XVI Seminário Nacional de Bibliotecas Universitárias
II Seminário Internacional de Bibliotecas Digitais - Brasil

GERÊNCIA DE REGISTROS DUPLOS EM BASE DE DADOS BIBLIOGRÁFICA

Zita Prates de Oliveira[1], Lais Borges Freitas[1], Caterina Groposo Pavão[1]
Janise Silva Borges da Costa[1], Margarida Cecília Schmidt[2], Zaida Horowitz[3],
Carla Metzler Saatkamp[4], Bruce Lucien Santos Notario[4]

[1] Bibliotecária, Comissão de Automação, UFRGS, Porto Alegre, RS

[2] Bibliotecária contratada

[3] Analista de Tecnologia da Informação, UFRGS, Porto Alegre, RS

[4] Técnico de Tecnologia da Informação, UFRGS, Porto Alegre, RS

Resumo

O texto aborda o tratamento de registros duplos de monografias identificados na base bibliográfica SABi/UFRGS. Duplicidade que reflete a dispersão geográfica dos acervos e as diferentes políticas de catalogação e *softwares* de entrada de dados adotados pelas bibliotecas da Universidade ao longo dos anos. Considerando o volume de registros na base, aproximadamente 700 mil, e os procedimentos de análise, comunicação e arbitragem envolvidos na identificação e exclusão de registros duplos, a metodologia adotada permitiu excluir a duplicidade e avaliar a política de catalogação cooperativa e descentralizada adotada pelo SBU.

Palavras-chave: Gerência de registros duplos; Biblioteca Universitária; Campos MARC; Detecção de registros bibliográficos duplos.

Abstract

The paper deals with the management of duplicate monographic records identified in the SABi/UFRGS library database. This duplication reflects the geographic dispersion of the collections and the different cataloging policies and data entry programs adopted by the University libraries through the years. Taking into account the 700 thousand records in the database and the analysis, communication, and arbitration procedures used for the identification and exclusion of duplicate records, the adopted methodology is able to exclude duplication and to evaluate the cooperative and decentralized cataloging policy adopted by the University library system.

Keywords: Duplicate records management; University libraries; MARC fields; Detection of duplicate bibliographic records.

1 Introdução

A existência de registros bibliográficos duplos é um problema em ambiente de rede de bibliotecas que adota política de catalogação cooperativa.

Em uma instituição que adote a catalogação cooperativa, todos os catalogadores utilizam os mesmos instrumentos de análise e de codificação do registro bibliográfico (código de catalogação, formato/campos de registro e tabelas de códigos internacionais, locais e institucionais), mas esses instrumentos não garantem a exclusividade de cada registro da base de dados, em relação ao documento que representa. Fatores externos como a origem dos registros, a interpretação particular de cada catalogador ao aplicar as regras/políticas de catalogação e a sua eventual inabilidade, ao pesquisar a base de dados, para identificar registro bibliográfico que corresponda à mesma descrição do documento a ser catalogado, favorecem a inserção de registros duplicados na base. Essa duplicidade acarreta problemas como: sobrecarga de informação para o usuário, redução da eficiência do sistema, redução da produtividade da catalogação e aumento do custo de manutenção da base (SITAS; KAPIDAKIS, 2008).

A identificação e exclusão de registros duplos em base de dados bibliográfica estão diretamente associadas à idéia de oferecer informação qualificada ao usuário e de melhorar a *performance* do sistema. Entretanto, não é menor o seu papel na monitoria à correta aplicação da política de catalogação cooperativa em rede de bibliotecas.

2 Sistema de Bibliotecas da UFRGS

O Sistema de Bibliotecas da Universidade Federal do Rio Grande do Sul (SBU) é constituído por 33 bibliotecas setoriais dispersas em 4 *campi*. Devido a essa dispersão geográfica nunca houve, por parte da Universidade, um projeto de reunir os acervos das bibliotecas em um único local. Ao contrário, a UFRGS sempre preferiu a sua fragmentação. Em 1999 foi criada a mais recente das bibliotecas, a da Escola de Administração, com a separação do acervo da biblioteca da Faculdade de Economia.

Descentralização de acervos significa, no caso da UFRGS, descentralização de atividade/serviços e duplicação dos mesmos títulos em mais de uma biblioteca. Na tentativa de homogeneizar procedimentos a Biblioteca Central, responsável pela coordenação técnica do SBU, juntamente com grupos de trabalho, desenvolveu nos anos 1980/90 uma série padrões de serviços bibliotecários (PSBUs) a serem observados pelas bibliotecas setoriais na execução de suas atividades e na prestação de serviços aos usuários (MACHADO et al., 1982). Os PSBUs foram o embrião do trabalho coordenado que seria proposto às 33 bibliotecas, a partir da implantação do SABi versão Aleph.

Quanto a duplicação de títulos, este era um problema transparente enquanto os catálogos das bibliotecas foram manuais, mas se tornou um problema bem visível com a reunião de todos os registros bibliográficos no catálogo *on-line* SABi e com a implantação do módulo de Circulação, responsável pelo controle de empréstimo, devolução, renovação e reserva dos itens associados aos registros bibliográficos do catálogo.

Embora a Universidade mantenha a dispersão geográfica dos acervos adota, desde 2000, política de catalogação cooperativa e descentralizada utilizando *software* em versão multiusuário para gerência da catalogação. Anterior a essa data, as versões mono e multiusuário do software eram utilizadas simultaneamente, conforme a infraestrutura de rede de dados disponível em cada Unidade da Universidade. Estas limitações contribuíram para o alto índice de inclusão de registros duplicados na base SABi.

3 Características do Sistema de Automação de Bibliotecas (SABi)

O SABi realiza a gerência integrada das atividades e serviços das 33 bibliotecas e do catálogo *on-line* da UFRGS através dos módulos de Aquisição, Catalogação, Itens, Periódicos e Circulação e adota padrões internacionais para registro de dados bibliográficos (MARC21) e intercâmbio de informações (ISO 2709 e ANSI Z39.2).

Utiliza o *software* Aleph 500 e o banco de dados Oracle, instalado em

equipamento HP Proliant ML370, com dois processadores Intel Xeon, 4Gb de memória e com sistema operacional RedHat.

O sistema compreende três bases de dados: bibliográfica - URS01, de autoridades - URS10 e administrativa - URS50 (registros de itens, transações de circulação e controle de aquisição).

O catálogo *on-line* do Sistema de Bibliotecas exibe os registros bibliográficos, informações sobre os itens a eles associados, bem como sobre o status de circulação desses itens.

4 Revisão de literatura

Adotar procedimentos para detecção e consolidação de registros duplicados é uma preocupação frequente na literatura sobre catálogos coletivos. Dado o volume de registros originados de diferentes instituições, a literatura tem enfatizado o desenvolvimento de programas de detecção automática da duplicidade. Ridley (1992) detalha a criação de um sistema especialista capaz de detectar duplos, desenvolvido na Universidade de Bradford. O sistema estabelece uma série de regras, aplicadas na identificação da duplicidade e na definição do melhor registro entre eles, mas o autor conclui pela necessidade de criar um registro único, composto pelas informações mais corretas de cada um dos duplos.

Hickey e Rypka (1979) desenvolveram um programa que detecta a duplicidade em registros de monografias em grandes sistemas *on-line*. Testado em amostra do catálogo da OCLC o programa detectou 5% da mesma como registros duplos, os quais foram eliminados.

Sitas e Kapidakis (2008) apresentam uma visão geral dos algoritmos de detecção de duplos, os tipos de documentos a que se aplicam, as etapas de sua criação e aplicação e o tratamento final dos registros duplos identificados (deleção, incorporação ou apresentação dos registros de forma composta no momento da visualização de uma pesquisa) em bases de dados de diferentes sistemas de bibliotecas. Concluem que não há um programa mais eficiente, mas sim que, cada ambiente tem suas necessidades e políticas próprias. O uso de um ou outro

programa demanda análise e estudo específicos para a sua adequação às necessidades da instituição.

Cousins (1998) aborda o tratamento da duplicidade no catálogo coletivo COPAC (*Consortium of University Research Libraries - Online Public Access Catalogue*), que registra o acervo das maiores bibliotecas acadêmicas do Reino Unido e Irlanda. Detalha os campos MARC utilizados para comparação da duplicidade e relata uma redução de 50% no número de registros individuais incluídos no COPAC, para novas bibliotecas que ingressam no mesmo.

5 Exclusão de registros duplos no SABi

Os tópicos que englobam o processo de exclusão de registros duplos na base SABi são abordados a seguir.

5.1 Definição de registro duplo

No ambiente da base de dados bibliográficos SABi, registros duplos são definidos como aqueles cujos campos correspondentes contêm conteúdo igual ou aproximado.

Levantamento inicial realizado no catálogo *on-line* do SBU utilizando o recurso de pesquisa Percorrer lista*,* no índice de título, possibilitou identificar registros duplos de diferentes níveis de descrição bibliográfica (monografias e analíticas) ou de documentos em diferentes suportes (impresso e digital) e formatos (monografia e periódico), bem como o ano de inclusão dos mesmos no SABi.

Registro bibliográfico de monografia sem analíticas vinculadas foi selecionado como o primeiro tipo de duplo a ser excluído, não considerando duplos os registros bibliográficos de reimpressões e das versões em brochura e encadernada.

O ano de inclusão do registro bibliográfico no SABi foi definido como elemento de controle, para avaliar a aplicação da política de catalogação cooperativa no SBU.

5.2 Definição dos campos MARC da estratégia de busca

A escolha dos campos do formato MARC21 utilizados nas estratégias de busca define o maior ou menor grau de precisão no resultado da identificação dos registros potencialmente duplos.

A Tabela 1 apresenta um resumo dos campos utilizados para detecção de duplos por autores citados neste trabalho. Sitas e Kapidakis (2008) identificam a presença dos campos de autor, título e ano de publicação em 90% dos programas desenvolvidos com esta finalidade. Já a paginação é utilizada em 70% e o ISBN em 60%.

Tabela 1 - Campos MARC utilizados para detecção de duplicidade

Autor / Campo	Cousins	Hickey/Rypka	Sitas/Kapidakis
Ano de publicação	✓	✓	✓
Autor	✓		✓
Edição	✓		
ISBN	✓		✓
Local de publicação		✓	✓
Paginação	✓	✓	✓
Publicador	✓	✓	✓
Série	✓		
Título	✓	✓	✓

Para detecção de duplos de monografias, sem analíticas, na base bibliográfica SABi foram selecionados os campos considerados mais estáveis (SITAS; KAPIDAKIS, 2008) em termos de presença e conteúdo nos registros bibliográficos: título, data, edição e paginação (Tabela 2). Os campos de autor e ISBN, embora considerados importantes no processo de identificação de duplos, não foram incluídos na estratégia. O campo de autor, devido às variações de entrada identificadas no levantamento inicial realizado no catálogo *on-line.* E o ISBN porque, embora seja um identificador único do documento e, em tese, um indicador preciso de duplicidade, só pode ser utilizado quando é de uso regular nos registros bibliográficos (RIDLEY, 1992), o que não ocorre no SABi. Com a implantação do

módulo de Aquisição do SABi, o campo passará a ser incluído regularmente nos registros bibliográficos, possibilitando sua futura utilização como parâmetro de detecção de duplicidade.

Tabela 2 - Campos selecionados para detecção de registros duplos no SABi

Campo/subcampo	Conteúdo
008/7	4 dígitos = data
245|a	título
|b	outras informações sobre o título
|h	designação geral do material
250|a	edição
260|c	data de publicação
300|a	paginação

5.3 Programa de identificação de duplos

Foi desenvolvido um programa em PHP que faz a análise dos dados de um arquivo que contém os campos dos registros selecionados a partir de uma pesquisa feita no SABi, no caso monografias sem analíticas. Para a análise foram desconsideradas todas as pontuações, partículas iniciais e hifenizações de palavras.

Os campos obrigatórios são verificados integralmente. São eles: campo codificado 008|7 com informação de data, campo de título - 245|a, e campo de data de publicação - 260|c.

Os campos opcionais são analisados somente se existirem em todos os prováveis registros candidatos a duplos, caso contrário, existindo num registro e em outro não, este registro não é considerado como provável duplo. São eles: outras informações sobre o título - 245|b, designação geral do material - 245|h, edição - 250|a e paginação - 300|a, neste caso será considerado como provável duplo quando a diferença entre os dígitos for de até 10.

Ao final lista, as informações de título, ano, edição, paginação, além do número de sistema e biblioteca dos prováveis registros duplos para análise manual.

5.4 Rotinas e procedimentos para exclusão de registros

A partir da listagem de possíveis duplos identificados pelo programa, uma bibliotecária contratada especificamente para este trabalho realizou o procedimento manual de análise e exclusão da duplicidade. A ela couberam as seguintes tarefas:

a) analisar os registros duplos e manter contato com as bibliotecas para dirimir dúvidas quanto a diferenças nas entradas e na descrição bibliográfica;
b) selecionar o registro a ser mantido na base considerando sua correção (grafia) e completeza (maior número de campos e presença dos campos de uso institucional: 090 - Classificação por área de conhecimento do CNPq; 909 - Produção intelectual da Universidade e 902 - Disciplina de curso de graduação);
c) transferir os campos 910 - Código da biblioteca, 6XX, de assunto, e informações de itens/aquisição dos registros a serem excluídos para o registro a ser mantido na base;
d) excluir os registros considerados duplos;
e) elaborar e preencher os formulários de controle, a serem encaminhados a cada biblioteca do SBU, informando os novos números de sistema dos registros transferidos e as alterações necessárias nas informações dos itens (p. ex. no número de chamada, devido a mudanças na entrada principal do registro bibliográfico), e
f) manter os controles estatísticos sobre o número de registros duplos avaliados e excluídos, ano de inclusão dos mesmos na base e número de itens transferidos para os registros mantidos na base SABi.

6 Resultados finais

Entre novembro de 2008 e maio de 2010 foram analisados 2.870 registros bibliográficos potencialmente duplos. Destes, 2.397 foram excluídos. Também foram transferidos 3.932 itens (contendo informações dos exemplares e de aquisição) para os registros mantidos na base SABi.

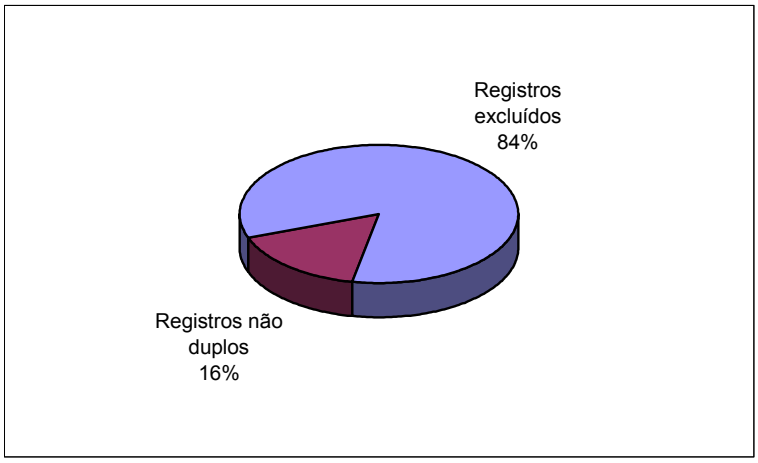

Figura 1 - Exclusão de registros duplos de monografias da base SABi, nov. 2008 - maio 2010.

O objetivo inicial da rotina era o de excluir registros duplos criados no SABi versão monousuário (registros bibliográficos de cada biblioteca do SBU incluídos em disquetes e posteriormente transferidos para o servidor do CPD) porém, foram identificados 734 registros duplos criados entre 2000 e 2010, quando já vigorava a política de catalogação cooperativa, visando à não duplicação de registros idênticos na base bibliográfica. Este número corresponde a 31% dos duplos excluídos da base.

A Figura 2 registra a tendência à duplicação de registros no SABi, a partir de 2000. O número de registros duplos que ingressam a cada ano sinalizam a dificuldade de manter uma política de catalogação cooperativa, baseada apenas em instrumentos e procedimentos comuns de descrição bibliográfica no SABi.

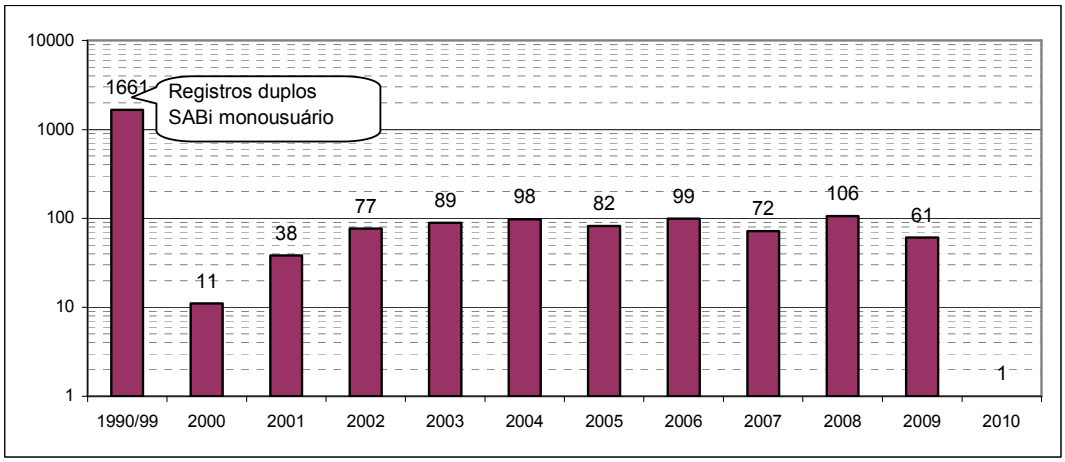

Figura 2 - Registros duplos por ano de inclusão no SABi, nov. 2008 - maio 2010.

7 Considerações finais

O procedimento de exclusão de registros duplos foi concebido para qualificar a recuperação da informação no SABi, catálogo *on-line* das bibliotecas da UFRGS, identificando e excluindo a duplicidade de registros criados pelo uso do *software* de entrada de dados monousuário. O elevado índice de duplicidade identificado no período entre 1990 e 1999 retrata a superposição de acervos (p. ex. nas áreas de referência e da saúde e nos assuntos cálculo e metodologia da pesquisa) para dar conta das demandas de informação de usuários que frequentam as bibliotecas das unidades universitárias dispersas pelos 4 *campi* da UFRGS.

Duplicidade que também representa sobrecarga de informação para o usuário do catálogo, adicionando ruído e reduzindo o número de documentos relevantes no resultado de uma pesquisa. Um subproduto do trabalho foi a identificação de falhas na aplicação da política de catalogação cooperativa, implantada no SBU a partir de 2000, considerando que 31% dos registros duplos excluídos foram criados entre 2000 e 2010.

A adoção de instrumentos e procedimentos comuns de análise e registro de dados bibliográficos, na base de dados, não garante a correta aplicação da catalogação cooperativa no SABi. Os catalogadores seguem alimentando a base com catalogações desnecessárias, que nada acrescentam em informação para o usuário, ocasionando sobrecarga de trabalho para excluí-las da base bibliográfica. Além do desperdício de tempo que isso representa, será necessário desenvolver um programa de detecção de duplicidade na entrada de dados, a ser associado ao módulo de catalogação do *software* Aleph utilizado pelas bibliotecas da UFRGS.

Os campos do formato MARC21, utilizados pelo programa teste de identificação de registros duplos, possibilitaram excluir e consolidar os duplos de monografias da base bibliográfica. Novos campos devem ser definidos para integrar as estratégias de identificação da duplicidade em registros de analíticas e em documentos em diferentes suportes e formatos de registro bibliográfico.

8 Referências

COUSINS, S. A. Duplicate detection and record consolidation in large bibliographic databases: the COPAC database experience. **Journal of information science**, v. 24, n. 4, p. 231-240, 1998.

HICKEY, T. B.; RYPKA, D. J. Automatic detection of duplicate monographic records. **Journal of library automation**, v. 12, n. 2, p. 125-142, 1979.

MACHADO, I. C. N. et al. Padrões para os serviços bibliotecários na UFRGS: relatório da implantação. In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 11., 1982, João Pessoa. 11f.

RIDLEY, M. J. An expert system for quality control and duplicate detection in bibliographic databases. **Program**: automated library and information systems, v. 26, n. 1, p. 1-18, Jan. 1992.

SITAS, A; KAPIDAKIS, S. Duplicate detection algorithms of bibliographic descriptions. **Library Hi Tech**, v. 26, n. 2, p. 287-301, 2008.