Evolution and Change in Data Management - Issues and Directions

John F. Roddick,* Lina Al-Jadir, Leopoldo Bertossi, Marlon Dumas, Florida Estrella, Heidi Gregersen, Kathleen Hornsby, Jens Lufter, Federica Mandreoli, Tomi Männistö, Enric Mayol, Lex Wedemeijer.

Abstract

One of the fundamental aspects of information and database systems is that they change. Moreover, in so doing they evolve, although the manner and quality of this evolution is highly dependent on the mechanisms in place to handle it. While changes in data are handled well, changes in other aspects, such as structure, rules, constraints, the model, etc., are handled to varying levels of sophistication and completeness.

In order to study this in more detail a workshop on Evolution and Change in Data Management was held in Paris in November 1999. It brought together researchers from a wide range of disciplines with a common interest in handling the fundamental characteristics and the conceptual modelling of change in information and database systems. This short report of the workshop concentrates on some of the general lessons that emerged during the four days.

Keywords: Schema Evolution, Data Evolution, Model Evolution, Change Management.

1 Introduction

Change is a fundamental but too often a neglected aspect of information and database systems. Moreover, while there are a number of fields that have to deal with change, each field has, to date, tended to develop its own conceptual framework and to deal with the problem of change separately. For example, changes in process modelling, data modelling, spatio-temporal modelling and the structural aspects of databases each have their own, largely independent literature sources (as evidenced by largely non-overlapping citation trees in the workshop papers).

The *First International Workshop on Evolution and Change in Data Management*, which was held with the International Conference on Conceptual Modelling (ER'99) in Paris in November 1999, was thus designed to bring together researchers from a variety of fields to discuss and exchange ideas and to learn from each other. In particular, in accord with its parent conference, the conceptual modelling aspects were examined. The 11 presented papers [1, 4, 6, 7, 8, 10, 11, 12, 13, 20, 21] are representative of areas such as process modelling, temporal databases, logic programming, spatio-temporal modelling, archive versioning, constraint maintenance and product evolution.

This paper represents a summary of the issues of concern. In so doing it concentrates on some of the fundamental lessons that emerged during the four day workshop. The full papers are available in [2].

The categorisation of issues could, of course, have been achieved in different ways, however, the manner chosen – to use *What?, Why?, Where?, When?, Who? and How?* – was felt to disconnect the categorisation from the topic and thus to encourage creative approaches. (An alternative summary is given in Figure 1.) Two further topics on *outstanding areas* and *directions for future research* were also discussed and are included here.

2 What?

Change implies difference over time and/or (conceptual or real) space. It is therefore hardly surprising that the domains within which change is being investigated vary widely. Within the workshop papers alone, change was discussed in data, rules, constraints, schemata, models and meta-models as well as in real world objects such as documentation, database indexing methods, physical instances, product lines and individual instances of products and the user's perceptions of situations and of the universe of discourse.

However, it is interesting to note that while the subject of the change may be quite different, many of the approaches for dealing with change have a remarkable similarity in terms of the issues that need to be addressed and in the principles adopted in the modelling of the problem, if not in the execution of the solution.

Furthermore, the types of change – incremental in comparison with wholesale, local in comparison with global, structural in comparison with semantic, gradual

*School of Computer and Information Science, University of South Australia, Mawson Lakes 5095, South Australia, E-mail roddick@cis.unisa.edu.au. The affiliation and e-mail addresses of all other authors may be obtained from the proceedings [2].

in comparison with immediate, partial in comparison with total, etc. – have many similar characteristics. For example, all have to consider a model for time, the granularity of phenomena, the links between cause and effect, and so on.

3 Why?

The reason why a change occurs fundamentally affects the response to it and the manner in which automated responses are implemented. Six categories of cause were identified during discussions.

- A change in the universe of discourse. The extent and nature of the changes will determine the reaction to the change.

- A change to the interpretation of facts about the universe of discourse and the manner in which the task is realised in a system. Changes for this reason probably account for a far larger proportion than is commonly acknowledged. Acknowledging that a one-to-many relationship is more properly a many-to-many, for example, could result in changes to database structure and possibly to system operation. Included in this category could be a change in the modelling paradigm or in the analysis / design tool, which in turn may result in a different view of the world being adopted.

- Changes in the form of updates to effect upgrades to the functionality or scope of a system.

- Changes in the form of updates to effect efficiency improvements.

- Changes caused by system operation. For example, the discovery of new information which is then fed back into the system or the abnormal behaviour of a component.

- Error correction. While similar to either the second or third points above, this is listed separately as the behaviour associated with the change is likely to differ. For example, amendments resulting from the need for millennium compliance resulted from a change to the interpretation of facts about the universe of discourse regarding the use of two digit years.

This list also highlights the differences between a planned or scheduled change and an unexpected, imposed change, which can cause very different procedures to be performed.

4 When?

The question of when can be divided into three smaller questions:

- what is the model of time in use?

- when did the change occur?

- when will the response to change be effected, if at all?

For the first, philosophical aspects about time including the model of time being considered – linear, cyclic, branching (qv. [16, 17]), the local linguistic and cultural conventions for recording time [5], granularity considerations, the way in which values at unrecorded points in time are calculated and the nature of *now* [3] all need to be considered. The significance of the change (see Section 8.2 below) is also an issue.

For the second, not only the real world (valid-time) time of occurrence of the change but also the time it *should* have occurred (sometimes referred to as *decision time* [15]), the recording time (transaction-time) as well as the duration of the event is of importance. This latter aspect is often neglected with *change* often being used synonymously with *event*.

Finally, for the latter question, the answers are immediately, delayed, later and never. For example, schema changes can result in the immediate conversion of data to the new format, a scheduled conversion at some later time, lazy conversion in which data are changed only when accessed or the use of filters to simulate change.

5 Where?

The location of change is often an important factor for many situations. Spatial considerations become an issue when, for example, the database schema must evolve differently according to *where* it is accessed rather than (or in addition to) *when*. For example, in [18] a model is developed that enables schema versions to apply spatially as well as temporally. This problem is closely related to that of schema integration.

One point of interest is that the pre-eminence of the temporal *now* has a corresponding (although crucially actor specific) analogue in *here*, which can be useful as the proper identification of time and location can allow appropriate emphases in the output provided to a user.

6 Who?

Except in terms of the determination of *where* above, the question of who initiates changes, who is the object of those changes and who is directly or indirectly affected is a largely neglected (and perhaps unimportant)

Subject
- Logical
  - Data
  - Rules
  - Constraints
  - Schemata
  - Models
  - Meta-models
- Physical
  - Documentation
  - Indexes
  - Instances
  - Products
- User Perceptions
- Modelling Paradigm

Type
- Incremental .v. wholesale
- Local .v. global
- Structural .v. semantic
- Gradual .v. immediate
- Partial .v. total
- Planned/scheduled .v. unexpected/imposed

Cause
- Change in UoD
- Change in interpretation of UoD
- Functionality upgrades
- Efficiency upgrades
- System induced
- Error correction

Effect
- Significance
- Immediate .v. Lazy .v. Delayed .v. Never
- Localised .v. System-wide

Response
- Unique .v. Part of a Series
- Patterns
- Sequences
- Online .v. Offline
- Correctness

Temporal Issues
- Structure of Time
  - Linear
  - Cyclic
  - Branching
- Linguistic and cultural conventions
- Granularity
- Nature of *now*
- Method of interpolation
- Temporal Dimension
  - Valid
  - Transaction
  - Decision

Spatial Issues
- Structure of Space
  - Euclidean
  - Non-euclidean
- Linguistic and cultural conventions
- Granularity
- Dimensionality
- Nature of *here*
- Method of interpolation

Figure 1: Aspects of Change

question with the common approach to simply indicate that *a change is required/performed* with a possible class of user inferred (user, DBA, etc.).

We are not aware of any research significantly concerned with identifying the actors involved, except in those cases that are domain-specific, however, action research in business environments may be useful.

7 How?

This question is last as the response to a change clearly depends on the answers to previous questions. Nevertheless, there are some common categorisations that can be observed.

- Patterns. The cause and therefore the response might be part of a pattern of behaviour. For example, rules activated as part of an event-condition pair in an active database. The identification of emerging patterns of change might be useful.

- Sequences. The cause and/or the response could be one part of a known sequence of activity. Indeed, often deviations from the accepted sequence might be the significant issue.

- One-off. The response is customised to the situation. For example, product line modifications may results from changes in fashions or environmental concerns.

A further issue here is the development of a correctness criteria for change. The development of a model for information capacity [14] for example has shown that most practical schema versioning activities cannot be lossless. An alternative idea of correctness may thus be appropriate.

8 Outstanding Areas

This section focuses on those areas that despite being active for a number of years, still represent an open area for research.

8.1 Analysis of Rule Change

Despite the comment in Section 2, there has been little research looking at the nature of changing rules. Schema evolution and versioning for deductive databases, for example, has been a largely neglected area. Even where adopted, the approach is often to implement change in a rule through a delete/insert pair rather than an amendment. For example, a change to the rule:

$$A(a_1, a_2), B(b) \rightarrow D(d), E(e) \qquad (1)$$

to

$$A(a_1, a_2), B(b), C(c) \rightarrow D(d), E(e) \qquad (2)$$

would not generally be dealt with by some form of *Add Antecedent Element* operation. This means that the strengthening/weakening of the database rule change and therefore the history of the business rules as represented by a set of database rules are harder to trace.

8.2 Quantification of the Significance of Change

A variety of research areas have encountered the problem of quantifying the significance of a change. At one extreme, a change, although monitored, is not considered significant enough to record while at the other, a change is of such magnitude that the object is arguably not the original object. Moreover the histories of the composition and decomposition of objects often needs to be kept as the behaviour and attribute values of its constituent parts may need to be retained. The ideas discussed in [9] may provide some clues here.

8.3 Ontologies of Change

While there have been some ontologies of change, there is still little in the way of a unifying framework and there is the opportunity for useful work in this area.

8.4 Experience in Real Systems Development

It is too often the case that despite the advantages, the development of real systems that include some of the features described here are hard to find. At present, such is the case in many of the fields represented at the workshop.

9 Directions for the Future

In this section we outline some of the areas we believe could be worthwhile areas for investigation in the future. The question asked in relation to this at the workshop was *What areas would you recommend to a new doctoral student?*

9.1 Spatial and Spatio-Temporal Change

The recognition of the value of spatial, particularly geo-referenced data has grown continually over the past decade. Research into the handling of spatial and spatio-temporal change and the evolution of objects/phenomena in spatial systems is expected to be a useful area.

9.2 Semi-structured Data

Interest in semi-structured data and the use of XML as an enabling standard provides a rich opportunity for future research. Monitoring change in the content of the web is clearly important and handling versions of schemata for semi-structured data is likely to be a rich, if difficult research field.

9.3 Conceptual Modelling Tools

The development of tools (or more precisely *components* of business process engineering tools) to handle change would be a challenging and worthwhile direction. That is, the development of tools that would help determine how the parts of a conceptual model would be affected by a change, either in the universe of discourse or in the user's requirements.

9.4 Workflow Management

Workflows as a tool for managing business processes, have become a popular and useful research area. Workflows effectively manage processes in an organisation and have the advantage of reacting to changes in procedure fairly easily. How well they react to substantial business rule change and how well they can be used in other domains remains an interesting problem.

9.5 Mining Change

Data mining and knowledge discovery systems are an important new area for research combining database systems, artificial intelligence and statistics. A subset of this research investigates the production of rules with a spatial and/or a temporal component [19]. Such rules that can thus describe change, and in some cases suggest linkages between cause and effect.

10 Future of the Workshop

It is planned to hold the *Second International Workshop on Evolution and Change in Data Management* at some point in the future, probably about two years after this first one. Further details will be able to be obtained, at the appropriate time, from the first author.

11 Acknowledgements

The authors of the paper include those that presented at the workshop. We would like to thank all other workshop attendees and co-authors for their valuable contribution to the papers and the debates.

References

[1] L. Al-Jadir and M. Leonard. Transposed storage of an object database to reduce the cost of schema changes. In *[2]*, pages 48–61, 1999.

[2] P.P. Chen, D.W. Embury, J. Kouloumdjian, S.W. Liddle, and J.F. Roddick, editors. *Advances in Conceptual Modelling*, volume 1727 of *Lecture Notes in Computer Science*. Springer, Berlin, 1999.

[3] J. Clifford, C. Dyreson, T. Isakowitz, C.S. Jensen, and R.T. Snodgrass. On the semantics of "now" in databases. *ACM Transactions on Database Systems*, 22(2):171–214, 1997.

[4] M. Dumas, M.-C. Fauvet, and P.-C. Scholl. Updates and application migration support in an odmg temporal extension. In *[2]*, pages 24–35, 1999.

[5] D.E. Duncan. *The Calendar*. Fourth Estate, London, 1998.

[6] F. Estrella, Z. Kovacs, J.-M. Le Goff, R. McClatchey, and M. Zsenei. Handling evolving data through the use of a description driven systems architecture. In *[2]*, pages 1–11, 1999.

[7] F. Grandi and F. Mandreoli. Odmg language extensions for generalised schema versioning support. In *[2]*, pages 36–47, 1999.

[8] H. Gregersen and C.S. Jensen. On the ontological expressiveness of temporal extensions to the entity-relationship model. In *[2]*, pages 110–121, 1999.

[9] K. Hornsby and M. Egenhofer. Identity-based change operations for composite objects. In T. Poiker and N. Chrisman, editors, *Eighth International Symposium on Spatial Data Handling*, pages 202–213, Vancouver, Canada, 1998.

[10] K. Hornsby, M.J. Egenhofer, and P. Hayes. Modeling cyclic change. In *[2]*, pages 98–109, 1999.

[11] T. Mannisto and R. Sulonen. Evolution of schema and individuals of configurable products. In *[2]*, pages 12–23, 1999.

[12] C.A. Mareco and L. Bertossi. Specification and implementation of temporal databases in a bitemporal event calculus. In *[2]*, pages 74–85, 1999.

[13] E. Mayol and E. Teniente. A survey of current methods for integrity constraint maintenance and view updating. In *[2]*, pages 62–73, 1999.

[14] R.J. Miller, Y.E. Ioannidis, and R. Ramakrishnan. The use of information capacity in schema integration and translation. In R. Agrawal, S. Baker, and D. Bell, editors, *Nineteenth International Conference on Very Large Databases*, pages 120–133, Dublin, Ireland, 1993. Morgan Kaufmann, Palo Alto, CA.

[15] M.A. Nascimento and M.H. Eich. Decision time in temporal databases. In *TIME'95*, pages 157–162, Melbourne Beach, FL, 1995.

[16] A. Prior. *Past, present and future*. Oxford University Press, New York, 1967.

[17] N.C. Rescher and A. Urquhart. *Temporal logic*. Springer-Verlag, New York, 1971.

[18] J.F. Roddick, F. Grandi, F. Mandreoli, and M.R. Scalas. Towards a model for spatio-temporal schema selection. In *Workshop on Spatio-Temporal Data Models and Languages (W06 - Tenth International Workshop on Database and Expert Systems Applications)*, pages 434–440, Florence, Italy, 1999. IEEE Computer Society Press.

[19] J.F. Roddick and M. Spiliopoulou. A bibliography of temporal, spatial and spatio-temporal data mining research. *SIGKDD Explorations*, 1(1):34–38., 1999.

[20] R. Schaarsschmidt and J. Lufter. Schema versioning for archives in database systems. In *[2]*, pages 86–97, 1999.

[21] L. Wedemeijer. Semantic change patterns in the conceptual schema. In *[2]*, pages 122–133, 1999.