**AHCIP Ingest Data File Format Specification -- Draft, 2014-08-14**

The data ingest format will be CSV files, with no headers.

Each line of the file will comprise the following (regular characters are literal, **BOLDCAPS** are symbolic references to a substring format, explained below). Ignore any line wrapping ... input lines will be one line terminated by newline characters.

"**PRIMSRCURN**","**CMNTSRCURN**","**CMNTTYPE**","**CMNTTEXT**"

Example:

"urn:cts:greekLit:tlg0013.tlg003:239-242","urn:cts:dmk:DouglasFrame.HippotaNestor:712-713","references","analysis of the one real exception to the hymn’s geographical coherence, the Kephisos River’s being out of place, as an expansion by the victors in the First Sacred War (to counter the earlier hymn’s Spartan propaganda following the Second Messenian War)"

**PRIMSRCURN** is the CTS greekLit URN for the [primary source] passage to which this commentary applies

**CMNTSRCURN** is the CTS ahcip URN for the source where this commentary is first made

**CMNTTYPE** is a brief notation of what type of commentary this is. Examples: "reference", "reference to scholia", "translation", "scholia quoted and translated", "reference to I.20.188-194", "Plato’s text of the verses quoted and translated as part of Hippias Minor 369d-370a", and [the winner of the prize for least brief notation]: "reference in the list of passages where the threat of fire to Achaeans from the Trojans is formalized in the theme of “Hektor will burn the ships of the Achaeans”" .

**Note:** While I suspect that this was originally intended to be one of a fixed set of values from a taxonomy of commentary types, our human annotators were not so instructed ... or they got carried away in their thoroughness. There are about 169 different variants in this field. This is probably something that will have to be fixed in the editing system maybe.

**CMNTTEXT** is a string of text that is "the commentary itself" [we all know that it's really a placeholder, but it's the starting point of the text of the commentary].

I believe that**CMNTTYPE** and **CMNTTEXT** will be ingested "as is"

**Caveat Programmer:*****if*** *it is desirable to actually have a taxonomy of commentary "types" in the system, that I recommend that a separate field be allocated for it in the data model and that what is uploaded as* ***CMNTTYPE****be stored as a "hint" for what the value of the commentary type should be set to when editing the commentary. If we have an initial list of what the initial items in that taxonomy should be, the ingest process can make an educated guess (e.g., if* ***CMNTTYPE****== "reference" and 'reference' is in the taxonomy, then set that value on ingest. Otherwise that value should be blank and the editor should set it from a controlled vocabulary menu (with the flexibility to add new types to the system as needed).*

The format of a **PRIMSRCURN** is as follows:

urn:cts:greekLit:**PSWORK**:**PSPASSAGE**

where **PSWORK** will come from a well-defined set of strings, such as tlg0012.tlg001 for Homer's *Iliad*  tlg0013.tlg003 for *Homeric Hymn to Apollo*, etc. There will be a complete list with correct Human-readable names.

**Caveat Programmer:** there are two optional components of **PSWORK** that we're not using presently that can be used to point to a specific version and instance (Chris and Neel say "exemplar" in the [CTS URN specification](http://www.homermultitext.org/hmt-docs/specifications/ctsurn/)). It is something to keep far in the back of our minds if AHCIP is going to actually refer to and/or retrieve text from other online sources, that we'll have to map between these abstract text references and a URI (with embedded full CTS URN) of the text.

**PSPASSAGE** format depends on what **PSWORK** is:

Where **PSWORK** == tlg0012.tlg001 or **PSWORK** == tlg0012.tlg002, the format will be either **BOOK**.**LINE** or a range **BOOK1**.**LINE1**-**BOOK2**.**LINE2** ... where **BOOK\*** is always an integer > 0 and **LINE\*** is either an integer > 0. In some cases, the integer is sufffixed with letters or numerals, such as "a" or "b" or "c1" or "f." or "ff."

**Note:** I'm waiting for an explanation of what the letters mean. I am also waiting to hear about whether we are going to keep the references to scholia in the **PRIMSRCURN** as an appended "@scholia"... but I'm not sure that should be in the input data.

The format of a **CMNTSCRURN** is as follows:

urn:cts:ahcip:**CSWORK:CSPASSAGE**

We are, essentially, creating CTS URNs in a namespace that will be named **ahcip**  **(dmk is going away)**.  We do need to maintain a catalogue of what CTS URNs we create in the ahcip namespace, although we might never stand up a proper CTS server that knows how to serve up the text in bite-size pieces of XML. There will be a complete list with correct Human-readable names.

**CSWORK** will come from a well-defined set of strings, such as GregoryNagy.HomerTheClassic for Greg's *Homer the Classic*, etc.