Богољуб Мазић

Универзитетска библиотека
"Светозар Марковић", Београд

Bogoljub MaziC

University LibrarY
"Svetozar Markovic", Belgrade

РАДИОНИЦА О КОНВЕРЗИЈАМА ФОРМАТА

WORKSHOP ON FORMAT CONVERSION

Алан Хопкинсон је водио радионицу о конверзијама формата, што је једна од кључних тема коју је требало покрити у оквиру Tempus пројекта "Изградња кооперативне мреже високошколских библиотека у Србији". Радионица је обухватила пет тема:

Историја различитих формата

Предавач је говорио о историји различитих формата. COBISS је користио UNIMARC, формат који је развијан од 1977. године, јер су до тада многе националне библиотеке развиле сопствене формате, пратећи Конгресну библиотеку, која је развила иницијални MARC формат - LCMARC 1966. Британска национална библиотека је успоставила UK MARC 1967, а већина осталих формата је била заснована или на једном или на другом. Па ипак било је потребно писати компјутерске програме да би се један формат конвертовао у други. На пример Британска библиотека је увек одржавала збирку програма за конверзију и редовно је конвертовала записе Конгресне библиотеке у новом MARC формату , како би били доступни и у UK MARC формату. Седамдесетих година то је постао проблем, јер је пролиферација националних формата значила да свака национална библиотека када жели да преузме записе из друге земље треба да направи програм за конверзију за сваку земљу посебно. Писање програма за конверзију је било прилично скупо у то време, у ери меинфреим компјутера (пре него што су се појавили микрорачунари). Зато је IFLA спонзорисала развој UNIMARC-а као формата за конверзију, како би свако могао своје записе да конвертује у UNIMARC и добије записе из других земаља у неутралном интернационалном UNIMARC формату. Напоменуо је да је тај формат касније прихваћен као национални у неким земљама као што су Јужна Африка, Јапан, Југославија и Португал. Португал се сада стара о UNIMARC-у у име IFLA-е и домаћин је сталног комитета за управљање UNIMARC-ом. Пошто се UNIMARC појавио после LC MARC-а, имао је нека побољшања у структури. Био је боље опремљен за процесирање општег каталога, док је LC MARC започет као начин аутоматизације продукције лисног каталога, па се то одражавало на његову структуру, на пример тако да су имена подељена између поља 100 и 700, зависно од тога да ли су главна или допунска одредница, што у аутоматизованом систему нема смисла, јер је неважно на којем се месту налази главна одредница.

- LC MARC је касније постао US MARC, како би било јасно да је то формат који користе све библиотеке у САД а не само Конгресна библиотека. Касније је постао MARC 21 (MARC за 21. век), у покушају да се овај формат интернационализује.

- UK MARC је већ поменут, али као део тренда да се пређе на MARC 21, јер ће се конверзија записа Конгресне библиотеке у UK MARC прекинути у наредне три године. Британска библиотека је управо набавила Aleph интегрисани библиотечки систем који ће користити MARC 21 као свој формат, а онима који то буду тражили испоручиваће и записе у UK MARC формату у наредне три године.

Предавач је поменуо и CCF формат (Common Communication Format) који је развио UNESCO у покушају да

Alan Hopkinson led a workshop on Format Conversion which was one of the key topics to be covered under the Tempus project UM_JEP-16059-2001 "Building a Cooperative Academic Library Network in Serbia" The session was divided into five sections.

History of different formats

He talked about the history of different formats. COBISS used UNIMARC which was a format developed from 1977 onwards because many national libraries had developed their own formats following the Library of Congress's initial development of LC MARC in 1966. The British Library had established UK MARC in 1967 and most other formats were based on one or the other. However, it needed the writing of computer programs to convert from one format to another. For example the British Library has always maintained a suite of conversion programs and has converted on a regular basis the Library of Congress's output of new MARC records making them available in the UK MARC format. In the 1970's there appeared to be a problem because the proliferation of national formats meant that each national library which wished to take records from other countries would have to write a conversion program for each. Conversion programs were quite expensive to write at that time in the era of mainframe computers (before microcomputers had come on the scene). Therefore IFLA sponsored the development of UNIMARC as a switching format or a conversion format so that everyone would convert their records into UNIMARC and receive records from other countries in the neutral international UNIMARC format. He mentioned that this later became adopted as a national format in some countries such as South Africa, Japan, Yugoslavia and Portugal. Portugal now look after UNIMARC on behalf of IFLA and host the Permanent UNIMARC Committee management. Because UNIMARC was later than LC MARC It had some improvements in its structure. It was geared more to general catalogue processing whereas LC MARC began as a way of automating catalogue card production and there were vestiges of this in its structure, such as the fact that names are split over 100 and 700 fields depending on whether they were main entry or added entry which did not make sense in automated systems where main entry was less important.

- LC MARC had later become US MARC to indicate it was the format of all libraries in the US and not just the Library of Congress. Later it became MARC21 (MARC for the 21st century) in an attempt to internationalise the format.

- UK MARC had already been mentioned but was part of the trend to switch to MARC 21 since it was mentioned that the conversion undertaken by the British Library of LC records into UK MARC would cease in the next three years, as the British Library had just taken delivery of the Aleph integrated library system and were going to use MARC21n as their format, though they would be providing UK MARC records for customers who wanted them for up to there years.

He mentioned the CCF (Common Communication Format) which was developed by UNESCO in an attempt to provide a switching format between abstracting and indexing services records and those of libraries who used UNIMARC. It took into account a number of existing formats including UNIMARC and

обезбеди формат за пренос података за апстрактне и индексне сервисе и за библиотеке које користе UNIMARC. Овај формат је узео у обзир велики број постојећих формата, укључујући UNIMARC, као и UNESCO UNISIST Reference Manual. Овај формат је промовисан у Југославији на једној радионици 1987. године, коју су спонзорисали UNESCO, Институт Винча и Национална и свеучилишна књижница из Загреба, а одржана је у Винчи и у Загребу. CCF је коришћен у свету у центрима за научне информације и у свим областима рада у Индији.

Главно је код свих тих формата да они сви користе исту структуру записа дефинисану стандардом ИСО 2709, што чини конверзију међу њима лакшом.

Постоје и други нестандардни формати. Базе података у Microsoft Access-у нису стандардне у смислу библиографске базе података. UNESCO-в пакет CDS/ISIS користи донекле нестандардну структуру записа, што ствара неке проблеме.

Конверзија података је сада велика тема јер многе организације у свету и многе земље, укључујући и Мидлсекс универзитет, прелазе на MARC 21. MARC 21 и UK MARC су врло слични, те методологија конверзије није сувише сложена, али конверзија из UNIMARC-а у MARC 21 је нешто сложенија.

Методе конверзије

Пре него што пређемо на конверзију треба се сложити око терминологије: извор се користи за оригинални запис а циљ за формат у који се конверзија обавља.

Постоји више начина за конвертовање и сви не захтевају да се пишу посебни рачунарски програми.

а Преклапање ISBN броја са екстерним подацима

Организација која жели да конвертује своје записе може склопити уговор са организацијом која има записе (на пр. COBISS.SR) да добије записе од ње. У случају организација које се укључују у кооперативну базу података то би био веома добар начин. Обично све што је потребно је да се екстрахује листа ISBN бројева из постојеће базе и њихов идентификациони број у бази и да се то дода MARC записима, а затим унесе у циљни сйстем.

б Компликованији начини преклапања екстерних података

У случају старих материјала који немају стандардни број, потребно је преклапања вршити преко аутора и наслова и слично.

в Конверзија квалитетних записа из једног формата у други

Ако имате квалитетне записе, који су обично у стандардном формату, онда је можда исплативо писати програм за конверзију из једног формата у други. Ако конверзију

the UNESCO UNISIST Reference Manual. This format had been promoted in Yugoslavia at a workshop in 1987 sponsored by UNESCO, Institute Vinca and the National and University Library of Croatia which had taken place in Vinca and Zagreb. The CCF was used around the world in the scientific information centres and across all sectors in India.

The main point about all these formats is that they all use the same record structure, ISO 2709 which makes conversion between them a little easier.

There are other non-standard formats. Microsoft Access databases are not standard in terms of bibliographic databases. Unesco's CDS/ISIS package uses a slightly non-standard record structure which causes some problems.

Data conversion is now a large issue as many organizations around the world in certain countries including Middlesex University are converting to MARC21. MARC21 and UK MARC are very similar so the methodologies need not be too complex but if you were considering converting from UNIMARC to MARC21 it could be more complex.

Methods of conversion

Before talking about conversion it is necessary to agree on the terminology: source is used for the original records and target is used for the format into which conversion takes place

There are a number of ways of doing conversion and they do not all require computer programs to be written.

a ISBN match on external data

The organization which wishes to convert its records can contract with an organization which holds records (for example COBISS.SR) to be provided with records. In the case of an organization joining a cooperative database this would be a very good way of proceeding. Usually all that is necessary is to extract from the existing database a list of ISBNs against their accession numbers and these can be added to MARC records and imported into the target system.

b More sophisticated matching on external data

In the case of old material which did not have standard numbers, more sophisticated matching on external databases by author and title etc is required.

c Conversion of quality records between formats

If you have good quality records, which would usually be in a standard format, then it may be worth writing programs to convert between the formats. If you are converting between well known formats there may be some users who have done this already before and who have devised the algorithms. There are also tools to do this such as USEMARCON which is available on the British Library's website. But note that most users of a for-

обављате између добро познатих формата, могуће је да постоје већ неки корисници који су то већ радили и који су урадили алгоритам. Постоје и алати који то раде као што је USEMARCON, који је доступан преко веб странице Британске библиотеке. Али имајте у виду да сви корисници формата креирају сопствене "дијалекте", тако да алгоритми који су већ направљени раније негде другде траже прилагођавања.

Коришћење CDS/ISIS-а за конверзију

На практичном делу радионице господин Хопкинсон је демонстрирао како CDS/ISIS, посебно Reformatting Field Select Tables, може да се користи за конвертовање записа у CDS/ISIS-у из једног MARC формата у други. Он је приказао и како је могуће конвертовати записе EXCEL-у у CDS/ISIS, да би приказао како се и најкраћи нестандардни записи могу ковертовати.

Конверзија сетова словних знакова

Пошто конверзија сетова словних знакова може бити проблем који мора бити решаван у земљама као што је Србија, где се користи проширени сет знакова латинице и ћирилица, предавач је нагласио најзначајније ствари на које треба обратити пажњу и показао је како CDS/ISIS може да понуди нека решења за конверзију словних знакова. Приказао је коришћење CDS/ISIS програма mx.exe, који се може добити са сајта BIREME (http://www.bireme.br), где је CDS/ISIS развијан заједно са UNESCO-м.

Уношење у циљну базу података: остали проблеми

Чак и ако сте произвели фајл са квалитетним MARC записима који желите да унесете у свој систем, треба да будете свесни неких проблема у специфичним областима.

а Одреднице

Многи системи аутоматски креирају одреднице из имена и предметне одреднице када се у њих унесе фајл са MARC записима и када запис са одредницом већ не постоји. Могуће је да ћете морати да радите доста исправки код тако аутоматски генерисаних записа. Они могу бити погрешни ако су рецимо засновани на грешци у куцању.

б Примерци

Многи системи држе податке о примерцима одвојено од библиографских записа. У сваком случају може се десити да нема стандардног MARC поља за тај тип података када се ради пренос из система у систем. У неким системима као што је Horizon, могуће је увести податке у претходно одређена поља која се односе на појединачне примерке и то се аутоматски преносе у регистар примерака. Други системи могу имати нешто слично.

ц Подаци о часописима

Исто се односи на податке о часописима, али мање система има ту могућност. Већина система су тек при првој генерацији развоја у односу на податке о холдинзима серијских публикација, али у будућности то ће сигурно постати један од захтева када институције мењају библиографски софтверски систем на којем раде и желе да пренесу податке о холдинзима из старог система у нови. Сада постоје хол-динг формати за MARC 21 и за UNIMARC.

д Други подаци

Други подаци треба такође да буду конвертовани, као што је фајл о корисницима, а велика је вероватноћа да су ти фајлови у нестандардним форматима и стога ће за њих морати да се праве посебни програми за конверзију.

mat create their own 'dialect' so any algorithms from elsewhere may need some editing.

Using CDS/ISIS for conversion

A practical session took place in which Mr Hopkinson demonstrated how CDS/ISIS specifically the Reformatting Field Select Tables, could be used to convert records in CDS/ISIS from one MARC format to another. He also reported, in order to illustrate how even the briefest non-standard records could be converted, how it was possible to convert EXCEL records to CDS/ISIS.

Character set conversions

Because conversion between character sets can be a problem which has to be solved in the context of a country like Serbia using as it does extended roman and Cyrillic, he outlined the main areas of concern and showed how CDS/ISIS can offer some solutions to the conversion of characters. He demonstrated the use of a CDS/ISIS utility, mx.exe, which is available from BIREME (http://www.bireme.br) who have developed CDS/ISIS along with UNESCO.

Importing into the target: other problem areas

Even if you have produced a file of quality MARC records for import into your system, you need to be aware of certain problems in specific areas.

Authorities a

Many systems create authority records automatically from name and subject headings when a file of MARC records is imported and where an authority record does not already exists. You may have to go and do a large amount of editing on these automatically generated records. They may for example be invalid if they were based on say a mis-print.

Items b

Many systems keep item or copy data apart from the bibliographic records. In any case there may not always be standard MARC fields for this kind of data when transferring between systems. In certain systems such as Horizon, it is possible to import data in a pre-specified field which relates to an individual copy and it is automatically transferred to the item tables. Other systems may have something similar.

Serials data c

The same applies for serials holdings as for item data but fewer systems have provision for this. Most systems are only in their first generation of serials holdings systems but in future it is likely to become more of a requirement when institutions replace a bibliographic software system that they will wish to transfer holdings system from the source to the target. There are now a MARC21 and a UNIMARC / Holdings format

Other data (d

Other data has to be converted such as borrower files and these are more likely to be in a non-standard format and will always need tailored conversion.