Scan for Marc: sintaxe e semântica de registros bibliográficos na conversão de dados analógicos para o Formato MARC21 Bibliográfico

Article · June 2013

CITATIONS
0

READS
10

1 author:

Zaira Regina Zafalon
Universidade Federal de São Carlos
117 PUBLICATIONS **72** CITATIONS

SEE PROFILE

Artigos

Scan for Marc[1]: sintaxe e semântica de registros bibliográficos na conversão de dados analógicos para o Formato MARC21 Bibliográfico[2]

Zaira Regina Zafalon

Docente na Universidade Federal de São Carlos (UFSCar), Doutora em Ciência da Informação pela Universidade Estadual Paulista (UNESP). Dados para correspondência: R. São Paulo, 260, Vila Monteiro, São Carlos, SP. CEP: 13560-340.
zaira@ufscar.br

Resumo: Estuda-se a conversão de registros bibliográficos e delimita-se o objeto de estudo pelo entendimento da conversão de dados analógicos para o formato MARC21 Bibliográfico, a partir da análise sintática e semântica de registros descritos segundo padrões de estrutura de metadados descritivos e padrões de conteúdo. Busca-se desenvolver um modelo teórico-conceitual de sintaxe e semântica em registros bibliográficos, a partir de estudos lingüísticos saussureanos e hjelmslevianos das manifestações da linguagem humana, que subsidie o desenvolvimento de um interpretador computacional, voltado à conversão de registros bibliográficos ao formato MARC21 Bibliográfico, passível de se confirmar tanto o valor semântico do recurso informacional representado quanto a fidedignidade da representação. O percurso metodológico da pesquisa baseia-se na abordagem qualitativa, de caráter exploratório, descritivo e experimental e recorre à pesquisa bibliográfica. Contribuições no plano teórico são vislumbradas diante do desenvolvimento de questões inerentes aos aspectos sintáticos e semânticos de registros bibliográficos, e por envolver, a um só tempo, a interdisciplinaridade entre a Ciência da Informação, a Ciência da Computação e a Lingüística. Contribuições no campo prático são identificadas pelo fato de o estudo abranger o desenvolvimento do Scan for MARC, um interpretador computacional de conversão de registros bibliográficos impressos para o formato MARC21 Bibliográfico.

Palavras-chave: Conversão de registros bibliográficos; Análise sintática de registros bibliográficos; Análise semântica de registros bibliográficos; Formato MARC21 Bibliográfico; ISBD. AACR2r; Padrão de estrutura de metadados descritivos; Padrão de conteúdo.

INTRODUÇÃO

Esta pesquisa apresenta como tema nuclear o estudo do processo de conversão de registros bibliográficos e delimita-se pelo entendimento da conversão de registros bibliográficos para o Formato MARC21 Bibliográfico[3], a partir de análise sintática e semântica. Diz respeito, portanto, ao estudo da representação de recursos informacionais e do compartilhamento e conversão de registros bibliográficos em ambiente digital, ao desenvolvimento tecnológico e de aspectos teóricos e metodológicos que envolvem tais processos, utilizando-se de ferramentas e métodos inerentes às tecnologias da informação e comunicação. Os recursos tecnológicos e midiáticos, por meio de estruturas computacionais, permeiam a produção, a organização, a distribuição, o acesso, o armazenamento, a preservação, o uso e o reuso dos recursos informacionais por meio de métodos de representação e de recuperação, construindo, consequentemente,

[1] Programa de computador registrado no Instituto Nacional de Propriedade Industrial (INPI) sob o nº 13192-0.
[2] Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (UNESP) como requisito parcial para a obtenção do título de Doutor em Ciência da Informação. Data da defesa: 29 jun. 2012. Banca examinadora: Profa. Dra. Plácida Leopoldina Ventura Amorim da Costa Santos (orientadora), Profa. Dra. Dulce Maria Baptista (UnB), Prof. Dr. Edberto Ferneda (UNESP), Profa. Dra. Elisa Campos Machado (UNIRIO) e Prof. Dr. Ricardo César Gonçalves Santana (UNESP).
[3] MARC é acrônimo de MAchine Readable Cataloging Record. O Formato MARC 21 agrega formatos destinados a dados bibliográficos, dados de autoridade, para holdings, para dados de classificação e para informações de comunidade. Nesta pesquisa serão estudados dados bibliográficos.

novo contexto sócio-cultural, educacional, econômico e tecnológico. Aliado a esses fatores há o crescimento exponencial de acervos documentais em unidades de informação, o que requer a adequação de processos biblioteconômicos que satisfaçam, de modo eficiente e eficaz, à demanda informacional por diferentes públicos.

Recorrer a mecanismos computacionais nas atividades do dia-a-dia das bibliotecas, como substitutos das atividades manuais e mecânicas, já foram indicados por Ranganathan e Gopinath (1967), para quem esse processo exige viabilidade econômica e, em última instância, depende do estágio de desenvolvimento de um país. No âmbito técnico-funcional-administrativo das bibliotecas, o emprego de tecnologias de informação e comunicação tornou-se constante, o que favorece tanto o surgimento quanto a inovação de várias atividades cooperativas.

Com as novas tecnologias de informação e comunicação surgem, também dentre as atividades cooperativas, as facilidades de compartilhamento de registros bibliográficos entre as mais variadas unidades de informação. Identificou-se na literatura, entretanto, que para se estabelecer o processo de conversão de dados bibliográficos para o Formato MARC21 Bibliográfico parte-se do estudo, do conhecimento e do delineamento da estrutura do banco de dados da base legada. Porém, diante da diversidade da estrutura das bases de dados, as ações de conversão são dificultadas.

Diante do exposto, e do cenário de atuação profissional do bibliotecário a partir de aspectos das novas tecnologias da informação e da comunicação, questiona-se: [1] é possível proceder a conversão de registros bibliográficos para o formato MARC21 Bibliográfico por meio de uma mesma metodologia que seja aplicável a diferentes bases?; [2] é possível estabelecer-se um conteúdo sintático e semântico do registro bibliográfico que possa nortear o processo de conversão ao Formato MARC21 Bibliográfico? [3] é possível aplicar-se um modelo teórico-conceitual de sintaxe e semântica de registros bibliográficos à uma ferramenta computacional que permita estabelecer a conversão para o Formato MARC21 Bibliográfico? Nesse sentido, estabelecem-se como premissas os fatos: a) não há padronização única adotada para a definição da estrutura de bases de dados em sistemas gerenciadores de bibliotecas e outras unidades de informação; b) no domínio bibliográfico, os registros bibliográficos são elaborados a partir de convenções advindas de comunidades de prática; e c) há esquemas tradicionais e internacionais de descrição e de visualização de registros bibliográficos, identificados nos padrões de estrutura de metadados descritivos e nos padrões de conteúdo.

Propõe-se, assim, o objetivo de desenvolver um modelo teórico-conceitual de sintaxe e semântica em registros bibliográficos, a partir de estudos lingüísticos saussureanos e hjelmslevianos das manifestações da linguagem humana, que subsidie o desenvolvimento de um interpretador computacional[4], voltado à conversão de registros bibliográficos ao formato MARC21 Bibliográfico, passível de se confirmar tanto o valor semântico do recurso informacional representado quanto a fidedignidade da representação.

[4] O conceito de interpretador computacional nesta pesquisa é adotado para refletir o processo, mediado por meios computacionais, que interpreta um registro bibliográfico analógico por meio de uma estrutura definida pelas marcações sintáticas, identificadas pelos sinais de pontuação presentes nos registros bibliográficos, de tal forma que permite inferir o valor semântico do recurso informacional representado de modo que seja garantida a fidedignidade da representação. Tendo em vista a busca pela garantia da interpretação dos dados e a simplificação do processamento optou-se pelo uso da linguagem Perl que, segundo Stockton (]2005]), é uma linguagem interpretada otimizada para examinar arquivos de textos e extrair informações de arquivos de textos. Considerou-se, também, o fato de estar sob os termos da Licença Pública Geral (GNU), caracterizando-se como sendo software livre (PERL, [2011]), e por permitir aliar programas desenvolvidos em diversificados ambientes, tais como UNIX, MSDOS, Windows, Macintosh, OS/2. Para o desenvolvimento dos testes recorreu-se ao ambiente Strawberry Perl.

Tendo em vista os objetivos definidos o percurso metodológico nesta pesquisa baseia-se na abordagem qualitativa, na qual é possível assumir uma relação dinâmica com o mundo real, por meio da interpretação e atribuição de significados aos fenômenos estudados, segundo Gonsalves (2011). De acordo com os objetivos, a pesquisa expõe caráter exploratório, uma vez que "se caracteriza pelo desenvolvimento e esclarecimento de ideias, com objetivo de oferecer uma visão panorâmica, uma primeira aproximação a um determinado fenômeno que é pouco explorado" e, por oferecer, por meio da pesquisa bibliográfica, "dados elementares que dão suporte para a realização de estudos mais aprofundados sobre o tema" (GONSALVES, 2011, p. 67). Por considerar, ainda, os objetivos, a pesquisa encarrega-se do caráter descritivo, por "descrever as características de um objeto de estudo." (GONSALVES, 2011, p. 68). O caráter experimental também está presente por referir-se a "um fenômeno que é reproduzido de forma controlada, submetendo os fatos à experimentação (verificação), buscando, a partir daí, evidenciar as relações entre os fatos e as teorias." (GONSALVES, 2011, p. 69).

Tal estudo, a nosso ver, deflagra contribuições tanto no plano teórico, por vislumbrar o desenvolvimento de questões sobre aspectos sintáticos e semânticos de registros bibliográficos, e por envolver, a um só tempo, a interdisciplinaridade entre a Ciência da Informação, a Ciência da Computação e a Lingüística, de modo a reafirmar a proposta de Borko (1968) e de Saracevic (1996); quanto no prático, por abranger o desenvolvimento de interpretador computacional que possa ser adotado por quaisquer instituições que queiram utilizar-se de procedimento de conversão de bases de dados de registros bibliográficos, para o formato MARC21 Bibliográfico a partir dos esquemas de descrição (AACR2) e de visualização de registros bibliográficos (ISBD), caráter que se considera inovador na pesquisa.

1. PADRÕES DE REPRESENTAÇÃO DE RECURSOS INFORMACIONAIS E CONVERSÃO DE REGISTROS BIBLIOGRÁFICOS

A relação intrínseca entre representação e recuperação de documentos exige que se leve em conta as ferramentas de descrição e, também, de estrutura de registros bibliográficos, o que promove consistência, precisão e relevância aos resultados obtidos em resposta a uma consulta. Em se tratando de estrutura e de descrição de conteúdo recorre-se a Foulonneau e Riley (2008) que apresentam os padrões de estrutura de metadados descritivos e os padrões de conteúdo.

Dentre os padrões de estrutura de metadados descritivos, que listam elementos considerados importantes para a descrição do recurso, incluindo características físicas e de conteúdo, destaca-se o Formato MARC Bibliográfico. Quanto aos padrões de conteúdo, que, por sua vez, fornecem regras para a sintaxe de uma entrada em um campo de metadado e tem o propósito de promover consistência em registros metadados para permitir uma melhor pesquisa e recuperação pelos usuários, encontram-se as AACR2.

Em bibliotecas é familiar o padrão de conteúdo e sua relação com o padrão de estrutura de metadados, como as AACR2 e sua relação com o MARC, ambas estudadas nesta pesquisa. As AACR2r, revisão de 2002 das Anglo-American Cataloging Rules, 2nd edition, apresentam, por meio de orientações, regras e exemplos, a descrição do conteúdo e, também, a escolha, a elaboração e a atribuição de pontos de acesso a um documento, o que permite que sejam criados os direcionamentos para a construção de catálogos bibliográficos. Pelo fato de as AACR2r apresentarem, inclusive historicamente, relação direta com as

ISBDs, pode-se dizer, de certo modo, que, para os catálogos manuais, elas assumem tanto aspectos do padrão de estrutura de metadados descritivos quanto do padrão de conteúdo.

Compreende-se a ISBD como padrão de estrutura de metadados descritivos a partir de Swanson (1973) e Langker (1974) para quem a ISBD especifica os elementos de uma descrição bibliográfica, prescreve a ordem em que devem ser apresentados, mas, principalmente, porque indica os sinais de pontuação pelo qual os elementos devem ser demarcados.[5] Desse modo, a ISBD visa três objetivos: fazer registros de diferentes fontes intercambiáveis, facilitar a sua interpretação acima das barreiras lingüísticas e facilitar a conversão de tais registros à forma legível por máquina.

Para a descrição e a recuperação de registros bibliográficos em meio automatizado, é requerida, aliada às AACR2r, a adoção de um padrão de estrutura de metadados descritivos, e, para esta pesquisa, interessa estudar o Formato MARC21 Bibliográfico, que abrange aspectos de leitura e de interpretação de dados disponíveis em registros bibliográficos por meio computacional. Aspectos inerentes à estrutura de um registro MARC podem ser notados pela flexibilidade da estrutura do arquivo e pelo número e tamanho ilimitados dos campos. O processo de leitura e interpretação computacional de um registro bibliográfico em Formato MARC é facilitado pelas marcações inerentes a ele.

Entende-se, por fim, que as convenções adotadas nas marcações de um registro bibliográfico, quer seja pela ISBD, quer seja pelo Formato MARC21 Bibliográfico, promovem, aliadas às regras de descrição de conteúdo, dadas pelas AACR2r, cada uma a seu modo, o desenvolvimento de catálogos e o atendimento aos objetivos bibliográficos.

Considerando-se que o tratamento técnico e o registro de informações bibliográficas, sem dúvida, são as atividades nas quais mais incidem os custos de um processo de automação, torna-se fundamental garantir que os dados em meio digital do acervo sejam (re)utilizados. Para tanto, é requerida a garantia da base tecnológica e metodológica fornecida pela adoção de padrões, que, por sua natureza, promovem compatibilidade e intercâmbio de registros bibliográficos. Em se tratando de intercâmbio de dados bibliográficos, uma das principais atividades envolve a conversão de dados bibliográficos. Porém, faz-se necessário esclarecer que os termos conversão e migração de dados bibliográficos, usados, algumas vezes, como sinônimos, são diferentes. No enfoque desta pesquisa, a expressão conversão é adotada para designar o processo de mudança do suporte no qual um registro bibliográfico está inscrito ou, ainda, o processo em que a alteração se dá no nível da estrutura do registro, o que não envolve a alteração da descrição de seu conteúdo. A conversão de dados bibliográficos é assumida, portanto, como meio de se proceder a mudança de padrão de estrutura de metadados descritivos dos recursos informacionais.

Tendo em vista a configuração da proposta teórica acerca dos registros bibliográficos, o próximo tópico apresenta a contribuição de Saussure e de Hjelmslev à representação de recursos informacionais.

---

[5] Considerar a ISBD como um padrão de estrutura de metadados descritivos por conta da pontuação assume caráter fundamental para o desenvolvimento desta pesquisa, uma vez que, segundo Trask (2008, p. 232), a pontuação é “um sistema convencional de marcas que representam informações sobre a estrutura de um texto escrito.” Por sua vez, Langker (1974) ressalta que a pontuação é usada para fins estruturais para delimitar os campos e subcampos (de modo a auxiliar um operador de máquina a registrar em formato legível por máquina). Entende-se que a pontuação prescrita nas ISBDs cumpra a dupla finalidade de proporcionar meios para precisar elementos bibliográficos, independentemente da linguagem, tanto para seres humanos quanto para máquinas.

2 A SINTAXE E A SEMÂNTICA DE REGISTROS BIBLIOGRÁFICOS A PARTIR DE SAUSSURE E HJELMSLEV

Recorre-se à contribuição lingüística da comunicação humana, feitos por Saussure, e à concepção semântica estruturalista, com Hjelmslev, para a elaboração do arcabouço teórico da sintaxe e da semântica dos registros bibliográficos.

Entende-se que, da mesma forma que a lingüística se forma pelas manifestações da linguagem humana (SAUSSURE, 2010, p. 13), o papel social das instituições de patrimônio cultural forma-se a partir da representação de tais manifestações, levando-se em conta aquelas registradas, independentemente do meio e do suporte em que isso seja feito. As manifestações da linguagem humana, por meio de seus registros, permitem a descrição, a identificação, o acesso, o uso, o reuso, a disseminação e o compartilhamento entre as mais diversificadas instituições de patrimônio cultural. O fenômeno da representação dos recursos informacionais integra duas faces que se correspondem e se complementam: a obra e a manifestação, ambas reconhecidas pela International Federation of Library Associations and Institutions (1998, 2005, 2009) como produtos do esforço intelectual ou artístico.

Obra é a criação intelectual ou artística que reflete o conteúdo e é identificada como uma entidade abstrata. Para Smiraglia (2002), obra é o conhecimento deliberadamente criado para representar um conjunto coordenado de idéias (ou seja, o conteúdo ideacional), que veiculado por meio do texto tem a finalidade de ser comunicado ao consumidor. Um documento pode conter uma ou mais obras, e uma obra pode existir em um ou mais documentos, o que significa que é possível que exista em várias instâncias.

Manifestação é a corporificação de uma obra, que só é possível de ser conhecida se manifestada, ou, de outro modo, a manifestação só existe a partir da concepção de uma obra; a obra só pode ser reconhecida por meio da manifestação. A manifestação assume a forma física. Assim, a representação da informação só é possível de ser realizada por meio da compreensão da correspondência entre obra e manifestação. Não se pode, portanto, reduzir a representação a uma ou a outra face: obra é resultado do pensamento, sem que, entretanto, tenha sido manifestada, vindo a público. Surge, por conseguinte, daí, a correspondência entre a obra, unidade complexa mental, e a manifestação, unidade complexa física. Entenda-se, portanto, que a manifestação é o “traje” que uma obra ocupa.

Compreende-se, portanto, que os recursos informacionais constituem-se de manifestações socializáveis de obras, estas, por sua vez, individuais ou coletivas. Partindo-se desta perspectiva, pode-se avistar um dilema sobre o que de fato se representa: ou envolve-se com a explicação de que é a obra, por ser concepção mental primeira; ou arrisca-se a perceber que é a manifestação, registro em suporte físico da concepção mental. Parte-se do ponto de vista de se observar a representação a partir da manifestação. Representação da informação é, portanto, o ato de articular formas de descrição a partir de instrumentos que permitam tornar cognoscível um recurso informacional sem que seja necessário recorrer ao documento original para identificá-lo. O enfoque entre obra e manifestação e sua relação com o processo comunicativo norteiam o estudo de uma possibilidade teórica advinda do campo da linguagem para se pensar os registros da informação, como o sugerido por Ferdinand de Saussure, em 1916, relacionando-o com a questão do significado e do significante. Para tanto, estuda-se a correspondência entre obra e manifestação, significado, significante.

Do mesmo modo que Saussure (2010, p. 81 et seq.) apresenta princípios como a arbitrariedade do signo e o caráter linear do significante, buscar-se-á explicitar a correlação entre obra e manifestação presentes nos princípios de Saussure. Quanto ao primeiro princípio, "O laço que une o significante ao significado é arbitrário" (SAUSSURE, 2010, p. 81), observa-se a arbitrariedade da manifestação em relação à obra. Neste sentido, a idéia de uma obra, como 100 cientistas que mudaram a história do mundo, de John Hudson Tiner, não apresenta relação direta com, tão somente, uma forma de manifestação; essa ligação é arbitrária e pode assumir outras tantas formas: um roteiro de uma peça teatral ou de um cinema; um musical; um livro; entre outras. Na prática, a forma de manifestação selecionada pode ser qualquer uma, desde que evoque a obra por meio da corporificação. Diante do segundo princípio, "O significante [...] desenvolve-se no tempo [...]" (SAUSSURE, 2010, p. 84), de tão notável relevância quanto o primeiro, entende-se que a linearidade que um registro assume requer uma sequência para o registro da obra na manifestação do produto mental (conteúdo ideacional), independentemente de sua forma.

A obra faz menção ao conceito mental, ou, para remeter a Saussure, ao significado, ao conceito; a manifestação, por sua vez, remete ao significante, à imagem acústica registrada. A obra, reduzida a um princípio essencial para a manifestação, apresenta correspondência entre tantas formas de expressão quantas forem possíveis.

O catalogador, assim sendo, precisa conhecer as manifestações para, delas, recorrer aos mecanismos de representação, o que possibilitará dar a conhecer aos outros os recursos informacionais. Dessa maneira, embora a manifestação seja, por si só, estranha à obra, é impossível abstraí-la das várias manifestações, processo pela qual a obra é constantemente passível de ser representada. Obra e manifestação, dois sistemas distintos, são, porém, complementares para a formação do objeto documental. A obra pode ter uma tradição oral, fixada diversamente da tradição escrita em um suporte, e, mesmo assim, ser transferida para outras gerações. Embora essas de tradição oral possam ser objeto das instituições de patrimônio cultural, conquanto não estiverem registradas, não são passíveis de representação.

Na área de lingüística, há estudos voltados à semântica estruturalista que, por sua vez, abordam a semântica de forma concreta e analisam a semântica lexical a partir da idéia central de que a linguagem deve ser vista como um sistema (GEERAERTS, 2010; TAMBA-MECZ, 2006). Assim, a linguagem natural pode ser entendida como um sistema simbólico, com propriedades e princípios próprios que determinam como um signo lingüístico funciona.

A contribuição hjelmsleviana advém da visão de estrutura na lingüística estruturalista. Recorre-se, também, a Hjelmslev (1991, p. 116), para compreender melhor a relação entre objeto, estrutura e descrição científica, quando o autor afirma, diante da lingüística, que "Não há nem conhecimento nem descrição científica possível de um objeto qualquer sem recurso a um princípio estrutural." Diante de tal afirmativa observa-se a adoção de princípios formais em uma relação parte-todo o que supõe uma afinidade intrínseca à representação documental na Ciência da Informação.

Denota-se um paralelo entre a semântica estruturalista e a Ciência da Informação, tendo em vista que ambas fundamentam-se em constructos simbólicos e com visão sistêmica, senão, veja-se: tal qual a linguagem, a representação também recorre a um sistema, no qual é possível estudar, analisar e representar unidades informacionais, simbólicas, presentes nas mais diferentes manifestações de obras. Nesse sentido, a representação documental, baseada em convenções, normas e padrões, iguala-se à língua. A representação permite a análise sincrônica do documento, com traços inerentes ao recurso

informacional e seu suporte, no qual estão presentes as relações entre obra e manifestação. A sincronia, na representação, significa a impossibilidade de dissociação entre representante e representado; entre obra e manifestação. Nesse sentido, o estudo sincrônico da representação propõe o estudo dos registros bibliográficos a partir de suas relações, tanto no próprio registro quanto em relação ao objeto descrito. A semântica estruturalista de registros bibliográficos volta-se, assim, ao estudo descritivo do funcionamento dos catálogos.

No estudo estruturalista dos registros bibliográficos percebe-se a intenção de identificar a estrutura do registro, sua relação com outros registros e a relação com o documento. Desse modo, a sintaxe do registro bibliográfico não dá conta do catálogo; é a semântica que permite o contexto e as sinapses entre os vários registros bibliográficos; é a semântica que dá conta dos processos mentais segundo os quais se produz, constitui, compreende e descreve a representação de um recurso informacional. Nesta pesquisa, denomina-se papel semântico o estudo das diferentes relações que podem ser estabelecidas entre os registros bibliográficos, entre o registro bibliográfico e o recurso informacional, e entre os elementos do próprio registro bibliográfico.

Nesse sentido, considera-se a possibilidade de se estudar a representação documental a partir de uma teoria dos níveis: do nível subsemântico (entre os elementos de um registro bibliográfico) ao nível supersemântico (a relação entre os vários registros bibliográficos, a partir de suas semelhanças e diferenças), passando pelo nível semântico (objeto que está sendo descrito e a descrição em si). Assim, o objetivo da ênfase estruturalista para a análise semântica dos registros bibliográficos pode ser definido como: estudar a descrição de registros bibliográficos efetivamente realizados, no qual se considera a influência do catálogo, como meio para o estabelecimento de mensagens presentes nos recursos informacionais e nas necessidades informacionais dos usuários. É possível admitir três planos de diferença semântica em registros bibliográficos: entre o referente e a representação, entre o todo e as partes da representação e, entre as representações presentes no catálogo. São estes aspectos semânticos que reduzem a alteridade de um registro bibliográfico que, em um catálogo, faz com que o disperso e o aparente sejam marcados por identidades que lhe são próprias. A semântica na Ciência da Informação é dada pela forma da e na representação da informação.

A semântica, pensada por este prisma, refere-se à estrutura de um sistema que relaciona significado e significante, obra e manifestação. Na semântica dos registros bibliográficos o significado é dado pelo valor do significante, ou, a manifestação é o valor da obra no processo de representação. Esses valores semânticos em um registro bibliográfico formam uma rede de relações estruturais com outros registros bibliográficos, a qual se denomina supersemântica do registro bibliográfico. A semântica requer a adoção de sintaxe para a definição dos valores semânticos; em outras palavras, a sintaxe está presente no padrão de estrutura de metadados descritivos e a semântica nos padrões de conteúdo.

A sintaxe, no âmbito desta pesquisa, diz respeito à ordem dos elementos dispostos para a representação dos recursos informacionais. Entende-se, portanto, que a sintaxe do registro bibliográfico seja parte da Ciência da Informação voltada ao estudo da forma, arranjo e disposição em que cada elemento deva ser descrito quando da representação do recurso informacional. Neste sentido, faz parte do sistema biblioteconômico que determina relações formais entre a representação de cada uma das partes do

documento representado. Esses elementos são organizados segundo padrões de estrutura de metadados estabelecidos. Os aspectos sintáticos de um registro bibliográfico podem remeter à estrutura semântica.

A linguagem bibliográfica ultrapassa o nível sintático e faz com que se entenda que um registro apresente níveis semânticos, necessário ao entendimento do documento representado sintática e semanticamente. Desse modo, cada elemento sintático, assume um conteúdo semântico diante de cada elemento definidor da representação e este elemento, por sua vez, quando contextualizado e contraposto ao documento representado, assume significado entre o registro e o objeto.

O registro bibliográfico cinge, então, tanto questões sintáticas, por referir-se aos padrões de estruturas dos metadados de cada elemento do documento ou objeto a ser descrito, quanto questões semânticas, por permitir analisar a coesão e o significado indicado entre elementos do representante e do representado e entre o próprio representante e representado. Cada elemento sintático, quando contextualizado e contraposto ao documento representado, assume um significado concreto entre o registro e o objeto.

Como se pode entender a conversão, a partir da análise sintática e semântica dos registros bibliográficos, a ponto de ser possível conduzir processos de conversão de registros bibliográficos ao Formato MARC21 Bibliográfico? Entende-se que ocorra pela utilização das marcações dadas pelo padrão de estrutura de metadados descritivos, presentes nas AACR2r e nas ISBDs, e, consequentemente, pela inferência semântica, prevista em uma aplicação computacional. Neste momento, vislumbra-se aplicar aspectos teórico-conceituais dos princípios sintáticos e semânticos de registros bibliográficos à conversão de registros bibliográficos ao Formato MARC21 Bibliográfico em um interpretador computacional.

3 SCAN FOR MARC: INTERPRETAÇÃO SINTÁTICA E SEMÂNTICA DE REGISTROS BIBLIOGRÁFICOS IMPRESSOS

A partir do pressuposto que considera os esquemas sintáticos e semânticos dos registros bibliográficos, e não a estrutura da base de dados legada, necessários para a conversão de registros bibliográficos ao Formato MARC21 Bibliográfico, discute-se o interpretador computacional sintático e semântico de registros bibliográficos, identificado como Scan for MARC, por aludir ao método de escaneamento dos registros bibliográficos e sua conseqüente conversão ao Formato MARC21 Bibliográfico.

O desenvolvimento do interpretador, em sua versão beta, envolveu, sumariamente, ações de captura de imagens de registros bibliográficos analógicos, apresentados em fichas catalográficas em meios eletrônicos; análise de resultados de processamentos de imagem em software de reconhecimento de caracteres, o que permite a conversão de imagem para texto editável; processamento das imagens de registros bibliográficos selecionados para teste (processo inicial da construção da representação); tratamento sintático e semântico dos registros bibliográficos digitais, checagem dos resultados e avaliação de ajustes.

A fase de testes para a captura de imagens dos registros bibliográficos dividiu-se em três etapas: em scanner de impressora funcional, em câmera de dispositivo móvel e em câmera fotográfica digital. Finda esta fase procederam-se testes para o processamento de imagens com reconhecimento de caracteres de registros bibliográficos em formato analógico, para o qual se adotou um software de reconhecimento de caracteres (OCR). Nesta fase de testes, dividida em dois momentos (o de análise de softwares freeware ou free software, e o de softwares proprietários, com análise em versões trial), analisou-se os seguintes

softwares: ABBYY FineReader 11, Cognitive Open OCR (Cuneiform) 0.1, FreeOCR, FreeOCR 3.1, Leadtools, OnlineOCR.net, ScreenOCR 9.1, Sci2ools (i2OCR), SimpleOCR 3.5, TopOCR 3.1, WeOCR Server.[6]

Os testes foram desenvolvidos a partir de cotejo e análise quanto à pontuação, aos sinais diacríticos, à troca de letras, ao espaçamento entre informações e margens. Dentre os aplicativos identificados, o software OnlineOCR.net ofereceu melhores resultados. Seguiu-se nova fase de testes na qual se buscava definir o método de processamento de imagens que oferecesse melhor resultado (diante das diversas formas de captura de imagens).

Como resultado geral apresentam-se os seguintes comentários: questões referentes à iluminação na captura das imagens são extremamente relevantes para o processamento em software de OCR dada a influência na qualidade da imagem; os melhores índices de sucesso no processamento de imagens de registros bibliográficos foram capturados diante das imagens capturadas em scanner de impressora multifuncional; problemas com a troca de letras no processo de reconhecimento de caracteres em imagens foram identificados nos testes dos três diferentes recursos; aspectos quanto à fixação da câmera, quer seja de dispositivo móvel quer seja de fotografia digital, mostraram-se mais relevantes do que a distância a ser considerada na captura da imagem; a qualidade do OCR está intrinsecamente vinculada à qualidade da imagem e não ao método pelo qual a imagem é capturada. Diante dos resultados, optou-se por trabalhar com as imagens capturadas em scanner de impressora multifuncional e com o OnlineOCR.net.

Efetuados os testes de processamento de imagens dos registros bibliográficos analógicos selecionados, apresentam-se os procedimentos adotados para o tratamento sintático e semântico dos dados bibliográficos do arquivo, à luz da proposta teórica apresentada nesta pesquisa.

Estudou-se os padrões de estrutura de metadados descritivos (ISBDs, AACR2r e Formato MARC21 Bibliográfico) e definiu-se a interferência dos sinais de pontuação no conteúdo semântico, presente nos padrões de conteúdo (AACR2r). Para a efetiva adoção destas marcações no script de conversão de registros bibliográficos ao MARC21 Bibliográfico foi necessário, entretanto, cuidado quanto à pontuação que faz parte do conteúdo e não da estrutura de metadados descritivos. Outra análise, em busca da definição de padrões, foi requerida quanto às informações dos pontos de acesso para o que se determinou que os pontos de acesso definidos por cabeçalhos de assunto iniciam-se por numerais indo-arábicos seguidos de ponto, e de que os outros pontos de acesso que se fizerem necessários iniciam-se por numerais romanos.

Sem dúvida alguma, um dos primeiros problemas identificados para o processamento das informações foi decorrente do tipo de codificação dos caracteres do arquivo texto de entrada (UTF-8/ISO, UFT-16/UNICODE, ASCII/ANSI), necessários para a correta interpretação dos sinais diacríticos. A fase de tratamento sintático e semântico dos registros bibliográficos digitais foi feita a partir de testes de processamento de roteiros de registros bibliográficos. Foram estabelecidos quatro roteiros, com diferentes

[6] Endereço para acesso: ABBYY FineReader 11 (http://www.abbyy.com.br/finereader/); Cognitive Open OCR (Cuneiform) 0.1 (http://cognitive-openocr-cuneiform.en.softonic.com/download); FreeOCR (http://www.free-ocr.com/); FreeOCR 3.1 (http://www.paperfile.net/freeocr.exe); Leadtools (http://www.leadtools.com/sdk/ocr/default.htm); OnlineOCR.net (http://www.onlineocr.net/default.aspx); ScreenOCR 9.1 (http://www.screenocr.com); Sci2ools (i2OCR) (http://www.sciweavers.org/free-online-ocr); SimpleOCR 3.5 (http://www.charactell.com/scanstore/); TopOCR 3.1 (http://www.brothersoft.com/topocr-download-47055-s1.html); WeOCR Server (http://ocr1.sc.isc.tohoku.ac.jp/e1/).

graus de complexidade, que subsidiassem as versões de scripts (desenvolveram-se quatro versões, cada qual com pequenos ajustes necessários após a checagem dos resultados).

Neste trabalho apresentam-se alguns resultados que não mostram, entretanto, todos os casos previstos e já em funcionamento no Scan for MARC (cf. Figura 1). Em todos os casos são indicados, alinhados à esquerda, o resultado do processamento de imagem em OCR, e, à direita, o resultado do Scan for MARC.

Figura 1 – Resultado do processamento de imagem em OCR e do tratamento sintático e semântico pelo Scan for MARC

```
830.1   Dubell, Richard
D111b      A bíblia do diabo : romance histórico / Richard
        Dubell ;  tradução Claudia Abeling. - São Paulo :
        Planeta do Brasil, 2011.
           512 p.

           Título do original: Die Terfelsbibel.

           1. ROMANCE HISTÓRICO 2. LITERATURA ALEMÃ I. Abeling,
        Claudia. II. Título.
```

```
090  $a830.1 $bD111b
1001 $aDubell, Richard
24512$aA bíblia do diabo $bromance histórico $cRichard
      Dubell ; tradução Claudia Abeling
260  $aSão Paulo $bPlaneta do Brasil $c2011
300  $a512 p.
500  $aTítulo do original: Die Terfelsbibel
65014$aROMANCE HISTÓRICO
65024$aLITERATURA ALEMÃ
7001 $aAbeling, Claudia
```

```
021.3   Leitura e escrita de adolescentes na internet e na
L21        escola / Organização Maria Teresa de Assunção
2.ed.      Freitas, Sérgio Roberto Costa. - 2. ed. - Belo
           Horizonte : Autêntica, 2006.
           138 p. - (Coleção leitura, escrita e oralidade)

           ISBN 85-7526-156-8.

           1. TECNOLOGIA DA INFORMAÇÃO 2. LEITURA I. Freitas,
        Maria Teresa de Assunção, org. II. Costa, Sérgio Ro-
        berto, org. III. Série.
```

```
020  $a8575261568
090  $a021.3 $bL21 $c2.ed.
24500$aLeitura e escrita de adolescentes na internet e na
      escola $cOrganização Maria Teresa de Assunção Freitas,
      Sérgio Roberto Costa
250  $a2. ed.
260  $aBelo Horizonte $bAutêntica $c2006
300  $a138 p.
4900 $aColeção leitura, escrita e oralidade
65014$aTECNOLOGIA DA INFORMAÇÃO
65024$aLEITURA
7001 $aFreitas, Maria Teresa de Assunção $eorg.
7001 $aCosta, Sérgio Roberto $eorg.
```

Fonte: Elaborado pelos autores.

Observe-se que nos resultados a codificação em Formato MARC21 Bibliográfico manteve-se correlacionada com aquela feita por catalogadores. Diante da análise de resultado entende-se que a diante da proposta de interpretação sintática e semântica de registros bibliográficos, o resultado tenha sido de qualidade.

Ao final das análises consagradas aos testes de processamento de imagens dos registros bibliográficos analógicos, para o qual se recorreu aos aspectos sintáticos e semânticos dos registros bibliográficos, o tópico seguinte destina-se às considerações finais da pesquisa.

CONSIDERAÇÕES FINAIS

Diante da temática definida para esta pesquisa, a conversão de registros bibliográficos para o Formato MARC21 Bibliográfico, desenvolveu-se a teoria sintática e semântica de registros bibliográficos, definida por padrões de estrutura de metadados descritivos e por padrões de conteúdo, consubstanciados a partir das ISBDs e das AACR2r.

Apresentou-se abordagem teórico-conceitual acerca da representação de recursos informacionais e do compartilhamento e conversão de registros bibliográficos analógicos em ambiente digital; o desenvolvimento tecnológico alcançado diante da proposta de garantir a fidedignidade de aspectos da representação do conhecimento; e, a análise e o desenvolvimento de aspectos teóricos e metodológicos que subsidiem atividades de conversão de dados, recorrendo-se a métodos inerentes às tecnologias de informação e comunicação.

Reafirmou-se a importância da adoção de normas, regras, padrões, formatos, metodologias e critérios para a representação de recursos informacionais em unidades de informação tendo em vista a aplicação de processos, permeados por aplicações tecnológicas e midiáticas que recorrem a estruturas computacionais que avalizem produção, organização, armazenamento, gerenciamento, tratamento, preservação, distribuição, provisão, recuperação, acesso, uso, reuso e compartilhamento de registros informacionais em diversos suportes.

Estabeleceu-se, a partir de Saussure, o modelo de representação dos recursos informacionais calcado na relação entre significado e significante, no qual se discute a arbitrariedade da manifestação em relação à obra, bem como o desenvolvimento da linearidade da manifestação em relação ao conteúdo ideacional da obra, fator determinante para a compreensão do documento e necessária à elaboração do registro bibliográfico. Com Hjelmslev discutiu-se, a partir dos princípios formais adotados na Linguística para o estudo da estrutura do sistema lingüístico, a representação documental na Ciência da Informação, haja vista que esta se fundamenta nos constructos teóricos e sistêmicos da análise sincrônica do documento, com traços inerentes da indissociação entre obra e manifestação, para o qual se propõe o estudo dos registros bibliográficos a partir das relações internas entre os elementos de um registro (subsemântica), entre os registros de um catálogo (supersemântica), e em relação ao documento descrito (semântica).

Em decorrência das questões apresentadas procurou-se formar o arcabouço teórico-metodológico da representação bibliográfica e seus aspectos sintáticos e semânticos dos objetos representados, com reflexo nos seguintes aspectos: [1] compreensão da obra, enquanto significado, e da manifestação, enquanto significante; [2] entendimento da representação bibliográfica como resultado da relação entre significante e significado e entre obra e manifestação e como definitiva para a semântica; [3] percepção da sintaxe para a definição da subsemântica, além de necessária à representação do recurso informacional; [4] apreensão do conceito de supersemântica, a partir de sua co-dependência com a subsemântica e a semântica, na relação, identificável nos catálogos, entre registros bibliográficos e documentos de um acervo e entre obra e manifestação, tomados como significante e significado.

Tendo em vista os resultados obtidos nos testes, embora considerados iniciais por requererem, ainda, ajustes e aprimoramentos, realizados em prol da definição do comportamento do Scan for MARC, entende-se que a contribuição quanto aos aspectos sociais ensejados apresentem repercussão teórico-prática

significativa na área da Ciência da Informação, bem como em sua interdisciplinaridade com a Ciência da Computação e a Lingüística.

Para estudos futuros, em se tratando do Scan for MARC retomem-se as necessidades identificadas para melhorias e aperfeiçoamento do interpretador computacional: [1] conjugação integrada das fases de reprodução e de representação, recorrendo-se à adoção de um OCR em linha de comando; [2] desenvolvimento de interface gráfica; [3] previsão de padrões para a interpretação de dados de classificação de assunto (CDU); [4] tratamento do conteúdo de termos qualificadores (subdivisão de forma, subdivisão cronológica, subdivisão geográfica e subdivisão geral) nos pontos de acesso de assunto tópico; [5] previsão de padrões para tratamento de pontos de acessos secundários de assuntos para nome pessoal, institucional, evento e título uniforme; [6] previsão de padrões para tratamento de pontos de acessos principais para instituições, eventos e título uniforme; [7] previsão para criação de dicionários semânticos para pontos de acesso e seus termos qualificadores; [8] implementação de processo de teste de consistência do arquivo convertido, uma vez que isso afere caráter qualitativo ao registro bibliográfico gerado; [9] integração do script com o OCR, o que exige a adoção de um que seja em linha de comando.

Enseja-se, ainda, dar continuidade aos estudos dos métodos sintáticos e semânticos de registros bibliográficos e pesquisar a validade deste método de conversão de dados bibliográficos analógicos quando aplicados à interpretação dos dados da catalogação na fonte do tipo documental livro.

REFERÊNCIAS

BORKO, H. Information science: what is it? American Documentation, v. 19, n. 1, p. 3-5, jan. 1968.

FOULONNEAU, M.; RILEY, J. Choosing metadata standards for a digital library project. In: _____. Metadata for digital resources: implementation, systems design and interoperability. Oxford: Chandos, 2008. p.13-28.

GEERAERTS, D. Theories of lexical semantics. New York: Oxford University Press, 2010.

GONSALVES, E. P. Conversas sobre iniciação à pesquisa científica. 5. ed. rev. e ampl. Campinas: Alínea, 2011.

HJELMSLEV, L. Ensaios lingüísticos. São Paulo: Perspectivas, 1991.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. Declaração de princípios internacionais de catalogação. 2009. Disponível em: http://www.ifla.org/files/cataloguing/icp/icp_2009-pt.pdf. Acesso em: 20 jun. 2011.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. Functional requirements for bibliographic records: final report. 1998. Disponível em: http://archive.ifla.org/VII/s13/frbr/frbr3.htm#6. Acesso em: 20 jun. 2011.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. Guidelines for Online Public Access Catalogue (OPAC) displays: final report: may 2005. München: K. G. Saur, 2005.

LANGKER, R. ISBD: another step in the right direction. The Australian Library Journal, v. 23, n. 3, p. 99-103, April, 1974.

PERL Programming Documentation. [2011]. Disponível em: http://perldoc.perl.org/perl.html. Acesso em: 24 maio 2012.

RANGANATHAN, S. R.; GOPINATH, M. A. Prolegomena to library classification. 3rd ed. New York: Asia Publishing, 1967.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. Perspectivas em Ciência da Informação, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

SAUSSURE, F. Curso de lingüística geral. São Paulo: Cultrix, 2010.

SMIRAGLIA, R. P. Further Reflections on the Nature of 'A Work': An Introduction. Cataloging & Classification Quarterly, v. 33, n. ¾, p. 1-11, 2002.

STOCKTON, R. PERL: practical extraction and report language. [2005]. Disponível em: http://www.stacken.kth.se/help/perl/. Acesso em: 24 maio 2012.

SWANSON, G. ISBD: standard or secret? Library Journal, n. 15, p. 124-130, Jan. 1973.

TAMBA-MECZ, I. A semântica. São Paulo: Parábola, 2006.

TRASK, R. L. Dicionário de linguagem e lingüística. São Paulo: Contexto, 2008.