AHCIP: MRBTP and Ingest Data File Format Specification

### Updated: 2014-08-14

# MRBTP = Modified Ryan Baumann Transform Process

The original input data to AHCIP are Microsoft Word documents, created by researchers Anita Nikkanen, Ian Tewksbury, etc. to document placeholders to commentary in secondary references written by Homeric scholoars (Nagy, Frame, Muellner, Elmer). These files were coded in a Stream Format (SF), not documented here. It is necessary to convert these stream format files to CSV files in the AHCIP Ingest Data File Format (AIDFF).

The transformation of SF files to AIDFF files comprises several steps:

1. Save the Word document in SF as a text file, characters encoded as UTF-8 Unicode, with lines terminated by line-feed characters (U+000A). In the git repository, these files are stored in folder **src\_txt/** as ***filename***.txt .
2. Correct any syntax errors in the SF text files. If such a correction is made, the unchanged text from the Word document is renamed ***filename***.txt.ORIGINAL . Changes due to syntax errors are not documented in the commentary placeholder.
3. Convert the text SF files to CSV files, using a Groovy script found in **scripts/**. The end result of this transformation is referred to as "phase1" in the testing folders.
4. Transform individual elements in each line from the notation used by the researchers to the AIDFF format. Most of the transformation is to convert the passage references in Column 1 to CTS URNs in the **greekLit** namespace, and to convert text references in Column 2 to CTS URNs in the **ahcip** namespace. The end result of this transformation is referred to as "phase2" in the testing folders.
5. Due to slight differences in the syntax used by researchers from one file to the next, it is necessary to examine the output files at this point, determine whether improperly converted lines were due to a syntax error, or the need to modify the transformation scripts, then repeat steps 2-4.

**The following notations in SF are not encoded in the URN.**

* Source passage reference followed by f. or ff. : Deleted from URN, noted as [MRBPT: f] or [MRBPT: ff] in column 4
* Source passage reference followed by (scholia): Deleted from URN, noted as [MRBPT: scholia] in column 4
* Source passage reference followed by any other parenthetic note, such as (P.Oxy. III.428) or (bT): Deleted from URN, noted as [MRBPT: **parenthetic note**] in column 4
* For whatever reason, there was no source type reference in the commentary line. If there are only three columns from the SF conversion, this will be noted as [MRBPT: missing type] in column 2

# AHCIP Ingest Data File Format Specification

The data ingest format will be UTF-8 Unicode text files, formatted as Comma-separated Values (CSV) lines. The field separator will be the comma (U+0027) character. Each field will be surrounded by double-quotation marks (U+0022).

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:ahcip:DouglasFrame.HippotaNestor:s\_712-s\_ 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 in the first 16 SF files. This is probably something that will have to be fixed in the editing system maybe.

**CMNTTEXT** is a string of text that is the initial commentary itself.

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 3* (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 ("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. It is also valid for a CTS URN to be suffixed with a **REFERENCE** component which begins with an @ and is followed by a word in the referenced passage.

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

* Where **PSWORK** begins with tlg0012.\* (*Iliad*  and *Odyssey)* 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 or, 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.

* Where **PSWORK** == tlg0013.\* (*Homeric Hymns*) the format will be **LINE** or a range **LINE1-LINE2** where **LINE\*** is always an integer > 0.

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

urn:cts:ahcip:**CSWORK:CSSECTION**

We are, essentially, creating CTS URNs in a namespace that will be named **ahcip**. We do need to maintain a catalogue of what CTS URNs we create in the ahcip namespace, for when we 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. All values that are valid are in section "Works in AHCIP with URNS."

**CSSECTION** format depends on what **CSWORK** is. The section reference depends on what type of work and how it is structured and how we plan to model it in the system ore represent that structure in the URN. We have chosen to use an encoding format that explicitly tags the name of the structural elements as well as the actual name of the element this part of the URN refers to. The abbreviation for the name of the element will be separated from the element value by the underscore (U+005F) character.

The following table of structural element tags will be used inside the **CSSECTION** of AHCIP URNs. Additional tags may be added as needed:

|  |  |
| --- | --- |
| **c** | **Chapter** |
| **s** | **Section** |
| **n** | **Note** |
| **t** | **Translation** |

The value for each element be numerical, or symbolic. For example, *Homer the Classic* (HC), comprises structurally the following "chapters:" Preface, Prolegomena, Chapter 1, Chapter 2, Chapter 3, Chapter 4, Conclusions and Bibliography. In order to keep URNs a manageable size, we will use a three-character abbreviation for each chapter type. So, the following URNs can be used to retrieve the chapters of HC:

* urn:cts:ahcip:GregoryNagy.HomerTheClassic:c\_Pre
* urn:cts:ahcip:GregoryNagy.HomerTheClassic:c\_Pro
* urn:cts:ahcip:GregoryNagy.HomerTheClassic:c\_1
* urn:cts:ahcip:GregoryNagy.HomerTheClassic:c\_2
* urn:cts:ahcip:GregoryNagy.HomerTheClassic:c\_3
* urn:cts:ahcip:GregoryNagy.HomerTheClassic:c\_4
* urn:cts:ahcip:GregoryNagy.HomerTheClassic:c\_Con
* urn:cts:ahcip:GregoryNagy.HomerTheClassic:c\_Bib

The values of other subsection identifiers are usually numeric (they don't have to be).

In the case where a note was indicated in the SF document, but no note number was provided, the MRBTP will use a value of 0 to act as a sentinel that the reference was to the note, associated with that section or translation. Some examples:

* urn:cts:ahcip:GregoryNagy.HomerTheClassic:c\_2.s\_344.n\_0 = A note (or "the" note) to Section 344 in Chapter 2 of HC
* urn:cts:ahcip:GregoryNagy.HomerTheClassic:c\_Pro.s\_127.t\_8 = Translation 8 in Section 127 in the Prolegomena of HC
* urn:cts:ahcip:GregoryNagy.HomerTheClassic:c\_2.s\_235.t\_76.n\_0 = A note to Translation 76 in Section 235 of Chapter 2 of HC
* urn:cts:ahcip:GregoryNagy.HomerTheClassic = *Homer the Classic* (in it's entirety)
* urn:cts:ahcip:GregoryNagy.HomerTheClassic:c\_Con = Entire Conclusion of *Homer the Classic*

No specific structural element tag is required. Whether the URN represents a meaningful reference to a section in the work depends on how the URNs are created. To refer to an entire work, leave off the **CSSECTION** and the colon after **CSWORK**

# Works in AHCIP with URNs

|  |  |  |  |
| --- | --- | --- | --- |
| **Abbreviation** | **URN** | **Author** | **Title** |
| HN | urn:cts:ahcip:DouglasFrame.HippotaNestor | Douglas Frame | *Hippota Nestor* |
| HC | urn:cts:ahcip:GregoryNagy.HomerTheClassic | Gregory Nagy | *Homer the Classic* |
| BA | urn:cts:ahcip:GregoryNagy.TheBestOfTheAchaeans | Gregory Nagy | *The Best of the Achaeans* |
| PH | urn:cts:ahcip:GregoryNagy.PindarsHomer | Gregory Nagy | *Pindar's Homer* |
| ME | urn:cts:ahcip:LeonardMuellner.TheMeaningOfHomericEYXOMAI | Leonard Muellner | *The Meaning of Homeric EYXOMAI Through Its Formulas* |
| AA | urn:cts:ahcip:LeonardMuellner.TheAngerOfAchilles | Leonard Muellner | *The Anger of Achilles :* [*Mēnis in Greek Epic*](http://books.google.com/books?id=--zPn_8LgL0C&printsec=frontcover&dq=muellner+anger+of+achilles&hl=en&sa=X&ei=-Xn6U_HpIoGkyQSA_4G4BQ&ved=0CB8Q6AEwAA) |
| MR | urn:cts:ahcip:DouglasFrame.TheMythOfReturn | Douglas Frame | *The Myth of Return in Early Greek Epic* |
| HPC | urn:cts:ahcip:GregoryNagy.HomerThePreclassic | Gregory Nagy | *Homer the Preclassic* |
| PP | urn:cts:ahcip:GregoryNagy.PoetryAsPerformance | Gregory Nagy | *Poetry as Performance : Homer and Beyond* |