Care and conservation of manuscripts 15

Care and conservation of manuscripts 15

Proceedings
of the fifteenth international seminar held at
the University of Copenhagen 2nd–4th April 2014

Edited by
M.J. Driscoll

Museum Tusculanum Press · 2016
University of Copenhagen

Care and conservation of manuscripts 15

© Museum Tusculanum Press and the authors, 2016
Proceedings of the fifteenth international seminar held at the University of Copenhagen 2nd–4th April 2014

Edited by M.J. Driscoll

Typeset by Florian Grammel

Published by Museum Tusculanum Press,
University of Copenhagen 2016
Printed by Frederiksberg Bogtrykkeri
ISBN 978-87-635-4433-7
ISSN 1602-1339

Published with financial support from
Nordisk Forskningsinstitut, University of Copenhagen

The articles in these Proceedings have all been subjected to peer-review.

The programme and abstracts from the 15th seminar can be viewed on the seminar website: nfi.ku.dk/cc/

Museum Tusculanum Press
Birketinget 6
DK-2300 Copenhagen S
Denmark
www.mtp.dk

The vignette on the cover is taken from the allegorical representation of Professor Arnas Magnæus/Árni Magnússon's scholarly activities which adorned the oldest series of publications of the Arnamagnæan Commission 1773–1809.

Contents

Errata (per oculos) corrige: Visual identification of meaningless data in database records of bookbinding structures

by Alberto Campagnolo

Ligatus Research Centre,
University of the Arts London

Introduction

The resiliency of any system has to take into consideration human reliability and the effects of the human factor on the desired outcome and its accuracy.[1] When inaccuracy is objectively determinable, it can be expressed as error.[2] Errors are indicated by the fact that a planned sequence of activities fails to achieve the intended outcome, without intervention of external factors. Human errors can be broadly classified into three main categories: a) skill-based errors or lapses linked to attention or selection failures; b) ruled-based mistakes, linked to the misapplication of rules; c) knowledge-based mistakes, linked to inaccurate or incomplete mental models.[3]

In electronic databases, the encoding schema at the base of their structure allows for immediate monitoring of data correctness and completeness during input.[4] Missing or inadmissible data is highlighted straightaway by the computer, prompting the compiler to add or correct it. Data validation yields to a reduction in errors and acts as quality control. Not all mistakes can be impeded through careful database design, however, and those errors that do occur are not easy to identify through automated means.

The task of checking for meaningfulness of data still lies with the compiler or a subsequent reader/editor. But due to the limitations of the human working memory system[5] and the fragmentation of the information within the dataset, the synchronic data analysis necessary to check for meaningfully correct data is unmanageable for the human mind alone.

Data validation

Data validation is the process of ensuring that a dataset is complete, correct and meaningful. Validation rules check for correctness or meaningfulness of data that are input by the user.[6]

Just as in language, a dataset can be considered valid when it satisfies the validation rules put in place in the system, but this does not necessarily mean that it is also meaningful, so one should not confuse the notion of 'grammatically correct' – or 'valid according to the validation routines' – with 'meaningful'.[7]

Consider the following sentences: (i) *colourless sewing passing through four stations*; (ii) *stations through passing four sewing colourless*. Both are nonsensical, but (i) would be recognised as grammatically correct by any English speaker. One can, in other words, distinguish between two senses of meaningfulness or validity. A statement that is 'valid in the first sense' is meaningful in as much as it follows the rules of the language in which it is expressed – e. g. it follows the rules of sentence formation set by English grammar. A statement that is 'valid in the second sense' is meaningful in as much as it makes sense in the context in which it is used.[8] A statement can therefore be meaningful in the first sense – i. e. make grammatical sense – but meaningless in the second sense – i. e. in the context in which it has been used.

In the same way, data within a database can be valid – grammatically correct – but nonetheless meaningless. Data that is not 'valid in the first sense' can be avoided through validation routines. Ambiguities due to human error that cause 'invalidity in the second sense' are, instead, not avoidable through validation routines.

Validation and visualisation of bookbinding structure records

The Ligatus Research Centre of the University of the Arts London has been involved in a project to record the bookbinding structures of the volumes from the Library of the Monastery of Saint Catherine at Mount

Sinai, Egypt.[9] In the first phase of the project, which covered the description of the manuscript holdings, the data was collected on paper forms and then automatically input into a relational database.[10] Because the data was first recorded on paper, it was not possible to implement data validation procedures during the survey, which inevitably resulted in errors in the dataset.[11]

In the second phase, a descriptive schema for bookbinding structures was developed utilising eXtensible Markup Language (XML)[12] technologies.[13] In 2007, this schema was used to study the bookbinding structures of the printed book collections.[14] During this part of the project, the surveyors input the data directly through electronic forms generated according to the XML schema. In addition to the electronic forms, the surveyors also sketched on an A3 paper form – subsequently scanned and added to the database – drawings of some of the structures described. The electronic forms allowed for the implementation of data validation protocols directly during the survey. This had the effect of reducing significantly the number of errors in the dataset compared to the previous paper-based survey.[15]

More recently, the present author has been working towards a methodology for automatic transformation of the XML bookbinding structure descriptions recorded during the 2007 survey into Scalable Vector Graphics (SVG)[16] diagrams. These automated visualisations have many advantages, among them standardised output and synchronic view of data for each structure. They can also offer better accuracy in the survey data, as they can provide verification of the meaningfulness of data during the survey.

Uncertainty is inherent in any dataset. As is often the case in the cultural heritage field, one cannot be certain that a binding structure is or was what it seems to be. It could be, in fact, that the structure to be described is not clearly visible – e.g. spine linings – or that the book is in such poor condition that it is now difficult to discern its original structure. The XML schema, which utilises multi-value logic[17] (true, false, unknown, not checked, other) permits the expression of the inherent uncertainty of the data. In this way uncertain and incomplete data can be accommodated for and flagged through graphical means in the automatically generated diagrams. The presence of data which is 'valid in the first sense', and therefore allowed by the validation routines, but 'meaningless in the second sense' is not easily identifiable through automated means, however.

It became apparent during the course of the visualisation project that these types of errors could however be identified if the automated visualisations were implemented directly during the survey.

Errors in the dataset

Based on the empirical experience accrued during the developing phase of the visualisation project, it is proposed here that a system capable of automatically generating diagrams can be used as a visual method to check for validity and meaningfulness.

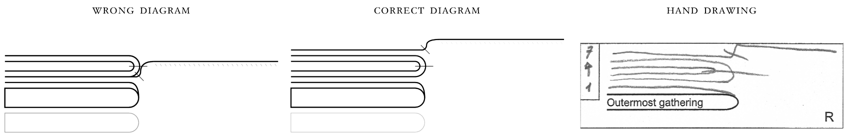
As we have seen, data can be valid but meaningless. One could foresee the surveyor or a subsequent reader/editor going through the data in the database to check for its correctness, one element at a time, diachronically and in sequence. However, the information describing each binding structure is divided into a series of elements and parameters, which can span multiple description levels. The amount of information needed to be kept in mind to visualise the data and analyse it synchronically exceeds the limited capacity of the human working memory.[18] Mistakes therefore easily slip through the control net and remain unchecked.

During the development of the visualisation algorithms, diagrams would occasionally show structures that are not possible in real life. Some of these were obviously due to coding problems in the algorithms which needed to be modified. Others were the result of something rather different, however: the coding algorithm was functioning properly showing exactly what had been encoded in the XML binding descriptions, but the dataset contained errors, and so resulted in odd-looking diagrams.

Error examples

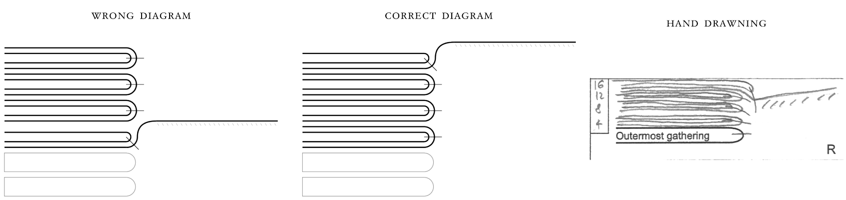
Let us consider some examples of errors found in the dataset. These can be divided into three main groups. There are cases in which the surveyor misinterpreted the description rules and conventions set by the schema. In other cases, typologies were not understood and were used inconsistently. In addition to these, there are obvious slips in which one option in a list was mistakenly chosen instead of the right one. Given below are four examples of errors; for each, the incorrect and the correct diagram as well as the hand drawing for the structure carried out during the survey are shown.

Fig. 31. – *example 1. Error due to schema convention misinterpretation. Right endleaves for volume 4725.3162.*



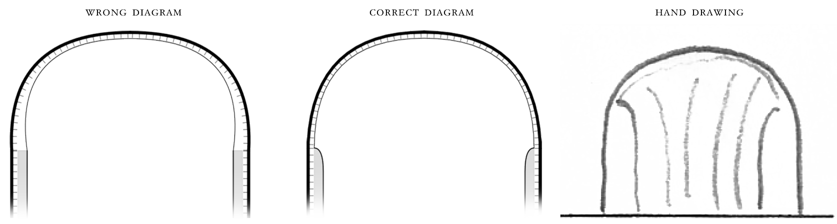
In example 1 (Fig. 31) the surveyor did not encode the structure correctly. According to the schema, an endleaf structure can be of two types – *integral* or *separate* – but then, different groupings of separate endleaves should constitute a series of units encoded within the same type of endleaves. In this instance, the endleaf structure was instead described as being composed of three types of endleaves, two of which are *separate* and composed of one unit each. The correct diagram in Fig. 31 was generated by encoding the description appropriately – i.e. one *integral* endleaf and one group of *separate* endleaves constituted by two units. Admittedly, this kind of problem could be avoided through data validation by rendering invalid the repetition of the same endleaf type within one structure.

Fig. 32. – *example 2. Error due to description convention misinterpretation. Right endleaves for volume 5365.3471d.*



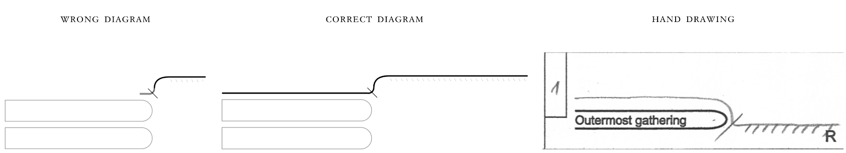
If an endleaf structure is formed by more than one unit and component, the surveyor was instructed to describe and number them counting from the outside towards the textblock at each end. This convention was not followed consistently. In example 2 (Fig. 32), for example, the surveyor opted to describe the structure starting from the textblock outward instead. The elements of the structures being described are all there, but the inconsistencies in their order create problems in the interpretation of their spatial arrangement. An incorrect unit order leads to a nonsensical diagram; tweaking the data to follow the numbering convention generates a correct diagram.

Fig. 33. – *example 3. Error due to inaccurate encoding, possibly because of typology misinterpretation. Spine shape and spine lining for volume 5.2β.*



Example 3 (Fig. 33) shows a spine lining diagram for which the spine shape joint typology were inaccurately selected. This is possibly due to a misunderstanding of the joint categories set by the description schema. In the dataset the joints are described as *slight*, which leads to an odd looking diagram. However, the hand drawing shows that the surveyor picked the wrong joint type. The distinction between the various abstract joint types can be difficult to appreciate. The volume had, in fact, *quadrant* joints. By changing the joint typology and regenerating the diagram, the shape coincides with what had been drawn, and this eliminates the abnormal gap between the bookblock and the lining at the joints.

Fig. 34. – *example 4. Slip error during encoding. Right endleaves for volume 236.115.*



Example 4 (Fig. 34) shows an obvious slip in the selection of the endleaf component type. The XML description indicates the endleaf as a *guard*. However, the hand drawing clearly shows a *fold* type.

Table 9 categorises the examples according to their cause and states whether they could have been avoided through more effective data validation. For each example, the table shows whether it is invalid 'in the first sense' – i.e. determined by the rules of its language – or 'in the second' – i.e. determined by its use in context. Examples 1, 2 and 3 are not 'valid in the first sense'. Their invalidity is ruled (and identifiable) by the grammar of the schema (example 1, which makes it possible to validate this error through automated routines), and by the fact that the diagrams do not show a configuration that would be possible in real life. Example 4, on the other hand, is not 'valid in the second sense', and its validity is therefore context-related: what the diagram shows is possible, but not true.

Table 9. *Table showing the cause for each error example, whether it would be avoidable through effective data validation, whether it is invalid in the first or second sense and the language or context ruling its invalidity.*

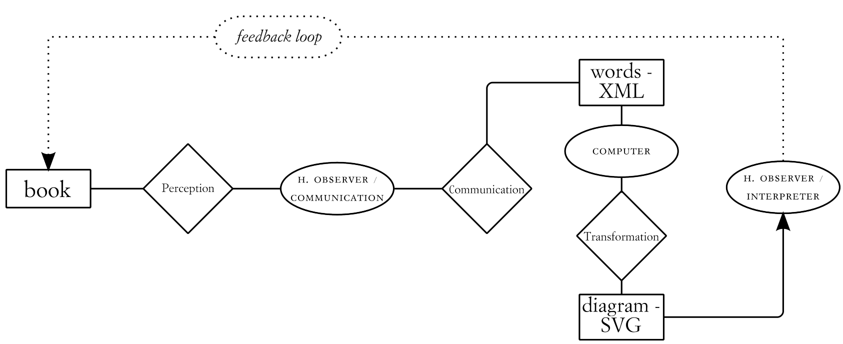
Example no.	Error cause	Avoidable through data validation	Validity sense (I or II)	Validity sense ruled by
1	Schema convention misinterpretation	Yes	I	Schema and object form
2	Description convention misinterpretation	No	I	Object form
3	Inaccurate encoding	No	I	Object form
4	Slip during encoding	No	II	Object context

These problems can be solved by resorting to visual means, if these are strictly linked to the recorded data, as in the case of diagrams that are automatically generated from it. More than a third of the human brain is devoted to vision, our primary way of gathering information about the world.[19] Diagrams, as visual communication systems, naturally offer information in a synchronic manner and can immediately highlight errors.

Exploitation of the feedback loop in the communication cycle

In any communication cycle there is a feedback loop connecting the receiver to the sender.[20] In this visualisation project, there is an extra step between the coding of the information and its delivery, as this is re-coded by the computer into a visual message (see Fig. 35). The feedback loop in this particular communication cycle can take on a different role and meaning. It could, in theory, connect not to a different person to whom information about the object is being communicated, but to the very person who encoded the information in the first place. This was obviously not the case for the project as it stands, since the information in the dataset was encoded in 2007 – and not by the present author – and the books described are in a remote location. It is not impossible, however, to envisage a system that integrates the kind of automated transformations described above directly within the surveying process. In such a case, the feedback loop would link the observer with the object being described, and the observer would receive the same information that was input into the system, but in a different form – a form that, being visual, is more immediate and synchronic.

Fig. 35. *Communication cycle: from the book being described to the human observer, through the automated transformation of the XML data into diagrams. Note the feedback loop linking the human observer back to the book.*



Had such a system been in place for the 2007 survey, it seems not unlikely that some of the mistakes registered in the dataset would not have occurred. Let us say that *p* is the structure to be recorded, but that, for some reason, *non-p* is instead input in the database. If *non-p* is a possibility allowed by the schema, the incorrect information will not be captured by data validation and will remain in the dataset. If, through an extra step, *non-p* is presented again to the encoder, it is probable that *non-p* would be corrected into *p* and the error amended.

Future work

It has not been possible to test in practice the feedback loop exploitation proposed here. Empirically, all of the errors listed above – and many more – have been identified thanks to the automated diagrams. It would, however, be interesting to test the efficacy of such a system in a project to see whether the reliability of the human surveyor does indeed increase and how much this affects the speed of the surveying process.

Conclusions

Data accuracy is an essential element for any database, but automated data validation systems cannot avoid some kinds of errors. A system capable of taking information from a bookbinding dataset and re-coding it into a visualisation can reveal errors that would otherwise be missed. If such a system were to be integrated within a survey input interface, data valid according to the schema, but incorrect, would be transformed into a diagram showing something that cannot represent reality. In other cases, the diagram would show something that is not relevant or consistent with the object being described. In both cases, the diagram would prompt the surveyer to check the data again and correct it accordingly.

This system would not prevent every kind of error. If an error is knowledge-based, resulting from a mistake in the interpretation of the evidence, the feedback would probably still produce the same incorrect interpretation. Based on the empirical experience accrued during the developing phase, it seems probable that most of the skill-based lapses and ruled-based mistakes presented here could be avoidable with the application of this system to the surveying process.

Notes

1. W. B. Rouse & S. H. Rouse, 'Analysis and classification of human error', *IEEE Transactions on systems, man and cybernetics SMC-13*, 4 (1983), pp. 539–49.
2. A. M. MacEachren et al.,'Visualizing geospatial information uncertainty: What we know and what we need to know', *Cartography & geography information science* 3 (2005), pp. 139–60.
3. J. Reason, *Human error* (Cambridge, 1990).
4. L. Mocean, 'Internet data validation', *Economy informatics* 1 (2007), pp. 96–99.
5. A. Baddeley, *Essentials of human memory* (New York/London, 2013).
6. Mocean, 'Internet data validation', pp. 96–99.
7. N. Chomsky, *Syntactic structures* (The Hague, Paris, 1957).
8. L. Wittgenstein, *On certainty*, ed. G. E. M. Anscombe & G. H. von Wright (Oxford, 1969); F. Stoutland, 'Wittgenstein: On certainty and truth', *Philosophical investigations* 23 (1998), pp. 203–21.
9. N. Pickwoad, 'The condition survey of the manuscripts in the monastery of Saint Catherine on Mount Sinai', *Paper conservator* 28 (2004), pp. 33–61.
10. A. Velios & N. Pickwoad, 'The database of the St Catherine's library conservation project in Sinai, Egypt, *IS&T archiving 2005 conference* (Washington, 2005), pp. 73–78. A. Velios & N. Pickwoad, 'Collecting digital data on paper', *Second international conference of museology – technology for cultural heritage 2004* (Mitilini, 2004), pp. 650–58; A. Velios & N. Pickwoad, 'Current use and future development of the database of the St Catherine's library conservation project', *Paper conservator* 29 (2005), pp. 39–53.
11. A. Velios & N. Pickwoad, 'Collecting and managing conservation survey data', *Care and conservation of manuscripts 10*, ed. G. Fellows-Jensen & P. Springborg (Copenhagen, 2008), pp. 172–88; A. Velios & N. Pickwoad, 'An optimised workflow for large-scale condition surveys of book collections', *Care and conservation of manuscripts 11*, ed. M. J. Driscoll & R. Mósesdóttir (Copenhagen, 2009), pp. 269–90.
12. T. Bray et al., 'Extensible Markup Language (XML) Version 1.0 (Fifth Edition)', W3C (2008). www.w3.org/TR/REC-xml [accessed 22.08.2014].
13. Velios & Pickwoad, 'Collecting and managing', pp. 172–88.
14. Velios & Pickwoad, 'Collecting and managing', pp. 172–88.
15. Velios & Pickwoad, 'An optimised workflow', pp. 269–90.
16. E. Dahlström et al., 'Scalable Vector Graphics (SVG) 1.1 (Second Edition)', W3C (2011). www.w3.org/TR/SVG [accessed 22.08.2014].
17. S. Gottwald & A. Matar, 'Many-valued logic', *The Stanford encyclopedia of philosophy*, ed. E. N. Zalta, Spring 2014 edition (Stanford, 2014). plato.stanford.edu/archives/spr2014/entries/logic-manyvalued [accessed 22.08.2014].
18. Baddeley, *Essentials*, 2013.
19. J. M. Findlay & I. D. Gilchrist, *Active vision: The psychology of looking and seeing* (Oxford/New York, 2003).
20. J. D. Rothwell, *In the company of others: An introduction to communication* (Oxford/New York, 2012).