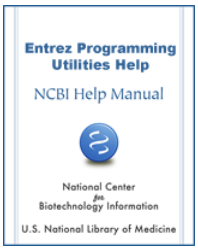
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Kans J. Entrez Direct Release Notes. 2013 Apr 23 [Updated 2022 Sep 20]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Entrez Direct Release Notes

Jonathan Kans, PhD[1]

Created: April 23, 2013; Updated: September 20, 2022.

EDirect was conceived in 2012, prototyped in Perl, and released to the public in 2014.

Xtract was subsequently rewritten in the compiled Go programming language, for a hundredfold speed improvement on modern multi-processor computers. Transmute and rchive are also provided as platform-specific Go executables.

A major refactoring, for ease of maintenance and long-term stability, was completed in 2020. EDirect now exists as a set of Unix shell scripts plus the trio of utilities implemented in Go. The original Perl code has been retired.

2022

Version 17.9: September 20, 2022

- Consolidated local archive cache maintenance functions.
- Moved first-level inverted index cache files out of Archive directory.
- Moved PubMed sentinel files to a new Sentinels directory.
- Above changes will make it feasible to have separate "live" and "next" copies of local archive files, eliminating nightly down-time if a 2 TB solid-state drive is available.

Version 17.8: September 6, 2022

- Refactored XML parser improves speed of "parent / star" processing.
- Restored proper archiving order for adjacent versions of the same publication.
- Adjusted code for downloading desc2022.xml MeSH descriptor file.

Version 17.7: August 29, 2022

- Citation matcher maps journal to standard name, checks for unresolvable ambiguity.
- Xtract -wrp delays reencoding if -replace, allowing -rep " & " instead of -rep " & ".

Version 17.6: August 15, 2022

- Recompiled with Go 1.19 for faster execution on all processors.

Author Affiliation: 1 NCBI; Email: kans@ncbi.nlm.nih.gov.

✉ Corresponding author.

- Download-ncbi-data "journals" choice generates journal title lookup files with and without a leading article, allowing both "Journal of Immunology" and "The Journal of Immunology" to match the expected "J Immunol" record regardless of original cataloging.
- Similar ampersand expansion (of encoded " & " to " and ") is done for journal title lookup keys, but not for the local archive JOUR index, so that it matches the current term list convention in PubMed.
- Precomputed journal tables attempt to resolve ambiguous name or alias collisions by favoring journals that are currently indexed in MEDLINE. This removes two competing entries with "Journal of Immunology" aliases.
- Journal sets file includes multiple choices for 2671 abbreviations that cannot be resolved automatically, separated by vertical bars, so "dmj" maps to "Danish medical journal | Diabetes & metabolism journal".
- Edict.go server adds support for remote journal title queries.
- Local archive citation matcher caches most recent unique query results to avoid repeatedly recalculating the same reference on all components of a pop/phy/mut/eco set.
- Xtract -reg and -exp, regular expression patterns for -replace, now protected by mutex.
- Xtract -aliases reads mapping file to support -classify argument, which uses multiple whole word or phrase substring matching to populate custom Entrez indices.
- Deprecated index-pubmed script now just calls archive-pubmed -index.
- Custom-index runs collect and expand steps, previously performed by index-pubmed.
- Added idx-affil custom indexing helper script for affiliation experiments.
- Added idx-journals to index Journal/Title in JRNL field, skipping the ISOAbbreviation, ISSN, ISSNLinking, MedlineTA, and NlmUniqueID elements also indexed in JOUR field.
- Transmute -gbf uses file of accessions for filtering a GenBank flatfile stream.
- Speed of the (rate-limiting) -gbf flatfile partitioning step (using an uncompressed GenBank release file on a dedicated machine) was just under 40,000 records per second. This is sufficient to support the possibility of applying the local PubMed archive approach to the orders-of-magnitude larger set of sequence records.
- Fixed bug in RepairUnicodeMarkup that produced unbalanced symbols when runs of Unicode superscripts or subscripts were immediately followed by another type of non-ASCII character, such as a Greek letter.

Version 17.5: August 2, 2022

- Added edict.go remote server for local archive (RESTful equivalent of phrase-search commands), with support for search, fetch, (compressed) stream, and (citation) match operations. The -host and -port arguments override the default "localhost:8080" address used for testing.
- Nquire -edict is a shortcut for access to a running edict server, defaulting to "localhost:8080".
- Set the NQUIRE_EDICT_SERVER environment variable to override the nquire -edict address.
- Nquire -preview is a shortcut for the upcoming PubMed SOLR server.
- Nquire -get and -url send explicit curl -X GET and -X POST arguments.
- Archive-pubmed -extras flag indexes chemical, disease, and gene references (extracted from article contents by NCBI text mining and NLM indexing groups) as CHEM, DISZ, and GENE fields. This supersedes the index-extras script, and only runs if any relevant data file is out of date or not yet downloaded.
- Archive-pubmed writes two independently-compressed blocks (xml+DOCTYPE header and PubmedArticle XML record data) to each file. When streaming sets of compressed files for efficient network transfer, the headers before each record are skipped by advancing a fixed number of bytes.
- New asn2ref script helps with citation matching from Seq-entry ASN.1 records.
- Elink internal chunk size lowered to 400 to avoid server timeouts.
- Test-eutils -preview runs the basic -alive tests on the SOLR server.

Version 17.4: July 18, 2022

- Efilter handles <Query> or <Id> items in the ENTREZ_DIRECT message.
- Archive-pubmed second-level inverted index cache moved to Increment folder.
- Downloading each PubMed ftp release file validates contents, retries on failure.
- Normalization removes combining accents in author affiliation field.
- Remaining functionality needed for full EDirect support of PubMed SOLR server includes combining queries (in esearch) and specifying a range of output records (in efetch).

Version 17.3: June 29, 2022

- Esearch -title queries individual words with [TITL] to bypass automatic term mapping.
- Elink supports <Query> or <Id> items in the ENTREZ_DIRECT message.
- Local index restores PAIR field to accelerate inexact citation matching.
- Removed TLEN and TNUM fields from local index.
- Local archive adds xml and DOCTYPE lines to every PubmedArticle record. This will allow biopython's Bio.Entrez subpackage to use the archive files directly.
- Fetch-pubmed removes the xml and DOCTYPE prefix from individual PubmedArticle records. A single pair precedes the <PubmedArticleSet> wrapper.
- Fetch-pubmed -turbo places <NEXT_RECORD_SIZE> objects in front of each XML record, to allow faster XML data extraction with xtract -turbo.
- Xtract -mirror reverses the characters in a string.

Version 17.2: June 13, 2022

- Esearch passes a <Query> field in the ENTREZ_DIRECT message when using the upcoming PubMed SOLR backend. This allows efetch to circumvent the 10,000 PMID-per-query limit by repetitive searching with a sliding window of dynamically-resized create date ranges.
- Efetch temporarily calls "transmute -mixed -normalize pubmed" twice, as a quick fix to compensate for the SOLR server's unexpected encoding of non-ASCII characters with the "&#x...;" construct.
- Internal esearch -count and -uids added to access PubMed SOLR without using history.
- Esearch -translate and -components now handle SOLR output variant.
- Local indexing adds bad date, future date, medline date, has abstract, and versioned to the PROP field.
- Local JOUR field now indexes MedlineTA, NlmUniqueID, and ISSNLinking values.
- Local YEAR field standardizes on 4 directory levels for all dates (e.g., /1/8/9/2/), which includes 104,824 citations from the eighteenth and nineteenth centuries.
- Run "archive-pubmed -clear -index" to refresh the inverted index cache after changes to indexing code.
- Phrase-search -totals prints the term list with document counts for the given field.
- Improved local citation matcher parsing of "in press" journal references.

Version 17.1: May 16, 2022

- Esearch -translate and -components, instead of -query, return the full query translation or the individual translation components, respectively. For -db pubmed they show the automatic term mapping expansions.
- Archive-pubmed adds a second level of caching for faster search index rebuilding. Use -clean to remove Inverted folder intermediate files, or -clear to also remove Archive inverted index cache files.
- Improved author name normalization and indexing of apostrophes and combining accents.
- New cit2pmid script calls "https://pubmed.ncbi.nlm.nih.gov/api/citmatch" service, or with "-local" flag performs citation matching using EDirect local archive and search system.
- Added ncbi::edirect:Execute function to control EDirect from NCBI C++ toolkit programs.

Version 17.0: April 18, 2022

- Removed STEM field (Porter2 algorithm) from standard local indexing.
- Run rchive -e2delete "${EDIRECT_PUBMED_MASTER}/Archive" to clear the incremental cache. This is needed to remove residual STEM postings.
- Next execution of archive-pubmed -daily will reinitialize the pre-indexed cache in under 90 minutes. Future daily updates should take around 3 minutes.
- Archive-pubmed -index can subsequently rebuild local search indices on demand in under 2 hours.
- Xtract adds -stemmed argument to index Porter2-processed sentences.
- Use custom-index $(which idx-stemmed) STEM to manually restore stemmed index.

Version 16.9: April 7, 2022

- Recompiled with Go 1.18, which should execute faster on ARM and Apple Silicon processors.
- Nquire -pubchem, previously a legacy alias for -pugrest, is now repurposed for more convenient access to PubChem Pathways, which replaces the retired Entrez BioSystems database.
- Xtract automatic detection of JSON input accidentally conflated the default record names for top-level objects and arrays. Now it uses the same policy as transmute -j2x ("opt" and "anon", respectively).
- Local search index adds PROP field for publication types.
- Using parallel pgzip (de)compression library for faster search index construction.
- Archive-pubmed -daily keeps an incremental cache of pre-indexed files up to date.
- Archive-pubmed -index uses the cache to repopulate all local retrieval system files in under 3 hours, superseding the slower index-pubmed process.

Version 16.8: March 17, 2022

- Phrase-search supports MESH queries by mapping to indexed TREE or CODE fields.
- Phrase-search reads an expandable file of JOUR field aliases (e.g., PNAS).
- Xtract -pairx extends -pairs to include isolated single words.
- Idx-pairs helper for custom-index reintroduces modified non-positional title word PAIR index.
- EDIRECT_PREVIEW environment variable allows testing of the upcoming PubMed API.
- Elink stores identifiers in ENTREZ_DIRECT message instead of history with new PubMed server.
- Efilter accepts ENTREZ_DIRECT instantiated identifiers when using the new PubMed API.

Version 16.7: March 1, 2022

- Added ds2pme script to convert PubMed DocumentSummary XML to Pubmed-entry ASN.1.
- Phrase-search adds -words (replacing -partial) and -pairs, for local index citation matching test.
- Filter-stop-words script adds optional -plus argument to support phrase-search -pairs.
- Index-pubmed script adds TLEN and TNUM indices, phrase-search can query those fields by range.
- Xtract -trim removes leading and trailing spaces, and leading zeros.
- Xtract -wct counts the number of words in a string, obeying -stops and -stems modifiers.
- Transmute -g2r (gbf2ref) tracks previously-encountered titles and authors, suppressing duplicate citations, which allows xtract -select STAT to create non-redundant inpress or unpub subsets.
- Transmute -r2p (and ref2pmid shortcut) reads gbf2ref subsets for local citation matching.

Version 16.6: February 14, 2022

- Nquire -pugwait polls asynchronous PubChem PUG-REST searches, returning ENTREZ_DIRECT structure with instantiated compound identifiers. (Support for this form was included in the 2020 EDirect redesign.)

- Nquire -timer prints milliseconds between initial service request and completion of network data transfer.
- Xtract warns that capitalized exploration arguments (e.g., -Block), which were needed for recursive data in the original (retired) Perl implementation, are now deprecated, and will be removed in the near future.
- Added pma2pme script to convert PubmedArticle XML to Pubmed-entry ASN.1.
- Archive-pubmed script takes optional -asn argument to save Pubmed-entry ASN.1 files in the local archive, coexisting with the PubmedArticle XML records. Use fetch-pubmed -asn to retrieve.
- Index-pubmed script now creates indices for author (AUTH, ANUM, FAUT, LAUT, CSRT, INVR) and citation (JOUR, VOL, ISS, PAGE, LANG) fields, eliminating the need for a separate custom-index step. To be used with TITL, TIAB, and YEAR fields for local citation matching experiments with phrase-search script.
- Xtract modifies -plain (removal of mixed-content sections) and adds -simple (normalization of accented letters), both derived from -basic (cleanup of superscripts and subscripts).
- Xtract repurposes -author and adds -prose, to correct commonly misused lookalike characters (sharp S and lower-case beta) and remove accents and markup, for use in generating search indices and ASN.1.
- Xtract -auth replaces commas, periods, and hyphens to allow queries with GenBank reference authors.
- Xtract -month "PubDate/*" finds first month name or abbreviation, returning a number from 1 to 12.
- Xtract -page extracts first page (digits and letters) from a page range.
- Xtract adds -hex (hexadecimal), -oct (octal), and -acc (running total accumulator) numeric arguments.
- Xtract -element "." (period) generates text ASN.1 from customized XML records. Underscores in XML tag names show unlabeled braces around SEQUENCE OF components (single), unquoted INTEGER or ENUMERATED values (trailing), or alternative CHOICE selections (internal).
- Xtract -element "%" (percent) generates JSON from customized XML records. Underscores in XML tag names show unlabeled brackets (single), named arrays in brackets (leading), or unquoted values (trailing).
- Transmute -g2r (and gbf2ref shortcut) extract reference fields from sequence flatfiles for citation matching.
- Phrase-search -partial queries title words individually and combines results for ranked matches.
- EDirect now uses a custom internal Unicode-to-ASCII conversion map.

Version 16.5: January 3, 2022

- Using efetch.fcgi instead of esearch.fcgi to retrieve UID lists, in preparation for new PubMed API.
- Nquire adds initial implementation of -puglist and -pugwait helper functions for PubChem.
- Xtract -verify raises maxDepth limit to 30 to avoid warning about depth of PMC XML records.
- Rchive -invert uses a separate map for each initial character, now runs in 1/3 less time.
- Added indexing of INVR investigators to idx-authors helper script.

2021

Version 16.4: December 20, 2021

- Nquire -pugrest -inchi silently adds "InChI=" prefix if it is missing in the argument value.
- Xtract -year "PubDate/*" construct, broken in recent refactoring, again returns a single 4-digit year.
- Separate EDIRECT_PUBMED_MASTER and EDIRECT_PUBMED_WORKING environment variables also apply to index-extras and custom-index scripts.

Version 16.3: December 15, 2021

- Esummary -format docsum has a clearer "redundant argument" message.
- Efilter -db assembly adds -status shortcut.
- Nquire adds -pugrest and -pugview shortcuts for PubChem Power User Gateway services.

- Nquire allows Windows version of curl to recognize Cygwin paths.
- Xtract supports multiple -insd feature clauses for output on a single line.
- Xtract -insd source taxid qualifier extracts integer from "taxon:###" db_xref.
- Xtract -accent no longer needs -strict or -mixed processing flags.
- Transmute -plain removes accents and diacritical marks from text.
- Local archive scripts use 2022 PubMed release files and 2022 MeSH data files.
- Local archive Extras directory added to store original MeSH data files and NLP downloads. Separate environment variables (EDIRECT_PUBMED_MASTER and EDIRECT_PUBMED_WORKING) can place Archive, Data, and Postings folders on the computer's internal drive, while keeping release files and indexing intermediates on the external SSD.
- Renamed alternative edirect-install.sh script to update-edirect.sh.

Version 16.2: October 28, 2021

- Xtract automatically detects and processes JSON, text ASN.1, and GenBank/GenPept formats. An explicit transmute command is only needed if you wish to inspect the intermediate XML, or to override the default conversion arguments.
- Transmute -j2x -nest "element" choice adds "_E" suffixes to multi-dimensional array components. This is now the default for both transmute and the xtract JSON converter, assigning a distinct tag name to each level.
- Transmute -g2x parses the DBLINK section to capture BioProject and BioSample identifiers.
- Using printf instead of echo to generate configuration commands in the install-edirect.sh script.
- Efetch -db pubmed -format uid retrieves PMIDs in chunks of 10,000 in preparation for new PubMed API.
- Added idx-metadata script for indexing JOUR, LANG, MAJR, MESH, and SUBH fields in the local archive.
- Added edirect.py module file for import by Python programs. Use edirect.execute to run individual EDirect commands, Unix tools, or shell scripts, controlling multiple steps by passing the previous result to the next command. Or use edirect.pipeline to launch a chain of several commands contained in a single argument. An edirect.efetch shortcut that takes named arguments is also provided.

Version 16.1: October 13, 2021

- Efetch -strand argument can accept "\+" or "\-". The leading backslash is required.
- Efetch -express flag works like -immediate, but on several sequence records at a time.
- Nucleotide accession lookup is only needed for master sequences.
- Epost looks up all nucleotide accessions to silently skip replaced records.
- Xtract -self applies a default value to allow detection of empty self-closing tags.
- Xtract uses both time interval and record count to trigger an output buffer flush.
- Minor refactoring of xtract argument parsing code.
- Removed obsolete setup.sh script.
- Updated cacert.pem file.

Version 16.0: October 4, 2021

- Xtract -insd feat_intervals is a version of feat_location that generates 0-based positions.
- Transmute -extract accepts -0-based and -1-based modifiers, defaulting to the 1-based feat_location form.
- The eutils.SequenceExtract library function has a new isOneBased boolean argument.
- Added fuse-segments to the set of scripts for post-processing magicblast -outfmt asn.
- Renamed uniq-columns script to uniq-table.

Version 15.9: September 27, 2021

- Redesigned install-edirect.sh script, which no longer calls setup.sh.
- Alternative edirect-install.sh script does not offer to edit configuration files.
- Installation prints PATH command to use for current terminal session.
- Successfully installed and ran test-eutils and test-edirect on Cloud.
- Minor refactoring of xtract -element variant code.
- Modified snp2hgvs script to remove records where the Id and SNP_ID do not match.

Version 15.8: September 16, 2021

- Efetch -db clinvar leaves VCV prefix for -format vcv.
- Added gene column to spdi2tbl component of snp2tbl.
- Renamed spdi2prod script to tbl2prod, now reads output of snp2tbl.
- Xtract -hgvs allows R and Y nucleotide ambiguity characters.
- Xtract -author replaces commas and periods in GenBank reference authors.
- Refactored eutils.FASTAConverter into tokenizer and streamer components.
- Added uniq-columns script.

Version 15.7: September 9, 2021

- Efetch -immediate flag avoids memory overflow on sets of extremely large sequences.
- Efetch -format fasta chunk size reduced to 50 records at a time.
- Xtract -backward presents elements in reverse order.
- Modified spdi2prod script to print "wild-type" protein and CDS translation.
- Renamed filter-table script to filter-columns.
- Moved all help text to a new help subfolder.
- Recompiled with Go 1.17, which generates smaller binaries that execute faster.

Version 15.6: September 1, 2021

- Added datasets and sra-toolkit choices to download-ncbi-software script.
- Added efilter -keyword and -purpose shortcuts for "purposeofsampling" keywords.
- Restored missing efilter -pub and -released shortcuts.
- Xtract -ncbi2na and -ncbi4na functionality available in common eutils library.
- Added vendor option to Go build.sh scripts to cache source code for all external library dependencies.

Version 15.5: August 16, 2021

- Documentation introduces Go compiled language, dependency management with modules, and EDirect's local eutils library package.
- Now including go.mod and go.sum module files for eutils and cmd project directories.
- Transmute -counts implements Go base count example.
- Added xml2fsa script for converting INSDSeq XML to FASTA.
- Xtract -fasta, used by xml2fsa script, now generates conventional 70 uppercase characters per line.
- Xtract -element "~" for object contents joins "?" for object name.
- Added download-ncbi-software script, initially for magic-blast on linux, macosx, and win64 platforms.
- Added blst2tkns, split-at-intron, fuse-ranges, and find-in-gene scripts for merging overlapping reference coordinate matches from magicblast -outfmt asn.
- Nucleotide accession lookups may use [PACC] or [ACCN] fields.
- Nquire supports NQUIRE_IPV4 environment variable for diagnosing suspected IPv6 problem.

Version 15.4: July 16, 2021

- Added efilter -source select as shortcut for RefSeq Select dataset.
- Efetch -id removes PMC prefix in front of PubMed Central identifiers.
- Transmute -align -a argument adds m choice for using commas to separate integers into groups of 3 digits.
- Added transmute -align -w argument to specify minimum width for all columns.
- Improved scripting introduction in automation section of documentation.

Version 15.3: June 21, 2021

- Transmute -separate replaced by -combine, default is not to remove internal top-level objects, which result from retrieving large sets of data in smaller chunks.
- Added transmute -search for finding positions of patterns in sequence data (e.g., restriction sites), -circular flag allows match to pattern spanning origin of circular molecule.
- Added disambiguate-nucleotides script to expand degenerate base letters for restriction enzymes like AccI.
- Added transmute -find for searching non-sequence text that includes digits, spaces, and punctuation, -relaxed flag ignores spacing differences, punctuation.
- Added idx-words helper script to split words at punctuation for indexing [WORD] field.
- Local indexing speed improvements.
- Removed deprecated rchive -phrase function, not needed with positional indices.

Version 15.2: June 3, 2021

- Expand-current prepares decompressed nonredundant PubMed archives with xtract -index, which places <NEXT_RECORD_SIZE> objects in front of each XML record.
- Xtract -turbo uses precomputed <NEXT_RECORD_SIZE> information to double the speed of the (rate-limiting) partitioning step, allowing additional CPU cores to participate in XML data extraction.
- Added custom-index, which calls a user-supplied helper script to build an initial PMID-term index, and then completes the inversion and posting steps to integrate the new field into the local search system.
- Added idx-author sample helper script, with boilerplate xtract instructions for generating an IdxDocument set, and specific commands for populating a novel [ANUM] index.
- Rchive -invert converts input strings to lower-case if that was not already done in the indexing step.
- Added xtract -includes, like -contains but requiring the substring match to align on word boundaries.
- Index-pubmed adds a new [TITL] field, which indexes only the article title.
- Phrase-search -title performs an exact search on the local [TITL] field.
- Index-pubmed replaces the [NORM] field with [TIAB], the conventional code for title and abstract. Legacy queries using [NORM] will be internally redirected to [TIAB].
- README and readme.pdf files are more concise, with several sections now residing only in the more detailed web documentation.

Version 15.1: May 20, 2021

- Xtract and transmit now run as native executables on Apple Silicon machines.
- Setup.sh script can add edirect folder to PATH in both .bash_profile and .zshrc for MacOS.
- Documented efetch and elink reading large lists of identifiers from stdin or file, bypassing need for epost.
- Nquire supports -ncbi, -eutils, and -pubchem URL shortcuts.
- Nquire uses --cacert instead of --capath in curl command.
- Efetch -db snp uses -format -self flag to keep self-closing attribute-free PAIRED or SINGLE items in XML.
- SNP processing adds Gene, Hgvs, and Spdi fields, and (when different from Id) adds OldId field.
- Added transmute -codons to display nucleotide codons above protein residues.

- Transmute -a2x, -c2x, -g2x, -j2x, and -t2x have shortcut scripts asn2xml, csv2xml, gbf2xml, json2xml, and tbl2xml, respectively.

Version 15.0: April 29, 2021

- Removed edirect.pl and setup-deps.pl scripts containing original Perl implementation.
- Deprecated -oldmode and -newmode arguments, and USE_NEW_EDIRECT environment variable.
- Simplified -hgvs output structure, and converted to 0-based positions for SPDI processing.
- Added snp2hgvs script as shortcut for -hgvs extraction from SNP docsum.
- Added hgvs2spdi script to adjust CDS-relative -hgvs offsets into sequence-relative positions suitable for use by transmute -replace.
- Added spdi2prod script to pair modified NM translation with modified NP sequence.
- Added snp2tbl script to produce tab-delimited table of adjusted SNP values in one step.
- Transmute -replace and -extract use -lower to specify lower-case output.
- Transmute -g2x writes INSDAltSeqItem_first-accn and INSDAltSeqItem_last-accn objects for WGS master.
- Xtract -pkg and -enc XML generators accept multiple slash-delimited object names.

Version 14.9: April 15, 2021

- EDirect runtime errors display an "ERROR:" tag in inverse bold red text on the terminal.
- Efetch -id now works on WGS master accessions in nucleotide databases.
- Lookup of rs/ss numbers supported in -id argument for -db snp.
- Epost uses chunks of 1000 to avoid server truncation.
- Transmute -hgvs parsing now supports ".g" genomic and ".c" coding sequences, in addition to ".p" proteins. The ".c" item location uses <Offset> rather than <Position>, since it is relative to the first base of the CDS initiation codon, not the sequence. Additional operations will be required to adjust this value for SPDI processing.
- The next release will retire the edirect.pl and setup-deps.pl scripts, and will ignore the old/new mode arguments and environment variable.

Version 14.8: March 24, 2021

- Esearch has improved logic to recognize [FILT], [PROP], and [ORGN] controlled-vocabulary phrases without the need for extra quotation marks to protect the query.
- Elink runs -cmd acheck to confirm empty history result, can detect stale links to deleted records.
- Elink writes structured message on empty history input to avoid breaking pipeline of multiple link operations.
- Nquire adds -dir command to show ftp directory listing with column of file sizes.
- Blank lines are ignored by sort-uniq-count, sort-uniq-count-rank, and sort-table scripts.
- Removed eblast script after URL-based BLAST service was disabled.
- Use of edirect.pl script (through -oldmode command-line argument or USE_NEW_EDIRECT=false environment variable setting) prints DEPRECATED warning message.

Version 14.7: March 8, 2021

- Xtract -insd handles qualifiers without values (e.g., pseudo, transgenic).
- Xtract -insd adds feat_location qualifier to print feature intervals.
- Transmute -extract reads xtract -insd feat_location interval format.
- Transmute -remove -first and -last arguments can use sequence letters instead of count.
- Efetch -format ipg processes potentially large records one at a time.

- Error message printed if non-integer -id argument is used with -db taxonomy.
- Nquire -dwn and -asp file download failure report improved.
- Refactored common.go and transmute.go functions into reusable local "eutils" package.
- Go compiler uses "replace eutils => ../eutils" line in go.mod file to import from local eutils package.
- Transmute -degenerate > gdata.go recreates library file with updated genetic code mapping tables.
- Removed obsolete and deprecated scripts.
- Added sort-table, filter-table, and print-columns scripts.
- Revised eblast script as prelude for running on cloud.

Version 14.6: February 3, 2021

- Transmute -cds2prot added to translate coding regions with nucleotide substitutions.
- Transmute -revcomp and -molwt added, sharing code with xtract functions.
- Transmute -remove, -retain, and -replace allow script-driven editing of sequences and insertion of SNP bases, with argument values obtained by -hgvs parsing of HGVS data.
- Transmute -diff simplifies visualization of sequence point mutations.
- Esearch and efilter enforce restriction of -country shortcut to sequence databases.

Version 14.5: January 19, 2021

- Added -hgvs support for genomic single nucleotide substitutions.
- Sort by RefSeq accession numbers within categories of -hgvs output.
- Updated efilter -db snp -class frameshift mapping to FXN entry.
- Updated test-eutils upon retirement of Entrez sparcle database.
- Updated Amino Acid Substitutions example.
- Added Reference Formatting example.
- Xtract -replace uses -reg and -exp values for regular expression substitution.

Version 14.4: January 13, 2021

- HGVS format parsed into XML by xtract and transmute -hgvs commands. Initial implementation supports amino acid missense and nonsense variations. More coding to follow.
- Transmute -align -a argument adds z choice for padding numbers with leading zeros.
- Fixed bug in efetch -start and -stop subrange arguments.
- Restored trailing newline lost in switch from echo to printf.

Version 14.3: January 7, 2021

- Major overhaul of EDirect documentation completed.
- Release notes and additional examples moved to separate web pages.
- Xtract -doi cleans DOI data and generates complete URL.
- Transmute -align -a argument adds n and N choices for aligning to decimal point.
- Added align-columns script as preferred front-end to transmute -align.
- Transmute -j2p works on a concatenated stream of JSON records.
- Transmute -aa1to3 and -aa3to1 amino acid abbreviation converters added for HGVS processing.
- Improved lookup of accessions in -id argument, now includes 21 Entrez databases with 10 fields.
- Improved code that maps PDB protein accessions with case-sensitive chain letters.
- Efetch -db bioproject no longer converts formats in order to handle accession input.
- Efetch -format fasta adds newline if missing before angle bracket.
- Uses printf "%s" instead of echo to avoid misinterpreting backslash in retrieved records.
- Nquire keeps error messages from leaking into output file.

- Confirmed that EDirect will run on Apple Silicon under Rosetta translation environment.

2020

Version 14.2: December 14, 2020

- Transmute -normalize now handles unexpected DocumentSummary attributes.
- Transmute -t2x -heading takes tags from columns in first row of file.
- Transmute -align converts a tab-delimited table to columns padded with spaces.
- Nquire -raw does minimal URL encoding for GeneOntology query.
- Efetch -db bioproject maps -format docsum to -format xml.
- Download-ncbi-data script updated to use desc2021.xml and supp20201.xml for mesh lookup table.

Version 14.1: November 30, 2020

- Nquire always uses new implementation.
- Scripts updated to use new nquire features.
- Efetch and esummary warn about -db mismatches and collisions between -format choices.
- Efetch supports PDB accessions with case-sensitive chain letters in the -id argument for -db protein.
- Efetch supports GCA and GCF accessions in the -id argument for -db assembly.
- Esearch -mindate and -maxdate can be used alone, with defaults filling in the missing argument.
- Xtract -set, -rec, -pkg, and -enc shortcuts join -wrp for creating XML objects at appropriate levels during XML generation.
- Xtract -head can read separate arguments for individual column headings.
- Xtract uses double-hyphen to append value into variable.

Version 14.0: November 12, 2020

- Redesigned EDirect active by default, deselected by running "export USE_NEW_EDIRECT=false" to set environment variable.
- Old implementation also chosen by adding -oldmode as the first argument to individual efetch, efilter, einfo, elink, epost, esearch, esummary, and nquire commands.
- Automatic retry supported on empty result for all formats, not just structured data.
- Esearch protects [PROP], [FILT], and [ORGN] queries in biological databases from parsing artifact on controlled vocabulary entries with embedded "or" and "not".
- Efetch adds -showgaps flag.
- Xtract -wrp causes angle brackets, ampersands, quotation marks, and apostrophes to be reencoded in the new XML, without the need for explicitly using -encode.
- Xtract formatting, modification, normalization, and conversion functions moved to transmute, with old calls automatically redirected to avoid breaking user scripts.
- Transmute -format takes optional -comment and -cdata flags to keep XML comments and CDATA blocks.
- Transmute -a2x converts text ASN.1 data to XML.
- Xtract -ncbi2na and -ncbi4na decompress nucleotide sequence converted from text ASN.1 hex representation.

Version 13.9: September 14, 2020

- Xtract -fasta splits long sequences into groups of 50 letters.
- Xtract -select restores -in to indicate name of identifier file.
- Redesigned EDirect selected by running "export USE_NEW_EDIRECT=true" to set environment variable.

- New implementation also chosen by adding -newmode as the first argument to individual efetch, efilter, einfo, elink, epost, esearch, esummary, and nquire commands.

Version 13.8: August 27, 2020

- Deprecated econtact and eproxy utilities.
- Deprecated -alias argument.
- Added accn-at-a-time and skip-if-file-exists scripts.
- Xtract -g2x converts GenBank and GenPept to INSDSeq XML.
- Xtract -c2x is a CSV (comma-separated values) variant of -t2x.

Version 13.7: May 28, 2020

- Removed xtract -repair shortcut for -unicode conversion.
- Xtract -decode supports direct Base64 decoding.
- Xtract -normalize added for Efetch post-processing.

Version 13.6: April 17, 2020

- Efilter -db snp -class shortcuts mappings updated to reflect changes to FXN index.
- Xtract adds -matches and -resembles string conditional tests, -order string processing command.
- Minor change to term granularity for MeSH [TREE] index creation.

Version 13.5: February 21, 2020

- Xtract -select -streaming uses case-insensitive test, concurrent processing.
- Minor change to term granularity for reenabled [CODE] index creation.

Version 13.4: February 4, 2020

- Local [CONV] term list indexes GNBR theme plus relationship plus identifier pair.
- Theme indexing splits genes at semicolons, adds M prefix to OMIM, H to ChEBI identifiers.
- Run "source theme-aliases" to load commands for navigating theme identifier connections.

Version 13.3: January 28, 2020

- Index-extras downloads newest version of Global Network of Biomedical Relationships theme data.
- Xtract -contour replaces -synopsis -leaf variant.

Version 13.2: January 7, 2020

- Expand-current can be run after archive-pubmed as well as index-pubmed.

2019

Version 13.1: December 30, 2019

- Local index only builds TREE fields for A, C, D, E, F, G, and Z categories.
- Stop word list now includes "studies" and "study" in addition to "studied".
- Xtract -select driver file command names changed to -retaining, -excluding, and -appending.
- New efetch -format types documented for -db clinvar.
- Updated MeSH tree category descriptions in phrase-search -help.

Version 13.0: December 16, 2019

- Xtract -histogram collects data for sort-uniq-count on entire set of records.
- Xtract -wrp with empty string or dash resets -sep, -pfx, -sfx, -plg, and -elg customization values.
- Xtract -fwd and -awd print once before and once after a set of object instances.
- Xtract -t2x uses asterisk before column name to indicate XML contents, will not escape angle brackets in string.

Version 12.9: December 9, 2019

- Xtract -plain, an -element variant that removes embedded mixed-content markup, now also converts Unicode subscripts/superscripts, cleans bad spaces.
- Xtract -select -in changed to -select -using.
- Xtract -select -adding matches by identifier and appends XML metadata.
- Xtract -select -merging requires original records and identifier-metadata file to be in same order.
- Moved xtract -examples text to separate hlp-xtract.txt file.
- Efetch -format asn maps to -format asn.1.
- Disabled building of [CODE] field in local index.
- Minor changes to term granularity for faster local index creation.
- Index-pubmed takes optional [-collect | -index | -invert | -merge | -promote] argument to resume processing in an intermediate step.

Version 12.8: December 3, 2019

- Xtract -molwt sequence processing command added.
- Xtract -len triggers -def instead of returning 0 for missing object.
- Xtract -pairs does not drop the first component of hyphenated terms (e.g., site-specific).
- Xtract -synopsis with optional -leaf argument only reports content nodes.
- Adjusted test-eutils -esummary dbvar test.
- Local archive asks user to check TRIM status on slow update.
- Local archive checks that volume is actually a solid-state drive.

Version 12.7: November 21, 2019

- Local archive detects slow performance, reminds user about antivirus scanning and content indexing.
- Local archive warns if Mac file system is not APFS, prints instructions on how to reformat drive.
- Xtract -examples adds namespace prefix and Base64 decoding example.
- Efetch -db nucleotide/nuccore -style withparts/conwithfeat processes -id accession list one record at a time.

Version 12.6: November 15, 2019

- Added support for automatic fallback to IPv4.
- Xtract -insd adds mol_wt as a synonym for calculated_mol_wt.
- Changed granularity of local index for [YEAR] field.
- If archiving is slow, ask user to ensure that antivirus scanning and content indexing are disabled.

Version 12.5: November 6, 2019

- Efetch -start and -stop subset retrieval arguments are now documented.
- Efetch -format docsum rescues uid attribute if no existing <Id> tag, now using case-sensitive test to prevent <id> from blocking.

- Added efetch -revcomp flag, sets -strand 2.
- Einfo sorts -fields and -links results by tag name.
- Fixed stdin-vs-argument bug in ftp-cp introduced when moving code into edirect.pl for Anaconda issue.
- Local search system [THME] field also indexes disambiguated themes - Jc (chemical-disease) and Jg (gene-disease) - under original code J (role in disease pathogenesis).

Version 12.4: October 28, 2019

- Expanded README file to cover more advanced features of potential interest to codeathon participants.
- Elink -released also accepts four-digit year.
- Local search system builds separate [CODE] and [TREE] indices for MeSH code and hierarchy values.

Version 12.3: October 23, 2019

- Added single-line copy-and-execute commands, in curl and wget flavors, as convenient options for EDirect installation.
- Elink -db pubmed supports -cited and -cites to follow reference connections from the NIH Open Citation Collection dataset.
- Added missing retry code to esearch and esummary.
- Xtract -t2x converts tab-delimited table to XML.
- Xtract -sort rearranges records by designated identifier.
- Xtract -split breaks up a large XML stream into multiple files.
- Xtract -chain changes_spaces_to_underscores.
- Efetch -db gtr -format docsum -mode json downloads in smaller groups to avoid timeouts.
- Nquire -get does not need -url command if followed immediately by URL argument.
- Consolidated code for perl-based commands (e.g., nquire, transmute) into master edirect.pl script.
- Wrappers to all perl-based commands now handle conflict with Anaconda installation.
- Moved external resource indexing code from xtract to rchive.
- Added phrase-search -filter command to pipe efetch -format uid results into local query.
- Download-ncbi-data script updated to use desc2020.xml and supp2020.xml for mesh lookup table.
- Experimental index-extras script loads natural language processing results into local retrieval system.
- Experimental fetch-extras scripts retrieves indexed NLP fields per PMID saved in local archive.

Version 12.2: September 27, 2019

- Added install-edirect.sh script, with download link in web documentation, for easier installation.
- Updated setup.sh script to ask permission to edit the PATH setting in the user's .bash_profile file.
- Xtract -is-before and -is-after tests compare order of strings.
- Xtract -mul, -div, and -mod numeric processing commands added.

Version 12.1: August 29, 2019

- Local query index now creates field-specific subdirectories immediately below Postings folder.
- Term position index files, needed for phrase and proximity searching in [NORM] and [STEM] fields, not made for [CODE] and [YEAR].
- Phrase-search -terms [field] prints complete term list for given field.
- Rchive -help gives example of how to sequentially ascend the MeSH [CODE] hierarchy index.
- Added expand-current script, to be run after index-pubmed, to prepare for fast scanning of all PubMed records.
- Added -repeat option to test-eutils monitoring script.

Version 12.0: August 14, 2019

- Efetch adds -format bioc for -db pubmed and -db pmc, retrieving annotated records from PubTator Central.
- Added download-ncbi-data script to consolidate and replace special-case scripts.

Version 11.9: August 5, 2019

- Updated test-eutils driver files due to retirement of unigene.
- Test-eutils progress line prints a period for success, x for failure.
- Test-eutils -timer prints response times in milliseconds for each query.
- Added exclude-uid-lists script.
- Added download-pmc-bioc script.
- Nquire supports -bioc-pubmed and -bioc-pmc shortcuts.
- Fetch-pubmed -fresh generates uncompressed PubMed files from local archive for fast data extraction.
- Index-pubmed retains compressed PubMed intermediate files for batch scanning.
- Xtract supports -select parent/element@attribute^version -in file_of_identifiers.

Version 11.8: July 23, 2019

- Xtract -j2x converts newline to space, compresses runs of spaces.
- Xtract -j2x supports JSON null tags.

Version 11.7: July 3, 2019

- Efetch -format docsum -mode json reads 500 records at a time, in conformance with the server limit.
- Fetch-pubmed -all sequentially streams all live records from the local cache.
- Xtract -plain removes embedded mixed-content markup tags.
- Added download-pmc-oa to fetch the open-access subset of PubMed Central.

Version 11.6: June 11, 2019

- Xtract -path implementation was simplified.
- Einfo -db all returns a combined eInfoResult XML, containing field and link names, and record and term counts, for all Entrez databases in one operation.

Version 11.5: June 7, 2019

- Xtract -is-equal-to and -differs-from conditional arguments compare values in two named elements.

Version 11.4: May 23, 2019

- Efetch -db snp -format json reads 10 records at a time for the initial server deployment.
- Updated SNP examples to use efetch -format json and xtract -j2x.
- Added Genes in Pathways example.
- Added xml2tbl script.

Version 11.3: May 3, 2019

- Xtract -j2x converts JSON stream to XML suitable for -path navigation.
- Xtract -format -self retains self-closing tabs with no attributes.
- Esample replaces xtract -samples.

Version 11.2: April 15, 2019

- Efetch -db snp only supports -format docsum and -format json.
- Efilter -db biosystems has -kind and -pathway shortcuts.

Version 11.1: April 3, 2019

- Xtract optimizes performance for 6 CPUs with hyperthreading.

Version 11.0: March 11, 2019

- Xtract -path generates exploration commands from dotted object path.
- Xtract -format -separate retains internal </parent><parent>.

Version 10.9: February 1, 2019

- Xtract -insd supports a sub_sequence qualifier that uses -nucleic and produces upper-case sequence.
- Xtract now has an -is-within string conditional test.

Version 10.8: January 20, 2019

- EDIRECT_DO_AUTO_ABBREV environment variable restores relaxed matching of command-line arguments.
- Efilter shortcut -journal added for -db pubmed.
- Efilter -pub last_* shortcuts moved to -released.
- Efilter -pub and -feature can take comma-separate list of choices.
- Transmute -docsum command added.
- Transmute -decode and -encode commands renamed to -unescape and -escape.
- Transmute -decode64 and -encode64 commands added.

Version 10.7: January 14, 2019

- Xtract -nucleic uses bracketed range direction to determine whether to reverse complement the sequence.

2018

Version 10.6: December 13, 2018

- Local archive script creates command for saving data path in configuration file.
- Xtract -reverse returns -words output in reverse order.
- Efilter shortcut added for -db snp (e.g., -class missense).
- Efetch -format gbc (INSDSeq XML) supports -style withparts and -style conwithfeat.

Version 10.5: December 4, 2018

- EDirect commands and pipelines support faster access with API keys.
- Xtract attributes can be delimited by quotation marks or apostrophes.
- Transmute -encode and -decode commands added.
- Simplified processing of local inverted index intermediate files.

Version 10.4: November 13, 2018

- Rchive local indexing code refactored for faster performance.

- Xtract -deq deletes and replaces queued tab separator after the fact.
- Efilter -organism queries in [ORGN] field if argument is not in shortcut list.

Version 10.3: November 1, 2018

- Rchive -invert, -merge, -promote, and -query steps make better use of multiple processor cores.
- New phrase-search script replaces local-phrase-search.

Version 10.2: October 15, 2018

- Transmute -x2j joins -j2x to simplify the use of JSON-based services.
- Efetch -json converts adjusted XML output to JSON as a convenience.
- Xtract tag alphabet expanded to accommodate converted JSON data.
- Nquire -ftp takes server, directory, and filename arguments, sends data to stdout.

Version 10.1: October 9, 2018

- Xtract -mixed improves support for mixed-content XML.

Version 10.0: September 27, 2018

- Efilter can search for sequence records by sample collection location (e.g., -country "canada new brunswick").
- Xtract parsing code was refactored in preparation for improvements in handling mixed-content XML data.
- Added transmute script for format conversions (e.g., -j2x for JSON to XML).

Version 9.90: September 17, 2018

- Normalized archive path for low-value PMIDs in preparation for incremental indexing.

Version 9.80: September 4, 2018

- Xtract XML block reader can run on separate thread for improved performance on computers with surplus processor cores.
- Fixed bug in string cleanup when text starts with a non-ASCII Unicode character.
- Efetch regular expression pattern for detecting mixed-content tags was adjusted.

Version 9.70: August 22, 2018

- Local archive builds parallel stemmed and non-stemmed indices of terms in the title and abstract.
- Rchive and local-phrase-search use -query for evaluation of non-stemmed terms, -search for evaluation using the stemmed index.

Version 9.60: August 9, 2018

- Local archive script removes newlines inside PubMed text fields.
- Efetch adds missing newline at end of PubmedArticleSet XML.

Version 9.50: July 30, 2018

- Local indexing scripts adjusted to accommodate projected range of PMID values.
- Fixed inconsistency in positional indexing of terms with embedded non-alphanumeric characters.

- EDIRECT_PUBMED_WORKING environment variable keeps local archive intermediate files on a separate volume.
- Rchive and local-phrase-search use -exact to round-trip ArticleTitle contents without interpretation as a query formula.

Version 9.40: July 18, 2018

- Xtract handles misplaced spaces in attributes.
- Xtract -format repairs misplaced spaces in attributes.

Version 9.30: July 9, 2018

- Local data indexing retains intermediate products, allows rapid streaming of non-redundant current records.
- Index preparation removes apostrophe in trailing 's possessives.
- Wildcard minimum varies with prefix-driven posting character depth.

Version 9.20: June 26, 2018

- Portability and efficiency improvements to local data cache scripts.
- Xtract handles misplaced spaces in self-closing tags.

Version 9.10: June 18, 2018

- Added Parent/* element exploration construct to xtract.
- Xtract -year reliably obtains the year from "PubDate/*".

Version 9.00: June 6, 2018

- Fetch-pubmed -path supplies missing Archive directory if root path is given.
- Efetch cleanup of MathML markup properly handles parentheses.

Version 8.90: June 4, 2018

- Xtract -transform and -translate allow data value substitution.
- Xtract -wrp simplifies wrapping of extracted values in XML tags.

Version 8.80: May 29, 2018

- Efetch removes MathML tags from PubmedArticle XML contents, unless the -raw flag is used.

Version 8.70: May 14, 2018

- Local phrase indexing now uses positional indices instead of adjacent overlapping word pairs.
- Xtract -select uses conditional expressions to filter records.

Version 8.60: April 26, 2018

- Efetch -format uid pauses between groups, retries on failure.
- Fetch delay drops from 1/3 to 1/10 second if API key is used.
- Local phrase indexing uses smaller files to avoid memory contention.
- Phrase index removes hyphens from selected prefixes.

Version 8.50: April 13, 2018

- Efetch markup tag removal modified after change in server.
- Xtract -phrase filter split into -require and -exclude commands.

Version 8.40: April 9, 2018

- Efetch removes markup tags in all PubMed XML.
- Xtract without -strict prints warnings if markup tags are encountered.
- Xtract proximity search moved from -matches to -phrase.

Version 8.30: April 4, 2018

- Xtract is now available for ARM processors.

Version 8.20: March 12, 2018

- Minor changes to local record archiving scripts.

Version 8.10: March 2, 2018

- Xtract -strict and -mixed support MathML element tags in PubmedArticle XML.

Version 8.00: February 26, 2018

- Efetch -raw skips database-specific XML modifications.
- Added local-phrase-search script.
- Xtract -strict, -mixed, and -repair flag speed improvements.

Version 7.90: February 1, 2018

- Minor change to installation commands for tcsh.

Version 7.80: January 12, 2018

- Updated setup.sh script with additional error checking.

2017

Version 7.70: December 27, 2017

- Added archive-pubmed script to automate local record archiving.

Version 7.60: November 15, 2017

- Epost -id numeric argument bug fixed.
- Xtract conditional tests can now use subrange specifiers.
- Xtract -strict and -mixed use separate -repair flag to normalize Unicode superscripts and subscripts.

Version 7.50: October 31, 2017

- Setup instructions now work with the tcsh shell.
- API key value is taken from the NCBI_API_KEY environment variable.
- Efetch -format gb supports -style withparts and -style conwithfeat.

- Xtract supports optional element [min:max] substring extraction.
- Xtract -position supports [first|last|outer|inner|all] argument values.
- Added prepare-stash script for local record archive.

Version 7.40: September 27, 2017

- Xtract -hash reports checksums for local record archiving.
- Initial support for API keys.

Version 7.30: September 6, 2017

- Modified stash-pubmed script to work around Cygwin artifact.
- Removed unpack-pubmed script.
- Xtract -archive replaces -stash for local record archiving.
- Xtract -gzip allows compression of archived XML records.

Version 7.20: August 28, 2017

- Added download-pubmed, download-sequence, unpack-pubmed, stash-pubmed, and fetch-pubmed scripts, for experimental local record storage.
- Xtract -flags [strict|mixed] added to support new local storage scripts.
- Removed obsolete, original Perl implementation of xtract.pl.

Version 7.10: August 10, 2017

- Xtract -ascii converts non-ASCII Unicode to hexadecimal numeric character references.
- Setup script recognizes Cygwin running under the MinGW emulator.

Version 7.00: July 10, 2017

- Xtract -mixed and -strict handle multiply-escaped HTML tags.
- Efetch removes normal and escaped HTML tags from PubMed fields.
- Esearch -field processes individual query terms using the designated field, also removing stop words.
- Esearch -pairs splits the query phrase into adjacent overlapping word pairs.

Version 6.90: July 5, 2017

- Xtract -mixed replaces -relaxed, and -accent replaces -plain.
- Efetch uses larger chunks for -format uid, url, and acc.
- Esearch -log shows constructed URL and QueryTranslation result.

Version 6.80: June 8, 2017

- Modified download instructions to use edirect.tar.gz archive.
- The ftp-cp script can now read from stdin without the need for xargs.
- Rerunning ftp-cp or asp-cp only attempts to download missing files.

Version 6.70: May 8, 2017

- Added asp-cp script for faster download of NCBI ftp files using Aspera Connect.
- Xtract -strict and -relaxed handle empty HTML tag variants (e.g., <b/> and <sup/>).

Version 6.60: April 25, 2017

- Xtract -strict replaces -degloss to remove HTML <i>, <b>, <u>, <sup> and <sub> tags from XML contents.
- Xtract -relaxed allows HTML tags in XML contents, to support current PubMed ftp release files.
- Xtract -plain removes Unicode accents.
- The setup.sh script prints an error message if it cannot fetch missing Perl modules.

Version 6.50: March 6, 2017

- Xtract -degloss replaces -html to remove HTML <i>, <b>, <u>, <sup> and <sub> tags.

Version 6.40: March 1, 2017

- Epost detects accession.version input for sequence databases and sets -format acc.
- Xtract -html [remove|encode] converts <i> and <b> tags embedded in XML contents.

Version 6.30: February 13, 2017

- Efetch -format docsum skips GI-less sequences without summaries.
- Xtract local indexing commands moved to -extras documentation.

Version 6.20: January 30, 2017

- Xtract -limit and -index allow extraction of selected records from XML file.

Version 6.10: January 19, 2017

- Added run-ncbi-converter script for processing ASN.1 release files.
- Xtract -format flush option added.
- Removed obsolete accession-dot-version conversion code.

2016

Version 6.00: December 27, 2016

- Efetch -format docsum removes eSummaryResult wrapper.
- Fixed content truncation bug when Xtract encounters very long sequences.

Version 5.90: December 21, 2016

- Efetch and Elink readied for switch to accession-dot-version sequence identifier.
- Xtract -insd recognizes INSDInterval_iscomp@value and other boolean attributes.
- Xtract adds experimental phrase processing commands for word index preparation.

Version 5.80: December 12, 2016

- Efilter adds shortcuts for -db gene (e.g., -status alive, -type coding).
- Xtract numeric conditional tests can use an element name for the second argument (e.g., -if ChrStop -lt ChrStart finds minus strand genes).

Version 5.70: November 30, 2016

- Xtract -format takes an optional [compact|indent|expand] argument. Processing compact XML is about 15% faster than indent form. Using expand places each attribute on a separate line for ease of reading.

Version 5.60: November 22, 2016

- Fixed bug in -datetype argument for Esearch and Efilter.
- Added optional argument to filter-stop-words script to indicate replacement.

Version 5.50: November 16, 2016

- Efetch -id allows non-numeric accessions only for sequence databases.
- Xtract element selection no longer considers fields in recursive sub-objects.
- Xtract introduces a double-star "**/Object" construct to flatten recursive child objects for linear exploration.
- Xtract conditional tests ignore empty self-closing tags.
- Xtract -else simplifies insertion of a placeholder to indicate missing data.

Version 5.40: November 7, 2016

- Added filter-stop-words and xy-plot scripts.

Version 5.30: October 31, 2016

- Added support for ecitmatch utility.
- Added amino-acid-composition and between-two-genes scripts.
- The sort-uniq-count and sort-uniq-count-rank scripts take an optional argument (e.g., -n for numeric comparisons, -r to reverse order).

Version 5.20: October 26, 2016

- Setup script no longer modifies the user's configuration file to update the PATH variable. Instead, it now prints customized instructions for the user to execute. The user may choose to run these commands, but is free to edit the .bash_profile file manually.
- Xtract deprecates -match and -avoid functions and the Element:Value conditional shortcut.
- Xtract -if and -unless commands use compound statements for conditional execution (e.g., -if Element -equals Value).
- Colon now separates namespace prefix from element name in xtract arguments (e.g., -block jats:abstract). Colon at start of element name matches any namespace prefix.
- Xtract -insd uses a dash as placeholder for missing field. Experimental -insdx command is deprecated.
- Precompiled versions of xtract are now provided for Darwin, Linux, and CYGWIN_NT platforms. The appropriate executable is downloaded by the setup script.

Version 5.10: October 13, 2016

- Xtract adds -0-based, -1-based, and -ucsc numeric extraction/conversion commands for sequence positions from several Entrez databases.

Version 5.00: September 26, 2016

- Efetch -format fasta removes blank lines between records.
- Xtract -insdx uses a dash to indicate a missing field.

- Xtract -insd no longer has blank lines between records.
- Xtract -input allows reading XML data from a file.

Version 4.90: September 14, 2016

- Epost -input allows reading from an input file instead of using data piped through stdin.
- Efilter now supports the -sort argument.
- Xtract -filter can recover information in XML comments and CDATA blocks.

Version 4.80: August 9, 2016

- Xtract -insd controlled vocabularies updated.

Version 4.70: August 4, 2016

- Einfo -db request can also display -fields and -links data summaries.
- Einfo -dbs prints database names instead of eInfoResult XML.

Version 4.60: July 18, 2016

- Elink -cmd acheck returns information on all available links for a record.
- Efilter -pub structured limits to articles with structured abstracts.

Version 4.50: July 1, 2016

- Esearch and Efilter detect and report -query phrase quotation errors.
- Efilter -pub shortcut adds last_week, last_month, and last_year choices.
- Efetch sets -strand 2 for minus strand if -seq_start > -seq_stop or if -chr_start > -chr_stop.

Version 4.40: June 21, 2016

- Transitioning to use of https for access to NCBI services.
- Epost -db assembly -format acc uses [ASAC] field instead of [ACCN].

Version 4.30: June 13, 2016

- Efilter -pub preprint limits results to ahead-of-print articles.
- Xtract -pattern Parent/* construct can now process catenated XML files.

Version 4.20: May 24, 2016

- Xtract command-line argument parsing improvements.
- Nquire -get supersedes -http get.

Version 4.10: May 3, 2016

- Xtract -format removes multi-line XML comments and CDATA blocks.

Version 4.00: April 4, 2016

- Esearch adds -spell to correct known misspellings of biological terms in the query string.
- Efilter adds -spell to correct query misspellings, and -pub, -feature, -location, -molecule, -organism, and -source shortcuts. Run efilter -help to see the choices available for each argument.

Version 3.90: March 21, 2016

- Code optimizations for increased Xtract speed.

Version 3.80: February 29, 2016

- Xtract can distribute its work among available processor cores for additional speed.

Version 3.70: February 8, 2016

- Xtract performance improvements.

Version 3.60: January 11, 2016

- The setup.sh configuration script now downloads a precompiled Xtract executable for selected platforms.

2015

Version 3.50: December 27, 2015

- Xtract reports error for element:value construct outside of -match or -avoid arguments.

Version 3.40: December 20, 2015

- Xtract -insd supports extraction from multiple features (e.g., CDS,mRNA).

Version 3.30: December 3, 2015

- Efetch -format docsum can accept a single sequence accession number in the -id argument.

Version 3.20: November 30, 2015

- Xtract supports -match conditional execution on values recorded in variables.

Version 3.10: November 18, 2015

- Efetch adds -chr_start and -chr_stop arguments to specify sequence range from 0-based coordinates in gene docsum GenomicInfoType object.

Version 3.00: October 30, 2015

- Xtract rewritten in the Go programming language for speed. The setup.sh configuration script installs an older Perl version (2.99) if a local Go compiler is unavailable.
- Efetch -format docsum only decodes HTML entity numbers in select situations.

Version 2.90: October 15, 2015

- Xtract warns on use of deprecated arguments -present, -absent, and -trim, in preparation for release of much faster version.

Version 2.80: September 9, 2015

- Xtract uses the "*/Child" construct for nested exploration into recursive structures, replacing the -trim argument.

Version 2.70: July 14, 2015

- Added entrez-phrase-search script to query on adjacent word pairs indexed in specific fields.

Version 2.60: June 23, 2015

- Xtract -match and -avoid support "Parent/Child" construct for BLAST XML.

Version 2.50: April 9, 2015

- Xtract capitalized -Pattern handles recursively-defined top-level objects.

Version 2.40: March 25, 2015

- EDirect programs use the http_proxy environment variable to work behind firewalls.

Version 2.30: March 11, 2015

- Cleaned up logic in setup.sh configuration script.
- EPost -format acc works properly on protein accessions.

Version 2.20: March 4, 2015

- Xtract -match and -avoid recognize "@attribute" without element or value.

Version 2.10: February 3, 2015

- Added ftp-ls and ftp-cp scripts for convenient access to the NCBI anonymous ftp server.

2014

Version 2.00: August 28, 2014

- Introduced copy-and-paste installation commands with setup.sh configuration script.

Version 1.90: August 8, 2014

- Xtract -format combines multiple XML results into a single valid object.
- Improved suppression of 0-count failure messages with -silent flag in scripts.

Version 1.80: July 15, 2014

- EPost -format acc accepts accessions in an -id argument on the command line.

Version 1.70: April 23, 2014

- EFetch -format docsum decodes HTML entity numbers embedded in the text.

Version 1.60: April 3, 2014

- Minor enhancements to xtract -insd.

Version 1.50: March 29, 2014

- Esearch -sort specifies the order of results when records are retrieved.
- Xtract exploration arguments (e.g., -block) now work on self-closing tags with data in attributes.

Version 1.40: March 17, 2014

- Xtract -format repairs XML line-wrapping and indentation.
- Implemented -help flag to display the list of command-line arguments for each function.

Version 1.30: March 3, 2014

- Xtract -insd partial logic was corrected to examine both 5' and 3' partial flags, and the location indicator recognizes "+" or "complete" and "-" or "partial".

Version 1.20: February 26, 2014

- Xtract -insd detects if it is part of an EDirect sequence record query, and dynamically executes the extraction request for specific qualifier values. When run in isolation it generates extraction instructions that can be incorporated (with modifications, if necessary) into other queries.

Version 1.10: February 10, 2014

- ESummary was replaced by "efetch -format docsum" to provide a single command for all document retrieval. The esummary command will continue to work for those who prefer it, and to avoid breaking existing scripts.
- Xtract processes each -pattern object immediately upon receipt, eliminating the need for using xargs and sh to split document retrieval into smaller units.

Version 1.00: February 6, 2014

- Initial public release.

2013

Version 0.00: April 23, 2013

- Initial check-in of web documentation page.

Version 0.00: March 20, 2013

- Initial check-in of edirect.pl script source code.