Tutorial: Languages for Collection Types

Val Tannen*
University of Pennsylvania
val@cis.upenn.edu

Abstract

Collection types such as sets, bags (multisets), lists, complex objects, nested relations, arrays, and certain kinds of trees are widely used in non-traditional database applications and must therefore be properly supported by modern database systems. The study of languages for these structures has only recently received serious attention. Can we find languages that naturally support these structures in the same way that relational languages support relational databases? Practical database languages are based on first-order logic, and much research has been devoted to augmentations of first-order logic to produce languages of increased expressive power, but with the added demands of dealing with collection types, we must ask whether we are stretching this computational paradigm beyond its breaking point.

This tutorial will suggest that we can cope with a rich variety of collection type semantics by looking at the operations that are "naturally" (in the sense of category theory) associated with the data structures involved. Surprisingly, category theory concepts that one might dismiss as too esoteric lead to a succession of useful languages for collection types, that includes equivalents of familiar languages for relational databases, such as relational algebra, nested relational algebra and datalog. As an extra bonus, this approach provides a basis for language syntax that is closely related to that of practical database query languages such as SQL. Moreover, this approach provides equational logics which serve to study certain classes of query optimizations. Finally, there are some intriguing connections between these languages and query complexity classes, both sequential and parallel.

1 Pointers

I am using this note only to communicate the abstract of the tutorial talk (above) and some bibliographical references, as well as to acknowledge the help of my colleagues. Copies of the talk slides are available by anonymous ftp or through World-Wide Web/Mosaic. Ftp users: look in `pub/papers/db-research` after connecting to `ftp.cis.upenn.edu`. Mosaic users: follow the hyperlink to the database page from the University of Pennsylvania's Department of Computer and Information Science home page whose URL is `http://www.cis.upenn.edu/`. Most of the papers co-authored by people at Penn and cited below are available in the same manner.

By necessity, this tutorial cannot be a comprehensive survey. My own work and related papers have influenced the choice of topics. The bibliography that follows is similarly affected.

2 A Bibliography

As early as in [Sch78] and then more recently in eg. [AB87, HS89, ADG$^+$89, SS89, ART90, ABD$^+$90, MS91, AL$^+$91, RS91, SP91, LMS$^+$94] the design and implementation of collection types (sometimes called—less ceremoniously—bulk types) is identified as a key issue in database programming languages.

The collection type constructs surveyed in this tutorial are related to forms of *structural recursion* (s.r.) on sets, bags and lists. The properties of well-defined programs using s.r. were examined in [BTS91]. Using s.r. for database languages was suggested in [BTBN91] where it was also observed that a limited form of always well-defined s.r. suffices to express the relational algebra. Limited forms of s.r., in particular those connected to the categorical notion of *monad* were studied in [BBW92]. Through the connection between monads and *comprehension syntax* [Wad92], this work ties in with [TW88, TW89, Tri91, WT91]. The need for array languages and the possibility of using comprehension syntax for arrays are examined in [Bun93, MV93].

An early form of divide-and-conquer recursion on sets (different from structural recursion) appears under the name "pump" in FAD [BBKV88] from which it is borrowed, under the name "hom", in Machi-

*Partially supported by NSF Grant CCR-90-57570 and ONR Contract NOOO14-93-11284

avelli [OBB89] where it fit nicely into the language's type system. Called (a form of) "transducer", it is part of SVP [PSV92]. Some limitations of its theoretical expressive power were examined in [IPS91]. In [SBT94] it plays a crucial role in a language that captures the complexity class NC.

The most thoroughly investigated collection type is that of *complex objects* (see [Hul87, ABGG89, AK89] and the references therein). Variations on the theme are nested relations [JS82] and V-relations [SAB+87, AB86]. Algebras and/or logical calculi (with or without fixpoints) are presented in [KV84, AB88, RKS88, GV91, KV93] as well as in [TF86, SS86, Col90] where they are called *nested relational algebras*. [BBW92] offers a rational reconstruction of the nested relational algebras ($\mathcal{NRA}$) starting from the monad constructs as well as connections between s.r. and the algebra with powerset of [AB88]. That $\mathcal{NRA}$ at relational types gives us a language equivalent to the relational algebra was shown in [PG92, Won93], even if nesting is used in intermediate results. This fails in the presence of powerset [HS91]. Another result of this kind [Suc94] shows that adding a bounded fixed-point construct to $\mathcal{NRA}$ gives us, again at relational types, inflationary datalog, and in [LW94a, LW94b] it is shown that nesting at intermediate types does not add expressiveness in presence of aggregate functions and certain generic queries. Other results on expressive power are to be found in [LW94a, LW93a, LW94b]. [SP94] shows that transitive closure, which is efficiently expressible using structural recursion, has a necessarily exponential implementation in the algebra with powerset of [AB88].

The s.r./monad approach can also be used for other collections: languages for or-sets [INV91] were studied in [LW93b] and bag languages in [LW94c] (see also [GM93]). [HKM93, HK94] show how to encode related database languages in the simply typed lambda calculus.

A central issue is the optimization of collection type queries, see [Sch86, Osb88, Col89, HN91, LD91, LD92, Sar92] and [BK93, ACM93, CV93, CS93, Feg94, FMS93]. Some connections between the s.r./monad and comprehension languages and equational transformations leading to optimizations are discussed in [BTBN91, BBW92, Won93].

The important extension to *object identifiers*, which also allows recursive type support is studied in [AK89].

3 Acknowledgements

It would be hard to find a place in this presentation in which the influence of Peter Buneman, Leonid Libkin, Dan Suciu, and Limsoon Wong is not prominently felt. I have also had many inspiring discussions with Scott Weinstein at Penn and with Shamim Naqvi and his group at Bellcore, in particular with Yatin Saraya.

References

[AB86] S. Abiteboul and N. Bidoit. Non-first normal form relations: An algebra allowing data restructuring. *Journal of Computer and System Sciences*, 33(3):361–371, 1986.

[AB87] M.P. Atkinson and P. Buneman. Types and persistence in database programming languages. *ACM Computing Surveys*, June 1987.

[AB88] Serge Abiteboul and Catriel Beeri. On the power of languages for the manipulation of complex objects. In *Proceedings of International Workshop on Theory and Applications of Nested Relations and Complex Objects*, Darmstadt, 1988. Also available as INRIA Technical Report 846.

[ABD+90] M. Atkinson, F. Bancilhon, D. DeWitt, K. Dittrich, D. Maier, and S. Zdonik. The object-oriented database system manifesto. In *Deductive and Object-oriented Databases*. Elsevier, Amsterdam, The Netherlands, 1990.

[ABGG89] S. Abiteboul, C. Beeri, M. Gyssens, and D. Van Gucht. An introduction to the completeness of languages for complex objects and nested relations. In S. Abiteboul, P. C. Fisher, and H.-J. Schek, editors, *LNCS 361: Nested Relations and Complex Objects in Databases*, pages 117–138. Springer-Verlag, 1989.

[ACM93] Serge Abiteboul, Sophie Cluet, and Tova Milo. Querying and updating the file. In *Proceedings of 19th International Conference on Very Large Databases*, 1993.

[ADG+89] A. Albano, A. Dearle, G. Ghelli, C. Marlin, R. Morrison, R. Orsini, and D. Stemple. A framework for comparing type systems for database programming languages. In *Proceedings of 2nd Workshop on Database Programming Languages*, pages 170–178, Gleneden Beach, Oregon, June 1989.

[AK89] S. Abiteboul and P. Kanellakis. Object identity as a query language primitive. In *Proceedings of ACM SIGMOD Conference on Management of Data*, pages 159–173, Portland, Oregon, 1989.

[AL+91] M. Atkinson, C. Lécluse, , P. Philbrow, and P. Richard. Design issues in a map language. In *Proceedings of 3rd International Workshop on Database Programming Languages, Naphlion, Greece*, pages 20–32. Morgan Kaufmann, August 1991.

[ART90] Malcolm Atkinson, Philippe Richard, and Phil Trinder. Bulk types for large scale programming. In J. W. Schmidt and A. A. Stogny, editors, *LNCS 504: Next Generation Information System Technology*, pages 229–250, Berlin, 1990. Springer-Verlag.

[BBKV88] F. Bancilhon, T. Briggs, S. Khoshafian, and P. Valduriez. A powerful and simple database language. In *Proceedings of International Conference on Very Large Data Bases*, pages 97–105, 1988.

[BBW92] Val Breazu-Tannen, Peter Buneman, and Limsoon Wong. Naturally embedded query languages. In J. Biskup and R. Hull, editors, *LNCS 646: Proceedings of 4th International Conference on Database Theory, Berlin, Germany, October, 1992*, pages 140–154. Springer-Verlag, October 1992. Available as UPenn Technical Report MS-CIS-92-47.

[BK93] Catriel Beeri and Yoram Kornatzky. Algebraic optimisation of object oriented query languages. *Theoretical Computer Science*, 116(1):59–94, August 1993.

[BTBN91] V. Breazu-Tannen, P. Buneman, and S. Naqvi. Structural recursion as a query language. In *Proceedings of 3rd International Workshop on Database Programming Languages, Naphlion, Greece*, pages 9–19. Morgan Kaufmann, August 1991. Also available as UPenn Technical Report MS-CIS-92-17.

[BTS91] V. Breazu-Tannen and R. Subrahmanyam. Logical and computational aspects of programming with Sets/Bags/Lists. In *LNCS 510: Proceedings of 18th International Colloquium on Automata, Languages, and Programming, Madrid, Spain, July 1991*, pages 60–75. Springer Verlag, 1991.

[Bun93] Peter Buneman. The fast Fourier transform as a database query. Technical Report MS-CIS-93-37/L&C 60, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, March 1993.

[Col89] Latha S. Colby. A recursive algebra and query optimisation for nested relations. In James Clifford, Bruce Lindsay, and David Maier, editors, *Proceedings of ACM-SIGMOD International Conference on Management of Data*, pages 273–283, Portland, Oregon, June 1989.

[Col90] Latha S. Colby. A recursive algebra for nested relations. *Information Systems*, 15(5):567–582, 1990.

[CS93] Surajit Chaudhuri and Kyuseok Shim. Query optimization in the presence of foreign functions. In *Proceedings of 19th International Conference on Very Large Data Bases*, Dublin, Ireland, 1993.

[CV93] Surajit Chaudhuri and Moshe Y. Vardi. Optimization of real conjunctive queries. In *Proceedings of 12th ACM Symposium on Principles of Database Systems*, pages 59–70, Washington, D. C., May 1993.

[Feg94] Leonidas Fegaras. Efficient optimization of iterative queries. In Catriel Beeri, Atsushi Ohori, and Dennis E. Shasha, editors, *Proceedings of 4th International Workshop on Database Programming Languages, New York, August 1993*, pages 200–225. Springer-Verlag, January 1994.

[FMS93] Leonidas Fegaras, David Maier, and Tim Sheard. Specifying rule-based query optimizers in a reflective framework. In *Proceedings of 3rd International Conference on Deductive and Object Oriented Databases*, Scottsdale, Arizona, December 1993.

[GM93] Stephane Grumbach and Tova Milo. Towards tractable algebras for bags. In *Proceedings of 12th ACM Symposium on Principles of Database Systems*, pages 49–58, Washington, D. C., May 1993.

[GV91] Stephane Grumbach and Victor Vianu. Tractable query languages for complex object databases. Technical Report 1573, INRIA, Rocquencourt BP 105, 78153 Le Chesnay, France, December 1991. Extended abstract appeared in PODS 91.

[HK94] Gerd G. Hillebrand and Paris C. Kanellakis. Functional database query languages as typed lambda calculi of fixed order. In *Proceedings of 13th ACM Symposium on Principles of Database Systems*, Minneapolis, Minnesota, May 1994. This volume.

[HKM93] Gerd G. Hillebrand, Paris C. Kanellakis, and Harry G. Mairson. Database query languages embedded in the typed lambda calculus. In *Proceedings of 8th IEEE Symposium on Logic in Computer Science*, pages 332–343, Montreal, Canada, June 1993.

[HN91] Michael L. Heytens and Rishiyur S. Nikhil. List comprehension in Agna, a parallel persistent object system. In J. Hughes, editor, *LNCS 523: 5th ACM Conference on Functional Languages and Computer Architecture*, pages 569–591. Springer-Verlag, August 1991.

[HS89] R. Hull and J. Su. On bulk data type constructors and manipulation primitives: a framework for analyzing expressive power and complexity. In *Proceedings of 2nd Workshop on Database Programming Languages*, pages 396–410, Gleneden Beach, Oregon, June 1989.

[HS91] Richard Hull and Jianwen Su. On the expressive power of database queries with intermediate types. *Journal of Computer and System Sciences*, 43:219–267, 1991.

[Hul87] Richard Hull. A survey of theoretical research on typed complex database objects. In Jan Paredaens, editor, *Databases*, chapter 5, pages 193–256. Academic Press, 1987.

[INV91] Tomasz Imielinski, Shamim Naqvi, and Kumar Vadaparty. Incomplete objects — a data model for design and planning applications. In James Clifford and Roger King, editors, *Proceedings of ACM-SIGMOD International Conference on Management of Data, Denver, Colorado, May 1991*, pages 288–297. ACM Press, 1991.

[IPS91] Neil Immerman, Sushant Patnaik, and David Stemple. The expressiveness of a family of finite set languages. In *Proceedings of 10th ACM Symposium on Principles of Database Systems*, pages 37–52, 1991.

[JS82] G. Jaeschke and H. J. Schek. Remarks on the algebra of non-first-normal-form relations. In *Proceedings ACM SIGACT/SIGMOD Symposium on Principles of Database Systems*, pages 124–138, Los Angeles, California, March 1982.

[KV84] Gabriel M. Kuper and Moshe Y. Vardi. A new approach to database logic. In *Proceedings of 3rd ACM Symposium on Principles of Database Systems*, pages 86–96, April 1984.

[KV93] Gabriel M. Kuper and Moshe Y. Vardi. On the complexity of queries in the logical data model. *Theoretical Computer Science*, 116(1):33–58, August 1993.

[LD91] Daniel F. Lieuwen and David J. DeWitt. Optimising loops in database programming languages. In *Proceedings of 3rd International Workshop on Database Programming Languages, Naphlion, Greece*. Morgan Kaufmann, August 1991.

[LD92] Daniel F. Lieuwen and David J. DeWitt. A transformation-based approach to optimizing loops in database programming languages. *SIGMOD Record*, 21(2):91–100, June 1992.

[LMS+94] Theodore W. Leung, Gail Mitchell, Bharathi Subramaniam, Bennet Vance, Scott. L. Vandenberg, and Stanley B. Zdonik. The AQUA data model and algebra. In Catriel Beeri, Atsushi Ohori, and Dennis E. Shasha, editors, *Proceedings of 4th International Workshop on Database Programming Languages, New York, August 1993*, pages 157–175, London, January 1994. Springer-Verlag.

[LW93a] Leonid Libkin and Limsoon Wong. A bounded degree property and finite-cofiniteness of graph queries. Technical Report MS-CIS-93-95/L&C 75, University of Pennsylvania, December 1993.

[LW93b] Leonid Libkin and Limsoon Wong. Semantic representations and query languages for orsets. In *Proceedings of 12th ACM Symposium on Principles of Database Systems*, pages 37–48, Washington, D. C., May 1993. See also UPenn Technical Report MS-CIS-92-88.

[LW94a] Leonid Libkin and Limsoon Wong. Aggregate functions, conservative extension, and linear orders. In Catriel Beeri, Atsushi Ohori, and Dennis E. Shasha, editors, *Proceedings of 4th International Workshop on Database Programming Languages, New York, August 1993*, pages 282–294. Springer-Verlag, January 1994. See also UPenn Technical Report MS-CIS-93-36.

[LW94b] Leonid Libkin and Limsoon Wong. Conservativity of nested relational calculi with internal generic functions. *Information Processing Letters*, 49(6):273–280, March 1994. See also UPenn Technical Report MS-CIS-93-60.

[LW94c] Leonid Libkin and Limsoon Wong. Some properties of query languages for bags. In Catriel Beeri, Atsushi Ohori, and Dennis E. Shasha, editors, *Proceedings of 4th International Workshop on Database Programming Languages, New York, August 1993*, pages 97–114. Springer-Verlag, January 1994. See also UPenn Technical Report MS-CIS-93-36.

[MS91] F. Matthes and J. Schmidt. Bulk types: Built-in or add-on? In *Proceedings of 3rd International Workshop on Database Programming Languages, Naphlion, Greece*, pages 33–54. Morgan Kaufmann, August 1991.

[MV93] David Maier and Bennet Vance. A call to order. In *Proceedings of 12th ACM Symposium on Principles of Database Systems*, pages 1–16, Washington, D. C., May 1993.

[OBB89] A. Ohori, P. Buneman, and V. Breazu-Tannen. Database programming in Machiavelli, a polymorphic language with static type inference. In James Clifford, Bruce Lindsay, and David Maier, editors, *Proceedings of ACM-SIGMOD International Conference on Management of Data*, pages 46–57, Portland, Oregon, June 1989.

[Osb88] S. L. Osborn. Identity, equality, and query optimization. In K. R. Dittrich, editor, *LNCS 334: Advances in Object Oriented Databases Systems*. Springer-Verlag, 1988.

[PG92] Jan Paredaens and Dirk Van Gucht. Converting nested relational algebra expressions into flat algebra expressions. *ACM Transaction on Database Systems*, 17(1):65–93, March 1992.

[PSV92] D. Stott Parker, Eric Simon, and Patrick Valduriez. SVP: A model capturing sets, streams, and parallelism. In Li-Yan Yuan, editor, *Proceedings of 18th International Conference on Very Large Databases, Vancouver, August 1992*, pages 115–126, San Mateo, California, August 1992. Morgan-Kaufmann.

[RKS88] M. A. Roth, H. F. Korth, and A. Silberschatz. Extended algebra and calculus for nested relational databases. *ACM Transactions on Database Systems*, 13(4):390–417, December 1988.

[RS91] S. Rozen and D. Shasha. Rationale and design of BULK. In *Proceedings of 3rd International Workshop on Database Programming Languages, Naphlion, Greece*, pages 71–85. Morgan Kaufmann, August 1991.

[SAB+87] Michel Scholl, Serge Abiteboul, Francois Bancilhon, Nicole Bidoit, Sophie Gamerman, Didier Plateau, Philippe Richard, and Anne Verroust. VERSO: A database machine based on nested relations. In S. Abiteboul, P. C. Fischer, and H.-J. Schek, editors, *LNCS 361: Nested Relations and Complex Objects in Databases*, pages 27–49, Berlin, April 1987. Springer-Verlag.

[Sar92] Yatin Saraiya. Fixpoints and optimizations in a language based on structural recursion on sets, December 1992. Manuscript available from `Yatin@Bellcore.Com`.

[SBT94] Dan Suciu and Val Breazu-Tannen. A query language for NC. In *Proceedings of 13th ACM Symposium on Principles of Database Systems*, Minneapolis, Minnesota, May 1994. To appear. See also UPenn Technical Report MS-CIS-94-05.

[Sch78] J. Schmidt. Type concepts for database definition. In B. Schneiderman, editor, *Databases: Improving Usability and Responsiveness*. Academic Press, 1978.

[Sch86] Marc H. Scholl. Theoretical foundation of algebraic optimisation utilising unnormalised relations. In Giorgio Ausiello and Paolo Atzeni, editors, *LNCS 243: International Conference on Database Theory, Rome, Italy, September 1986*, pages 380–396, Berlin, 1986. Springer-Verlag.

[SP91] Carol Small and Alexandra Poulovassilis. An overview of PFL. In *Proceedings of 3rd International Workshop on Database Programming Languages, Naphlion, Greece*, pages 96–110. Morgan Kaufmann, August 1991.

[SP94] Dan Suciu and Jan Paredaens. Any algorithm in the complex object algebra needs exponential space to compute transitive closure. In *Proceedings of 13th ACM Symposium on Principles of Database Systems*, Minneapolis, Minnesota, May 1994. To appear. See also UPenn Technical Report MS-CIS-94-04.

[SS86] H.-J. Schek and M. H. Scholl. The relational model with relation-valued attributes. *Information Systems*, 11(2):137–147, 1986.

[SS89] Tim Sheard and David Stemple. Automatic verification of database transaction safety. *ACM Transaction on Database Systems*, 14(3):322–368, September 1989.

[Suc94] Dan Suciu. Fixpoints and bounded fixpoints for complex objects. In Catriel Beeri, Atsushi Ohori, and Dennis Shasha, editors, *Proceedings of 4th International Workshop on Database Programming Languages, New York, August 1993*, pages 263–281. Springer-Verlag, January 1994. See also UPenn Technical Report MS-CIS-93-32.

[TF86] S. J. Thomas and P. C. Fischer. Nested relational structures. In P. C. Kanellakis and F. P. Preparata, editors, *Advances in Computing Research: The Theory of Databases*, pages 269–307, London, England, 1986. JAI Press.

[Tri91] P. W. Trinder. Comprehensions, a query notation for DBPLs. In *Proceedings of 3rd International Workshop on Database Programming Languages, Nahplion, Greece*, pages 49–62. Morgan Kaufmann, August 1991.

[TW88] P. W. Trinder and P. L. Wadler. List comprehensions and the relational calculus. In *Proceedings of 1988 Glasgow Workshop on Functional Programming*, pages 115–123, Rothesay, Scotland, August 1988.

[TW89] P. W. Trinder and P. L. Wadler. Improving list comprehension database queries. In *Proceedings of TENCON'89, Bombay, India*, pages 186–192, November 1989.

[Wad92] Philip Wadler. Comprehending monads. *Mathematical Structures in Computer Science*, 2:461–493, 1992.

[Won93] Limsoon Wong. Normal forms and conservative properties for query languages over collection types. In *Proceedings of 12th ACM Symposium on Principles of Database Systems*, pages 26–36, Washington, D. C., May 1993. See also UPenn Technical Report MS-CIS-92-59.

[WT91] David A. Watt and Phil Trinder. Towards a theory of bulk types. Fide Technical Report 91/26, Glasgow University, Glasgow G12 8QQ, Scotland, July 1991.