

Australian Standard™

**Information technology—Database
languages—SQL multimedia and
application packages**

Part 2: Full-Text



**STANDARDS
AUSTRALIA**

This Australian Standard was prepared by Committee IT-027, Data Management and Interchange. It was approved on behalf of the Council of Standards Australia on 21 April 2005. This Standard was published on 30 May 2005.

The following are represented on Committee IT-027:

Association of Superannuation Funds of Australia
Australia Post
Australia Bankers Association
Australian Bureau of Statistics
Australian Computer Society
Australian Customs Service
Australian Electoral Commission
Australian Electric and Electronic Manufacturers Association
Australian Industry Group
Australian Information Industry Association
Australian Institute of Health and Welfare
Australian Taxation Office
Centrelink
Data Management Association Australia
Department of Defence
Department of Immigration, Multicultural and Indigenous Affairs
Health Insurance Commission

Keeping Standards up-to-date

Standards are living documents which reflect progress in science, technology and systems. To maintain their currency, all Standards are periodically reviewed, and new editions are published. Between editions, amendments may be issued. Standards may also be withdrawn. It is important that readers assure themselves they are using a current Standard, which should include any amendments which may have been published since the Standard was purchased.

Detailed information about Standards can be found by visiting the Standards Web Shop at www.standards.com.au and looking up the relevant Standard in the on-line catalogue.

Alternatively, the printed Catalogue provides information current at 1 January each year, and the monthly magazine, *The Global Standard*, has a full listing of revisions and amendments published each month.

Australian Standards™ and other products and services developed by Standards Australia are published and distributed under contract by SAI Global, which operates the Standards Web Shop.

We also welcome suggestions for improvement in our Standards, and especially encourage readers to notify us immediately of any apparent inaccuracies or ambiguities. Contact us via email at mail@standards.org.au, or write to the Chief Executive, Standards Australia, GPO Box 5420, Sydney, NSW 2001.

Australian Standard™

**Information technology—Database
languages—SQL multimedia and
application packages**

Part 2: Full-Text

First published as AS ISO/IEC 13249.2—2005.

COPYRIGHT

© Standards Australia

All rights are reserved. No part of this work may be reproduced or copied in any form or by any means, electronic or mechanical, including photocopying, without the written permission of the publisher.

Published by Standards Australia GPO Box 5420, Sydney, NSW 2001, Australia

ISBN 0 7337 6720 6

PREFACE

This Standard was prepared by the Standards Australia Committee IT-027, Data Management and Interchange.

This Standard is identical with, and has been reproduced from, ISO/IEC 13249-2:2003, *Information technology—Database languages—SQL multimedia application packages—Part 2: Full Text*.

The objective of this Standard is to define full text types for use in SQL databases by software developers.

This Standard is Part 2 of AS 13249, *Information technology—Database languages—SQL multimedia and application packages*, which is published in parts as follows:

Part 1: Framework

Part 2: Full-Text (this Standard)

Part 3: Spatial

Part 5: Still image

Part 6: Data mining

As this Standard is reproduced from an international standard, the following applies:

- (a) Its number appears on the cover and title page while the international standard number appears only on the cover.
- (b) In the source text ‘this part of ISO/IEC 13249’ should read ‘this Australian Standard’.
- (c) A full point substitutes for a comma when referring to a decimal marker.

References to International Standards should be replaced by references to Australian or Australian/New Zealand Standards, as follows:

<i>Reference to International Standard</i>		<i>Australian Standard</i>	
ISO/IEC		AS ISO/IEC	
9075	Information technology—Database languages—SQL (all parts)	9075	Information technology—Database languages—SQL (all parts)
13249	Information technology—Database languages—SQL multimedia and application packages	13249	Information technology—Database languages—SQL multimedia and application packages
13249-1	Part 1: Framework	13249.1	Part 1: Framework

Only international references that have been adopted as Australian Standards have been listed.

CONTENTS

	<i>Page</i>
1	Scope 1
2	Normative references 2
3	Terms and definitions, notations and conventions 5
3.1	Terms and definitions..... 5
3.1.1	Terms and definitions provided in ISO/IEC 13249-1:2002..... 5
3.1.2	Terms and definitions provided in this part of ISO/IEC 13249..... 5
3.1.3	Terms and definitions taken from ISO/IEC 9075 (all parts)..... 6
3.1.4	Terms and definitions taken from ANSI/NISO Z39.19:1993..... 6
3.2	Notations..... 7
3.3	Conventions
4	Concepts..... 9
4.1	Concepts taken from ISO/IEC 9075 (all parts)..... 9
4.2	Text model
4.3	Text identification facilities..... 11
4.3.1	Single word patterns (patterns of the form <word>)..... 12
4.3.2	Single phrase patterns (patterns of the form <phrase>)..... 12
4.3.3	Patterns representing sets of single words..... 12
4.3.4	Patterns formed by sets of single phrases..... 14
4.3.5	Patterns specifying context conditions..... 15
4.3.6	Patterns involving Boolean operators..... 16
4.3.7	Identification of FullText values which are pertinent to a given text
4.4	Text scoring facilities
4.5	Language aspects..... 19
4.5.1	Multilingual texts and patterns..... 19
4.5.2	Treatment of stop words..... 19
4.6	Word normalization..... 21
4.7	Types and routines provided in this part of ISO/IEC 13249..... 22
4.7.1	Types and routines intended for public use
4.7.2	Types and routines for definition..... 22
4.7.3	Technique for defining the semantics of Category 1 Contains methods
4.7.4	Complementary SMI invoked regular functions
4.8	The Full-Text Information Schema..... 23
5	Full-Text Types 25
5.1	FullText Types and Routines
5.1.1	FullText Type..... 25
5.1.2	Contains Methods..... 28
5.1.3	Score Methods
5.1.4	NumberOfMatches Methods
5.1.5	Tokenize Method..... 32
5.1.6	TokenizePosition Method..... 33
5.1.7	Segmentize Method
5.1.8	TokenizeAndStem Method..... 36
5.1.9	TokenizePositionAndStem Method..... 37
5.1.10	FullText Methods..... 38
5.1.11	Contains Function..... 39
5.1.12	Score Function..... 40

5.1.13	NumberOfMatches Function.....	41
5.1.14	FullText_to_Character Function.....	42
5.1.15	StrctPattern_to_FT_Pattern Function.....	43
5.2	FT_TokenPosition Type and Routines.....	44
5.2.1	FT_TokenPosition Type.....	44
5.3	FT_Pattern Type and Routines.....	45
5.3.1	FT_Pattern Type.....	45
5.3.2	FT_Pattern Key Words.....	59
6	Structured Search Pattern Types.....	61
6.1	FT_Any Type and Routines.....	61
6.1.1	FT_Any Type.....	61
6.1.2	Contains Method.....	63
6.1.3	FT_Any Method.....	65
6.2	FT_Primary Type and Routines.....	66
6.2.1	FT_Primary Type.....	66
6.2.2	Contains Method.....	67
6.2.3	StrctPattern_to_FT_Pattern Method.....	68
6.3	FT_WordOrPhrase Type and Routines.....	69
6.3.1	FT_WordOrPhrase Type.....	69
6.3.2	Contains Method.....	70
6.3.3	StrctPattern_to_FT_Pattern Method.....	71
6.3.4	getWordArray Method.....	72
6.4	FT_TextLiteral Type and Routines.....	73
6.4.1	FT_TextLiteral Type.....	73
6.4.2	Contains Method.....	75
6.4.3	NumberOfMatches Method.....	77
6.4.4	StrctPattern_to_FT_Pattern Method.....	79
6.4.5	matches Method.....	80
6.4.6	Tokenize Method.....	81
6.4.7	getWordArray Method.....	82
6.4.8	FT_TextLiteral Methods.....	83
6.4.9	EliminateDQS Function.....	84
6.4.10	InsertDQS Function.....	85
6.5	FT_StemmedWord Type and Routines.....	86
6.5.1	FT_StemmedWord Type.....	86
6.5.2	Contains Method.....	88
6.5.3	StrctPattern_to_FT_Pattern Method.....	90
6.5.4	TokenizeAndStem Method.....	91
6.5.5	FT_StemmedWord Methods.....	92
6.6	FT_Phrase Type and Routines.....	93
6.6.1	FT_Phrase Type.....	93
6.6.2	Contains Method.....	95
6.6.3	NumberOfMatches Method.....	98
6.6.4	StrctPattern_to_FT_Pattern Method.....	101
6.6.5	getWordArray Method.....	102
6.6.6	TokenizePosition Method.....	103
6.6.7	FT_Phrase Methods.....	104
6.6.8	matches Function.....	105
6.6.9	prune Function.....	107
6.7	FT_StemmedPhrase Type and Routines.....	108
6.7.1	FT_StemmedPhrase Type.....	108
6.7.2	Contains Method.....	110
6.7.3	StrctPattern_to_FT_Pattern Method.....	113
6.7.4	TokenizePositionAndStem Method.....	114
6.7.5	FT_StemmedPhrase Methods.....	115
6.8	FT_Proxi Type and Routines.....	117
6.8.1	FT_Proxi Type.....	117
6.8.2	Contains Method.....	118
6.8.3	StrctPattern_to_FT_Pattern Method.....	121

6.8.4	FT_Proxi Method	122
6.9	FT_Soundex Type and Routines	123
6.9.1	FT_Soundex Type	123
6.9.2	Contains Method	124
6.9.3	StrctPattern_to_FT_Pattern Method	125
6.9.4	FT_Soundex Method	126
6.9.5	GetSoundsSimilar Function	127
6.10	FT_Fuzzy Type and Routines	128
6.10.1	FT_Fuzzy Type	128
6.10.2	Contains Method	129
6.10.3	StrctPattern_to_FT_Pattern Method	130
6.10.4	FT_Fuzzy Method	131
6.10.5	GetSpelledSimilar Function	132
6.11	FT_BroaderTerm Type and Routines	133
6.11.1	FT_BroaderTerm Type	133
6.11.2	Contains Method	134
6.11.3	StrctPattern_to_FT_Pattern Method	135
6.11.4	FT_BroaderTerm Method	136
6.11.5	GetBroaderTerms Function	137
6.12	FT_NarrowerTerm Type and Routines	139
6.12.1	FT_NarrowerTerm Type	139
6.12.2	Contains Method	140
6.12.3	StrctPattern_to_FT_Pattern Method	141
6.12.4	FT_NarrowerTerm Method	142
6.12.5	GetNarrowerTerms Function	143
6.13	FT_Synonym Type and Routines	145
6.13.1	FT_Synonym Type	145
6.13.2	Contains Method	146
6.13.3	StrctPattern_to_FT_Pattern Method	147
6.13.4	FT_Synonym Method	148
6.13.5	GetSynonymTerms Function	149
6.14	FT_PREFERREDTERM Type and Routines	151
6.14.1	FT_PREFERREDTERM Type	151
6.14.2	Contains Method	152
6.14.3	StrctPattern_to_FT_Pattern Method	153
6.14.4	FT_PREFERREDTERM Method	154
6.14.5	GetPreferredTerms Function	155
6.15	FT_RelatedTerm Type and Routines	157
6.15.1	FT_RelatedTerm Type	157
6.15.2	Contains Method	158
6.15.3	StrctPattern_to_FT_Pattern Method	159
6.15.4	FT_RelatedTerm Method	160
6.15.5	GetRelatedTerms Function	161
6.16	FT_TopTerm Type and Routines	163
6.16.1	FT_TopTerm Type	163
6.16.2	Contains Method	164
6.16.3	StrctPattern_to_FT_Pattern Method	165
6.16.4	FT_TopTerm Method	166
6.16.5	GetTopTerms Function	167
6.17	FT_IsAbout Type and Routines	169
6.17.1	FT_IsAbout Type	169
6.17.2	Contains Method	170
6.17.3	StrctPattern_to_FT_Pattern Method	171
6.17.4	FT_IsAbout Method	172
6.18	FT_Context Type and Routines	173
6.18.1	FT_Context Type	173
6.18.2	Contains Method	174
6.18.3	StrctPattern_to_FT_Pattern Method	176
6.18.4	FT_Context Method	177

6.19	FT_ParExpr Type and Routines.....	178
6.19.1	FT_ParExpr Type.....	178
6.19.2	Contains Method.....	179
6.19.3	StrctPattern_to_FT_Pattern Method.....	180
6.19.4	FT_ParExpr Method.....	181
6.20	FT_Term Type and Routines.....	182
6.20.1	FT_Term Type.....	182
6.20.2	Contains Method.....	183
6.20.3	StrctPattern_to_FT_Pattern Method.....	184
6.20.4	FT_Term Method.....	185
6.21	FT_Expr Type and Routines.....	186
6.21.1	FT_Expr Type.....	186
6.21.2	Contains Method.....	187
6.21.3	StrctPattern_to_FT_Pattern Method.....	188
6.21.4	FT_Expr Method.....	189
6.22	FT_PhraseList Type and Routines.....	190
6.22.1	FT_PhraseList Type.....	190
6.22.2	Contains Method.....	191
6.22.3	StrctPattern_to_FT_Pattern Method.....	193
6.22.4	FT_PhraseList Method.....	194
7	FullText_Token Type and Routines.....	195
7.1	FullText_Token Type.....	195
8	SQL/MM Full-Text Thesaurus Schema.....	197
8.1	Introduction.....	197
8.2	FT_THESAURUS Schema.....	198
8.3	TERM_DICTIONARY base table.....	199
8.4	TERM_HIERARCHY base table.....	200
8.5	TERM_SYNONYM base table.....	201
8.6	TERM_RELATED base table.....	201
9	SQL/MM Full-Text Information Schema.....	203
9.1	Introduction.....	203
9.2	FT_FEATURES view.....	204
9.3	FT_SCHEMATA view.....	204
10	SQL/MM Full-Text Definition Schema.....	205
10.1	Introduction.....	205
10.2	FT_FEATURES base table.....	206
10.3	FT_SCHEMATA base table.....	209
11	Status Codes.....	211
12	Conformance.....	213
12.1	Requirements for conformance.....	213
12.2	Features of ISO/IEC 9075 required in this part of ISO/IEC 13249.....	214
12.3	Claims of conformance.....	214
Annex A	Implementation-defined Meta-variables.....	215
Annex B	Implementation-dependent Meta-variables.....	225
B.1	Implementation-dependent Meta-variables.....	226
Index	227

TABLES

Page

Table 1 — Method and function name correspondences.....	23
Table 2 — SQLSTATE class and subclass values.....	211

Currently in preview, click buy full version

INTRODUCTION

The purpose of this International Standard is to define multimedia and application specific types and their associated routines using the user-defined features in ISO/IEC 9075.

This document is based on the content of ISO/IEC International Standard Database Language (SQL).

The organization of this part of ISO/IEC 13249 is as follows:

- 1) Clause 1, "Scope", specifies the scope of this part of ISO/IEC 13249.
- 2) Clause 2, "Normative references", identifies additional standards that, through reference in this part of ISO/IEC 13249, constitute provisions of this part of ISO/IEC 13249.
- 3) Clause 3, "Terms and definitions, notations and conventions", defines the notations and conventions used in this part of ISO/IEC 13249.
- 4) Clause 4, "Concepts", presents concepts used in the definition of this part of ISO/IEC 13249.
- 5) Clause 5, "Full-Text Types", defines the full-text user-defined types and associated routines.
- 6) Clause 6, "Structured Search Pattern Types", defines user-defined types to provide for the construction of structured search patterns.
- 7) Clause 7, "FullText_Token Type and Routines", defines the user-defined FullText_Token type.
- 8) Clause 8, "SQL/MM Full-Text Thesaurus Schema", defines the SQL/MM Full-Text thesaurus schema used to define the thesaurus related routines.
- 9) Clause 9, "SQL/MM Full-Text Information Schema", defines the SQL/MM Full-Text Information Schema.
- 10) Clause 10, "SQL/MM Full-Text Definition Schema", defines the SQL/MM Full-Text Definition Schema.
- 11) Clause 11, "Status Codes", defines the SQLSTATE codes used in this part of ISO/IEC 13249.
- 12) Clause 12, "Conformance", defines the criteria for conformance to this part of ISO/IEC 13249.
- 13) Annex A, "Implementation-defined elements", is an informative Annex. It lists those features for which the body of this part of ISO/IEC 13249 states that the syntax or meaning or effect on the database is partly or wholly implementation-defined, and describes the defining information that an implementer shall provide in each case.
- 14) Annex B, "Implementation-dependent elements", is an informative Annex. It lists those features which the body of this part of ISO/IEC 13249 states explicitly that the syntax or meaning or effect on the database is implementation-dependent.

In the text of this part of ISO/IEC 13249, Clauses begin a new odd-numbered page, and in Clause 5, "Full-Text Types", and in Clause 12, "Conformance", Subclauses begin a new page. Any resulting blank space is not significant.

AUSTRALIAN STANDARD

Information technology—Database languages—SQL multimedia and application packages

Part 2: Full-Text

1 Scope

This part of ISO/IEC 13249:

- a) introduces the Full-Text part of ISO/IEC 13249 (all parts);
- b) gives the references necessary for this part of ISO/IEC 13249;
- c) defines notations and conventions specific to this part of ISO/IEC 13249;
- d) defines concepts specific to this part of ISO/IEC 13249;
- e) defines the full-text user-defined types and their associated routines.

The full-text user-defined types defined in this part of ISO/IEC 13249 adhere to the following.

- A full-text user-defined type is generic to text handling. It addresses the need to search and retrieve information based on aspects of full-text data using patterns such as words, phrases, proximity expansion, fuzzy expansion, and thesaurus based expansions. It also addresses the need to construct such search patterns for text identification facilities and text ranking facilities.
- A full-text user-defined type does not redefine the database language SQL directly or in combination with another full-text data type.

An implementation of this part of ISO/IEC 13249 may exist in environments that also support information and content management, decision support, data mining, and data warehousing systems.

Application areas addressed by implementations of this part of ISO/IEC 13249 include, but are not restricted to, library, newspaper, multimedia, and scientific research applications.