona em etapas iterativas, em que o reconhecimento de propriedades contribui na caracterização da natureza da planilha e seu respectivo padrão de construção. E este padrão de construção retroalimenta o reconhecimento de campos.

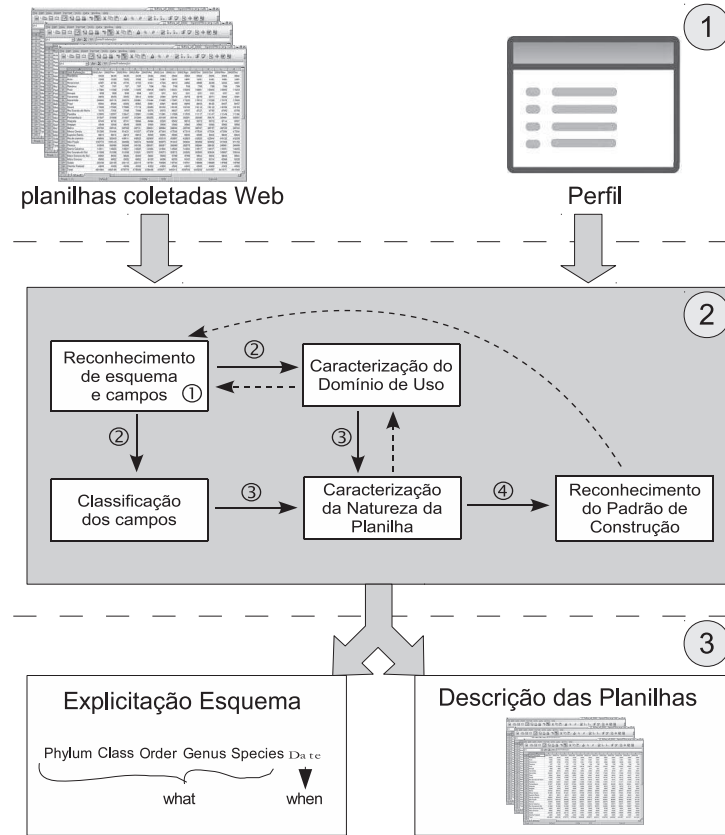


Figura 3.7: Arquitetura do Sistema.

Como cada termo do dicionário possui um peso-relevância, quando algum termo é encontrado, contabilizamos os respectivos pesos levando em consideração a natureza da planilha. O processo continua até encontrar um esquema que satisfaça as condições ou, caso o sistema não se consiga explicitar o esquema, ou seja, não se reconheça nenhuma estrutura que se enquadre nos padrões esperados, a respectiva planilha é classificada como não reconhecida.

Como resultado de saída, o sistema gera o esquema reconhecido e uma categorização da planilha de acordo com a ontologia SUMO.

A versão do sistema apresentada em Bernardo et al. [4] embute a lógica de reconhecimento das relações espaciais e de ordem entre campos – para reconhecimento de categorias e padrões de construção de planilhas – dentro do código. Na medida em que ampliamos o universo de análise de planilhas, tornou-se crescente a necessidade de se criar uma representação que expressasse a forma como autores e usuários pensam as planilhas. Tal representação deveria ser passível de interpretação por máquinas, de modo que pudesse

guiar o processo de reconhecimento. Neste ponto, foi formulada nossa abordagem pautada na representação de padrões de construção.

Como apresentaremos na seção a seguir, os dados obtidos a partir da execução do sistema em milhares de planilhas coletadas na Web formaram uma base de evidências, que dão suporte à nossa tese de que a construção de planilhas segue padrões compartilhados por comunidades. A análise estatística de dados coletados evidencia padrões de construção, bem como modelos conceituais subjacentes, projetados a partir das mentes de seus autores.

3.5 Evidências de Padrões de Construção

O processo de coleta e análise das planilhas foi dividido em quatro etapas principais:

1. Análise manual de uma amostra inicial de 42 planilhas.
2. Análise automática das mesmas planilhas que foram analisadas manualmente.
3. Confronto dos resultados obtidos nas etapas anteriores e refinamento do sistema;
4. Análise automática realizada com amostras crescentes contendo 1.914, 5.633 e 11.150 planilhas.

Na etapa 1, coletamos inicialmente 9 planilhas do Instituto de Biologia, utilizadas para diferentes fins. A análise dessas planilhas foi direcionada ao aspecto da identificação de padrões de construção seguidos pela comunidade. Os resultados obtidos nesta análise direcionaram a coleta de mais 33 planilhas na Web, que estivessem nesses padrões. A partir da análise manual destas planilhas, constatamos um padrão de construção nessas 42 planilhas, em relação à: disposição do esquema (labels das colunas), disposição dos dados, ordenação dos campos e seu agrupamento, título da planilha etc. Desenvolvemos uma versão inicial do sistema de reconhecimento automático, para validarmos as hipóteses levantadas.

Os resultados obtidos a partir da análise automática foram: das 42 planilhas analisadas, o sistema reconheceu 33 planilhas (78,6%). O sistema não reconheceu 9 planilhas (21,4%). Verificou-se que se tratavam de planilhas da categoria abstrata, às quais o sistema ainda não está preparado para reconhecer. Essa funcionalidade será implementada em trabalhos futuros.

Após esta primeira etapa de validação e depuração do sistema, coletamos 1.914 planilhas da Web de forma aleatória. Utilizamos o mecanismo da busca da Google para localizar planilhas a partir de palavras chave encontradas nas planilhas anteriores: *kingdom*, *phylum*, *order*, *biodiversity*, *species*, *identification key*. Das 1.914 planilhas coletadas,

o sistema reconheceu 137 planilhas (7%). A análise manual destas planilhas demonstrou que o sistema identificou corretamente 116 planilhas e identificou incorretamente (falsos positivos) 21 planilhas. Estas últimas apesar de apresentarem o padrão de construção esperado, não satisfazem ao foco de estudos, que são planilhas utilizadas para gerenciamento de dados.

Aumentamos a nossa amostragem para 5.633 planilhas e novamente o sistema reconheceu 7%. Posteriormente, aumentamos pra 11.150 planilhas e o sistema reconheceu 1.151 planilhas (10%). Esses resultados, embora pareçam insuficientes, foram satisfatórios se considerarmos que essas planilhas foram coletadas através de ferramentas de pesquisa da Web. Segundo Venetis et al. [19], essas ferramentas de pesquisa tratam estruturas tabulares como qualquer fragmento de texto, sem considerar a semântica implícita de sua organização e consequentemente causando prejuízos aos resultados da busca. Adicionalmente, os resultados atendem ao propósito do trabalho deste artigo, cujo foco é a coleta de evidências que dão suporte à nossa tese. Os gráficos a seguir são resultantes da análise das 1.151 planilhas reconhecidas em um universo de 11.151.

O gráfico da Figura 3.8 demonstra que os esquemas das planilhas se concentram nas linhas iniciais. A redução do número de planilhas que têm esquemas afastados das primeiras linhas decresce exponencialmente na medida que nos distanciamos das linhas iniciais. Há, portanto, uma tendência no posicionamento de esquemas na parte superior ($q\uparrow$) e instâncias logo em seguida ($q\downarrow$).

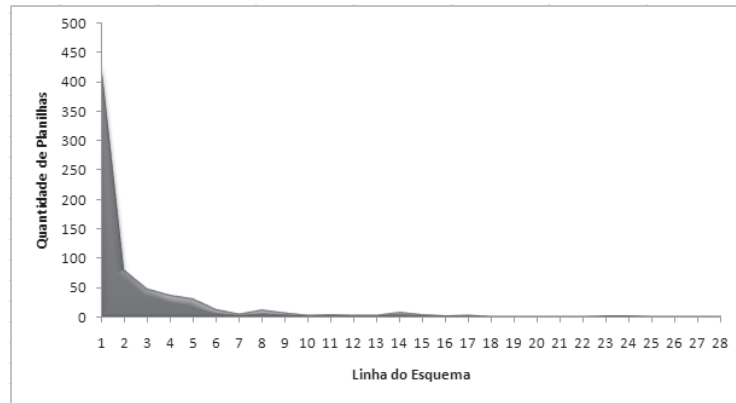


Figura 3.8: Localização do esquema.

A Figura 3.9 representa a localização (número da linha) dos termos contidos no dicionário referentes ao esquema dentro das planilhas coletadas. Verificamos que a maioria dos termos se localizam nas linhas iniciais. O resultado desta análise complementa a evidência apresentada na Figura 3.8 de que o esquema se posiciona na parte superior da planilha.

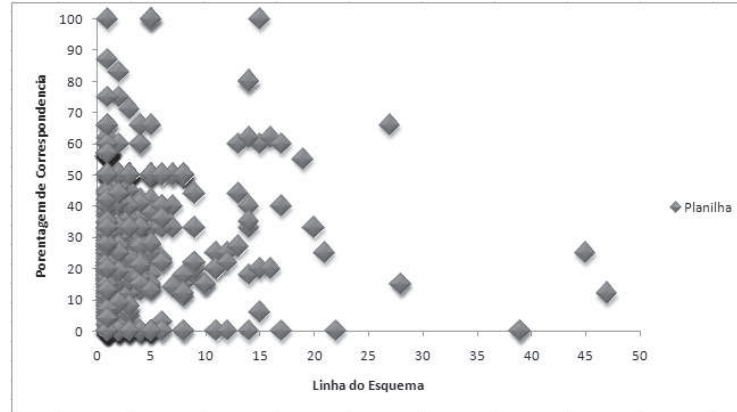


Figura 3.9: Termos por esquema nas linhas iniciais.

A representação do padrão de construção apresentado na Figura 3.5 para planilhas de espécimes estabelece que campos *what* identificam planilhas e se posicionam à esquerda. Para evidenciar tal padrão, produzimos diagramas que ilustram a incidência de termos – relativos a uma das seis questões – dentro da organização espacial de uma planilha. Como mostra a Figura 3.10, o diagrama é organizado em células dispostas em 9 colunas e 6 linhas. Cada célula representa proporcionalmente um quadrante da planilha. As células da primeira coluna representam o quadrante de todas as células mais à esquerda; as células mais à direita o último quadrante. A proporção de cada planilha foi ajustada para o diagrama. As linhas mostram a distribuição vertical das células no esquema. Foram consideradas as primeiras linhas da planilha. A graduação de cinza indica o percentual de células respondendo à respectiva pergunta localizadas naquela posição relativa da planilha.

A Figura 3.10 mostra a distribuição de elementos *what* em uma planilha de objetos. Como mostra a figura, a incidência destes itens é nas primeiras colunas, mais acima. Foram encontrados elementos em linhas abaixo da primeira, mas a quantidade foi insipiente.

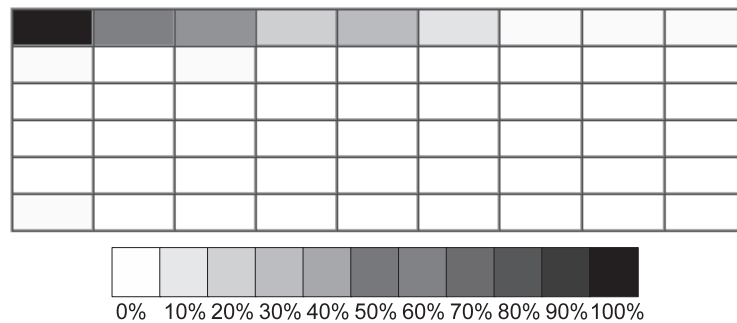


Figura 3.10: Distribuição do *what* - planilhas de objetos.

Em planilhas de objetos, as demais perguntas aparecem em proporções menores. Com o objetivo de realizar uma análise comparativa entre os campos das planilhas, foi construído um gráfico na forma de radar, que confronta as respectivas proporções, conforme mostra a Figura 3.11. Através desta figura, podemos notar que planilhas categorizadas como objeto tendem a possuir muitos termos que respondem à pergunta *what* e alguns termos que respondem à pergunta *who*. Já as outras perguntas não apresentaram quantidade significativa, o que reforça nossa hipótese que planilhas do tipo objeto tendem a ter campos destinados a identificação e detalhamento.

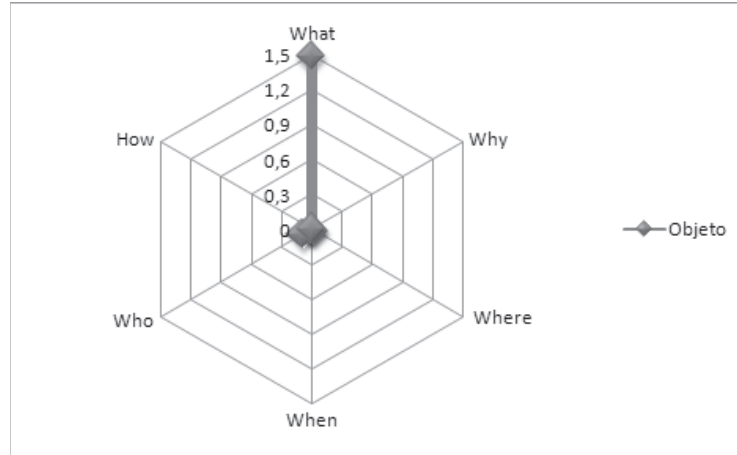


Figura 3.11: Relação entre os campos - planilhas objeto.

As Figura 3.12 e Figura 3.13 mostram a distribuição das perguntas exploratórias *what* e *where* nas planilhas de eventos. A partir delas é possível observar como a distribuição espacial caracteriza a planilha.

Figura 3.12: Distribuição do *what* - planilhas de eventos.

Através da Figura 3.14 podemos visualizar as proporções nas planilhas categorizadas como evento. Verificamos que nessas planilhas há grande quantidade de termos que respondem as perguntas *where* e *what* e uma quantidade significativa de *when*.

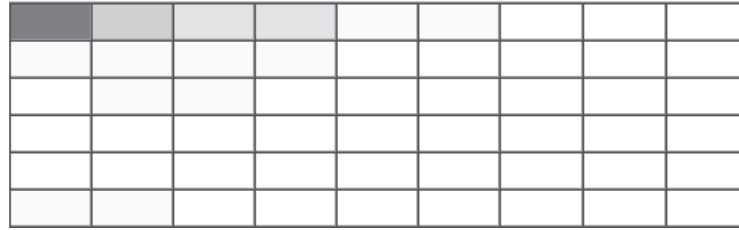


Figura 3.13: Distribuição do *where* – planilhas de eventos.

Se confrontarmos os resultados da Figura 3.14 e Figura 3.15, percebemos que considerar a organização posicional dos termos é de extrema relevância. O diagrama da Figura 3.15 atribui pesos que decrescem exponencialmente na medida que o termo se afasta das primeiras colunas. Deste modo, a quantidade de ocorrências é combinada com a sua proximidade da posição à esquerda. Analisando a Figura 3.14 de forma isolada, tendemos a pensar que as perguntas *what* apresentam maior relevância que as *when*, no entanto, através da Figura 3.15, percebemos que os termos que respondem as perguntas *when*, por estarem em sua maioria localizados nos campos iniciais, são os responsáveis por guiar o objetivo que está implícito na planilha.

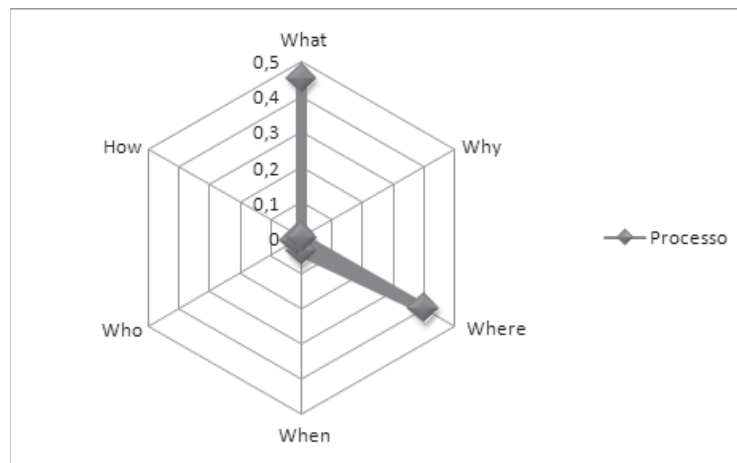


Figura 3.14: Relação entre os campos sem peso posicional - planilhas tipo processo

3.6 Trabalhos Relacionados

Tal como apontamos na Seção Modelo de Reconhecimento, uma característica fundamental de planilhas usadas para gerência de dados é a separação esquemas/instâncias, em que

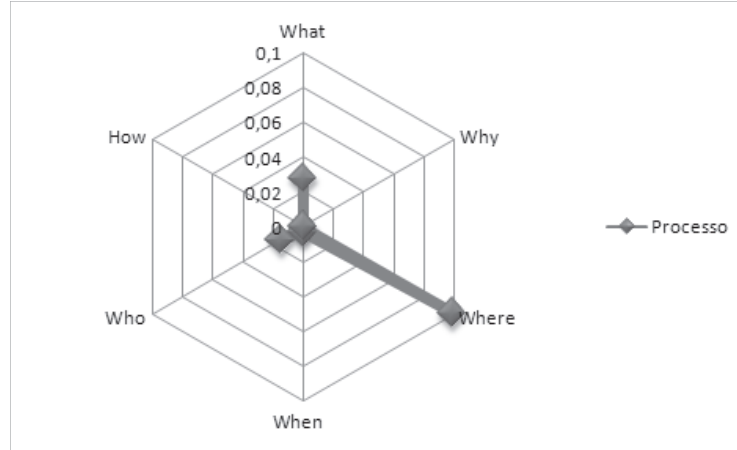


Figura 3.15: Relação entre os campos com peso posicional - planilhas tipo processo

o esquema se apresenta acima ($q \uparrow$) ou à esquerda ($q \leftarrow$) e as instâncias abaixo ($q \downarrow$) ou à direita ($q \rightarrow$).

Esta constatação aparece em todos os trabalhos, cujo primeiro desafio é reconhecer o esquema implícito em planilhas. Syed et al. [18] destaca que este desafio remete a um problema mais genérico de extrair esquemas implícitos de fonte 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.

Han et al. [5] adotam a abordagem mais simples de reconhecimento de esquema e separação esquema/instância chamada *entity-per-row*. Nesta abordagem, cada linha da tabela deve descrever uma entidade diferente e cada coluna um atributo para essa entidade. A planilha da Figura 3.1, 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). Como acontece em muitos trabalhos relacionados, a partir desta consideração inicial, Han et al. [5] adotam um 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 3.1 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.

Langegger and Wöb [7] possuem uma solução similar àquela de Han et al. [5] 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 3.1, os campos País, Município e Localidade se referem ao local onde a 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. Esta perspectiva hierárquica sobre a abstração dos campos é representada em nosso modelo pelo qualificador $q\#$.

Por ser um padrão aberto que possibilita a interoperabilidade sintática e semântica de dados, o RDF tem sido adotado como formato de saída para a integração de dados de várias planilhas. Langegger and Wöb [7] propõem o acesso a estes dados através do uso da linguagem SPARQL [13] – uma linguagem de query para acesso a RDF. Oconnor and Halaschek-Wiener [12] propõem uma solução semelhante à de Langegger and Wöb [7], mas utilizam OWL.

Uma segunda abordagem para o problema é desvincular os dados da planilha de sua estrutura tabular, pois segundo Zhao et al. [22] 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. [22] 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.

Abraham and Erwig [1] identificam que muitas planilhas não são criadas e sim reutilizadas, porém devido a sua flexibilidade e sua abstração, a reutilização dessas por pessoas que não estão inseridas no domínio gera erros de interpretação e portanto inconsistência.

Assim ele define o ciclo de vida da planilha em duas fases: desenvolvimento e utilização. Com essa definição, sua proposta é separar o esquema dos dados, em que o esquema será desenvolvido no primeiro ciclo e uma vez definido, não poderá ser mais alterado e os dados serão inseridos e manipulados no segundo ciclo de forma guiada pelo modelo desenvolvido.

Outra maneira de resolver o problema é automatizar o mapeamento semântico dos dados utilizando *Linked Data*. Syed et al. [18] 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.

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. Os dados destas bases também podem apresentar inconsistências. Deste modo, há trabalhos que identificam a importância de considerar um escopo antes de se tentar alcançar uma semântica mais completa.

Venetis et al. [19] consideram utilizar a semântica existente nas tabelas para guiar as operações de manipulação de dados que podem ser realizadas com elas. Sua proposta descreve um sistema que analisa pares de termos das colunas e a sua relação, a fim de encontrar uma semântica mais apropriada, pois identifica que o principal problema de interpretação de dados tabulares é a análise dos termos de forma independente. Além disso, destacam o problema de heterogeneidade dos dados que pode existir se um escopo não for definido. Dessa forma, este trabalho tenta identificar o escopo através da consolidação do resultado da análise da relação dos termos.

Dietmar et al. [6] também levantam a problemática da interpretação dessas tabelas, verificando que sua estrutura compacta e a forma precisa de apresentar os dados são direcionadas principalmente para leitura humana e não para sua manipulação. Sua proposta é um sistema para extrair informação de tabelas Web e associá-las a uma ontologia.

Dietmar et al. [6] desenvolvem o processo de mapeamento semântico dos dados com três tipos de ontologias:

1. core: modelo desassociado de qualquer conceito
2. core + domínio: informação que se deseja recuperar
3. instância da ontologia: ontologia + contexto

Essas ontologias têm por meta ir mapeando a informação para a sua respectiva representação semântica de forma gradual, e direcionada pelo objetivo do usuário. Dentre as soluções apresentadas, notamos que algumas elegem os dados das planilhas individuais – destituídas de contexto – e outras que apesar de não trabalharem com planilhas diretamente e sim com informação tabular, verificam a importância de se identificar e caracterizar um contexto. Todas as abordagens caracterizam padrões de construção de planilhas, mas nenhuma delas propõe um modelo para a sua representação, como é feito neste trabalho.

3.7 Conclusão e Trabalhos Futuros

Planilhas eletrônicas se tornaram muito populares em diversos meios por fornecerem liberdade e autonomia a seus diversos usuários. Este artigo buscou reconhecer, mapear e representar o modo como usuários estabelecem padrões de construção, que se refletem na organização dos esquemas e dados nas planilhas.

Neste trabalho partimos do pressuposto de que tal reconhecimento e representação podem guiar um processo automático de reconhecimento e explicitação de esquemas dirigido por padrões de construção que expressam modelos conceituais subjacentes. Nosso processo também envolve o reconhecimento automático do esquema e associação entre campos/registros das planilhas a conceitos disponíveis em ontologias. Nenhuma das abordagens analisadas parte da caracterização de modelos conceituais subjacentes e sua associação com padrões de construção para categorizar as planilhas, conforme a natureza da informação que representam, e para reconhecê-las. 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 espécimes em um museu (objetos) podem ser associados a registros de suas coletas (eventos) de uma maneira específica.

Embora nossos estudos estejam focado a área da biologia, pretendemos, em trabalhos futuros, realizar a generalização para outras áreas.

Capítulo 4

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

4.1 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 não 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 de planilhas em padrões abertos aptos ao reúso e integração [5, 7, 12, 18, 19]. 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 demasiadamente amplo de possibilidades de construção, em que não há restrição de um contexto ou domínio específico, que possibilite o direcionamento do 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 trabalhos anteriores demonstramos que muitos destes padrões são passíveis de serem reconhecidos por programas de computador. Em [4] introduzimos nossa estratégia para reconhecimento automático de tais padrões e em [3] apresentamos nossa abordagem para modelar tais padrões a partir da caracterização de modelos mentais de seus usuários.

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. Nosso processo inclui o reconhecimento automático do esquema e associação entre campos/registros das planilhas 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 [21, 14] 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 4.2 apresenta uma visão geral da literatura correlata. A Seção 4.3 introduz nosso processo de explicitação esquema das planilhas. A Seção 4.4 apresenta nosso sistema que integra dados de planilhas a partir do reconhecimento de sua natureza. A Seção 4.5 apresenta a conclusão e trabalhos futuros.

4.2 Revisão da Literatura

Há diversas pesquisas que se propõem a alcançar interoperabilidade semântica para dados tabulares. 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 4.1 apresenta uma planilha que registra informações sobre coleta de espécimes de abelhas, que será usada como exemplo para ilustrar a análise dos trabalhos relacionados.

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 4.2 e 4.4. Trata-se, portanto, de dados em um formato tabular para um grafo.

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

Figura 4.1: Exemplo de planilha de registro de coleta [siscom.ibama.gov.br].

Han et al. [5] utiliza uma abordagem de mapeamento manual *entity-per-row* [12] 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 4.1 apresenta o grafo RDF resultante do mapeamento semântico de Han et al. [5] em uma das linhas da planilha apresentada na Figura 4.1. 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 o reconhecimento de atributos, em detrimento da caracterização do propósito mais amplo da planilha – 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 Wöb [7] vai além do *entity-per-row* e propõe esquemas de mapeamento de hierarquias implícitas encontradas em planilhas. Sua abordagem é capaz de interpre-

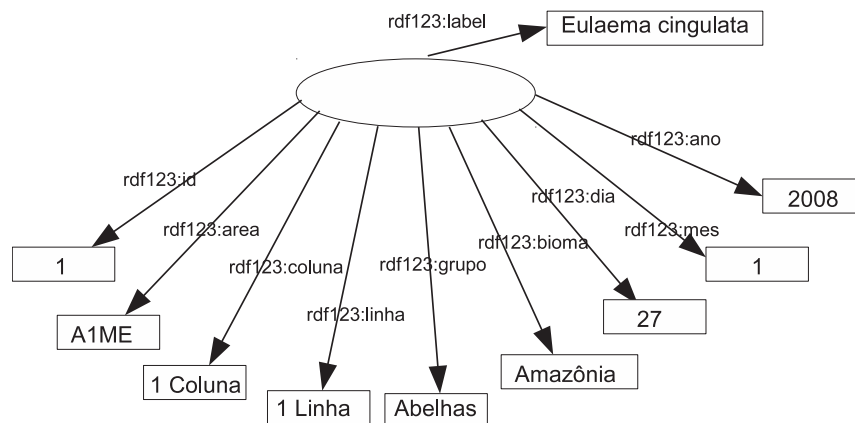


Figura 4.2: Exemplo mapeamento semântico realizado por [Han et al. 2008].

tar o agrupamento de células “número da unidade de coleta”, transformando-o em uma hierarquia de objetos.

Syed et al. [18] consideram que mapear os dados semanticamente 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 rótulos 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 desta abordagem está no fato de que ela pode gerar ligações ambíguas e inconsistentes. Além disto, como abordado anteriormente, seu enfoque está nos atributos.

Aplicando esta estratégia no exemplo da Figura 4.1, poderia ser gerada uma inconsistência ao se analisar a coluna **Grupo**, que pode ter diferentes interpretações em contextos distintos. Venetis et al. [19] aborda 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. [19] tentaria resolver a ambiguidade do termo **Grupo** relacionando-o com **Espécie** ou **Bioma**. Apesar de aprimorar a associação entre atributos e termos em ontologias, a interpretação continua com o enfoque fragmentados 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 4.1, 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 4.1 com duas outras planilhas, conforme mostrada na Figura 4.3. Tal como a planilha da Figura 4.1, a planilha da Figura 4.3(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 4.3(b) é de outra natureza, trata-se de um catálogo de espécimes. Neste caso não faz sentido um merge, mas dados das planilhas da Figura 4.1 e Figura 4.3(a) podem se articular com aqueles da Figura 4.3(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.

							POLINIZADOR									
1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
2	Filo	Subfilo	Classe	Orden	Superfamília	Família	Subfamília	Gênero	Especie	Eta de Vida	Função / At					
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								

1	Filo	Subfilo	Classe	Orden	Superfamília	Família	Subfamília	Gênero	Especie	Eta de Vida	Função / At
73	Arthropod	Hexapoda	Insecta	Hymenoptera	Apoidea	Apidae	Apinae	Eulaema	Eulaema bombiformis	Adulto	Polinizadores de vario
74	Arthropod	Hexapoda	Insecta	Hymenoptera	Apoidea	Apidae	Apinae	Eulaema	Eulaema cingulata	Adulto	Polinizadores de vario
75	Arthropod	Hexapoda	Insecta	Hymenoptera	Apoidea	Apidae	Apinae	Eulaema	Eulaema meriana	Adulto	Polinizadores de vario

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

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

Figura 4.3: Exemplo de planilhas.

Esse processo segue a mesma linha utilizada na Semântica In Loco [16, 17], pois interpreta padrões de organização e comportamento do usuário, a fim de automatizar parte do processo envolvido na identificação da semântica e mapeamento. A Semântica In Loco refere-se a uma metodologia que subsidia a identificação de esquemas implícitos associados a recursos digitais, explicitando-os pelo uso de semântica interoperável. Ela 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 concomitantemente à produção de conteúdo (in loco). *Integração da Metáfora*: 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 in loco são projetadas a fim de possibilitar identificação automática dos esquemas implícitos e conversão para padrões abertos da web semântica. *Persistência Semântica*: elementos da anotação in loco 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.

4.3 Explicitação de Esquema dirigida pela Natureza da Planilha

Conforme mencionado anteriormente, a proposta envolve a implantação de um processo em que os dados das planilhas são extraídos e transformados em RDF/OWL, para ser 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.

Em Bernardo et al. [4] 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 [4], 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.

A versão do sistema apresentada em Bernardo et al. [4] embute no código a lógica de reconhecimento das relações espaciais e de ordem entre campos, para reconhecimento de categorias e padrões de construção de planilhas. Na medida em que ampliamos o universo de análise de planilhas, tornou-se crescente a necessidade de se criar uma representação, que expressasse a forma como autores e usuários pensam as planilhas, explicitando os padrões compartilhados por comunidades. Tal representação deveria ser passível de interpretação por máquinas, de modo que pudesse guiar o processo de reconhecimento. Neste ponto, foi formulada nossa abordagem pautada na representação de modelos mentais [3].

Em trabalhos relacionados, a caracterização de modelos mentais tem sido usada para guiar o projeto de interfaces de sistemas, de modo que elas reflitam as expectativas dos usuários [15]. Há sucesso na interface do sistema, quando o modelo mental concebido pelo projetista é similar ao pensado pelo usuário. Nossa pesquisa [3] propôs que o mesmo

vale para o processo inverso, ou seja, máquinas podem interpretar com sucesso dados produzidos por usuários, se forem capazes de decifrar padrões de modelo mentais destes usuários, usados ao criarem esses dados.

4.4 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 4.4 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 4.1 (evento) e o lado (B) representa o RDF da planilha da Figura 4.3(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 espécie no museu e materializada em RDF como instância da classe `bio:Specimen`.

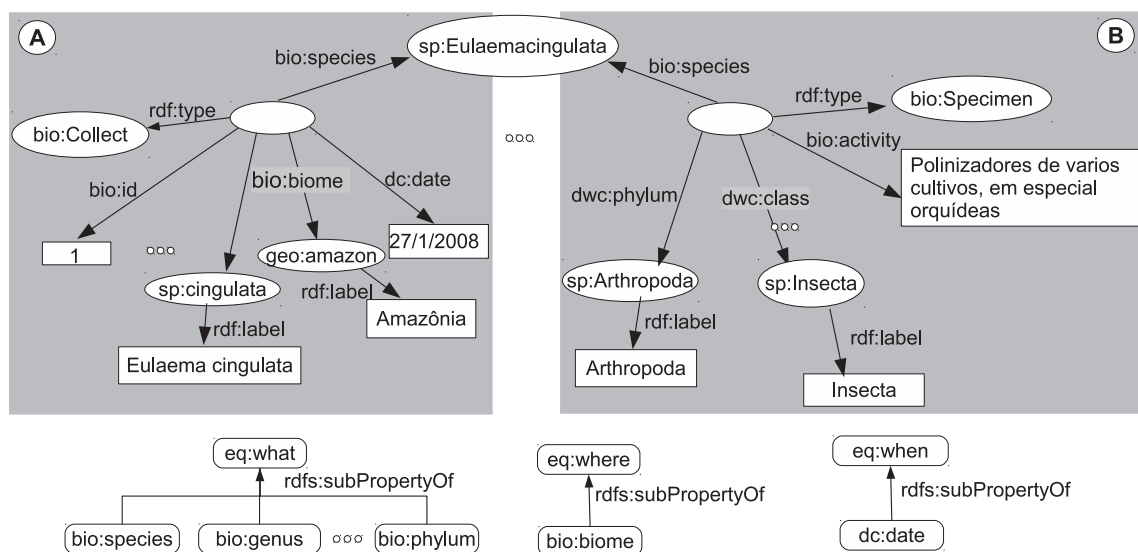


Figura 4.4: Mapeamento semântico das planilhas Fig.2 e Fig.3.(b).

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 1151, 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. 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 só 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 da planilha.

A Figura 4.5 apresenta as etapas de execução do nosso sistema. Na etapa 1 é realizado o reconhecimento da natureza da planilha e seu respectivo esquema. Na etapa 2 o módulo de mapeamento transforma os dados em RDF. Na etapa 3 estes dados são armazenados em um banco de dados RDF chamado Virtuoso (<http://virtuoso.openlinksw.com>), que permite o acesso por uma interface web.

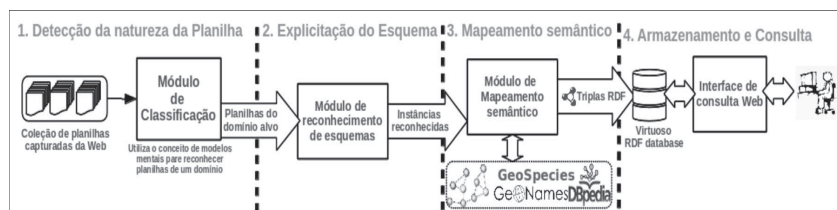


Figura 4.5: Etapas de execução do sistema de reconhecimento e mapeamento de planilhas.

Como ilustra a Figura 4.4, ao contrário dos trabalhos relacionados, em nossa abordagem o valor atribuído a cada propriedade não se limita a rótulos. Na instância da espécie, por exemplo, que está no lado (B) da Figura 4.4, é possível verificar que o valor da propriedade `dwc:phylum` – que indica o filo do animal representado usando o vocabulário Darwin Core (<http://rs.tdwg.org/dwc/>) – é por sua vez uma instância de um objeto. Neste caso, trata-se de uma instância que representa o filo Arthropoda (`sp:Arthropoda`) 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 congrega todos 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.

Como ilustra a parte inferior da Figura 4.4, as propriedades mapeadas em RDF são categorizadas como sub-propriedades das seis perguntas exploratórias. Por exemplo, as propriedades de caracterização da espécie (`bio:species`, `bio:genus`, `bio:phylum` 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 espécie coletada de um lado é a espécie da coleção de outro.

A Figura 4.6 apresenta uma cópia de tela do protótipo de consulta e visualização dos dados extraídos das planilhas. Esta interface – correspondente à Etapa 3 da Figura 4.5. 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 (<http://lod.geospecies.org/>). Foi desenvolvida uma interface interativa em JavaScript, sobre o framework de mapas OpenLayers (<http://openlayers.org/>), na qual podem ser visualizados interativamente os registros.



Figura 4.6: Cópia de tela da interface de consulta do protótipo desenvolvido.

4.5 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 em questão, 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 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.

Capítulo 5

Conclusões e Extensões

As planilhas eletrônicas obtiveram grande aceitação entre usuários de vários segmentos. Dentre outros fatores, está seu poder de expressão para atender as mais diversas necessidades dos usuários finais associado ao seu formato intuitivo.

Há uma grande demanda pela integração dos dados contidos em diversas planilhas. 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. Entretanto, tais abordagens não consideram dados de contexto, ou seja, não se preocupam em caracterizar a comunidade a qual a planilha pertence, seu domínio de uso e por conseguinte sua natureza. Num processo de interpretação, uma vez identificado esses dados de contexto, pode-se gerar resultados mais condizentes com a realidade do usuário e consequentemente semanticamente mais ricos.

Neste sentido, esta pesquisa contribuiu:

[(i)]na elaboração de um processo de reconhecimento de planilhas baseado na sua natureza; na concepção de uma estratégia para interpretação automática baseada em práticas compartilhadas de autores na criação de planilhas; na implementação de um protótipo de sistema para reconhecimento e explicitação de esquemas em planilhas eletrônicas, bem como o mapeamento para RDF/OWL.

Os resultados obtidos nesta pesquisa abriram novas possibilidades de investigação:

[(i)]a expansão do processo de reconhecimento e identificação da natureza da planilha para outros domínios; a ampliação do modelo para representar padrões de construção, bem como sua transformação em um artefato digital que possa ser compartilhado e reusado; a investigação de novas possibilidades de articulação dos dados de planilhas distintas, a partir da representação semântica produzida; a exploração de algoritmos de similaridade para associação entre elementos da planilha e ontologias; a investigação de técnicas de reconhecimento de esquemas em outras áreas que

possam ser aplicadas no contexto das planilhas.

Referências Bibliográficas

- [1] Robin Abraham and Martin Erwig. Inferring templates from spreadsheets. In *Proceedings of the 28th international conference on Software engineering*, ICSE '06, pages 182–191, New York, NY, USA, 2006. ACM.
- [2] Ivelize Rocha Bernardo, Matheus Silva Mota, and André Santanchè. Extrair e integrando semanticamente dados de múltiplas planilhas eletrônicas a partir do reconhecimento de sua natureza. In *Simpósio Brasileiro de Banco de Dados (SBBDB)*, pages 256–263, 2012.
- [3] Ivelize Rocha Bernardo and André Santanchè. Interpreta automática de planilhas baseada no reconhecimento de padrões de construção. Trabalho em desenvolvimento.
- [4] Ivelize Rocha Bernardo, André Santanchè, and Maria Cecília Calani Baranauskas. Reconhecendo padrões em planilhas no domínio de uso da biologia. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 360–371, 2012.
- [5] Lushan Han, Tim Finin, Cynthia Parr, Joel Sachs, and Anupam Joshi. RDF123: from Spreadsheets to RDF. In *Seventh International Semantic Web Conference*. Springer, October 2008.
- [6] Dietmar Jannach, Kostyantyn Shchekotykhin, and Gerhard Friedrich. Automated ontology instantiation from tabular web sources-the allright system. *Web Semant.*, 7(3):136–153, September 2009.
- [7] Andreas Langegger and Wolfram Wöß. Xlwrap – querying and integrating arbitrary spreadsheets with sparql. In *Proceedings of the 8th International Semantic Web Conference*, ISWC '09, pages 359–374, Berlin, Heidelberg, 2009. Springer-Verlag.
- [8] Frank Manola and Eric Miller. RDF Primer – W3C Recommendation. [w3.org/TR/2004/REC-rdf-primer-20040210](http://www.w3.org/TR/2004/REC-rdf-primer-20040210), 2004.

- [9] Varish Mulwad, Tim Finin, Zareen Syed, and Anupam Joshi. Using linked data to interpret tables. In *Proceedings of the the First International Workshop on Consuming Linked Data*, November 2010.
- [10] Ian Niles and Adam Pease. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9, New York, NY, USA, 2001. ACM.
- [11] D. A. Norman. Human-computer interaction. chapter Some observations on mental models, pages 241–244. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987.
- [12] Martin J. Oconnor and Christian Halaschek-Wiener. Mapping master: a flexible approach for mapping spreadsheets to owl. In *9th International Semantic Web Conference (ISWC2010)*, November 2010.
- [13] Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. Semantics and complexity of sparql. *ACM Trans. Database Syst.*, 34(3):16:1–16:45, September 2009.
- [14] W. F. Ponder, G. A. Carter, P. Flemons, and R. R. Chapman. Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology*, 15(3):648–657, 2001.
- [15] Sandra P. Roth, Peter Schmutz, Stefan L. Pauwels, Javier A. Bargas-Avila, and Klaus Opwis. Mental models for web objects: Where do users expect to find the most frequent objects in online shops, news portals, and company web pages? *Interacting with Computers*, 22(2):140–152, 2010.
- [16] A. Santanchè. Otimizando a anotação de objetos de aprendizagem através da semântica in loco. In *Anais do XVIII Simp. Brasileiro de Informática na Educação*, pages 526–535, 2007.
- [17] A. Santanchè and L. A. M. Silva. Document-centered learning object authoring. In *IEEE Learning Technology Newsletter*, volume 12, pages 58–61, 2010.
- [18] Zareen Syed, Tim Finin, Varish Mulwad, and Anupam Joshi. Exploiting a Web of Semantic Data for Interpreting Tables. In *Proceedings of the Second Web Science Conference*, April 2010.
- [19] Petros Venetis, Alon Halevy, Jayant Madhavan, Marius Paşca, Warren Shen, Fei Wu, Gengxin Miao, and Chung Wu. Recovering semantics of tables on the web. *Proc. VLDB Endow.*, 4(9):528–538, June 2011.

- [20] W3C OWL WG. OWL 2 Web Ontology Language – Document Overview. <http://w3.org/TR/2009/REC-owl2-overview-20091027>, 2009.
- [21] Song Yang, Sourav S. Bhowmick, and Sanjay Madria. Bio2x: a rule-based approach for semi-automatic transformation of semi-structured biological data to xml. *Data Knowl. Eng.*, 52(2):249–271, February 2005.
- [22] Chong-chon Zhao, Li yong Zhao, and Hui ling Wang. A spreadsheet system based on data semantic object. In *Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on*, pages 407 –411, april 2010.