

The `expl3` package and L^AT_EX3 programming^{*}

The L^AT_EX3 Project[†]

2009/08/05

Abstract

This document gives an introduction to a new set of programming conventions that have been designed to meet the requirements of implementing large scale T_EX macro programming projects such as L^AT_EX. These programming conventions are the base layer of L^AT_EX3.

The main features of the system described are:

- classification of the macros (or, in L^AT_EX terminology, commands) into L^AT_EX functions and L^AT_EX parameters, and also into modules containing related commands;
- a systematic naming scheme based on these classifications;
- a simple mechanism for controlling the expansion of a function's arguments.

This system is being used as the basis for T_EX programming within the L^AT_EX3 project. Note that the language is not intended for either document mark-up or style specification. Instead, it is intended that such features will be built on top of the conventions described here.

1 Introduction

The first step to develop a L^AT_EX kernel beyond L^AT_EX 2_ε is to address how the underlying system is programmed. Rather than the current mix of L^AT_EX and T_EX macros, the L^AT_EX3 system provides its own consistent interface to all of the functions needed to control T_EX. A key part of this work is to ensure that everything is documented, so that L^AT_EX programmers and users can work efficiently without needing to be familiar with the internal nature of the kernel or with plain T_EX.

The `expl3` bundle provides this new programming interface for L^AT_EX. To make programming systematic, L^AT_EX3 uses some very different conventions to L^AT_EX 2_ε or plain TeX. As a result, programmers starting with L^AT_EX3 will need to become familiar with the syntax of the new language.

The next section shows where this language fits into a complete T_EX-based document processing system. We then describe the major features of the syntactic structure of command names, including the argument specification syntax used in function names.

^{*}This file has version number 1439, last revised 2009/08/05.

[†]Frank Mittelbach, Denys Duchier, Chris Rowley, Rainer Schöpf, Johannes Braams, Michael Downes, David Carlisle, Alan Jeffrey, Morten Høgholm, Thomas Lotze, Javier Bezos, Will Robertson, Joseph Wright

The practical ideas behind this argument syntax will be explained, together with the expansion control mechanism and the interface used to define variant forms of functions.

As we shall demonstrate, the use of a structured naming scheme and of variant forms for functions greatly improves the readability of the code and hence also its reliability. Moreover, experience has shown that the longer command names which result from the new syntax do not make the process of *writing* code significantly harder.

2 Languages and interfaces

It is possible to identify several distinct languages related to the various interfaces that are needed in a T_EX-based document processing system. This section looks at those we consider most important for the L^AT_EX3 system.

Document mark-up This comprises those commands (often called tags) that are to be embedded in the document (the `.tex` file).

It is generally accepted that such mark-up should be essentially *declarative*. It may be traditional T_EX-based mark-up such as L^AT_EX 2_ε, as described in [3] and [2], or a mark-up language defined via HTML or XML.

One problem with more traditional T_EX coding conventions (as described in [1]) is that the names and syntax of T_EX's primitive formatting commands are ingeniously designed to be 'natural' when used directly by the author as document mark-up or in macros. Ironically, the ubiquity (and widely recognised superiority) of logical mark-up has meant that such explicit formatting commands are almost never needed in documents or in author-defined macros. Thus they are used almost exclusively by T_EX programmers to define higher-level commands, and their idiosyncratic syntax is not at all popular with this community. Moreover, many of them have names that could be very useful as document mark-up tags were they not pre-empted as primitives (e.g. `\box` or `\special`).

Designer interface This relates a (human) typographic designer's specification for a document to a program that 'formats the document'. It should ideally use a declarative language that facilitates expression of the relationship and spacing rules specified for the layout of the various document elements.

This language is not embedded in document text and it will be very different in form to the document mark-up language. For L^AT_EX, this level was almost completely missing from L^AT_EX 2.09; L^AT_EX 2_ε made some improvements in this area but it is still the case that implementing a design specification in L^AT_EX requires far more 'low-level' coding than is acceptable.

Programmer interface This language is the implementation language within which the basic typesetting functionality is implemented, building upon the primitives of T_EX (or a successor program). It may also be used to implement the previous two languages 'within' T_EX, as in the current L^AT_EX system.

The last layer is covered by the conventions described in this document, which describes a system aimed at providing a suitable basis for coding L^AT_EX3. Its main distinguishing features are summarised here:

- A consistent naming scheme for all commands, including \TeX primitives.
- The classification of commands as \LaTeX functions or \LaTeX parameters, and also their division into modules according to their functionality.
- A simple mechanism for controlling argument expansion.
- Provision of a set of core \LaTeX functions that is sufficient for handling programming constructs such as queues, sets, stacks, property lists.
- A \TeX programming environment in which, for example, all white space is ignored.

3 The naming scheme

$\text{\LaTeX}3$ does not use \@ as a ‘letter’ for defining internal macros. Instead, the symbols $_$ and : are used in internal macro names to provide structure. In contrast to the plain \TeX format and the $\text{\LaTeX}2_{\epsilon}$ kernel, these extra letters are used only between parts of a macro name (no strange vowel replacement).

The naming conventions for this programming language distinguish between *functions* and *variables*. Functions can have arguments and they are executed. Variables can be assigned values and they are used in arguments to functions; they are not directly executed but are manipulated by functions. Functions and variables with a related functionality (for example accessing counters, or manipulating token lists, etc.) are collected together into a *module*.

3.1 Examples

Before giving the details of the naming scheme, here are a few typical examples to indicate the flavour of the scheme; first some variable names.

_l_tmpa_box is a local variable (hence the _l_ prefix) corresponding to a box register.
 _g_tmpa_int is a global variable (hence the _g_ prefix) corresponding to an integer register (i.e. a \TeX count register).
 _c_empty_toks is the constant (_c_) token register variable that is always empty.

Now here is an example of a typical function name.

_seq_push:Nn is the function which puts the token list specified by its second argument onto the stack specified by its first argument. The different natures of the two arguments are indicated by the :Nn suffix. The first argument must be a single token which ‘names’ the stack parameter: such single-token arguments are denoted _N . The second argument is a normal \TeX ‘undelimited argument’, which may either be a single token or a balanced, brace-delimited token list (which we shall here call a *braced token list*): the _n denotes such a ‘normal’ argument form. The name of the function indicates it belongs to the *seq* module.

3.2 Formal naming syntax

We shall now look in more detail at the syntax of these names. A function name in L^AT_EX3 will have a name consisting of three parts:

$$\backslash\langle module\rangle_ \langle description\rangle : \langle arg-spec\rangle$$

while a variable will have (up to) four distinct parts to its name:

$$\backslash\langle scope\rangle_ \langle module\rangle_ \langle description\rangle_ \langle type\rangle$$

The syntax of all names contains

$$\langle module\rangle \text{ and } \langle description\rangle$$

these both give information about the command.

A *module* is a collection of closely related functions and variables. Typical module names include **int** for integer parameters and related functions, **seq** for sequences and **box** for boxes.

Packages providing new programming functionality will add new modules as needed; the programmer can choose any unused name, consisting of letters only, for a module. In general, the module name and module prefix should be related: for example, the kernel module containing **box** functions is called **l3box**.

The *description* gives more detailed information about the function or parameter, and provides a unique name for it. It should consist of letters and, possibly, `_` characters. In general, the description should use `_` to divide up ‘words’ or other easy to follow parts of the name. For example, the L^AT_EX3 kernel provides `\if_cs_exist:N` which, as might be expected, tests if a command name exists.

As a semi-formalized concept the letter **g** is sometimes used to prefix certain parts of the *description* to mark the function as ‘globally acting’, e.g., `\int_set:Nn` is a local operation while `\int_gset:Nn` is a global operation. This of course goes hand in hand with when to use `\l_` and `\g_` variable prefixes.

3.2.1 Variables: scope and type

The *scope* part of the name describes how the variable can be accessed. Variables are classified as local, global or constant. This *scope* type appears as a code at the beginning of the name; the codes used are:

- c** constants (global variables whose value should not be changed);
- g** variables whose value should only be set globally;
- l** variables whose value should only be set locally.

The $\langle type \rangle$ will be in the list of available *data-types*,¹ these include the primitive T_EX data-types, such as the various registers, but to these will be added data-types built within the L^AT_EX programming system.

The data types in L^AT_EX3 are:

int integer-valued count register;

tl ‘token list variables’: placeholders for token lists;

toks token register;

box box register;

skip ‘rubber’ lengths;

dim ‘rigid’ lengths;

num A ‘fake’ integer type using only macros. Useful for setting up allocation routines;

seq ‘sequence’: a data-type used to implement lists (with access at both ends) and stacks;

prop property list;

clist comma separated list;

stream an input or output stream (for reading from or writing to, respectively);

bool either true or false (the L^AT_EX3 implementation does not use `\iftrue` or `\iffalse`).

When the $\langle type \rangle$ and $\langle module \rangle$ are identical (as often happens in the more basic modules) the $\langle module \rangle$ part is often omitted for aesthetic reasons.

The name ‘token list’ may cause confusion, and so some background is useful. T_EX works with tokens and lists of tokens, rather than characters. It provides two ways to store these token lists: within macros and as token registers (toks). L^AT_EX3 retains the name ‘toks’ for the later, and adopts the name ‘token list variables’ (tl vars) for macros used to store tokens. In most circumstances, the tl var data type is more convenient for storing token lists.

3.2.2 Functions: argument specifications

Function names end with an $\langle arg-spec \rangle$ after a colon. This gives an indication of the types of argument that a function takes, and provides a convenient method of naming similar functions that differ only in their argument forms (see the next section for examples).

The $\langle arg-spec \rangle$ consists of a (possibly empty) list of letters, each denoting one argument of the function. The letter, including its case, conveys information about the type of argument required.

All functions have a base form with arguments using one of the following argument specifiers:

¹Of course, if a totally new data type is needed then this will not be the case. However, it is hoped that only the kernel team will need to create new data types.

- n** Unexpanded token or braced token list.
This is a standard \TeX undelimited macro argument.
- N** Single token (unlike **n**, the argument must *not* be surrounded by braces).
A typical example of a command taking an **N** argument is `\cs_set`, in which the command being defined must be unbraced.
- p** Primitive \TeX parameter specification.
This can be something simple like `#1#2#3`, but may use arbitrary delimited argument syntax such as: `#1,#2\q_stop#3`. This is used when defining functions.
- T,F** These are special cases of **n** arguments, used for the true and false code in conditional commands.

There are two other specifiers with more general meanings:

- D** This means: **Do not use**. This special case is used for \TeX primitives and other commands that are provided for use only while bootstrapping the \LaTeX kernel. Programmers outside the kernel team should not use these functions!
- w** This means that the argument syntax is ‘weird’ in that it does not follow any standard rule. It is used for functions with arguments that take non standard forms: examples are \TeX -level delimited arguments and the boolean tests needed after certain primitive `\if...` commands.

In case of **n** arguments that consist of a single token the surrounding braces can be omitted in nearly all situations—functions that force the use of braces even for single token arguments are explicitly mentioned. However, programmers are encouraged to always use braces around **n** arguments, as this makes the relationship between function and argument clearer.

Further argument specifiers are available as part of the expansion control system. These are discussed in the next section.

4 Expansion control

Let’s take a look at some typical operations one might want to perform. Suppose we maintain a stack of open files and we use the stack `\g_ior_file_name_seq` to keep track of them (`io` is the file reading and writing module: it uses the module prefix `ior` for reading functions and variables). The basic operation here is to push a name onto this stack which could be done by the operation

```
\seq_gpush:Nn \g_ior_file_name_seq {#1}
```

where `#1` is the filename. In other words, this operation would push the file name as is onto the stack.

However, we might face a situation where the filename is stored in a variable of some sort, say `\l_ior_curr_file_tl`. In this case we want to retrieve the value of the variable. If we simply use

```
\seq_gpush:Nn \g_io_file_name_seq \l_io_curr_file_tl
```

we will not get the value of the variable pushed onto the stack, only the variable name itself. Instead a suitable number of `\exp_after:wN` would be necessary (together with extra braces)² i.e.

```
\exp_after:wN
\seq_gpush:Nn
\exp_after:wN
\g_io_file_name_seq
\exp_after:wN
{ \l_io_curr_file_tl }
```

The above example is probably the simplest case but already shows how the code changes to something difficult to understand. Furthermore there is an assumption in this: that the storage bin reveals its contents after exactly one expansion. Relying on this means that you cannot do proper checking plus you have to know exactly how a storage bin acts in order to get the correct number of expansions. Therefore L^AT_EX3 provides the programmer with a general scheme that keeps the code compact and easy to understand.

To denote that some argument to a function needs special treatment one just uses different letters in the arg-spec part of the function to mark the desired behaviour. In the above example one would write

```
\seq_gpush:NV \g_io_file_name_seq \l_io_curr_file_tl
```

to achieve the desired effect. Here the V (the second argument) is for ‘retrieve the value of the variable’ before passing it to the base function.

The following letters can be used to denote special treatment of arguments before passing it to the base function:

c Character string used as a command name.

The argument (a token or braced token list) must, when fully expanded, produce a sequence of characters which is then used to construct a command name (via `\csname ... \endcsname`). This command name is the single token that is passed to the function as the argument. Hence

```
\seq_gpush:cV { g_file_name_stack } \l_tmpa_tl
```

is equivalent to

```
\seq_gpush:NV \g_file_name_stack \l_tmpa_tl.
```

V Value of a variable.

This means that the register in question is returned, be it an integer, a length-type register, a token list variable or similar. The value is returned to the function as a braced token list.

v Value of a register, constructed from a character string used as a command name.

This is a combination of **c** and **V** which first constructs a control sequence from the argument and then returns the value.

²`\exp_after:wN` is the L^AT_EX3 name for the T_EX `\expandafter` primitive.

- x** Fully-expanded token or braced token list.
This means that the argument is expanded as in the replacement text of an `\edef`, and the expansion is passed to the function as a braced token list. This means that expansion takes place until only unexpandable tokens are left.
- o** One-level-expanded token or braced token list.
This means that the argument is expanded one level, as by `\expandafter`, and the expansion is passed to the function as a braced token list. Note that if the original argument is a braced token list then only the first token in that list is expanded. In general, using **V** should be preferred to using **o** for simple variable retrieval.
- f** Almost the same as the **x** type except here the token list is expanded fully until the first unexpandable token is found and the rest is left unchanged. Note that if this function finds a space at the beginning of the argument it will gobble it and not expand the next argument.

4.1 Simpler means better

Anyone who programs in \TeX is frustratingly familiar with the problem of arranging that arguments to functions are suitably expanded before the function is called. To illustrate how expansion control can bring instant relief to this problem we shall consider two examples copied from `latex.ltx`.

```
\global\expandafter\let
\csname\cf@encoding \string#1\expandafter\endcsname
\csname ?\string#1\endcsname
```

This first piece of code is in essence simply a global `\let` whose two arguments firstly have to be constructed before `\let` is executed. The `#1` is a control sequence name such as `\textcurrency`. The token to be defined is obtained by concatenating the characters of the current font encoding stored in `\cf@encoding`, which has to be fully expanded, and the name of the symbol. The second token is the same except it uses the default encoding `?`. The result is a mess of interwoven `\expandafter` and `\csname` beloved of all \TeX programmers, and the code is essentially unreadable.

Using the conventions and functionality outlined here, the task would be achieved with code such as this:

```
\cs_gset_eq:cc
{ \cf@encoding \token_to_str:N #1 } { ? \token_to_str:N #1 }
```

The command `\cs_gset_eq:cc` is a global `\let` that generates command names out of both of its arguments before making the definition. This produces code that is far more readable and more likely to be correct first time. (`\token_to_str:N` is the \LaTeX 3 name for `\string`.)

Here is the second example.


```

\expandafter
\in@
\csname sym#3%
\expandafter
\endcsname
\expandafter
{%
\group@list}%

```

This piece of code is part of the definition of another function. It first produces two things: a token list, by expanding `\group@list` once; and a token whose name comes from ‘`sym#3`’. Then the function `\in@` is called and this tests if its first argument occurs in the token list of its second argument.

Again we can improve enormously on the code. First we shall rename the function `\in@` according to our conventions. A function such as this but taking two normal ‘`n`’ arguments might reasonably be named `\seq_test_in:Nn`; thus the variant function we need will be defined with the appropriate argument types and its name will be `\seq_test_in:cV`. Now this code fragment will be simply:

```

\seq_test_in:cV { sym #3 } \l_group_seq

```

Note that, in addition to the lack of `\expandafter`, the space after the `}` will be silently ignored since all white space is ignored in this programming environment.

4.2 New functions from old

For many common functions the L^AT_EX3 kernel will provide variants with a range of argument forms, and similarly it is expected that extension packages providing new functions will make them available in the all the commonly needed forms.

However, there will be occasions where it is necessary to construct a new such variant form; therefore the expansion module provides a straightforward mechanism for the creation of functions with any required argument type, starting from a function that takes ‘normal’ T_EX undelimited arguments.

To illustrate this let us suppose you have a ‘base function’ `\demo_cmd:Nnn` that takes three normal arguments, and that you need to construct the variant `\demo_cmd:cnx`, for which the first argument is used to construct the *name* of a command, whilst the third argument must be fully expanded before being passed to `\demo_cmd:Nnn`. To produce the variant form from the base form, simply use this:

```

\cs_generate_variant:Nn \demo_cmd:Nnn { cnx }

```

This defines the variant form so that you can then write, for example:

```

\demo_cmd:cnx { abc } { pq } { \rst \xyz }

```

rather than ... well, something like this!

```

\def \tempa {{pq}}%
\edef \tempb {\rst \xyz}%
\expandafter
\demo@cmd:nnn
\csname abc%
\expandafter
\expandafter
\expandafter
\endcsname
\expandafter
\tempa
\expandafter
{%
\tempb
}%

```

Another example: you may wish to declare a function `\demo_cmd_b:xcxcx`, a variant of an existing function `\demo_cmd_b:nnnnn`, that fully expands arguments 1, 3 and 5, and produces commands to pass as arguments 2 and 4 using `\csname`. The definition you need is simply

```

\cs_generate_variant:Nn \demo_cmd_b:nnnnn { xcxcx }

```

This extension mechanism is written so that if the same new form of some existing command is implemented by two extension packages then the two definitions will be identical and thus no conflict will occur.

5 The distribution

At present, the `expl3` modules are designed to be loaded on top of $\text{\LaTeX} 2_{\epsilon}$. In time, a $\text{\LaTeX} 3$ format will be produced based on this code. This allows the code to be used in $\text{\LaTeX} 2_{\epsilon}$ packages *now* while a stand-alone $\text{\LaTeX} 3$ is developed.

While `expl3` is still experimental, the bundle is now regarded as broadly stable. The syntax conventions and functions provided are now ready for wider use. There may still be changes to some functions, but these will be minor when compared to the scope of `expl3`.

New modules will be added to the distributed version of `expl3` as they reach maturity. At present, the `expl3` bundle consists of a number of modules, most of which are loaded by including the line:

```

\RequirePackage{expl3}

```

in a $\text{\LaTeX} 2_{\epsilon}$ package, class or other file. The `expl3` modules regarded as stable, and therefore suitable for basing real code on, are as follows:

l3basics This contains the basic definition modules used by the other packages.

l3box Primitives for dealing with boxes.

l3calc A re-implementation of the L^AT_EX 2_ε package `calc` package in `expl3` that provides extended methods for numeric and dimensional calculations and assignments.

l3clist Methods for manipulating comma-separated token lists.

l3expan This is the argument expansion module discussed earlier in this document.

l3int This implements the integer data-type `int`.

l3intexpr For carrying out integer expression calculations and comparisons, whether the input is stored in `int`, `tl` or other variable types, or is given literally.

l3io A module providing low level input and output functions.

l3keyval Low-level implementation for processing lists of the form `{ key1=val1 , key2=val2 }`.

l3keys Higher-level implementation of key–value methods: intended to work as a L^AT_EX 3 version of `xkeyval/kvoptions`, although with input syntax more like that of `pgfkeys`.

l3msg Communicating with the user: includes low-level hooks to allow messages to be filtered (higher-level interface for filtering to be written!).

l3names This sets up the basic naming scheme and renames all the T_EX primitives.

l3num This implements the ‘fake integer’ datatype `num`.

l3prg Program control structures such as boolean data type `bool`, generic do-while loops, case-switches, sorting routines and stepwise loops.

l3prop This implements the data-type for ‘property lists’ that are used, in particular, for storing key/value pairs.

l3quark A ‘quark’ is a command that is defined to expand to itself! Therefore they must never be expanded as this will generate infinite recursion; they do however have many uses, e.g. as special markers and delimiters within code.

l3seq This implements data-types such as queues and stacks.

l3skip Implements the ‘rubber length’ datatype `skip` and the ‘rigid length’ datatype `dim`.

l3tl This implements a basic data-type, called a *token-list variable* (`tl var.`), used for storing named token lists: these are essentially T_EX macros with no arguments.

l3token Analysing token lists and token streams, including peeking ahead to see what’s coming next and inspecting tokens to detect which kind they are.

l3toks A data-type corresponding to T_EX’s primitive token registers.

6 Moving from L^AT_EX 2_ε to L^AT_EX 3

To help programmers to use L^AT_EX 3 code in existing L^AT_EX 2_ε package, some short notes on making the change are probably desirable. Suggestions for inclusion here are welcome! Some of the following is concerned with code, and some with coding style.

- `expl3` is mainly focussed on programming. This means that some areas still require the use of L^AT_EX 2_ε internal macros. For example, you may well need `\@ifpackageloaded`, as there is currently no native L^AT_EX 3 package loading module.
- User level macros should be generated using the mechanism available in the `xparse` package, which is available from CTAN or the L^AT_EX 3 SVN repository.
- At an internal level, most functions should be generated `\long` (using `\cs_new:Nn`) rather than ‘short’ (using `\cs_new_nopar:Nn`). The exceptions are:
 - Functions which take no arguments;
 - Functions which are used with pre-set arguments which therefore cannot be `\long`; this is mainly the case with auxilliary functions.
- Where possible, declare all variables and functions (using `\cs_new:NN`, `\tl_new:N`, etc.) before use.
- Prefer ‘higher-level’ functions over ‘lower-level’, where possible. So for example use `\cs_new:Nn` in preference to `\cs_new:Npn`, and `\cs_if_exist:N(TF)` over `\if_cs_exist:N`.
- Use space to make code readable. In general, we recommend a layout such as:

```
\cs_new:Nn \foo_bar:Nn {  
  \cs_if_exist:NTF #1 {  
    \foo_bar_aux_i:n {#2}  
  }{  
    \foo_bar_aux_ii:nn {#2} { literal }  
  }  
}
```

where spaces are used around `{` and `}` except for isolated `#1`, `#2`, etc.

- Put different code items on separate lines: readability is much more useful than compactness.
- Use long, descriptive names for functions and variables, and for auxilliary functions use the parent function name plus `aux`, `aux_i`, `aux_ii` and so on.
- If in doubt, ask the team via the LaTeX-L list: someone will soon get back to you!

References

- [1] Donald E Knuth *The T_EXbook*. Addison-Wesley, Reading, Massachusetts, 1984.
- [2] Goossens, Mittelbach and Samarin. *The L^AT_EX Companion*. Addison-Wesley, Reading, Massachusetts, 1994.
- [3] Leslie Lamport. *L^AT_EX: A Document Preparation System*. Addison-Wesley, Reading, Massachusetts, second edition, 1994.
- [4] Frank Mittelbach and Chris Rowley. ‘The L^AT_EX3 Project’. *TUGboat*, Vol. 18, No. 3, pp. 195–198, 1997.

6.1 expl3 implementation

Well, it’s not complicated :)

<*package>

```
1 \ProvidesExplPackage
2   {\filename}{\filedate}{\fileversion}{\filedescription}
```

We already loaded l3names at the beginning of the dtx file.

The l3chk package is omitted since it is only used for conditional processing with full error-checking turned on. Most users will generally not need to do this, and we haven’t set it up yet, anyway.

Fundamentals:

```
3 \RequirePackage{
4   l3basics,
5   l3expan,
6   l3tl,
7   l3num,
8   l3intexpr,
9   l3quark,
10  l3seq,
11  l3toks,
12  l3int,
13  l3prg,
14  l3clist,
15  l3token,
16  l3io,
17  l3prop,
18  l3msg,
19  l3skip,
20 }
```

All the rest:

```
21 \RequirePackage{
22   l3box,
23   l3keyval,
24   l3keys,
```

```
25  l3precom,  
26  l3calc,  
27  l3xref,  
28  l3file  
29 }  
  
</package>
```