

### ARB specific fields and entries

ARB field name	owned by	description
aligned	user	user defined entry, e.g. name and date of the person who aligned the sequence
ambig	ARB	ambiguities calculated in ARB using 'count ambiguities'
ARB_color	ARB	stores the information about sequence colors
name	ARB	internal ARB database ID, do not change!
nuc	ARB	number of nucleotides; calculated by ARB using 'count nucleotides'
nuc_term	ARB	number of nucleotides coding for the respective rRNA gene; calculated by 'count nucleotides gene'
remark	user	field for remarks
tmp	ARB	used by diverse ARB modules

### Fields and entries imported from ENA/EBI

More information about the fields at <http://www.ebi.ac.uk/ena/WebFeat/>

ARB field name	EMBL field	description
acc	AC	accession number
ali_xx/data	sequence	sequence information
author	RA	reference author(s)
bio_material	FT /bio_material	identifier for the biological material from which the nucleic acid sequenced was obtained
clone	FT /clone	clone from which the sequence was obtained
clone_lib	FT /clone_lib	clone library from which the sequence was obtained
collected_by	FT /collected_by	name of the person who collected the specimen
collection_date	FT /collection_date	date that the sample/specimen was collected
country	FT /country	geographical origin of sequenced sample
culture_collection	FT /culture_collection	institution code and identifier for the culture from which the nucleic acid sequenced was obtained, with optional collection code
date	DT	entry creation and update date separated by ;
description	DE	description
embl_class	EMBL files, relnotes.txt	describes the data class in EMBL, e.g. CON: Constructed, WGS: Whole Genome Shotgun
embl_division	EMBL files,	describes the taxonomic division in EMBL, e.g.

	relnotes.txt	ENV: Environmental Samples, PRO: Prokaryotes
env_sample	FT /environmental_sample	identifies sequences derived by direct molecular isolation from a bulk environmental DNA sample (by PCR with or without subsequent cloning of the product, DGGE, or other anonymous methods) with no reliable identification of the source organism. Indicated by 'yes' in the ARB files
full_name	OS	organism species
gene	FT /gene	symbol of the gene corresponding to a sequence region
haplotype	FT /haplotype	name for a specific set of alleles that are linked together on the same physical chromosome.
identified_by	FT /identified_by	name of the taxonomist who identified the specimen
insdc	PR	the International Nucleotide Sequence Database Collaboration (INSDC) Project Identifier that has been assigned to the entry
isolate	FT /isolate	individual isolate from which the sequence was obtained
isolation_source	FT /isolation_source	describes the physical, environmental and/or local geographical source of the biological sample from which the sequence was derived
journal	RL	reference location
lab_host	FT /lab_host	scientific name of the laboratory host used to propagate the source organism from which the sequenced molecule was obtained
lat_lon	FT /lat_lon	geographical coordinates of the location where the specimen was collected
nuc_region	FT source	identifies the biological source of the specified span of the sequence
nuc_rp	RP	reference positions
pcr_primers	FT /PCR_primers	PCR primers that were used to amplify the sequence.
plasmid	FT /plasmid	name of naturally occurring plasmid from which the sequence was obtained, where plasmid is defined as an independently replicating genetic unit that cannot be described by /chromosome or /segment.
product	FT /product	name of the product associated with the feature
publication_doi	RX	cross-reference DOI number
pubmed_id	RX	cross-reference Pubmed ID
host	FT /host	natural host from which the sequence was obtained.

specimen_voucher	FT /specimen_Voucher	an identifier of the individual or collection of the source organism and the place where it is currently stored, usually an institution
start	FT rRNA	start of the ribosomal RNA gene
stop	FT rRNA	stop of the ribosomal RNA gene
strain	FT /strain	strain from which the sequence was obtained. (t) or [T]: typestrains, [C]: cultivated, [G]: genomes
submit_author	RL	submission authors from reference location
submit_date	RL	submission date from reference location
sub_species	FT /sub_species	name of sub-species of organism from which sequence was obtained
tax_embl	OC	organism classification according to EMBL
tax_embl_name	OC	organism name taken from the classification field
tax_xref_embl	FT /db_xref	database cross-reference: pointer to related information in another database
title	RT	reference title
version	ID SV	subversion from identification line

### SILVA specific fields and entries

ARB field name	description
align_bp_score_slv	calculates the number of bases in helices in the aligned sequence taken into account canonical and non canonical basepairing. The cost matrix is taken from ARB Probe_Match 2
align_cutoff_head_slv	unaligned bases at the beginning of the sequence
align_cutoff_tail_slv	unaligned bases at the end of the sequence
align_family	shows the accession numbers of the sequences used for alignment
align_log_slv	indicates if the sequence was reversed and/or complemented
align_quality_slv	maximal similarity to reference sequence in the seed
aligned_slv	data and time of alignment by Silva
alternative_name_slv	synonyms or basonyms of the species according to the DSMZ 'nomenclature up to date' catalogue
ambig_slv	Calculated percent ambiguities in the sequences, a maximum of 2% is allowed
ann_src_slv	additional sources of sequence information is indicated in this field. Current identifiers: RNAmmer and RDP
clustered_slv	members of an OTU (not yet available)
depth_slv	depth
habitat_slv	habitat description according to EnvO-Lite

homop_slv	Calculated percentages repetitive bases with more than four bases, a maximum of 2% is allowed
homop_events_slv	absolute number of repetitive elements with more than four bases
nuc_gene_slv	aligned bases within gene boundaries
pintail_slv	information about potential sequence anomalies detected by Pintail (1); 100 means no anomalies found.
replicates_slv	replicates in on OTU (not yet available)
seq_quality_slv	summary sequence quality value calculated based on values from vector, ambiguities and homopolymers, 100 means very good
tax_gg	taxonomy mapped from greengenes
tax_gg_name	organism name in greengenes
tax_rdp	nomenclatural taxonomy mapped from RDP II
tax_rdp_name	organism name in RDP II
tax_slv	SILVA taxonomy path
vector_slv	percent vector contamination, a maximum of 5% is allowed

**With release 104 fields for additional environmental parameters have been removed from the SILVA datasets.**

**Green: New fields in SILVA 106**

**If you are interested in extending your environmental sequences with additional contextual (meta)data please have a look at the Minimum Information for Environmental Sequences (MIENS) checklist at [http://gensc.org/gc\\_wiki/index.php/MIMARKS](http://gensc.org/gc_wiki/index.php/MIMARKS).**

Release: 17.02.2011

1. Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman. 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* **71**:7724-7736.

2. Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. König, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K. H. Schleifer. 2004. ARB: a software environment for sequence data. *Nucleic Acid Res.* **32**:1363-1371.