CONTENTS

REGULAR EXPRESSIONS USED ON FDSYS CHRG TEXT FILES

STAR PRINT NUMBER

```
starprintP =

   "(\\s*star print\\s*)|(\\s*<star>\\s*)"

   CASE_INSENSITIVE | MULTILINE
```

COVER PAGE

```
preCoverPageT = "\r?\n(?:--+|__+)\r?\n"
```

```
pre2CoverPageT = "<spaces>-{5,}<spaces>"
```

```
coverSubCommitteePageP =

   "(^\\s*(?:(?:before (?:the|a))|(?:hearing\\s+of the))\r?\n\r?\n?[\\W\\w]+
(?:CONGRESS|SESSION)\r?\n)"

   CASE_INSENSITIVE | MULTILINE | PRESERVE_WHITESPACE
```

```
coverPageP =

   "(.*\r?\n\r?\n\\s*(?:(?:Printed (?:for the use|by authority
)of)|(?:U\\.S\\. GOVERNMENT (PRINTING|PUBLISHING) OFFICE\r?\n)|(?:-
{8,}\r?\n\\s*For sale by the Superintendent of Documents, )|(?:-
{8,}\r?\n\\s*For sale by the U\\.S\\. Government (Printing|Publishing)
Office\r?\n)))|(.*\r?\n\\s*-{5,})|(.*\r?\n\\d{2}-\\d{3}|(.*\r?\nWashington,
D.C.\r?\n))"

   CASE_INSENSITIVE | MULTILINE | DOTALL | PRESERVE_WHITESPACE)
```

```
coverPage2P =

   "(U\\.S\\. GOVERNMENT (PRINTING|PUBLISHING) OFFICE\r?\n|House of
representatives)"

   CASE_INSENSITIVE | MULTILINE | DOTALL | PRESERVE_WHITESPACE
```

```
coverPageEnd1P =
   "(\\s*Printed (?:for the use|by authority) of)"
   CASE_INSENSITIVE | MULTILINE

coverPageEnd2P =
   "(^\\s*U\\.S\\. GOVERNMENT (PRINTING|PUBLISHING) OFFICE$)"
   CASE_INSENSITIVE | MULTILINE

coverPageEnd3P =
   "(-{8,}\r?\n\\s*For sale by the U\\.S\\. Government (Printing|Publishing) Office\r?\n|C O N T E N T S)"
   CASE_INSENSITIVE | MULTILINE

coverPageEnd4P =
   "(-{8,}\r?\n\\s*For sale by the Superintendent of Documents, U\\.S\\. GOVERNMENT (PRINTING|PUBLISHING) OFFICE\r?\n)"
   CASE_INSENSITIVE | MULTILINE

coverPageEnd5P =
   "(\r?\n\\d{2}-\\d{3})"
   CASE_INSENSITIVE | MULTILINE

coverPageEnd6P =
   "(\r?\n\\s*Washington, D\\.?C\\.\r?\n)"
   CASE_INSENSITIVE | MULTILINE

coverPageEnd7P =
   "(\r?\n\\s*-{5,})"
   CASE_INSENSITIVE | MULTILINE

coverPageEnd8P =
   "(House of representatives)"
   CASE_INSENSITIVE | MULTILINE
```

TITLE

```
/**
 * Following expression matches a title followed by a . and finished by a
 * sequence of 2 break lines.
 */
titleP =
   "^\\.\r?\n\\s*(.+)","[=_-]{2,}|(hearing\\s*\\n)"
   MULTILINE | DOTALL

titleP2 =
   "^\\.(?:\\s+S\\. +Hrg\\. +\\d+-\\d+)?\r?\n\\s*((?:.+\r?\n)*)"
   MULTILINE
```

APPROPRIATION TITLE

```
appropriationTitleP =
   "(Appropriations?)"
   CASE_INSENSITIVE
```

APPROPRIATION COMMITTEE

```
appropriationCommitteeP =
   "\\s*(Committee on Appropriations),?\\s*"
   CASE_INSENSITIVE
```

TYPE

```
/**
```

```
 * Matches following entry:
 * Field Hearing Field Briefing
 * Only "H" character is returned as "type".
 */
typeFP =
   "\n\\s+(F)ield (?:Hearing|Briefing)\r?\n"
   CASE_INSENSITIVE

/**
 * Used to search in coverPage. Matches the following entries:
 * Hearings and Markup Mark up Markup Markups Compilation of Markups
 * Only "M" character is returned as "type."
 */
typeMP =
   "\n\\s+(?:Hearings and |Compilation of )?(M)ark ?ups?"
   CASE_INSENSITIVE

/**
 * Used to search in title only. Matches the following entries:
 * Mark-up on Markup on Markup of
 * Only "M" character is returned as "type."
 */
typeM2P =
   "(M)ark-?up (?:on|of)"
   CASE_INSENSITIVE

/**
 * Used to search in coverPage only. Matches the following entries:
 *         Oversight Hearing
 *         Oversight Hearings
 *          Oversight and Legislative Hearing
 *          Oversight and Legislative Hearings
 *          Oversight Field Hearing Oversight
 *          Field Hearings
 * Only "O" character is returned as "type."
 */
typeOP =
   "\n\\s+(O)versight (?:and Legislative |Field )?Hearings?\r?\n"
   CASE_INSENSITIVE

/**
 * Used to search in title only. Matches the following entries:
 * Oversight Hearing
 * Only "O" character is returned as "type."
 */
typeO2P =
   "\n\\s+(O)versight Hearing"
   CASE_INSENSITIVE

/**
 * Used to search in title only. Matches the following entries:
 *          Authorization
 *          Authorization on
 *          Reauthorization
 *          Reauthorization on
 * Only "AU" character is returned as "type.
 */
typeAUP =
   "(?:Re)?(au)thorization(?: on)?"
   CASE_INSENSITIVE
```

```
/**
 * Used to search in title only. Matches the following entries:
 *          Treaty on
 *          Treaty act
 * Only "T" character is returned as "type.
 */
typeTP =
   "(T)reaty (?:on|act)"
   CASE_INSENSITIVE
```

CHAMBER

```
/**
 * Extracting chamber.
 * Matches following text in one single line:
 * "House of representatives" "U.S. House of representatives"
 */
chamberHouseP =
   "\n\\s*(?:U.S. )?(HOUSE)(?: OF)? REPRESENTATIVES?\\s*\r?\n"
   CASE_INSENSITIVE

/**
 * Extracts chamber.
 * Matches following text in one single line:
 * "United States Senate"
 */
chamberSenateP =
   "\n\\s*UNITED STATES (SENATE)\r?\n"
   CASE_INSENSITIVE
```

```
chamberSenateAltP =
   " {3,}S\\. +Hrg\\."
   CASE_INSENSITIVE


/**
 * Extracts chamber.
 * Matches following text in one single line:
 * "JOINT COMMITTEE HEARING" "JOINT HEARING"
 */
chamberJointP =
   "(JOINT) (?:COMMITTEE )?HEARING"
   CASE_INSENSITIVE


chamberFallBackP =
   "(?:(Senate)|(House) of Representatives?)"
   CASE_INSENSITIVE
```

HELPER PATTERNS EMBEDDED IN OTHER REGULAR EXPRESSIONS

```
decimalsT = "ten|twenty|thirty|forty|fifty|sixty|seventy|eighty|ninety"

teensT =
"eleven|twelve|thirteen|fourteen|fifteen|sixteen|seventeen|eighteen|nineteen"

numbers1T = "one|two|three|four|five|six|seven|eight|nine"

ordinals1T = "first|second|third|fourth|fifth|sixth|seventh|eighth|ninth"

ordinals2T = "(?:" + teensT + ")th"

ordinals3T = "(?:" + decimalsT + ")th"

ordinals4T = "(?:" + decimalsT + ")-(?:" + ordinals1T + ")"

ordinalsT = ordinals1T + "|" + ordinals2T + "|" + ordinals3T + "|" +
ordinals4T
```

CONGRESS, HEARING NUMBER, PART NUMBER

```
/**
 * Extracts congress and chamber number.
 * Gets the "S. Hrg. #-#." First number is Congress. Second is chamber
 * number.
 */
congressNumberP =
   "\\s*S\\. Hrg\\. (\\d{3})-(\\d+)(?:, Pa?r?t\\.? ([\\dIVX]+\\w?))?\r?\n"
   CASE_INSENSITIVE | MULTILINE
```

CONGRESS LETTERS

```
/**
 * Extracts numbers in letters in Congress.
 * Matches following expressions:
 * One hundred first congress Two hundred eleventh congress
 */
congressNumberLettersP =
   "\n\\s*((?:" + numbers1T + ") hundred (?:AND )?(?:" + ordinalsT + "))
CONGRESS"
   CASE_INSENSITIVE
```

CONGRESS ORIGINAL, CONGRESS

```
congressNumber2P =
   "\n\\s*(\\d{3})TH CONGRESS\r?\n"
   CASE_INSENSITIVE | MULTILINE
```

CONGRESS, NUMBER, DETAIL

```
/**
 * Extracts the Congress number and serial number.
 * Matches following expressions:
```

```
 * Serial #-# Serial Number #-# Serial No. #-# Serial No. #-#, Part 1
 * First number is the Congress number. Second number is the serial number.
 */
serialNumberP =
   "\\s*Serial (?:Number |No\\. )?(\\d+)-([\\d\\w]+)(?:, Part (\\d))?"
   CASE_INSENSITIVE
```

CONGRESS, NUMBER

```
/**
 * Extracts the Congress number serial number.
 * Matches following expressions:
 * #-#
 * (#-#)
 * First number is the Congress number. Second number is the serial number.
 */
serialNumberWithParanP =
   "\n\\s*\\(?(\\d{3})-([\\d\\w]+)\\)?\r?\n"
   CASE_INSENSITIVE
```

NUMBERS

```
/**
 * Extracts the serial numbers as string. Needs post-process functionality.
 * Matches following expressions:
 * Serial Nos. #-#, #-# Serial Nos. #-#, #-#, and #-#
 * First numbers are the Congress number. Second numbers are the serial
 * number.
 */
serialNumbersP =
   "\n\\s*Serial (?:Numbers |Nos\\. )?((?:(?:and )?\\d+-
\\d+(?:,\\s*)?)+)\r?\n"
```

```
   CASE_INSENSITIVE


/**
 * Extracts #-#.
 * Matches two numbers separated by dash "-". Only second number is returned
 * as "number". This pattern is used in post-process.
 */
numbersP =
   "\\d+-(\\d+)"
```

HEARING NUMBER

```
/*
 * To extract number from S. Hrg. 110-161
 */
numbers2P =
   "S\\.\\s*Hrg\\.\\s*\\d+-(\\d+)"
   CASE_INSENSITIVE
```

VOLUME NUMBER

```
volumeNumberP =
   "volumeNumber" }, "\n\r?\n *Volume ([\\dIV]+)"
   CASE_INSENSITIVE | MULTILINE


volumeNumber2P =
   "S\\.\\s*Hrg\\.\\s*\\d+-\\d+\\.?\\s*(?:Vol|Volume)\\.?\\s*([\\dIV]+)"
   CASE_INSENSITIVE
```

PART NUMBER

```
partNumberP =
   "partNumber" }, "\r?\n\\s*Part (\\d)\\s*\r?\n"
   CASE_INSENSITIVE | MULTILINE

partNumber2P =
   "S\\.\\s*Hrg\\.\\s*\\d+-\\d+\\.?\\s*(?:Part|Pt)\\.?\\s*(\\d)"
   CASE_INSENSITIVE
```

SESSION

```
/**
 * Extracts if hearing is First Session.
 * Matches text:
 * First Session
 */
firstSessionP =
   "\n\\s*(FIRST) SESSION\r?\n"
   CASE_INSENSITIVE

/**
 * Extracts if hearing is Second Session.
 * Matches text
 * Second Session
 */
secondSessionP =
   "\n\\s*(SECOND) SESSION\r?\n"
   CASE_INSENSITIVE
```

HELD DATE

```
/**
 * Extracts dates.
 * Matches following expressions:
 * January 23, 2007
 * MARCH 1, 2003
 */
heldP =
   "((?:" + monthsAllT + ") \\d{1,2}, \\d{4}),?\r?\n"
   CASE_INSENSITIVE

heldP_Centered =
   "^\\s+((?:" + monthsAllT + ") \\d{1,2}, \\d{4})\\s*$"
   CASE_INSENSITIVE | MULTILINE
```

DATE

```
/**
 * Extracts dates.
 * Matches following expressions:
 * February 8, April 17, and May 15, 2007
 */
held3P =
   "((?:(?:" + monthsAllT + ") \\d{1,2},? (?:and )?){2,}\\d{4})"
   CASE_INSENSITIVE

held3P_Centered =
   "^\\s+((?:(?:" + monthsAllT + ") \\d{1,2},? (?:and )?){2,}\\d{4})\\s*$"
   CASE_INSENSITIVE | MULTILINE

/**
 * Extracts dates.
 * Matches following expressions:
 * APRIL 8, 9, 15 AND 17, 1997
 * JANUARY 30 AND 31, 2007
 */
held2P =
   "((?:" + monthsAllT + ") (?:\\s|,|\\d{1,2}|and){5,}\\d{4})"
   CASE_INSENSITIVE

held2P_Centered =
   "^\\s+((?:" + monthsAllT + ") (?:\\s|,|\\d{1,2}|and){5,}\\d{4})\\s*$"
   CASE_INSENSITIVE | MULTILINE

/**
 * Extracts dates.
 * Matches following expressions:
 *                THURSDAY, MARCH 13, 1997
 *                Friday, June 21, 2002.
 */
held4P =
   "^\\s*(?:" + daysAllT + "), ((?:(?:" + monthsAllT + ") \\d{1,2},
\\d{4}))\\.?\\s*$"
   CASE_INSENSITIVE | MULTILINE

/**
 * Extracts Dates
 * Matches the following expression:
 *                  February 6, 2009.--Ordered to be printed
 */

held6P =
   "((?:" + monthsAllT + ") \\d{1,2}, \\d{4}),?(?:\\.--Ordered to be
printed)?\r?\n"
   CASE_INSENSITIVE
```

```
/**
 * Extracts dates.
 * Matches following expressions:
 *                MARCH 1997
 */
held5P =
   "\\s*((?:" + monthsAllT + ") \\d{4})"
   CASE_INSENSITIVE


/**
 * Extracts Dates
 * Matches the following expression:
 *                  JANUARY 16-19, 2001
 */


held7P =
   "((?:" + monthsAllT + ") \\d{1,2}-\\d{1,2}, \\d{4})"
   CASE_INSENSITIVE
```

MONTHS

```
monthsAllT =
"january|february|march|april|may|june|july|august|september|october|november
|december"
```

DAYS

```
daysAllT = "monday|tuesday|wednesday|thursday|friday|saturday|sunday"
```

HELD DATE ALL

```
heldDateP0 =
   {"={3,}","\\n\\s*(?:(?:" + daysAllT + ")[, ]+)?((?:" + monthsAllT +
").+\\d{4}\\.?\\s*\\n)"}

   CASE_INSENSITIVE


heldDateP =

   "\\n\\s*(?:(?:" + daysAllT + ")[, ]+)?((?:" + monthsAllT +
").+\\d{4}\\.?\\s*\\n)"

   CASE_INSENSITIVE


heldDateP1 =

   "\\n\\s+(?:(?:" + daysAllT + ")[, ]+)?((?:" + monthsAllT +
").+\\d{4}\\.?\\s*\\n)"

   CASE_INSENSITIVE


heldDateP2 =

   "\\n\\s+(?:hearing held[\\w\\s,]*)(?:(?:" + daysAllT + ")[, ]+)?((?:" +
monthsAllT + ").+\\d{4}\\.?\\s*\\n)"

   CASE_INSENSITIVE


heldDateOrderedP =

   "\\n\\s*(?:(?:" + daysAllT + ")[, ]+)?((?:" + monthsAllT +
").+\\d{4}\\.(?:\\-\\-Ordered to be printed)?\\s*\\n)"

   CASE_INSENSITIVE
```

HELD MONTH

```
heldMonthP =

   "(" + monthsAllT + ")"

   CASE_INSENSITIVE
```

HELD DAY

```
heldDayP =

   {"[^0-9](\\d{1,2})[^0-9]",monthsAllT}
   CASE_INSENSITIVE
```

HELD YEAR

```
heldYearP =
   "(\\d{4})"
   CASE_INSENSITIVE
```

ERRATA

```
/**
 * Matches following entry:
 * ERRATA This errata sheet is being assigned because the original serial
 * number was incorrect. The correct serial number is 106-57.
 */
errataP =
   "[^\\.]\r?\n *\\[?errata\\]?\r?\n\\s*([^\\.](?:.+\r?\n)*)"
   CASE_INSENSITIVE | MULTILINE
```

ERRATA NUMBER

```
errataNumberSP =
   "\\d+err(\\d*)\\."
   CASE_INSENSITIVE
```

TABLE OF CONTENTS

```
contentsOldP =
   "^\\s+c ?o ?n ?t ?e ?n ?t ?s\r?\n$"
   CASE_INSENSITIVE | MULTILINE | PRESERVE_WHITESPACE | DOTALL


contentsEndOldP =
   "(\r?\n){3}"
   CASE_INSENSITIVE | MULTILINE | PRESERVE_WHITESPACE | DOTALL

contentsP =
   "\\n\\s+C *O *N *T *E *N *T *S\\s*[\\n\\s]+\\-*"
   CASE_INSENSITIVE

witnessContentsEndP =
   "\\n\\.\\s*\\n"
   CASE_INSENSITIVE

contentsEndP =
   "\\n\\s+\\-{8,}\\s+(?:\\-{2})?\\n"
   CASE_INSENSITIVE

contentsEndP2 =
   "\\.{5,}\\s*\\d+\\s*\\n"
   CASE_INSENSITIVE
```

AGENCY

```
/**
 * Used to search only in title.
 * Matches the following prefix in the title
 * Department of Departments of
 */
agencyTitleP =
   "\\.\r?\n\\s*(Departments? of [^\\W]+)"
   CASE_INSENSITIVE

/**
 * Used to search in the coverPage.
 * Matches following prefix in the coverPage

 * Department of <until new brake line>

 */

agencyCoverP =

   "\n\\s*(Department of [^\n]+)"

   CASE_INSENSITIVE
```

FISCAL YEAR

```
fiscalYearP =

   "(?:Fiscal\\s+year|Appropriations\\s+for)\\s+(\\d{4})"

   MULTILINE|CASE_INSENSITIVE
```

CONGHASC, CONGRESS, NUMBER

```
congHASCP =

   "\\[H\\.A\\.S\\.C\\. No\\. (\\d+)-(\\d+)\\]"

   CASE_INSENSITIVE
```

STATE NAMES PATTERNS

```
stateNamesPatternS =

"(A|a)labama|(A|a)laska|(A|a)rizona|(A|a)rkansas|(C|c)alifornia|(C|c)olorado|
(C|c)onnecticut|(D|d)elaware|" +
"(D|d)istrict\\s*(O|o)f\\s*(C|c)olumbia|(F|f)lorida|(G|g)eorgia|(H|h)awaii|(I
|i)daho|(I|i)llinois|(I|i)ndiana|(I|i)owa|(K|k)ansas|" +

"(K|k)entucky|(L|l)ouisiana|(M|m)aine|(M|m)aryland|(M|m)assachusetts|(M|m)ich
igan|(M|m)innesota|(M|m)ississippi|" +

"(M|m)issouri|(M|m)ontana|(N|n)ebraska|(N|n)evada|(N|n)ew\\s*(H|h)ampshire|(N
|n)ew\\s*(j|J)ersey|(N|n)ew\\s*(M|m)exico|(N|n)ew\\s*(Y|y)ork|" +

"(N|n)orth\\s*(C|c)arolina|(N|n)orth\\s*(D|d)akota|(O|o)hio|(O|o)klahoma|(O|o
)regon|(P|p)ennsylvania|(R|r)hode\\s*(I|i)sland|" +

"(S|s)outh\\s*(C|c)arolina|(S|s)outh\\s*(D|d)akota|(T|t)ennessee|(T|t)exas|(U
|u)tah|(V|v)ermont|(V|v)irginia|(W|w)ashington|" +

"(W|w)est\\s*(V|v)irginia|(W|w)isconsin|(W|w)yoming|North\\s{2,}|South\\s{2,}
|West\\s{2,}|New\\s{2,}"
```

STATE CODES PATTERNS

```
stateCodesPatternS =

"AL|AK|AZ|AR|CA|CO|CT|DE|DC|FL|GA|HI|ID|IL|IN|IA|KS|KY|LA|ME|MD|MA|MI|MN|MS|M
O|" +

"MT|NE|NV|NH|NJ|NM|NY|NC|ND|OH|OK|OR|PA|RI|SC|SD|TN|TX|UT|VT|VA|WA|WV|WI|WY"
```

CONGRESS MEMBER COMMITTEE

```
// original, name, state

stateP =

   "(?:\n|\\s{3,})\\s*((?:[A-Z][a-z]?[A-Z]+)\\.?\\s*(?:[A-Z][a-z'-]?[A-Z'`-
]+|[A-Z]\\.[A-Z])\\.?\\s*(?:(?:``)?[A-
Z]+(?:'')?\\.?\\s*){0,3}\\s*(?:,?\\s*(?:Jr|Sr|jr|sr|junior|senior|ii|iii|II|I
II)\\.)?),\\s*("+stateNamesPatternS+")"



state2P =

   "(?:\\n|  )([A-Z][a-zA-Z'-]*[A-Z]\\.? (?:\\b[A-Z][a-zA-Z'-]*[A-
Z]\\b|\\b[A-Z]\\b|[\\.\\(\\)]| (?! )|,
?(?:Jr|Sr|jr|sr|junior|senior|ii|iii|II|III))+),\\s*("+stateNamesPatternS+")"



state3P =

   "(?:\\n|  )([A-Z][a-zA-Z'-]*[a-z]\\.? (?:\\b[A-Z][a-zA-Z'-]*[a-
z]\\b|\\b[A-Z]\\.|[\\.\\(\\)]| (?! )|,
?(?:Jr|Sr|jr|sr|junior|senior|ii|iii|II|III))+),\\s*("+stateNamesPatternS+")"
```

OTHER HOSTING ORGANIZATION

```
otherHostingOrgP =

   "(?:before +(?:the|a))\\s+(.+?)(?:\\s*(?:" + numbers1T + ")\\s+hundred)"

   CASE_INSENSITIVE + DOTALL



otherHostingOrg2P =

   "(CONGRESSIONAL-EXECUTIVE COMMISSION ON\\s*[^\n]+)"

   CASE_INSENSITIVE
```

SUBCOMMITTEE

```
subCommitteeP3 =

   {"(CONGRESS\\s*\\n)|(SESSION\\s*\\n)","(subcommittee on.+?)[\\n\\s]{2,}" }

   CASE_INSENSITIVE + DOTALL


subCommitteeP =

   { "(subcommittee.+?)(?:\\s*\\n){2,}","(?<=\\n) +committee"}

   CASE_INSENSITIVE + DOTALL


subCommitteeP2 =

   {"(before|and|with) the","(.+?subcommittee.*?)(?:of|and|(?:(?:joint
)?with)) the\\s*\\n","\\n +committee"}

   CASE_INSENSITIVE + DOTALL
```

DAY, MONTH, YEAR

```
// extract date from filename

fileP =

   "(\\d{2})(ja|fe|mr|ap|my|jn|jy|au|se|oc|no|de)(\\d{4})"

   CASE_INSENSITIVE
```

NOMINEE INFORMATION

```
// last name, first name, position

nomineeTOC =

   "\\n(\\w[\\w\\- ]+?), +([\\w .]+?), +Nominee (.+?)\\.{5,}\\s*\\d+\\s*\\n"

   CASE_INSENSITIVE + DOTALL


nomineeTitle1 =

   "Nomination of (\\w+ (?:\\w\\.)?) ?(.+?), of ([\\w ]+), (to be .+)"
   CASE_INSENSITIVE


nomineeTitle2 =

   "Nomination of (?:\\w{2,4}\\. )?(\\w+ (?:\\w\\.)?) ?(.+?),? (to be .+)"

   CASE_INSENSITIVE


nomineeTitle3 =

   "Nomination of (\\w+ (?:\\w\\.)?) ?(\\w+)"

   CASE_INSENSITIVE
```

WITNESS TITLE

```
witnessTitleP =

   "\\n\\s*((?:Statements? of )?WITNESS(?:ES\\:?)?)\\s*(?:\\n[a-zA-
Z].+?\\:\\s*)?(?=\\n)"

   CASE_INSENSITIVE
```

WITNESS INFORMATION

```
witnessP =

   { "\\n(\\w.+?)(?:oral
(?:statement|testimony))?\\.*\\s+\\d+(?=\\s*\\n)","(\\n[a-zA-
Z][^\\n]+?\\:\\s*\\n)|((?<=\\n)\\s*\\n {10,40}[a-zA-Z][^\\n]*)"}

   CASE_INSENSITIVE + DOTALL


witnessP2 =

   {"\\n *(\\w.+?)(?:oral
(?:statement|testimony))?\\.*\\s+\\d+(?=\\s*\\n)","((?<=\\n)\\s*\\n[a-zA-
Z][^\\n]+?\\:\\s*\\n)|((?<=\\n)\\s*\\n {10,40}[a-zA-Z]+[^\\n]*)"}

   CASE_INSENSITIVE + DOTALL
```

COMMITTEE NAME

```
committeeP =

   {"\\n\\s+(committee.+?)(?:\\s*\\n){2,}","\\s_{2,}\\s+"}
   CASE_INSENSITIVE + DOTALL


committeeP2 =

   {"(of|and|before) the","(?<=\\n)\\s+(committee on.+)","(\\n\\s+and
the\\s*committee)|(HOUSE OF REPRESENTATIVES)|(UNITED STATES SENATE)"}

   CASE_INSENSITIVE + DOTALL


committeeP3 =

   {"\\n\\s+(committee on.+)","(\\n\\s+and the\\s*committee)|(HOUSE OF
REPRESENTATIVES)|(UNITED STATES SENATE)"}

   CASE_INSENSITIVE + DOTALL


committeeP4 =

   {"before the","\\n\\s+(.*?committee.+)","(CONGRESS OF THE UNITED
STATES)|(HOUSE OF REPRESENTATIVES)"}

   CASE_INSENSITIVE + DOTALL
```

JACKETID

```
File name matches: ("\\A\\d{5}([a-zA-Z]+[0-9a-zA-Z]*)?\\Z")

File name matches: ("\\A\\d{5,8}\\Z")
```

STANDARD REFERENCE REGULAR EXPRESSIONS USED ON FDSYS CHRG TEXT FILES

CONGRESSIONAL COMMITTEES

```
/** Routine to parse a committee name string. Returns a list of matching

 * committees.

 *

 * The result of running this utility is the standardized references for the

 * committee names.

 *

 * Parameters:
```

```
 *  Committee Name / Appropriation Committee: the committee name string
 *  Date: the date of the document being parsed
 *  Chamber: the default chamber in which to look for committees
 * returns a list of matching committee names.
*/

chamberP =
  "(house|representatives|senate|joint|jt)"
  CASE_INSENSITIVE
```

CONGRESS MEMBERS

```
/** Parse a Congress member name and then look it up in the authority files.
 * This routine will parse out all of the components of a name such as "Mr.
 * H. James Saxton Jr. of Ohio" and will then look it up in the authority
 * files.
 *
 * Parameters:
 *    extractedName:  The full name, as parsed, of the Congress member
 *    date: The issue date of the document which contains the reference
 *    chamber: The chamber to which the Congress member belongs and returns
 *    the <congMember> element object.
 */

prefixesP =
   "(?<!\\()\\b(ms|mr|miss|mrs)\\b"
   CASE_INSENSITIVE

prefixRemoveP =
   "(?<!\\()\\b(ms|mr|miss|mrs)\\b\\.?"
   CASE_INSENSITIVE
```

```
suffixesP =

   "\\b(jr|iii|sr|junior|senior)\\b"

   CASE_INSENSITIVE


suffixRemoveP =

   ",?\\s*\\b(jr|iii|sr|junior|senior)\\b\\.?"

   CASE_INSENSITIVE


stateNamesPatternS =
"ALABAMA|ALASKA|ARIZONA|ARKANSAS|CALIFORNIA|COLORADO|CONNECTICUT|DELAWARE|" +

"DISTRICT\\s*OF\\s*COLUMBIA|FLORIDA|GEORGIA|HAWAII|IDAHO|ILLINOIS|INDIANA|IOWA|KANSAS|" +

"KENTUCKY|LOUISIANA|MAINE|MARYLAND|MASSACHUSETTS|MICHIGAN|MINNESOTA|MISSISSIPPI|" +

"MISSOURI|MONTANA|NEBRASKA|NEVADA|NEW\\s*HAMPSHIRE|NEW\\s*JERSEY|NEW\\s*MEXICO|NEW\\s*YORK|" +

"NORTH\\s*CAROLINA|NORTH\\s*DAKOTA|OHIO|OKLAHOMA|OREGON|PENNSYLVANIA|RHODE\\s*ISLAND|" +

"SOUTH\\s*CAROLINA|SOUTH\\s*DAKOTA|TENNESSEE|TEXAS|UTAH|VERMONT|VIRGINIA|WASHINGTON|" +

"WEST\\s*VIRGINIA|WISCONSIN|WYOMING"


stateCodesPatternS =

"AL|AK|AZ|AR|CA|CO|CT|DE|DC|FL|GA|HI|ID|IL|IN|IA|KS|KY|LA|ME|MD|MA|MI|MN|MS|MO|" +

"MT|NE|NV|NH|NJ|NM|NY|NC|ND|OH|OK|OR|PA|RI|SC|SD|TN|TX|UT|VT|VA|WA|WV|WI|WY"


stateNameP =

   "\\bof\\s+(" + stateNamesPatternS + ")"

   CASE_INSENSITIVE


stateCodeP =

   "\\(\\s*(" + stateCodesPatternS + ")"
   CASE_INSENSITIVE


isStateCodeP =

    stateCodesPatternS

    CASE_INSENSITIVE


stateRemoveP =

   "(\\([^\\)]*\\)|\\bof\b.*)"

   CASE_INSENSITIVE


lnfNameP =

   "([\\p{L}][\\p{L}'\\.-]*)\\s*,\\s*([\\p{L}][\\p{L}'\\.-]*\\b)"

   CASE_INSENSITIVE


fnfNameP =

   "([\\p{L}][\\p{L}'\\.-]*)\\s+(?:[\\p{L}][\\p{L}'\\.-
]*\\s+)*([\\p{L}][\\p{L}'\\.-]*)"

   CASE_INSENSITIVE


lastNameOnlyP =

   "([\\p{L}][\\p{L}'-]*)"

   CASE_INSENSITIVE
```

PUBLIC LAWS

```
lawContentP =

"(\\b(?:Public|Private|Pub|Priv|Pvt|P)\\.*\\s*(?:Law|L|R)\\.*)\\s*(?:No\\.)?\
\s*(\\d+)[-\\xAD]+\\s*(\\d+)"

   CASE_INSENSITIVE | MULTILINE


multiLawContentP =
"(\\b(?:Public|Private|Pub|Priv|Pvt)\\.?\\s*(?:Laws|L)\\.?)\\s*(?:Nos?\\.|Numbers?)?(\\s*\\d+[-\\xAD]+\\s*\\d+(?:\\b\\d+[-\\xAD]+\\s*\\d+|,|and|\\s+)+)"

   CASE_INSENSITIVE | MULTILINE


multiLawNumbersP =

   "(\\d+)[-\\xAD]*\\s*(\\d+)")
```

UNITED STATES CODE

```
uscT =

   "U\\.?\\s*S\\.?\\s*C(?:\\.|ode)?\\s*"


postAUscT =

   "app\\.|Appendix"


singleSectionNoCaptureRegex =

   "\\d[a-z0-9-]*\\b(?:\\([a-z0-9]+\\))*(?:\\s+note|\\s+et seq\\.?)?"


singleChapterNoCaptureRegex =

 "\\d[a-z0-9-]*\\b"


/**

 * Matches following formats: chapter 8 of title 212, United States Code

 * Section 1477 of title 10, United States Code

 */

usCodeLargeP =

   "(?:sections?\\s*(\\w+)\\s*(?:of\\s*))?CHAPTERS?\\s*(\\d+[a-z]*) of title (\\d+),\\s*UNITED\\s*STATES\\s*CODE"

   CASE_INSENSITIVE | MULTILINE


/**
```

```
 * Matches the following formats: 42 USC 1526 42 U.S.C. 1526 42 U.S.
 * Code 1526 42 US Code 1526. All previous formats plus the following appendix
 * and details 42 USC app. 1526 42 USC appendix 1526 42 USC app. 1526, 1551,
 * 1553, 1555, and 1561
 *
 */
usCodeShortA2P =
   "([0-9]+)\\s*" + uscT + "(?:" + postAUscT + ")\\s*(?:((?:(?:and )?\\d+[a-
z]*(?:,\\s*)?)+(?:-[\\w]+)?)((?:\\([\\w]+\\))*\\s*(?:note|et seq\\.)?))"
   CASE_INSENSITIVE | MULTILINE


usCodeLarge2P =
   "CHAPTER\\s*(\\d+[a-z]*)(?: \\([^\\)]*\\)) of title
(\\d+),\\s*UNITED\\s*STATES\\s*CODE"
   CASE_INSENSITIVE | MULTILINE


usCodeMultiLargeSectionsBP =
"sections?\\s+(.{1,100}?)\\s+of\\s+title\\s+(\\d+)(?:,|\\sof\\s+the)?\\s+unit
ed\\s+states\\s+code"
   CASE_INSENSITIVE | DOTALL


usCodeMultiShortSectionsP =
   "([0-9]+)\\s*" + uscT + "(?:sections?|sec\\.?)?\\s*" + "((?:" +
singleSectionNoCaptureRegex + "(?!\\s+"  + uscT + ")" +
"|and|through|,|\\s)+)"
   CASE_INSENSITIVE | MULTILINE


usCodeMultiShortChaptersP =
   "([0-9]+)\\s*" + uscT + "(?:chapters?|ch\\.?)\\s*" + "((?:" +
singleChapterNoCaptureRegex + "(?!\\s+"  + uscT + ")" +
"|and|through|,|\\s)+)"
   CASE_INSENSITIVE | MULTILINE
```

STATUTES AT LARGE

```
statuteAtLargeP =

   "([0-9]+)\\s*STAT\\.\\s*(?:L\\.)?\\s*(\\d+[a-z]*(?:-[0-9]+)?(?: et
seq\\.)?)"

   CASE_INSENSITIVE | MULTILINE
```

CONGRESSIONAL BILLS

```
/**

 * Matches following formats: "bill_type" "volume"

 *

 * Where "bill_type" could be:

 *

 * House Bill | H. R. | H.R. | HR Senate Bill |

 * S. House Resolution | H.Res. | H. Res | H. Res. | H Res | House Res. | H.

 * Resolution Senate Resolution | Senate Res. | S. Resolution | S.Res. | S.

 * Res | S. Res. House Joint Resolution | H. J. Resolution | H.J. Resolution

 * | H.J.Res. | H. J. Res. | H.J. Res Senate Joint Resolution | S. J.

 * Resolution | S.J. Resolution | S.J.Res. | S. J. Res. | S.J. Res House

 * Concurrent Resolution | H. Con. Resolution | H.Con. Resolution | H. Con

 * Res.| H. Con Res Senate Concurrent Resolution | S. Con. Resolution |

 * S.Con. Resolution | S. Con Res.| S. Con Res

 *

 * And "volume" could be: 1094 949-950 No. 119 23 (107th Congress) 213 of

 * the 107th Congress

 *

 */

billContentP =

   \\b(?:( + billTypeT + ")\\s*(?:No\\.)?\\s*(\\d+)(?:-
\\s*(\\d+))?)\\s*(?:(?:of\\sthe|\\()\\s*(\\d+)(t|th|nd|d|rd)\\sCONGRESS\\)?)?
"

   CASE_INSENSITIVE
```

```
billContent2P =

   "(" + billTypeT + ")\\s*(?:No\\.)?\\s*(\\d+),?(?:-
\\s*(\\d+))?\\s+(?:of\\s+the|\\()?\\s*(?:One\\s+Hundred\\s+([a-z]+)|(One
Hundredth))\\s+Congress"

   CASE_INSENSITIVE
```

```
billContent3P =

   "(" + billTypeT + ")\\s*(?:No\\.)? *(\\d+),?(?:- *(\\d+))? +(?:of
+the|\\()? *(Ninety-(?:[a-z]+)|(Ninetieth)) +Congress"

   CASE_INSENSITIVE
```

```
billContent4P =

   "(" + billTypeT + ")\\s*(?:No\\.)? *(\\d+),?(?:- *(\\d+))? +(?:of
+the|\\()? *(Eighty-(?:[a-z]+)|(Eightieth)) +Congress"

   CASE_INSENSITIVE
```

```
billContent5P =

   "(" + billTypeT + ")\\s*(?:No\\.)? *(\\d+),?(?:- *(\\d+))? +(?:of
+the|\\()? *(Seventy-(?:[a-z]+)|(Seventieth)) +Congress"

   CASE_INSENSITIVE
```

```
billContent6P =

   \\b(?:( + billTypeT + ")\\s*(?:No\\.)?\\s*(\\d+)(?:-
\\s*(\\d+))?),\\s*(\\d{2,3})-[1-2], +"

   CASE_INSENSITIVE
```

CONGRESSIONAL REPORTS

```
reportVariationsT =

   "(?:Rept|Rpt|Report)\\.?\\s*(?:(?:Number|No)\\.?)?"
```

```
congReportHouseT =

   "\\b((?:House|H)\\.?\\s* + reportVariationsT + ")"
```

```
referenceDocT =

   "\\s+(\\d+)\\s*-\\s*(\\d+)\\s*(,\\s*((?:Pt\\.|Part|Volume|Vol\\.|and|[0-9])*\\s*(?:Pt\\.|Part|Vol\\.|Volume|[0-9])(\\([^)]+\\)\\s*)))?"


congReportSenateT =

   \\b((?:Senate|S)\\.?\\s* + reportVariationsT + ")"


congReportExecutiveT =

   "((?:Ex|Exec|Executive)\\.?\\s*" + reportVariationsT + ")"


congReportConferenceT =

   "((?:Conf|Conference)\\.?\\s*" + reportVariationsT + ")"


congReportHouseP =

   congReportHouseT + referenceDocT

   CASE_INSENSITIVE | MULTILINE


congReportSenateP =

   congReportSenateT + referenceDocT

   CASE_INSENSITIVE | MULTILINE


congReportExecutiveP =

   congReportExecutiveT + referenceDocT

   CASE_INSENSITIVE | MULTILINE


congReportConferenceP =

   congReportConferenceT + referenceDocT

   CASE_INSENSITIVE | MULTILINE


congReportHouseConferenceP =
   "(Conference Report \\(?:H\\. Rept\\.\\))" + referenceDocT

   CASE_INSENSITIVE | MULTILINE
```

CONGRESSIONAL DOCUMENTS

```
referenceDocT =

   "\\s+(\\d+)\\s*-\\s*(\\d+)\\s*(,\\s*((?:Pt\\.|Part|Volume|Vol\\.|and|[0-9])*\\s*(?:Pt\\.|Part|Vol\\.|Volume|[0-9])(\\([^)]+\\)\\s*)))?"


docVariationsT =

   "(?:Doc|Document|Documentation)\\.?\\s*(?:(?:No|Number)\\.?)?"


congDocumentHouseT =

   "((?:H|House)\\.?\\s*" + docVariationsT + ")"


referenceT =

   "(?:\\.?\\s+([\\dA-Z]+),\\s*(?:(\\d+)(?:th|nd|rd|st)\\s*(?:Cong|Congress)\\.?))"


congDocumentSenateT =

   "((?:S|Senate)\\.?\\s*" + docVariationsT + ")"


congDocumentHouseP =

   congDocumentHouseT + referenceDocT

   CASE_INSENSITIVE | MULTILINE


congDocumentHouse3P =

   congDocumentHouseT + referenceT

   CASE_INSENSITIVE | MULTILINE


congDocumentSenateP =

   congDocumentSenateT + referenceDocT
   CASE_INSENSITIVE | MULTILINE
```

CONGRESSIONAL HEARINGS

```
referenceDocT =

   "\\s+(\\d+)\\s*-\\s*(\\d+)\\s*(,\\s*((?:Pt\\.|Part|Volume|Vol\\.|and|[0-9])*\\s*(?:Pt\\.|Part|Vol\\.|Volume|[0-9])(\\([^)]+\\)\\s*)))?"


senateCongressHearingsT =

   "(S)(?:enate)?\\.?\\s*(?:Hrg|Hearing)\\.?"


senateCongressHearingsP =

   senateCongressHearingsT + referenceDocT

   CASE_INSENSITIVE | MULTILINE
```

CONGRESSIONAL COMMITTEE PRINTS

```
referenceDocT =

   "\\s+(\\d+)\\s*-\\s*(\\d+)\\s*(,\\s*((?:Pt\\.|Part|Volume|Vol\\.|and|[0-9])*\\s*(?:Pt\\.|Part|Vol\\.|Volume|[0-9])(\\([^)]+\\)\\s*)))?"


senateCongCommPrintsT =
   "(S)(?:enate)?\\.?\\s*(?:Prt|Print)\\.?\\s*(?:(?:No|Number)\\.?)?"


senateCongCommPrintsP =

   senateCongCommPrintsT + referenceDocT

   CASE_INSENSITIVE | MULTILINE
```

CODE OF FEDERAL REGULATIONS

```
cfrContentPNew =

   "([1-50])\\s*CFR\\s*(Chapters?|Ch\\.|Parts?|Sec\\.|sections?)*\\s*"

   // number:
```

```
+ "([\\d]+|\\d+|[ILMVX]+),?\\s*"
// detail:
+ "((?:(?:et (seq|al)\\.)|"
+ "(?:\\s*(and|through|or|,)\\s+)|"
+ "(?:(\\d+,?\\s)+)|"
+ "(?:\\-?\\d+(?:\\.\\d+)?(?:\\([0-9a-z]+\\))*)|"
+ "(?:\\.[0-9a-z]+(?:\\([0-9a-z]+\\))*))+)?"
CASE_INSENSITIVE
```