

Data quality – what is it and does it matter?

Cathie Jilovsky
CAVAL Collaborative Solutions

ABSTRACT: The information systems used in libraries and information centres today range from websites and general databases to specialized software applications such as Integrated Library Management systems. Although a number of factors contribute to the success of these systems, their usefulness ultimately depends on the reliability of the data being stored, processed and displayed. This paper will examine a range of data issues, including standards, data conversion, storage and transport formats; the relationship of the data layer to other parts of the system architecture; the usability of the system and data quality.

There are a great variety of information systems, from simple to extremely complex, used in libraries and information centers today. It is not uncommon for people to say that “it doesn’t work” or “the system is no good” or the “data is of poor quality”, at times with considerable frustration and emotion. These are sweeping statements and what I would like to attempt to do in this paper is to look ‘under the bonnet’ of such systems and to provide some analyses of what is really happening. My experience is that such analysis combined with an understanding of the system components and the interaction between them is an important first step in toning down the emotion and addressing the frustration. There are times when the data is indeed of very poor quality so that even the most wonderfully clever system in the world can’t produce the expected results or even any useful results, but there are other times when the data is just fine but the system simply can’t find it or get to it. It can make a considerable difference to understand what is going on, and then move on to determine what can be changed and what might be better addressed through other means.

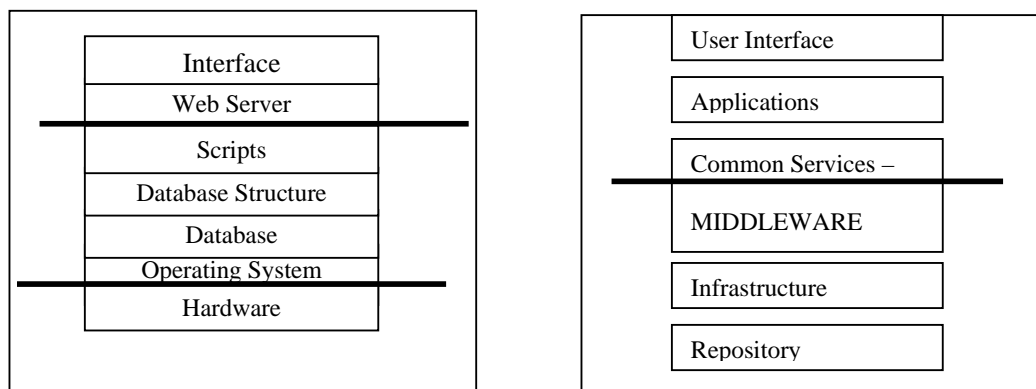
The information systems available today range from websites and general databases to specialized library software applications, commonly known as Integrated Library Management (ILMS) systems. Early library systems, running on micro-computers were described as turnkey systems. The expectation was that a library bought a system, turned it on, loaded some data and off it went (Jilovsky, 2003, p. 298). The current generation of Integrated Library Management systems cannot be classified as turnkey systems, rather they can be more accurately described as a sophisticated set of bibliographic tools which can be configured and implemented to suit local library requirements in almost limitless permutations. This wonderful sophistication has a downside – complexity. Our twenty-first century systems are powerful and extremely flexible – but they are often complex to understand, complex to specify, complex to configure and complex to operate. In this environment a number of factors will contribute to the usefulness of the system, however ultimately success will depend on the reliability of the data being stored, processed and displayed. The environment itself has also broadened considerably, incorporating Google and other search engines, and with users expectations of ‘finding it all on the web’.

The last decade has seen a steady move away from ‘stand-alone’ systems to highly sophisticated systems that inter-operate with a variety of other similar and dissimilar systems. These may be authentication systems, e-commerce systems, e-learning systems, repositories, scientific data, amongst others. The development and implementation of standards used both within and between these systems has become essential. Computers are becoming faster, smaller in physical size but much greater in capacity and more robust for the same price or even less. Whereas only a few years ago allowing for storage capacity limitations was an important part of database design, now it is relatively easy and cheap to simply increase the disc capacity so that not only size, but also the necessity to design

space-efficient structures, is no longer an issue. Add a smart search engine and what may have previously been regarded as deficiencies in the data no longer matter.

System architecture

Computer system architecture is commonly represented as 'layers'. The process of determining why a user is not able to obtain meaningful results can usefully begin with an examination of the system components. Here are two representations, showing typical components, presented in 'layers' (Kochtanek, 2002, pp 89-92).



The left hand diagram represents a typical ILMS. A user interacts with the system by using the interface, which is commonly a web browser. In this diagram there are 4 layers between the user and the data – so even on the assumption that the user interacted with the interface appropriately, if the communication between the interface and the web server, or between the web server and the scripts, or between the scripts and the database structure, or between the database structure and the database, is not right then there is no way the user will get those expected results. On the other hand the system layers may all perform and interact as required, but if the data is not in the expected location or format then the system may appear 'not to work'.

The right-hand diagram includes a "Middleware" layer that sits between applications and operating systems and network. Middleware is the infrastructure that allows construction of scalable, reliable, secure IT systems; that are widely recognized as essential ingredients for a seamless internet-based application environment (Dalziel, 2004).

Quality Issues

The quality of the underlying data is certainly a significant factor in the successful operation of systems. In library systems alone this may include bibliographic data, holdings data, financial data and patron data. Within system architecture models, such as the ones illustrated above, this is described as the data layer or it may be described as a repository. If this data sits within several un-connected systems, then these are often called data silos e.g. the same information about customers or products may be repeated in several different systems, making maintenance and interoperability difficult (Webster 2004).

Often quality issues can be traced to gaps between user expectations and actual search results. The Centre for Information Quality Management (CIQM) which was set up in the UK about 10 years ago, attempted to bridge this gap by offering a clearing house for database users who identified faults and methodologies for providing database quality assurance (Armstrong 1996).

The National Library of Australia announced a Quality Improvement Plan in its report to the Kinetica Annual Users Group Meeting in 2004. This plan will address data quality issues for

the National Bibliographic Database and the Kinetica CJK database. Components include the removal of duplicates, cleaning up new records added to the databases, fixing of incorrect codes and reviewing factors that will improve the success of searching for all users. The NLA expects that this quality improvement will improve access to Australian collections and assist the whole library network (NLA, 2004).

Data Quality components

Xu nominates the components of data quality as accuracy, timeliness, completeness, and consistency. He suggests that factors that influence data quality include training, senior management support, organisational structure and communication, the management of change, and employee/personnel relations. Finally, he makes a number of recommendations, including the importance of understanding data quality issues when implementing systems, and providing sufficient training in the content and usage of the system (Xu 2002).

Scalability has always been important in the wider IT environment but has not always recognized as an issue by libraries. This is a component in the development of middleware services that are common across a number of systems, some of which may be library or information systems.

Data issues – standards

Standards are important for reasons that include the optimization of interoperability between systems, the facilitation of information exchange across domains, the sharing of common services, the development of scalable solutions, the provision of an orderly base for infrastructure developments, the ensuring of migration strategies, and the maximization of return on IT investment (McLean 2004).

There are many standards that may be incorporated into library and information systems today. These include the OpenURL, DOI, Standard Address Number - SAN, Serial Item and Contribution Identifier - SICI, Holdings Statements for Bibliographic Items, Dublin Core Metadata Element Set, PAMS, ONIX, MARC21, AACR2r, ISO-ILL, ISSN and many others (NISO, 2004). Maxine Brodie, Chair of the Standards Australia IT-19-S0003 Committee suggests that in 2005 the ones to watch will be Directories (ISO 2146), Web services – VIEWS (Vendor Initiative for Enabling Web Services), SRU/SRW, ONYX, METS, FRBR, DOI, OAI, URI and Sharable Content Object Reference Model (SCORM) (Brodie 2004).

Vendors implementation of standards into their products varies in timing and degree, and some of the more complex standards have a range of levels at which they can be implemented. The commitment to implement standards into software products also varies - there is potentially a tension between full interoperability and maintaining differentiation between commercial products.

Standards will also evolve over time. For example, the International Standard Book Number (ISBN) (ISO Standard 2108) has been, since 1970, the identification system for the book trade, publishing industry as well as libraries. It has generally proven a reliable and efficient method of searching for known items. The ISBN standard has now been revised, increasing the length of the ISBN from 10 to 13 digits. All software used will need to be updated to allow for the storage and validation of 13 digit ISBNs.

The Z39.50 standard is evolving. ZING (Z39.50-International: Next Generation) covers a number of initiatives by Z39.50 implementors to make the intellectual/semantic content of Z39.50 more broadly available and to make Z39.50 more attractive to information providers, developers, vendors, and users (Zing 2004).

Data Issues - Matching

When entered accurately and in consistent formats ISBNs and ISSNs (and other numerical fields) usually work very well as finding or matching fields in bibliographic data. However when one of these numbers is not available then searching and matching is more difficult. For example, the VDX Document delivery/inter-library loan software searches for a bibliographic record with which to populate an inter-library loan request. The system is commonly configured to match on ISBN or ISSN. If one of these numbers is found in the record the match is very accurate, thus facilitating un-mediated services, however if neither is available then other matching options, such as author or title fields are much less reliable.

Accurate holdings data and item availability status are also key components of successful auto-authorisation of inter-library loans requests. The 2001 Australian Interlibrary Loan and Document Delivery Benchmarking study showed that where holdings are accurate and the supplier is selected on the basis of the holdings, then 80% of items are supplied within the first 2 suppliers (NRSWG, 2001). The holdings data in Kinetica does not always exist or is in summary form only for many libraries e.g. CARM records contain "held" without any details. An automated system cannot determine whether or not to request an inter-library loan without more details. Often the individual library catalogue will contain more details (the CARM catalogue certainly does), so this will be more accurate.

Authority control is a library device for reconciling variant forms, but for it to work well it is essential that records are consistently as complete as possible. Librarians, especially cataloguers, automatically remove punctuation when searching. Many library systems assume this. But users don't know this and why should they? Dublin Core metadata contains around fifteen fields and is the basis of many usable systems, such as Picture Australia (Picture Australia, 2004). The MARC format, on the other hand, extremely comprehensive with 80 plus fields, was developed when searching was entirely dependent on structured data.

Duplicates often create problems for library staff and for users. These may be duplicate bibliographic records, duplicate item or holdings records, or duplicate transactions or inter-library loan requests. e.g. a library being charged for articles not needed is a waste of resources.

If records in a bibliographic utility contain typographical errors, then these errors may spread into the catalogues of any libraries that use these records for copy cataloguing. Beall and Kafadar undertook a study of such records in 2002, and found that 35.8% of records were corrected i.e. the other 64.2% still contained the errors. For example, if a user is looking for a particular work by Shakespeare, and the author heading is erroneously spelled as "Shkespeare, William, 1564 – 1616", then the work will not be found. This builds on earlier work done by Beall – particularly the "Dirty Data test" (Beall, 1991). A website listing common mis-spellings grouped by the probability of being mis-spelled is maintained by Terry Ballard (Ballard, 2004).

Future Trends

The 2003 OCLC Environmental Scan: Pattern Recognition report was produced for OCLC's worldwide membership to examine the significant issues and trends impacting OCLC, libraries, museums, archives and other allied organizations, both now and in the future. The scan provides a high-level view of the information landscape, intended both to inform and stimulate discussion about future strategic directions. The "Technology Landscape" section identifies several trends that are being driven not only by the library and information community but by business and government communities worldwide. These include bringing structure to unstructured data, distributed, component-based software and the development of techniques that aim to help searchers find 'what they really want' (OCLC, 2003).

The creation of a standardized, cross-searchable, and interlinked online environment will provide a platform for users to simply and directly search the widest range of material to find relevant and good quality content. It is becoming clear that, perhaps paradoxically, in order to provide what appears as a simple search environment for a user, a very complex technical infrastructure must sit behind it (Webster, 2004).

Most people regard software as a tool for individual users, such as Word, Powerpoint, or Photoshop. These tools treat the computer as a box, or a self-contained environment in which the user does things. However when users are surveyed about what they actually do with their computers, social interaction activities top the list – including conversation, collaboration, playing games, and so on. “The practice of software design is shot through with computer-as-box assumptions, while our actual behavior is closer to computer-as-door, treating the device as an entrance to a social space” (Shirky 2004).

Gerry McGovern states that “understanding where technology is strong and where people are strong is an essential skill of the modern manager. Too often today, technology is doing things that would be better done by people”. He endorses using technology to automate mundane and repetitive tasks so that we can allocate our time to do things people do really well, such as human relationships. He quotes Porter Goss, the head of the CIA, who is after less focus on technological wizardry, and more on “humint” (human intelligence). (McGovern 2004)

Does Quality Matter?

The title of this paper contains two questions. A number of aspects of and answers to the first question, what is data quality, have been explored in depth. The simple answer to the second question, does it matter, can be simply answered as ‘it depends’. Data quality in some senses is much less critical than it was in the past, as the tools to access and use the data have become much more sophisticated. The focus has moved from good search results being dependent on underlying data structures, to the provision of user-friendly interfaces that use clever search engines i.e. from a backstage perspective to a front-of-house view.

REFERENCES:

Armstrong, C.J. (1996), “Databases: fit for use or fit for us?” *Library Management* Vol. 17, No. 2, pp. 40–42.

Ballard, Terry (2004). “Typographic errors in library databases”. <http://faculty.quinnipiac.edu/libraries/tballard/typoscomplete.html>. Accessed 10 December 2004.

Beall, Jeffrey (1991). “Dirty Data Test.” *American Libraries* Vol. 22, No. 3, March 1991, pp 97.

Beall, Jeffrey and Kafadar, Karen (2004), “The Effectiveness of Copy Cataloging at Eliminating Typographical Errors in Shared Bibliographic Records.” *Library Resources & Technical Services*; April 2004, Vol. 48, No. 2, pp 92 – 101.

Brodie, Maxine (2004). “Where to now with Standards?” Presentation at Standards Australia IT-019-S0003: AS/NZS Seminar Computer Applications - Information and Documentation, Sydney, 25th November 2004.

<https://committees.standards.org.au/COMMITTEES/IT-019/S0003/IT-019-S0003.HTM>
Accessed 10 December 2004.

Dalziel, James (2004). "Towards an evolving middleware framework for Australia"
<http://www.aarnet.edu.au/events/middle/2004/ref/Dalziel.EvolvingMiddlewareFramework.doc>
Accessed 10 December 2004.

McGovern, Gerry (2004). Web content management solutions
<http://www.gerrymcgovern.com>. Accessed 10 December 2004.

International ISBN Agency (2004)
www.isbn-international.org/en/download/implementation-guidelines-04.pdf. Accessed 10
December 2004.

Jilovsky, Cathie (2003). "Systems Librarianship in Australia: a historical perspective". Library
hi-Tech, Vol. 21, No. 3, pp 297- 308.

Kochtanek, Thomas and Matthews, Joseph (2002). 'Library Information Systems: From
Library Automation to Distributed Information Access Solutions'. Westport, Conn.: Libraries
Unlimited, 2002.

McLean, Neil (2004). "E-Learning and Libraries: Standards in Context". Presentation at
Standards Australia IT-019-S0003: AS/NZS Seminar Computer Applications - Information
and Documentation, Sydney, 25th November 2004.
<https://committees.standards.org.au/COMMITTEES/IT-019/S0003/IT-019-S0003.HTM>
Accessed 10 December 2004.

National Information Standards Organisation (NISO) website. <http://www.niso.org/standards>.
Accessed 10 December 2004.

National Library of Australia, Kinetica Quality Improvement Plan (2004)
<http://www.nla.gov.au/kinetica/manuals/nbdqual.html>. Accessed 10 December 2004.

National Library of Australia Kinetica Cataloguing Standards website
<http://www.nla.gov.au/kinetica/standards.html>. Accessed 10 December 2004.

National Resource Sharing Working Group (NRSWG) Interlibrary Loan and Document
Delivery Benchmarking Study (2001).
http://www.nla.gov.au/initiatives/nrswg/illdd_rpt_sum.html. Accessed 10 December 2004.

OCLC Environmental Scan (2003). <http://www.oclc.org/membership/escan/>. Accessed 2
December 2004.

Picture Australia (2004). <http://www.pictureaustralia.org/metadata.html> Accessed 10
December 2004.

Shirky, Clay (2004). "Clay Shirky's Writings About the Internet"
http://shirky.com/writings/group_user.html Accessed 9 Dec 2004.

Webster, Peter (2004). "Breaking Down: Information Silos: Integrating Online Information",
Online. Medford: Nov/Dec 2004, Vol. 28, No 6, pp 30-34.

Zing (Z39.50 International: Next Generation) (2004)
<http://www.loc.gov/z3950/agency/zing/zing-home.html>. Accessed 2 December 2004.

Xu, Hongjiang, Nord, Jeretta Horn, Brown, Noel and Nord, G. Daryl (2002). "Data quality issues in implementing an ERP". *Industrial Management & Data Systems*, Vol. 102, No. 1, pp. 47-58.