Traditional Communication Formats v SGML, Metadata, Dublin Core

Traditional Communication Formats: MARC is far from dead

Alan Hopkinson
Middlesex University
London, UK

1. **The subject for debate**

This section is entitled Traditional Communication Formats v SGML, Metadata, Dublin Core. Is there a subject here that needs to be debated?

I hope to show that the two, the traditional and the more recent formats, are complimentary not adversary. But to do this I need to explain why there is even a feeling that the traditional and the new might be opposed. I need to explain why the question has been asked.

I am going to try and answer two questions: *Do we still need MARC?*; and *Are MARC and the other 'formats' competing for the same role?*

But first, I would like to say a few words about what we mean by 'format' in this context. Format implies a carrier for data between automated systems. The systems may be two library automation systems (perhaps one of which holds a distributed database), or they may be a webserver and a personal computer accessing that webserver, or they may be a personal computer producing a bibliography and a remote computer, a server with bibliographic records.

The format is the set of rules that govern the data structure and content. Some formats govern only the structure, others the content and the structure. MARC uses for its structure ISO 2709: *Format for bibliographic information interchange on magnetic tape* [1]. Because MARC requires this structure as part of its definition, MARC is not MARC without its structure. However, many cataloguers believe they are MARC experts without knowing about the structure! So here is an anomaly. In fact there are a large numbers of anomalies of this kind involved in MARC. I have just said 'interchange on magnetic tape', but almost no one uses magnetic tape today. To be fair, the name of the standard was changed in 1996 so that the latest version of the standard is entitled *Format for information interchange* but most MARC formats have not yet recognised this and because the leading players in the game, the national libraries and the data large catalogue record utilities such as OCLC in the US or BLCMP in the UK, still use tape to transfer records between each other, the methods for dealing with tape are not insignificant.

The term MARC is used generically to cover formats such as MARC, UNIMARC and any other formats which have that common record structure based on the ISO 2709 standard which means they also have 3-digit tags and subfields.

NISO, the US National Information Standards Organization, is in fact working on the preparation of a new standard Z39.80 *Standard format for downloading records*, which will provide a definition for structured bibliographic records to export them from one computer system and import them to another [2]. This will resemble an exchange format without the legacy of tape processing.

We have introduced the term format. The more recent 'formats' are generally called DTDs 'document type definitions'. This term is used in SGML, the *Standard generalized markup language*, ISO 8879 [3]. SGML is a different kind of format and not a traditional library catalogue format. HTML, the 'format' or coding beneath web pages (which is well known because your World Wide Web browser will actually let you view it as the underlying format) is based on SGML. As well as its structure, marked by the 'less than' (<) and 'greater than' (>) signs there are 'codes' such as HTML, HREF= NAME= and conventions such as entering data between quotation marks "mailto:someone@univ.ac.uk".

There is a very important difference between the traditional and the other subjects under consideration: the 'traditional' based on the ISO 2709 record structure are used almost exclusively for bibliographic data, not even including full text of bibliographic documents; in the case of the others, bibliographic data is only one application among many. They are just as likely to be used for full text as for bibliographic references

2 **Do we still need MARC?**

Many years ago in computing there was a fashion to believe that if some equipment was more than ten years old it was outdated. This was particularly true of hardware. In the mid-1970's the minicomputer began to replace the mainframe and by 1995 the Pentium PC was more powerful than the mainframes of the early 1970's. So now a PC will be outdated after a couple of years.

MARC was developed in 1967 and by 1985 there were experts saying it was outdated and would be replaced. In 1998 it is still with us despite new generations of computers. Can we draw any conclusions from this?

It is clear that we need MARC because we use it; nevertheless, we can ask if it could be better, could it be improved? Is it outdated? To answer these questions we need to ask why it is like it is and what could replace it.

2.1 **Introduction to MARC**

MARC was developed as a record structure for exchanging records between systems using the exchange medium available at the time, half-inch magnetic tape. Though today many records are transferred by other media such as floppy disk or transferred across the internet, most individual records are still exchanged in the original record structure though they do not always implement the original record spanning structure which is needed by tape files. Very few systems hold their records internally in the MARC structure though in the early days when most files were kept off-line on tape rather than online in disc systems, records were customarily stored in MARC

In addition to the structure, MARC uses codes for identifying different elements of the catalogue record based in fact on the 'data elements' of the catalogue card, with main entry, title, imprint, date and with added entry for persons, corporate bodies and title, with data elements relating also to subject and classification. This is what the expert MARC cataloguer believes MARC to be.

The universal use of these codes is possible only because there are strict, widely applied standards for the recording of bibliographic data, such as the ISBDs [4]. In fact over the years as exchange of data has increased it has become apparent that even stricter control of data is required and organisations such as the Library of Congress and British Library have worked on projects to harmonise their authority files. The Library of Congress also worked on an early project using the Z39.50 standard [5], the Linked Systems project [6], which enabled automatic updating of records from updated authority data overnight using the Z39.50 protocols.

Since 1985, when librarians began to wonder what would replace MARC, a huge amount of work has gone into the development of library automation systems, originating for the most part in the United States. Similar systems were under development in the UK and elsewhere around that time. However, in the late 1980's, UK suppliers in general when asked at conferences and exhibitions, had not heard of MARC. If they had heard of MARC, they claimed their systems were MARC compatible even though this might mean they could import MARC data into their databases but not extract it. There was a feeling amongst systems suppliers, excluding those which were themselves providers of MARC records, that many librarians did not want the complexity of MARC even though they might want to store the content though not the detail or the record structure of MARC records in their databases. They also tended to suggest that MARC was outmoded and would soon be dead.

In the United States, it was a different story. Almost all systems, even those intended for small libraries, were MARC compatible; they could extract data from MARC records and recreate MARC records from the data in their system to send to other systems. In the UK today, most systems used by libraries are fully compatible with MARC. Many allow the cataloguers to be ignorant of some of the detail of MARC but they tend to be based on MARC underneath. So most systems do not force the use of MARC but they do facilitate it.

Systems based on shared cataloguing in the UK such as BLCMP as indeed OCLC in the US have always required quality MARC cataloguing of their participants and in those circles MARC has never been questioned.

MARC is of course a record structure and a set of identifiers for records, but it relies on a cataloguing code for its data element definitions. US MARC, UK MARC and many other national formats rely on Anglo-American Cataloguing Rules (AACR) [7] for the description of their data elements; those that don't rely on AACR use other cataloguing codes based on the International Standard Bibliographic Descriptions (ISBD's) for the definition of their data elements. MARC in the Anglo-American world would not be like it is without AACR. Similarly, the existence of MARC has ensured that AACR and other cataloguing formats can be represented in library automation systems. So AACR has had a very strong influence on MARC and MARC in turn has had a very strong influence on the development of systems that support AACR, ensuring that their databases are compatible with each other. The other part in the equation is the record structure; the defined record structure which means that the records, however they are held in the system, can be output in a particular structure which can be 'read in' by other systems. Interestingly enough, MARC originally did not cover everything in the bibliographic record as it was originally devised for national bibliographies: in many national formats it is still weak on a standard for items, i.e. copies, as opposed to works and this is evident when data are transferred between systems.

So, it is interesting that the use of MARC has increased since people began to prophesy its death.

It is interesting also to report that the California State University recently set up a project to establish a Unified Information Access System across all member universities and their separate sites. 'After extensive deliberations, the project team concluded that a system based on the open standards of TCP/IP for telecommunications in the Internet environment, HTTP/HTML for Web client to Web server communications, USMARC for bibliographic records and Z39.50 for computer to computer information search and retrieval is the optimum current technology for Universal Online Bibliographic Access. The team received very positive input for this approach from prospective vendors in the RFI [Request for Information] process.' [8]

2.2 MARC advantages and disadvantages

2.2.1 Advantages

- MARC uses the ISO 2709 structure which is a strictly defined record structure with choices of whether to implement subfields and indicators: MARC implements its structure in one of the standard ways; the other three are not used and are mentioned only for the sake of completeness.

- Systems designers have a fixed structure to deal with and they know exactly what to implement, even though it looks unusual to programmers not familiar with bibliographic data processing.

- There are now a huge number of bibliographic records around the world in the MARC format. Most of these systems can exchange records in this format because they include an appropriate software interface. Records can easily be transferred from one system to another because they all use the ISO 2709 format. When a supplier of automation systems upgrades its software, the records can be exported from one and imported into the new system. If anyone wishes to switch from one supplier's hardware to another it is also possible through this means.

- The format is not proprietary, though, as it was designed for the Library of Congress MARC project, it has features which were included because of the way that system was designed and the hardware that was used for it.

- The format takes into account the need for variable length and repeatable fields, which most rigid (and therefore easy to program) record structures do not entertain.

2.2.2 Disadvantages

Some of the disadvantages are perceived though not in fact problematic.

- The record structure is difficult to 'read' compared with other structures which do not have the complex directory structure. It is almost impossible for this structure to be created direct by a cataloguer: a computer program is required.

  This is not a problem as the data structure is intended only for exchange purposes. Users see the records in their systems; the record structure is only used to transfer between systems.

- The record structure is difficult to program; in fact it is not difficult so long as the programmer has the ISO 2709 specification as published in MARC format documentation.

- The record structure requires control characters which are difficult to use (e.g. ASCII 31 for the subfield identifier): internally systems can use other characters which are not used elsewhere such as $, @ or ^ for ascii 31, or even provide data entry facilities (e.g. different screen input windows) to obviate the need for these if it makes usage easier.

- Large files are difficult to manipulate

- MARC does not facilitate links between bibliographic records. Interestingly, a facility was added to the second edition of the standard, ISO 2709-1981, to do just this, but its use has not been taken up because it cuts across cataloguing practices. It was implemented by the UNESCO Common Communication Format [9], a MARC-liek format, but most users of that format have not implemented it. One reason is that records relate to works but links are made through authority files and indexes created from them based on certain data elements in the MARC record which are standardised by intellectual effort. Records stand alone and the intellectual linkages between them are created by extracting from them data elements and placing them in indexes. The UNIMARC authority format which is covered elsewhere in this conference illustrates this. Since the links required by records represent ultimately the links between objects, other methods have been found of doing this such as the use of the 856 field which allows access to other resources such as full text (e.g. the actual object the catalogue record represents or a copy of it in another form or a proxy of it).

- Is MARC really a standard? There are many different MARC formats around the world. This is perhaps the most serious problem with MARC and so I devote a complete section to this.

2.2.3 Is MARC a standard?

The large number of MARC formats negates the value of MARC as a standard; this is a valid criticism. The structure was agreed originally as an American (ANSI) standard (ANSI Z39.2, now NISO Z39.2) then as an ISO standard. When the American standard was agreed, it was also agreed to include the definitions of the data elements as an annex rather than as part of the standard recognising that different sectors of the information community would have different requirements as far as details of data elements were concerned. Of course, within a particular sector whether a format is a formal standard or not does not necessarily matter. But it is interesting that the structure has been agreed on and used universally, whilst the data element definition has not been agreed on to the same extent and has caused problems as follows.

Initially, when MARC was developed in the late 1960's, different national libraries decided they needed to make slight readjustments to the original Library of Congress MARC.

Most of the differences are subtle and not very problematic if one wants to transfer data between a system using one format and one using another: differences such as whether UK MARC '513 Summary note' is the same as US MARC '520 Summary etc'. If they are agreed to be the same, importing data in one format into a system which uses another requires conversion of a tag. Others differences are more subtle, so that for example a personal name in UK MARC is coded $aShakespeare$hWilliam, in UNIMARC $aShakespeare$bWilliam or $aShakespeare,$bWilliam and in US MARC $aShakespeare, William. (Note that the dollar is actually used to represent a special character ASCII 31.) This means that a two character code with its unique meaning in one MARC format is equivalent to a far from unique two character string in the other: 'comma space'. There are other examples of the second kind. There have been moves to harmonise this between the US, Canadian and British formats. Everyone in the UK hoped that harmonisation would mean conformity. In music, from where it originates, the term harmony means 'a simultaneous and successive combination of accordant sounds'

(OED). Harmonisation then means 'to bring into agreement', not to make uniform. We are talking about data of which many users demand a very high level of uniformity. In some cases that is really only justifiable for aesthetic reasons: but in others, such as the alternatives of comma and codes mentioned above, the data has to be very highly specified as it will be used for producing indexes. Conformity and uniformity are required rather than agreement. It will be disastrous for a catalogue if there are multiple sequences of Shakespeare, William in an index caused by the alternative underlying differences:

Shakespeare$bWilliam
Shakespeare, William
Shakespeare,$bWilliam

Automated systems can get round this but they are not usually developed with a view to resolving problems from the use of alternative data standards.

As is stated by R.W. Hill in *Setting the record straight*, [10] 'MARC harmonisation has raised many complex issues concerning the requirements of user groups.' It has been agreed that while US MARC and Canadian MARC have been fully harmonised, there has been only partial harmonisation with UK MARC.

Most of the problems have been caused by differences similar to the ones outlined above. The additional subfield identifier in personal names was introduced into UK MARC at a very early date to give greater flexibility in processing the data. To convert all UK MARC records and the software packages which contain them would be very costly and the flexibility would be lost. In fact the differences are very small compared with the differences between forms of data that you might find if you compared references from different sources or non-AACR catalogues with each other and they could be converted automatically. Nevertheless, the differences are ironic when you consider that the data from US and UK MARC are both AACR compatible and so should be completely compatible with each other.

When data are exchanged only between comparable organisations there should be no problems for data transfer and usually within one country, because all libraries have adopted the national format, there is no problem. It is at the international level where there is a problem. In an attempt to resolve this problem, UNIMARC has been developed as a bridge between national formats. In international projects supported by the European Communities Telematics for Libraries Program, it has been agreed that these differences are important enough to get in the way of the efficiency of the projects, so generally everyone is required to use UNIMARC.

Other speakers will cover UNIMARC so I will not talk about it other than to say that the idea is that if everyone sends data between different countries in UNIMARC, each country only needs one set of programs to convert into and one to convert out of its format.

This would not help the user at home with access to the Internet who finds a record in an unfamiliar format on the World Wide Web and whose computer program does not know that format. Data transfer at that level is better accomplished by using the newer formats.

Such a user, however, is also inconvenienced when doing cross systems searching. This kind of searching is facilitated by systems using standards such as the Z39.50 standard. The Z39.50 standard depends on attribute sets which are concerned with indexes. Different catalogue systems enable or necessitate different selections of MARC fields for the indexes. To search across more than one system with different methods of indexing requires a standard and one of the functions of Z39.50 is to do standardise just this. Having obtained the bibliographic records you are looking for you need to see the records. Z39.50 requires that the system sends back a MARC record. The searcher's software therefore needs to know which MARC format is being used to process the record. However, because the record structure is standard, a program that can read one MARC format can read another and should be able to *display* all the data though it might not be able to give it the correct label or punctuate it correctly. Consequently, a large amount of development work is going into the parsing of records in different MARC formats for systems implementing Z39.50. It is worth mentioning as an aside that even if MARC had been more consistently employed but systems had still created indexes from different MARC fields, we would still have needed the coded attribute sets of Z39.50. This is an illustration that the data (though standardisation has been achieved there at a very high level) is not the only part of a system that needs to be considered and that the indexes for which there has been less standardisation also have to be taken into account.

So here we have uncovered the biggest weakness of MARC, its shortcoming of having different national versions which, bearing in mind the universality of AACR (today if not when MARC was developed), is not excusable and ought to be resolved. But it is not a necessary weakness of MARC. Philosophers talk about possible worlds and it is conceivable that there could have been a world where MARC uniformity had been achieved across the library world.

This leads me to believe that though the MARC format has its faults and, like many standards, is not used as it should, the situation would have been much worse had we not had any MARC format at all. It is tempting to think that we can replace it with another standard format. But since the record structure is not the culprit and is indeed one of its strengths, changing that should make no difference. If we cannot get agreement on the data element definitions and representation, is there any reason to believe that we would gain agreement on any other standard or that if we did we would continue to use it 'harmoniously'?

2.2.4 **Exchanging data in the 21st Century**

MARC has been put in place to standardise data because standardisation is needed for data exchange and MARC is an exchange format. In fact, libraries need standards for their bibliographic records for another reason: to provide a reasonable level of internal consistency to ensure that they are comprehensible by different users. This is facilitated by MARC but provided ultimately by the use of a cataloguing code. Exchange used to mean cooperative cataloguing, but today data are also transferred when systems are migrated. Migration is not the prime reason for an exchange format but the ease of migration because of the existence of such formats is a good by-product of the exchange format. When systems are migrated the catalogue data though the most complex compared with say borrower files are usually the data which is most successfully and efficiently converted.

However, users are now wanting to access data from different systems in a different way. People are talking about standards for INTERNET, and the World Wide Web. Is this a reason for not using MARC any more? That MARC is an unsuitable structure for the World Wide Web is a common concern of technical people who are not really sure of the art of cataloguing. Other proposals are made but none really addresses the development of an underlying standard which is close enough to a data element directory to serve in the way MARC has done to standardise millions of bibliographic records around the world. If we were developing a standard now, we would probably not have three digit tags and subfield identifiers. We could have something more complex, but we would need to remember that data entry has to be done in a way that is reasonable for the cataloguers of this world, and what we have, which is what has evolved, is a flattish structure based on a catalogue card which has had added to it links to the bibliographic records and to authorities.

However, this structure and data element definition has put us in good stead to be able to present data in these other formats which are not communications formats in the same sense.

I asked earlier whether MARC can be improved: the different national formats are evolving all the while but major developments such as logical methods for record linking have always evolved more slowly. Stability has been given prominence over change, but as I say in my conclusion, this may be no bad thing.

3. **Are MARC and the other formats competing for the same role?**

This brings us to the next question that I posed at the start: Are MARC and the other formats competing for the same role?

In this paper, given the time available, I can take only one, and I have chosen Dublin Core as an example of a 'non-traditional' format.

I have not needed to use the word metadata so far though I have been talking exclusively about metadata. To find out what it means, what better place to start then David Stoker's *A beginner's guide to metadata* [11]. He writes: 'According to the ADAM Quick guide to metadata <URL http://adam.ac.uk/adam/metadata.html> "the most common definition of the term ... is data about data - information that describes other information". Thus any form of catalogue, contents list, inventory, review, abstract or index is a collection of metadata, so long as it may be used to describe other sources or documents'. He could have included MARC records. Note that the 'title page' is not mentioned. Dublin Core is used as a kind of identification for machine-readable material in the same way as a title page is used for printed materials. The interesting point to note is that because it is describing data on the internet and is itself accessible to index creators accessing the Internet, data in the same form (though not in the same logical or even I suppose physical location) can constitute an index to that data. This is no different than taking the old full CIP entries from the title page verso and gluing it on a catalogue card. Data in Dublin Core on a web page may not actually be filling the role of metadata, but its purpose is to make the generation of metadata easier.

Michael Day of UKOLN has prepared a study [12] for the European Communities Telematics Applications Programme Project 'Biblink: Mapping Dublin Core to UNIMARC'. The main conclusion of this for us (as we compare Dublin Core with UNIMARC) is that the definition of UNIMARC is such that Dublin Core data cannot be converted into UNIMARC. Main problems concern the absence of a distinction between personal and corporate author in Dublin Core, and the creation of coded data. He concludes that some form of manual intervention would be required but this would be costly. My view is that if we want to put into a catalogue records describing computer resources deriving them exclusively from html metadata, we are trying to do a better job than cataloguers of traditional materials. The html metadata headers are akin to the 'description' which ISBD regards as the data transcribed from the item in the form and content as on the item. It has always been the case that cataloguers have wanted to index collections by subject, according to local requirements. It is unlikely that the subject terms added to electronic documents by their authors would be universally adequate for their indexing. Of course Dublin Core can be extended and may have already been extended by now to increase compatibility with UNIMARC. But UNIMARC and the other MARC formats are tools for holding metadata of two kinds: that which has been created from the original and that which has been created with intellectual assistance, what librarians call authority data. Though librarians subject-indexing material to which they intend to give access will no doubt take account of subject indexing they find in Dublin Core data, they will not wish to use it uncritically unless it has been created by a central cataloguing agency such as a national library.

Day's study was intended to show to what extent Dublin Core data could be converted to UNIMARC in order to enable institutions wishing to catalogue their resources of all kinds to take advantage of the Dublin Core 'identifiers' to produce catalogue records which would be incorporated in a catalogue of records of more general materials. The reverse exercise is probably of no practical use, though I wrote a program to do this as part of my preparation for writing this paper: there is no reason to convert (meta)data in UNIMARC to HTML metadata. No one is going to design the title page of a book from its catalogue card. No one is going to find it convenient to retrieve a record from a catalogue, convert it to html and place it in a record of a document as a Dublin Code header.

4. **Conclusion**

One has to note first that MARC has led to some fossilisation of bibliographic data processing which has provided a stable environment for the development of systems; this has only been possible because of the continued use of AACR and a lack of change in that area. But if we had not had a standard so universally accepted (albeit with slight variations) we would not have had a stable environment for the local, national and international networks, systems and hardware packages that we see today which enable us to exchange data and which we probably take for granted. Any standard has a built in anomaly: standards stifle innovation but they also enable all users to reach the same platform or plateau from which they can all jump together to a higher level. We have not yet jumped but we are on the same level!

Thus I think I have shown firstly that:

1. MARC still serves a useful purpose in promoting building of catalogues, despite its failings. Its failings have come about partly because the standard has been used for a long time and scarcely changed, whilst the media for and modes of data transfer have changed. The standard could have been changed but this would have probably led to greater variation of practices than what we have now. Also it would have caused problems that we do not have now that we have huge files of bibliographic records from which anyone can draw. The alternative could have created separate systems at different levels of development. Library automation system suppliers would have had to take into account many different formats. Now they can develop their system

interfaces using ISO 2709 and tailor slightly their systems to take into account national differences of MARC format.

As an aside, despite the problems I have outlined which to some might seem serious problems, most humble systems librarians working to provide a system for their libraries' end users and providing access for their cataloguing colleagues to external resources need know very little about the problems I have outlined. However, the rigidity of the MARC structure internationally and the MARC data definition nationally have ensured that data transfer between systems for system migration and cooperative cataloguing have been much more successful than they would have been without a standard like MARC.

2. MARC serves a different but complementary purpose from the later generation of formats such as Dublin Core.

References

1. International Organization for Standardization. *Format for Bibliographic Information Interchange on Magnetic Tape.* Geneva, ISO, 1981 (ISO 2709-1981). 3rd edition: *Format for Information Interchange.* Geneva, ISO, 1996 (ISO 2709-1996).

2. NISO. *Standard format for downloading records (NISO Standards Committee AJ).* [WWW] http://www.niso.org/commitaj.html (27.04.98).

3. International Organization for Standardization. *Standard generalized markup language*. Geneva, ISO, 1986

4. ISBD. The International Standard Bibliographic Descriptions are cataloguing codes dealing with the description of documents. The main one on which the others are modelled is *ISBD(G): General International Standard Bibliographic Description*. Rev. ed. Munich, K.G. Saur, 1991

5. *Information retrieval (Z39.50): application service definition and protocol specification*. Washington DC, Z39.50 Maintenance Agency, July 1995.

6. Dempsey, Lorcan. *Libraries, networks and OSI.* Westport, Meckler, 1992

7. *Anglo-American Cataloguing Rules.* 2nd ed., 1988 rev. London, Library Association, 1988. Looseleaf, updated.

8. Pollard, Marvin. California State University Unified Information Access System: project report. [WWW] http://uias.calstate.edu/uiasproject.html, 9 July 1997 (27/04/98)

9. The latest version of CCF is defined in: *CCF/B: the Common Communication Format for Bibliographic Information.* Paris, UNESCO, 1992

10. Hill, R.W. *Setting the record straight: a guide to the MARC format.* 3rd ed. Boston Spa, British Library National Bibliographic Service, 1998

11. Stoker, David. *A beginner's guide to metadata. CIQM Newsletter* 3(1) January 1998

12. Day, Michael *Mapping Dublin Core to UNIMARC.* [WWW] http://www.ukoln.ac.uk/metadata/interoperability/dc_unimarc.html (25/04/98)