VII Encuentro de Catalogación y Metadatos 2012
Memoria

Compilador
Filiberto Felipe Martínez Arellano

Tabla de Contenido

Conferencias

Ponencias

Mesa Redonda

Scan for MARC: Syntax and Semantics of Bibliographical Records in the Conversion of Analogic Data to the MARC21 Bibliographic Format

Zaira Regina Zafalon
Plácida L. V. A. da Costa Santos
Jairo da Silva
Universidade Federal de São Carlos (UFSCar) São Carlos – São Paulo, Brasil

Abstract

This study presents the conversion of bibliographic records and delimits the object for understanding the conversion of analogic data to the MARC 21 Bibliographic Format, from the syntactic and semantic analysis of records described according to descriptive metadata structure standard and content standards. We aim to develop a theoretical-conceptual model of syntax and semantics in bibliographic records, from saussurean and hjelmslevian linguistic studies of human language manifestations, which underpins the development of a computational interpreter, focussed on the conversion of bibliographic records to MARC21 Bibliographic Format, which can confirm both the semantic value of the information resource represented as well as the reliability of the representation. The methodological approach of the research is based on qualitative, exploratory, descriptive and experimental methods, besides the use of the literature from the relevant areas. Contributions at the theoretical level are envisaged for the development of issues related to syntactic and semantic aspects of bibliographic records, and together entail an interdisciplinarity approach between Information Science, Computer Science and Linguistics. Contributions in the practical field are identified by the fact that the study includes the development of Scan for MARC, a computational interpreter for the conversion of bibliographic records for printed MARC21 Bibliographic Format.

Introduction

This research presents as its central theme the study of the bibliographic record conversion process and is framed by an understanding of bibliographic record conversion to the MARC21 Bibliographic Format[1], based on syntactic and semantic analysis. It therefore pertains to the study of information resources representation and to the sharing and conversion of bibliographic records in digital environments, to technological development and to theoretical and methodological aspects of such processes, using tools and methods inherent to information and communication technologies.

[1] MARC is an acronym for Machine Readable Cataloguing Record. The MARC 21 Format aggregates formats for bibliographic data, authority data, data for holdings, for classification and for community information. In this research, bibliographic data are studied.

Technological and media resources, through computational structures, permeate the production, organization, distribution, access, storage, preservation, use and reuse of information resources through representation and recovery methods, building, consequently, new socio-cultural, educational, economic and technological contexts. Along with these factors, there is the exponential growth of document collections in information units, which requires adequate librarianship processes that can satisfy, efficiently and effectively, the informational demands of different audiences.

The resorting to computational resources in the daily life of libraries, as substitutes for manual and mechanical activities, has already been commented on by Ranganathan and Gopinath (1967), for whom these processes require economic viability and, ultimately, depend on the development stage of a given country. In the technical-functional and administrative scope of libraries, the use of information and communication technologies has become constant, which favours both the emergence of, and innovation in, various cooperative activities.

Together with the new information and communication technologies emerge, also within the context of cooperative activities, facilities for the sharing of bibliographic records between the most varied types of information units. It has been identified in the literature, however, that to establish the process of converting bibliographic data to the MARC21 Bibliographic Format, it is necessary to start from the study, knowledge and design of the structure of the legacy database. However, given the great variation between database structures, conversion actions can be hampered.

Given the above, together with the professional librarian scenario in the context of aspects regarding the new information and communication technologies, we ask: [1] is it possible to carry out the conversion of bibliographic records to MARC21 Bibliographic Format using just one methodology that would be applicable to different bases?; [2] is it possible to establish a syntactic and semantic content of the bibliographic record which can guide the conversion process to MARC21 Bibliographic Format?; [3] is it possible to apply a theoretical-conceptual model of syntax and semantics of bibliographic records to a computational tool which permits the conversion to MARC21 Bibliographic Format? In this sense, we establish as premises the following facts: a) there is no unique standard adopted for the definition of the database structure of management systems in libraries and other information units; b) in the bibliographic domain, bibliographic records are elaborated from the conventions that come from communities of practice; and c) there are traditional and international description and visualization schemes for bibliographic records, identified in the descriptive metadata structure standards and content standards.

The objective therefore proposed is that of developing a theoretical-conceptual model of syntax and semantics in bibliographic records, from Saussure and Hjelmslevian linguistic studies of human language manifestations, which can then underpin the

development of a computational interpreter[2], for the conversion of bibliographic records to MARC21 Bibliographic Format, and which can confirm both the semantic value of the information resource represented as well as the reliability of the representation.

Given these objectives, the methodological idea in this research is based on a qualitative approach, in which a dynamic relationship with the real world is assumed, through the interpretation and attribution of meaning to the phenomena studied, according to Gonsalves (2011). According to the objectives, the research presents an exploratory view, given that "it is characterized by the development and clarification of ideas, in order to offer a panoramic overview, a first approximation to a given phenomenon that is little explored", and through the literature, offers "elementary data which support the carrying out of further studies on the subject" (op. cit., p. 67, free translation). The research is also of a descriptive nature, "describing the characteristics of an object of study" (op. cit., p. 68, free translation). An experimental dimension to the research is also present in that it refers to "a phenomenon which is reproduced in a controlled way, by submitting the facts to test (checking), and from then on, seeking to point to the relationships between facts and theories" (op. cit., p. 69, free translation).

Such a study, in our view, gives rise to contributions both to the theoretic field, by envisaging the development of questions about syntactic and semantic aspects of bibliographic records, and involving, at the same time, interdisciplinarity between Information Science, Computer Science and Linguistics, in order to reiterate what Borko (1968) and Saracevic (1996) proposed; and to the practical field, by including the development of a computational interpreter which can be implemented by any institution that may wish to make use of the procedure of database conversion of bibliographic records to the MARC21 Bibliographic Format from the description schemes (AACR2) and the visualization of bibliographic records (ISBD), characteristics considered innovative in the research.

Representation standards of informational resources and bibliographical record conversion

The intrinsic relationship between representation and retrieval of documents requires taking into account the description tools and also the structure of bibliographic records, which promotes consistency, accuracy and relevance of the results obtained in response to a query. In terms of structure and content description, we turn to Foulonneau and Riley (2008), who show the descriptive metadata and content standards.

Among the descriptive metadata structure standards, which list elements considered important for resource description, including physical and content characteristics, we

2 The concept of computational interpreter in this research is adopted to reflect the process, mediated by computational resources, which interpret an analogue bibliographic record through a structure defined by the syntactic markings, identified by the punctuation present in bibliographic records, in such a way that it allows for the inference of the semantic value of the represented information resource so that the reliability of representation is guaranteed. Given the search for security in data interpretation and simplification of the processing, we chose to use the Perl language that, according to Stockton ([2005]), is an interpreted language optimised for scanning text files and extract information from text files. It was considered also, the fact of being available under the terms of the General Public License (GNU), characterised as free software (PERL, [2011]), and to allow the combination of programs written in diverse environments such as UNIX, MSDOS, Windows, Macintosh, OS/2. For the development of tests we used the environment Strawberry Perl.

highlight the MARC Bibliographic Format. As for the content standards, which in turn provide the syntax rules of an entry in a metadata field and are intended to promote consistency in metadata records to allow more efficient search and retrieval by users, AACR2 is indicated.

In libraries, the content standard and its relation to descriptive metadata structure standard is familiar, such as AACR2, along with its relationship with MARC, which are both studied in this research. The AACR2r, 2002 revision of the Anglo-American Cataloguing Rules, 2nd edition, presents, by means of guidelines, rules and examples, the description of the content and also the choice, preparation and allocation of access points to a document, allowing the directions to be set for the construction of bibliographic catalogues. Because the AACR2R presents, even historically, a direct relationship with ISBDs, in a certain sense it can be said that for manual catalogues, they take on both aspects of the structure standard of descriptive metadata, as well as the content standard.

The ISBD is understood as a structure standard of descriptive metadata from Swanson (1973) and Langker (1974) for whom the ISBD specifies the elements of a bibliographic description, prescribes the order in which they should be presented, but mainly, because it indicates the punctuation by which elements should be marked.[3] Therefore, the ISBD has three objectives: to make records from different sources interchangeable, facilitating interpretation beyond language barriers, and facilitating the conversion of such records to machine-readable form.

For the description and retrieval of bibliographic records in an automated medium, however, the adoption of a structure standard of descriptive metadata is required, together with AACR2R, and for this research, the interest is in studying the MARC21 Bibliographic Format, which covers reading aspects and interpretation of available data in bibliographic records by computational means. Inherent aspects of the structure of a MARC record can be discerned by the flexibility of the file structure and the number and size unlimited of the fields. The process of reading and computational interpretation of a bibliographic record in the MARC format is facilitated by the markings inherent to it.

Finally, it is understood that the conventions adopted in bibliographic record markings, either by ISBD, either by the MARC21 Bibliographic Format, promote, together with the rules for content description given by AACR2R, each in its own way, the development of catalogues and bibliographic service objectives.

Considering that technical processing and bibliographic information recording are, without doubt, the activities most affected by the cost of automation processes, it becomes essential to ensure that data in a digital medium be (re)used. Therefore, it is required to guarantee the technological and methodological base provided by the adoption of standards,

[3] To consider the ISBD as a descriptive metadata structure standard because of the punctuation takes on an essential character for this research, since, according to Trask (2008, p. 232), the punctuation is "a conventional system of marks that represent information on the structure of a written text." In turn, Langker (1974) points out that the score is used for structural purposes to delimit the fields and subfields (in order to assist a machine operator to record in machine readable form). It is understood that the prescribed points on ISBDs fulfill the dual purpose of providing means to specify bibliographic elements, regardless of the language, for both humans and machines.

which by their nature, promote compatibility and exchange of bibliographic records. In the case of exchange of bibliographic data, one of the main activities involves the conversion of that bibliographic data.

However, it is necessary to clarify that the terms "conversion" and "migration" of bibliographic data, used sometimes synonymously, are different. In the focus of this research, the term "conversion" is adopted to describe the process of changing the media on which a bibliographic record is inscribed, or even the process in which the change occurs in the structure of the record, which does not involve changing the description of its contents. The conversion of bibliographic data is assumed, therefore, as a way to alter the descriptive metadata structure standard of information resources.

Given the configuration of the theoretical proposal regarding bibliographic records, the next section presents the contribution of Saussure and Hjelmslev to the representation of information resources.

A syntax and semantics of bibliographical records from Saussure and Hjelmslev

The elaboration of the theoretical framework of the syntax and semantics of bibliographic records is based on the linguistic contribution of human communication, made by Saussure, and the structuralist semantic conception, from Hjelmslev.

It is understood that, in the same way that linguistics is formed from the manifestations of human language (Saussure, 2010, p. 13), the social role of cultural heritage institutions is formed from the representation of such events, taking account of those registered, regardless of the environment and the medium in which this is done. The manifestations of human language, through their records, allow the description, identification, access, use, reuse, dissemination and sharing between the most diverse cultural heritage institutions. The phenomenon of representing information resources incorporates two faces that match and complement each other: the work and its manifestation both recognized by the International Federation of Library Associations and Institutions (1998, 2005, 2009) as products of intellectual or artistic endeavor.

The work is the intellectual or artistic creation which reflects the content and is identified as an abstract entity. For Smiraglia (2002), a work is the knowledge intentionally created to represent a coordinated set of ideas (i.e. ideational content), which, propagated through text, is intended to be communicated to the consumer. A document can contain one or more works, and a work can exist in one or more documents, which means that it might exist in several instances.

Manifestation is the embodiment of a work, which can only be known if it is manifested, or, in other words, the manifestation only exists from the conception of a work, the work can only be recognised through the manifestation. The manifestation assumes the physical form. Therefore, information representation can only be done through an

understanding of the correspondence between the work and its manifestation. It is not possible, thus, to reduce the representation to one or other face: work is the result of thought, even without being expressed and made public. From this rises the correlation between the work, mental complex unit, and the manifestation, physical complex unit. It is understood, therefore, that the manifestation is the "suit" that a work occupies.

It is thereby understandable that informational resources constitute socializable manifestations of works, which, in turn, are individual or collective. Starting from this perspective, a dilemma can be perceived regarding what is actually being represented: Either we engage with an explanation of what the work is, by being first and foremost a mental conception, or we risk perceiving that it is the manifestation, a record of mental conception in a physical medium. The point of departure is that the representation is observed from the manifestation. Information representation is, therefore, the act of articulating description forms from tools which enable making an informational resource knowable without recourse to the original document to identify it.

The focus between work and manifestation and its relationship with the communication process, direct the study of a theoretical possibility that comes from the field of language for thinking about information records, as suggested by Ferdinand de Saussure in 1916, linking it to the question of the signifier and the signified. To this end, we study the correlation between work and manifestation, signifier and signified.

In the same way that Saussure (2010, p. 81 et seq.) presents principles such as the arbitrariness of the sign and the linear character of the signifier, we will seek to clarify the correlation between work and manifestation present in the principles of Saussure. In the first principle, "The bond that unites the signifier to the signified is arbitrary" (op. cit., p. 81, free translation), we observe the arbitrariness of manifestation in relation to the work. In this sense, the idea of a work, such as "100 scientists who changed the history of the world", by John Hudson Tiner, has no direct relationship with just one form of manifestation; that connection is arbitrary and can take on many other forms: a script for a play or a movie, a musical, a book, among others. In practice, the manifestation can be of any type, as long as it resembles the work by means of embodiment. Regarding the second principle, "The signifier [...] unfolds in time [...]" (op. cit., p. 84, free translation), just as important as the first, it is understood that the linearity that a record assumes requires a sequence for the registration of the work in the manifestation of mental product (ideational content), regardless of its form.

The work makes mention of the mental concept, or, to refer to Saussure, to the signified, to the concept; as the manifestation, in turn, refers to the signifier, the recorded acoustic image. The work, reduced to an essential principle for the manifestation, presents the correspondence between many forms of expression as possible.

The cataloguer, therefore, needs to understand the manifestations in order to use representation mechanisms, making it possible to make the information resources known to others. Thus, although the manifestation is, in itself, foreign to the work, it is impossible to abstract it from the various manifestations, which make the work constantly possible of being represented. However, work and manifestation, two different systems, are

complementary to the formation of the documental object. The work may have an oral tradition, fixed differently from the written tradition on a support, and even then, be transferred to other generations. Although these oral traditions can be an object of cultural heritage institutions, while they are not registered, they are not eligible for representation.

In the field of Linguistics, there are studies focused on structuralist semantics that, in turn, address semantics in a concrete way and analysis the lexical semantics from the central idea that language should be seen as a system (GEERAERTS, 2010; TAMBA-MECS, 2006). Thus, natural language can be understood as a symbolic system with its own properties and principles that determine how a linguistic sign functions.

The Hjelmslevian contribution comes from the view of the structure in structuralist linguistics. We also use Hjelmslev (1991, p. 116, free translation), to better understand the relationship between object, structure and scientific description, when the author declares in the face of linguistics, that "There is neither knowledge nor scientific description of any possible object without recourse to a structural principle." In the light of such a statement, the adoption of formal principles in a part-whole relationship is observed, which implies an intrinsic affinity for document representation in Information Science.

A parallel is denoted between structuralist semantics and Information Science, given that both are based on symbolic constructs and systemic views: just as with language, representation also uses a system, in which it is possible to study, analysis and represent pieces of information, which may be symbolic, and present in the most diverse manifestations of works. In this sense, document representation, based on conventions, norms and standards, is equal to language. Representation allows the synchronous analysis of the document, with the inherent traits to the information resources and their supporting medium, in which the relationships between work and manifestation are present. In representation, synchrony implies the impossibility of differentiating between representatives and represented, between work and manifestation. In this way, the synchronic study of representation proposes the study of bibliographic records from their relationships, as much in the record itself, as in relation to the object described. The structuralist semantics of bibliographic records is thus directed to the descriptive study of the functioning of catalogues.

In the structuralist study of bibliographic records, there exists the intention of identifying the record structure, its relationship with other records and the relationship with the document. Thus, the syntax of the bibliographic record cannot account for the catalogue; it is the semantics that allows the context and the synapses between the various bibliographic records; it is the semantics that can account for the mental processes by which the representation of an informational resource is produced, constituted, understood and described. In this research, the study of the various relationships that can be established between bibliographic records, between the bibliographic record and the information resource, and between the elements of its own bibliographic record is called the semantic role.

In this way, the possibility of studying document representation from a theory of levels is considered: from the subsemantic level (between the elements of a bibliographic record)

to the supersemantic level (the relationship between the various bibliographic records, based on their similarities and differences), through the semantic level (the object being described and the description itself). Thus, the aim of the structuralist emphasis for the semantic analysis of bibliographic records can be defined as: the study of the description of bibliographic records actually made, in which one considers the influence of the catalogue as a means for establishing messages contained in the information resources and in the information needs of users. It is possible to admit three planes of semantic difference in bibliographic records: between the referent and the representation, between the whole and parts of the representation, and between the representations present in the catalogue. These are the semantic aspects that reduce the alterity of a bibliographic record, which in a catalogue, cause the dispersed and the apparent to be marked by identities of their own. Semantics in Information Science is given by the form and representation of the information.

Semantics, conceived in this way, refers to the structure of a system that links signifier and signified, work and manifestation. In the semantics of bibliographic records, the signified is given by the value of the signifier, or, the manifestation is the value of the work in the representation process. These semantic values in a bibliographic record form a network of structural relationships with other bibliographic records, which is called the bibliographic record supersemantics. Semantics requires the adoption of syntax to define semantic values; in other words, syntax is present in the descriptive metadata structure standard and in the semantics of the content standards.

Syntax, in the context of this research, concerns the order of elements arranged for the representation of information resources. Therefore, it is understood that the syntax of the bibliographic record is that part of Information Science dedicated to the study of form, arrangement and layout in which each element must be described when the representation of the information resource is done. In this sense, it forms part of the librarianship system that determines formal relationships between the representations of each of the parts of the represented document. These elements are organized according to established metadata structure standards. The syntactic aspects of a bibliographic record may refer to the semantic structure.

Bibliographic language surpasses the syntactic level and enables the understanding that a record presents semantic levels, needed to understand the represented document syntactically and semantically. Thus, each syntactic element assumes a semantic content for a given defining element of the representation, and this element, in turn, as opposed to the contextualized and represented document, assumes significance between the record and the object.

The bibliographic record thus binds both syntactic issues, by referring to standards of structural metadata for each element of the document or object to be described, as semantic issues, by allowing the analysis of the cohesion and meaning indicated between elements of the representative and of the represented and between the representative itself and the represented. Each syntactic element, when contextualized and when contrasted with the represented document, assumes a concrete meaning between the record and the object.

How can conversion be understood from the syntactic and semantic analysis of bibliographic records, so as to make possible the conversion processes of bibliographic records to MARC21 Bibliographic Format? It is understood to occur through the use of the marking given by the descriptive metadata structure standards, present in the AACR2r and in the ISBDs, and therefore, by semantic inference, provided in a computer application. At this time, the application of theoretical and conceptual aspects of syntactic and semantic principles of bibliographic records to the conversion of bibliographic records to MARC21 Bibliographic Format in a computer interpreter is envisioned.

Scan for MARC: syntactic and semantic interpretation of printed bibliographical records

Departing from the assumption that considers the syntactic and semantic schemes of bibliographic records, and not the structure of the legacy database, necessary for the conversion of bibliographic records to MARC21 Bibliographic Format, the syntactic and semantic computational interpreter of bibliographic records, identified as Scan for MARC, is discussed referring to the scanning method of bibliographic records and their subsequent conversion to the MARC21 Bibliographic Format.

The development of the interpreter, in beta version, involved, briefly, the capturing of images of analogue bibliographic records, presented in catalogue cards in electronic media; the analysis of the results of image processing in character recognition software, which allows converting image into editable text; the image processing of bibliographic records selected for testing (initial process of representation building); the syntactic and semantic processing of digital bibliographic records, the checking of results and evaluation of adjustments.

The testing phase for capturing images of bibliographic records was divided into three steps: in a functional printer scanner, mobile device camera and digital camera. At the end of this phase, tests were conducted for the processing of images with character recognition of bibliographic records in analogue format, for which we adopted character recognition software (OCR). In this test phase, divided in two parts (the analysis of freeware software or free software and proprietary software, with analysis in trial versions), we analysed the following software: 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.[4]

The tests were developed from the collating and analysis of punctuation, diacritical marks, the exchange of letters, the spacing between information and margins. Among the identified applications, the software OnlineOCR.net offered the best results. This was

[4] Access link: 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/).

followed by a new testing phase in which we sought to define the method of image processing to provide better results (in the various forms of image capture).

As a general result, the following comments can be presented: issues related to lighting in the capture of images are extremely relevant to OCR software processing, given its influence on image quality; the best success rates in image processing of bibliographic records were attained through capturing the images on a multifunctional printer scanner; problems with the exchange of letters in the process of character recognition in images were identified in the tests of the three different devices; aspects concerning the setting of the camera, whether from a mobile device or through digital photography, were more relevant than the distance to be considered in image capture; OCR quality is intrinsically linked to the quality of the image and not the method by which the image is captured. From the results, we chose to work with the images captured in a multifunctional printer scanner and with OnlineOCR.net.

Having completed the image processing tests of the selected analogue bibliographic records, procedures used for the syntactic and semantic treatment of bibliographic data in the file were then followed, which are now discussed in the context of the theoretical proposal presented in this research.

Standards of descriptive metadata structure (ISBD, AACR2r and MARC21 Bibliographic Format) were studied, and the interference of punctuation marks in semantic content, present in the content standards (AACR2r), was determined. For the effective implementation of these markings in the conversion script of bibliographic records to MARC21 Bibliographic, it was necessary, however, to take care regarding the punctuation that is part of the content and not of the descriptive metadata structure. In the search for the identification of standards, a detailed analysis of the presentation of the subject headings access points was also required, for those components that would be determined as beginning with Indo-Arabic numerals followed by a full-stop, and that, for the other access points, Roman numerals would be adopted.

Undoubtedly, one of the first problems identified for the information processing was due to the type of character encoding of the input text file (UTF-8/ISO, UFT-16/UNICODE, ASCII/ANSI) necessary for the correct interpretation of diacritic markings. The treatment phase of syntactic and semantic digital bibliographic records was made from tests for processing scripts of bibliographic records. We established four itineraries, with varying degrees of complexity, on which the versions of scripts could be based (developed in four versions, each with minor adjustments needed after checking the results).

In this paper, some results are presented that do not show, however, all cases foreseen, but which are already in operation in Scan for MARC (cf. Figure 1). In all the cases the result of the OCR image processing is indicated, aligned to the left, and to the right, the result from Scan for MARC.

Figure 1 - Results of OCR image processing and syntactic and semantic treatment by Scan for MARC

```
658.022 Oliveira, Antonio Carlos de
O48d        Desenvolvimento de produtos e inovação tecnológica
5. ed.  em pequenas e médias empresas do Estado de São Paulo /
        Antonio Carlos de Oliveira, Paulo Carlos Kaminski. -
        São Paulo : FAPESP, 2005.
           85 p.

           Inclui bibliografia.

           1. ADMINISTRAÇÃO 2. PEQUENAS E MÉDIAS EMPRESAS 3. INOVAÇÃO
        TECNOLÓGICA I. Kaminski, Paulo Carlos. II. Titulo.
```

```
090  $a658.022 $bO48d $c5. ed.
1001 $aOliveira, Antonio Carlos de
24510$aDesenvolvimento de produtos e inovação tecnológica em pequenas
       e médias empresas do Estado de São Paulo $cAntonio Carlos de
       Oliveira, Paulo Carlos Kaminski
260  $aSão Paulo $bFAPESP $c2005
300  $a85 p.
504  $aInclui bibliografia
65014$aADMINISTRAÇÃO
65024$aPEQUENAS E MÉDIAS EMPRESAS
65024$aINOVAÇÃO TECNOLÓGICA
7001 $aKaminski, Paulo Carlos
```

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

           Titulo do original: Die Terfelsbibel.

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

```
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  $aTitulo do original: Die Terfelsbibel
65014$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.
```

Source: The authors

Note that in the results, the encoding in MARC21 Bibliographic Format remained correlated with that made by cataloguers. Given the analysis of the outcome, it is understood that in light of the proposed syntactic and semantic interpretation of bibliographic records, the outcome was of good quality.

Having described above the analysis dedicated to testing the image processing of analogue bibliographic records based on the syntactic and semantic aspects of bibliographic records, we now turn to the final considerations of the research.

Final considerations

Given the topic set out for this research, namely, the conversion of bibliographic records to the MARC21 Bibliographic Format, a theory of the syntax and semantics of bibliographic records was developed, defined by structure standards of descriptive metadata and content standards, based on those of ISBD and AACR2r.

We presented a theoretical-conceptual framework for the representation of information resources and sharing and conversion of analogue bibliographic records in digital environments; technological development achieved before the proposal to ensure the reliability of knowledge representation aspects; and the analysis and development of theoretical and methodological aspects that support data conversion activities, resorting to methods inherent to information and communication technology.

The importance of adopting norms, rules, standards, formats, methodologies and criteria for the representation of information resources in information units was highlighted, with a view to implementing processes, permeated by technology and media applications that rely on computational structures that endorse the production, organization, storage, management, treatment, preservation, distribution, supply, recovery, access, use, reuse and sharing of informational records in various media.

Following Saussure, the representation model of information resources based on the relationship between signifier and signified was established, in which the arbitrariness of the manifestation in relation to the work is discussed, as well as the development of the linearity of the manifestation in relation to the work's ideational content, a major factor for understanding the document and necessary for creating the draft of the bibliographic record. Following Hjelmslev, and based on the formal principles adopted in Linguistics for the study of the linguistic system structure, the representation of documents in Information Science was discussed. It was observed that this is based on the theoretical and systemic constructs of the synchronic analysis of the document, with inherent traces of the in dissociation between work and manifestation, for which the study of bibliographic records is proposed, from the internal relations between elements of a record (subsemantics), among the records of a catalogue (supersemantics), and in relation to the described document (semantics).

Departing from the questions presented, we sought to form a theoretical and methodological framework for bibliographic representation, and the syntactic and semantic features of the represented objects, reflecting the following: [1] an understanding of the work, as the signified, and the manifestation, as the signifier; [2] an understanding of bibliographic representation as a result of the relationship between signifier and signified and between work and manifestation, and as definitive for the semantics; [3] the perception of syntax for the definition of subsemantics, and as being necessary for the representation of the information resource; [4] an understanding of the concept of supersemantics, from its co-dependency with subsemantics and semantics, in the relationship, identifiable in catalogues, between bibliographic records and documents of a collection, and between work and manifestations, taken as signifier and signified.

Considering the results obtained in the tests, although considered preliminary because they still require adjustments and improvements for defining the behavior of Scan for MARC, it is believed that the contribution of the social aspects sought present significant theoretical and practical impacts in the field of Information Science, as well as in its interdisciplinary interfaces with Computer Science and Linguistics.

For future studies, in the case of Scan for MARC, the identified needs for improvement and refinement of the computational interpreter are recognised, namely: [1] the integrated combination of the phases of reproduction and representation, appealing to the adoption of an OCR command line; [2] development of a GUI - graphical user interface; [3] prescription of standards for the interpretation of subject data classification (CDU); [4] treatment of qualifier terms content (form subdivision, chronological subdivision, geographical subdivision and general subdivision) in access points to the topic subject; [5] prescription of standards for the treatment of secondary subject access points for personal name, institution, event and uniform title; [6] prescription of treatment standards for access points for major institutions, events and uniform title; [7] prescription for building semantic lexicons for access points and their qualifier terms; [8] the implementation of a consistency test of the converted file process, since it measures quality in the generated bibliographic record; [9] integration of the script with the OCR, which requires the adoption of a command line.

Finally, there is a need to continue studies of syntactic and semantic methods of bibliographic records, and to investigate the validity of this conversion method for analogic bibliographic data when applied to the interpretation of cataloguing in publication data of the book document type.

References

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. Acessoem: 20 jun. 2011.

International Federation of Library Associations and Institutions. Functional requirements for bibliographic records: final report. 1998. Disponívelem: http://archive.ifla.org/VII/s13/frbr/frbr3.htm#6. Acessoem: 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. Acessoem: 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/. Acessoem: 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.