NCI SEER Edits Engine

Interoperable Approach to Data Validation

Fabian Depry
Information Management Services
NAACCR Conference, June 2011

Today's Presentation

- Introduction to the SEER Edits Engine
  - What is it?
  - Why do we need it?
  - How was it implemented?
  - How is it used?
  - What are our plans for the future?

What is the SEER Edits Engine?

- Java framework to validate incidence data
  - Used by Java applications that cannot use SEER and other edit sets via GenEdits
  - A library of Java source code that is used by Java programs maintained by the SEER program (SEER*DMS, SEER*Abs, SEER*Edits)
- Executes edits against data in any format
  - Text files, e.g., NAACCR Abstract files
  - Values entered on data entry forms
  - Values stored in proprietary database structures

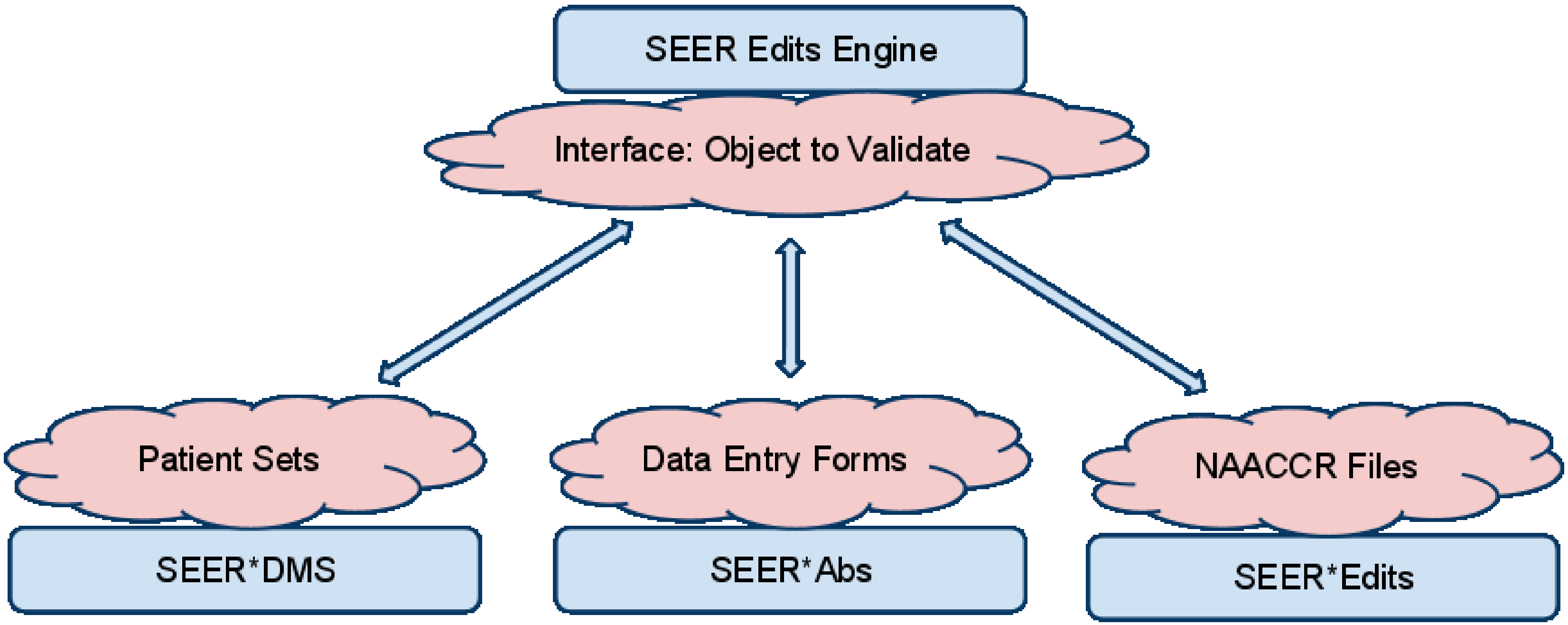
Why build a new engine?

- SEER*DMS has specific needs:
  - Validate data stored in a relational database
  - Allow registry-specific edits to be maintained through the application
  - Support a mechanism to easily test the edits logic
- SEER*Abs and SEER*Edits:
  - Validate data in data entry forms (SEER*Abs)
  - Validate data in text files (SEER*Edits)
  - Import edits written in other software (both)

System Requirements

- **For validating cancer data:**
  - Reliability
  - Speed
- **For maintaining edits:**
  - Simple but powerful syntax
  - Testing framework to verify the edits logic
  - Graphical tools to assist writing and updating edits
- **For maintaining the edits engine:**
  - Operability: running edits on different data types

Supporting Different Data Types



Edits Syntax - Groovy

- Groovy is a scripting language based on Java:
  - Anything that can be done in Java can be done in Groovy
  - Groovy also has unique syntax to allow for small, elegant scripts
  - Groovy scripts can also use readily available Java libraries
- Edits are based on Boolean logic.
  - If an edit returns "true", it passes. Examples:
    - *return line.primarySite ==~ /C\d\d\d/*
    - *return Hist_ICD_O_3_Table.contains(line.histologyIcdO3)*

Edits storage - XML

- The engine provides an API to read and write eXtensible Markup Language (XML) files:

```
<rule id="primary-site" name="Primary Site“ ruleset="field">
        <expression>return line.primarySite ==~ /C\d\d\d/</expression>
        <message>Primary site is not valid.</message>
</rule>
```

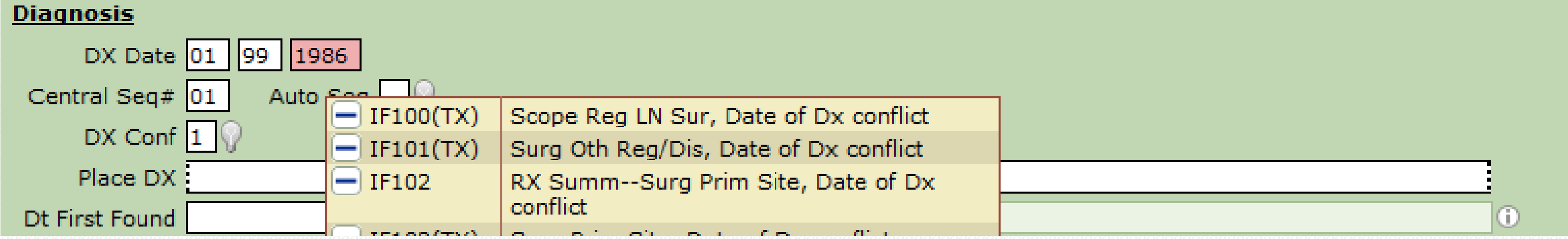
- The edits can actually be stored in any format
  - SEER*DMS persists edit source code in a database table
  - SEER*Abs and SEER*Edits – edits are stored in XML files

Engine Speed

- The key to super-speed: multi-threading
  - The more resources available, the faster the edits will run
- In SEER*DMS:
  - 1,294 edits ran on data for 269,727 patients using a Linux server
  - 58 minutes → 78 patients/second (may include multiple tumors)
- In SEER*Edits:
  - 582 SEER edits ran on 9,422,096 records using a 64-bit Windows desktop with two dual-core processors
  - 3 hours 37 minutes → 723 records/second
- In SEER*Abs:
  - Edits are executed on one record at a time, not in batches
  - Validates data entry form each time the user exits a field

Use case – SEER*DMS

SEER*DMS Editor:



Diagnosis

DX Date 01 99 1986

Central Seq# 01 Auto Seq

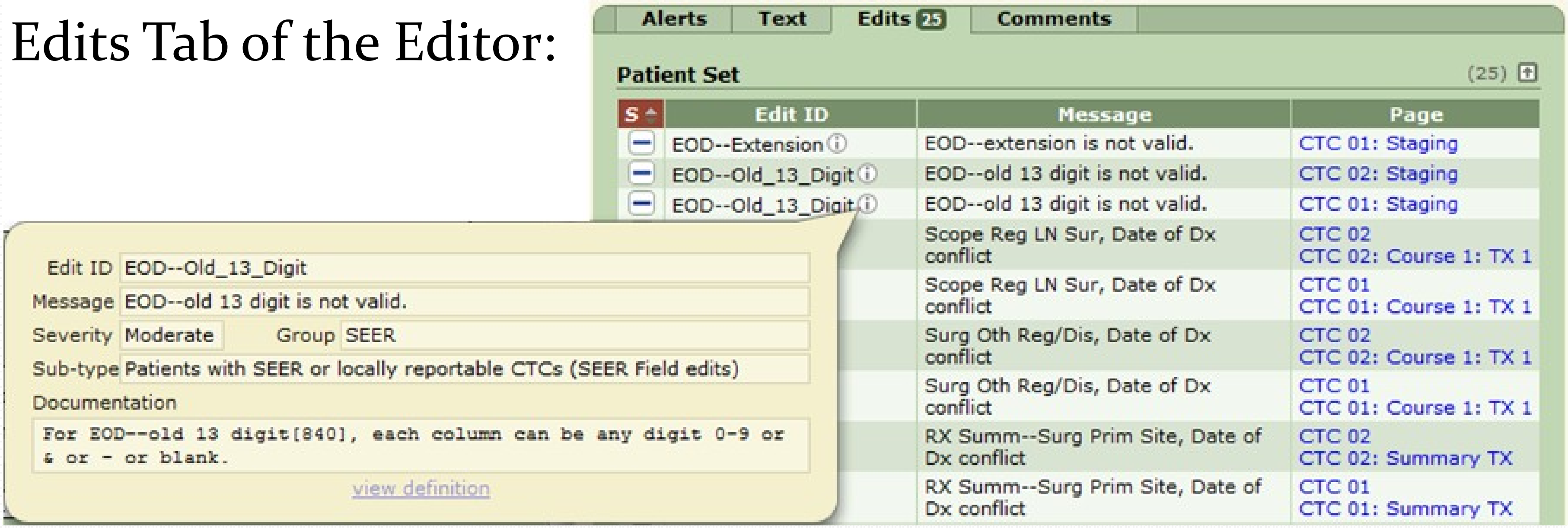
DX Conf 1

Place DX

Dt First Found

−	IF100(TX)	Scope Reg LN Sur, Date of Dx conflict
−	IF101(TX)	Surg Oth Reg/Dis, Date of Dx conflict
−	IF102	RX Summ--Surg Prim Site, Date of Dx conflict

Edits Tab of the Editor:



Patient Set (25)

S	Edit ID	Message	Page
−	EOD--Extension	EOD--extension is not valid.	CTC 01: Staging
−	EOD--Old_13_Digit	EOD--old 13 digit is not valid.	CTC 02: Staging
−	EOD--Old_13_Digit	EOD--old 13 digit is not valid.	CTC 01: Staging
		Scope Reg LN Sur, Date of Dx conflict	CTC 02 CTC 02: Course 1: TX 1
		Scope Reg LN Sur, Date of Dx conflict	CTC 01 CTC 01: Course 1: TX 1
		Surg Oth Reg/Dis, Date of Dx conflict	CTC 02 CTC 02: Course 1: TX 1
		Surg Oth Reg/Dis, Date of Dx conflict	CTC 01 CTC 01: Course 1: TX 1
		RX Summ--Surg Prim Site, Date of Dx conflict	CTC 02 CTC 02: Summary TX
		RX Summ--Surg Prim Site, Date of Dx conflict	CTC 01 CTC 01: Summary TX

Edit ID EOD--Old_13_Digit

Message EOD--old 13 digit is not valid.

Severity Moderate Group SEER

Sub-type Patients with SEER or locally reportable CTCs (SEER Field edits)

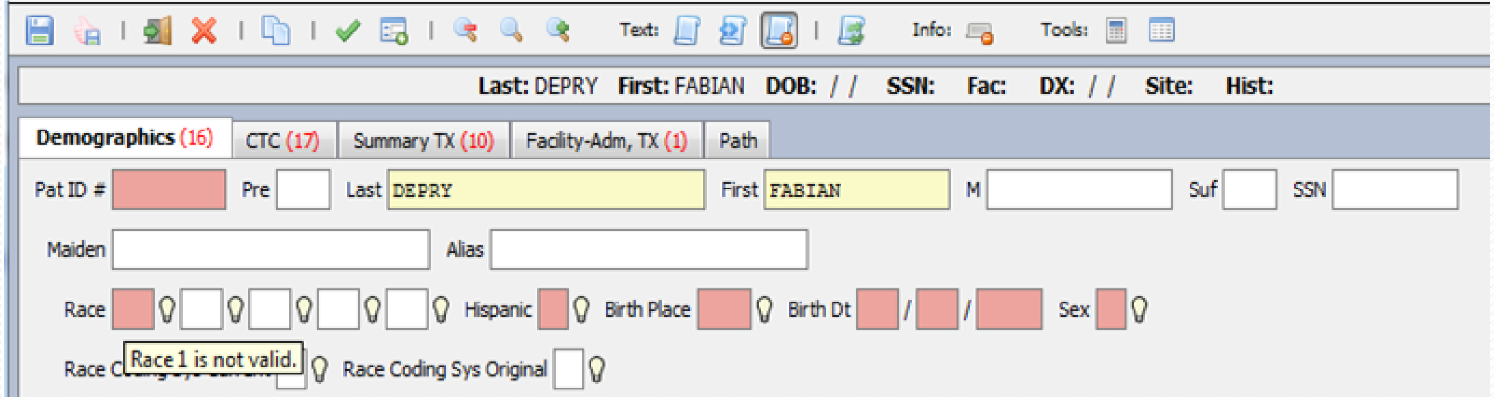
Documentation

```
For EOD--old 13 digit[840], each column can be any digit 0-9 or
& or - or blank.
```

view definition

Use case – SEER*Abs

Data Entry Form:



Last: DEPRY First: FABIAN DOB: / / SSN: Fac: DX: / / Site: Hist:

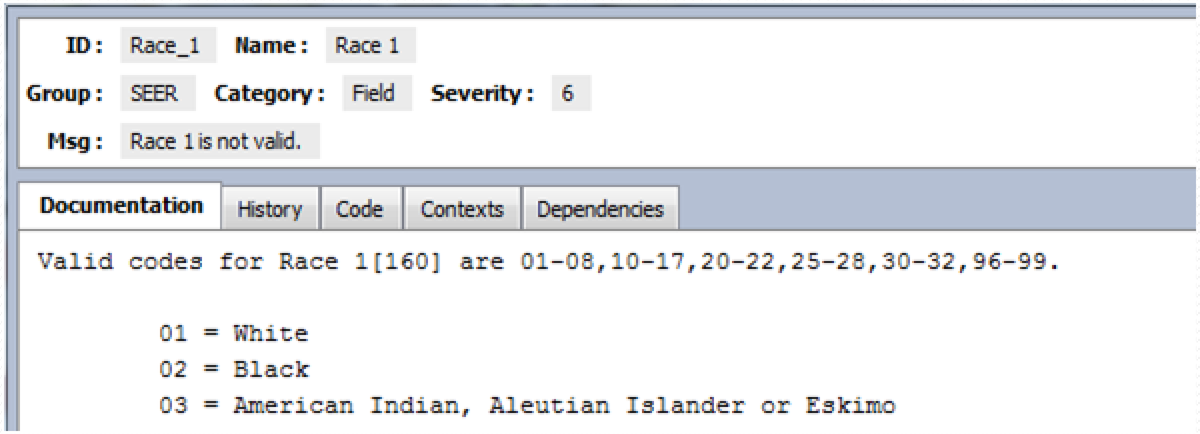
Pat ID # Pre Last DEPRY First FABIAN M Suf SSN

Maiden Alias

Race Hispanic Birth Place Birth Dt / / Sex

Race C Race 1 is not valid. Race Coding Sys Original

Edit Documentation:



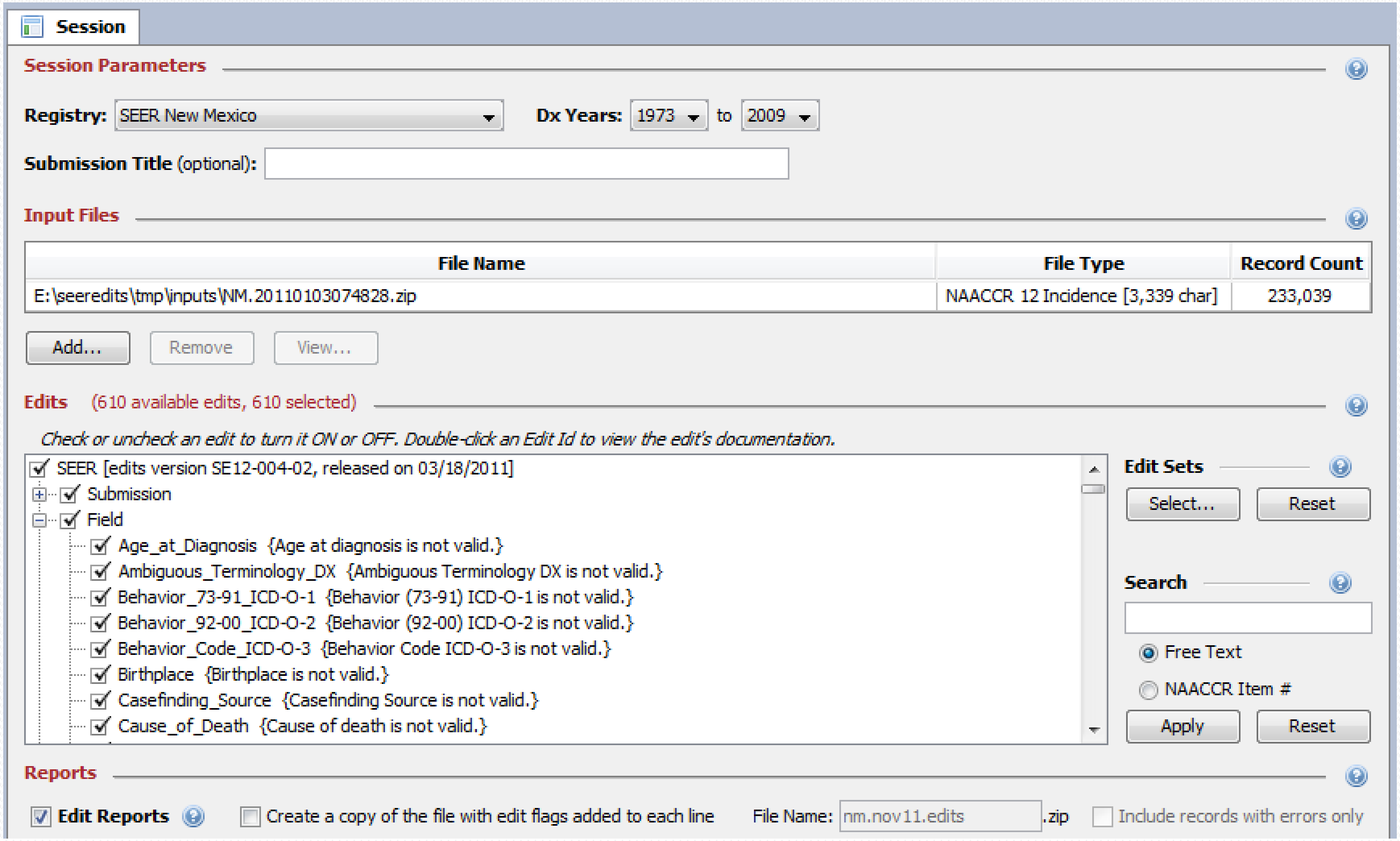
ID : Race_1 Name : Race 1

Group : SEER Category : Field Severity : 6

Msg : Race 1 is not valid.

```
Valid codes for Race 1[160] are 01-08,10-17,20-22,25-28,30-32,96-99.

        01 = White
        02 = Black
        03 = American Indian, Aleutian Islander or Eskimo
```

Use case – SEER*Edits (Session)



Session

Session Parameters

Registry: SEER New Mexico ▾ **Dx Years:** 1973 ▾ to 2009 ▾

Submission Title (optional):

Input Files

File Name	File Type	Record Count
E:\seeredits\tmp\inputs\NM.20110103074828.zip	NAACCR 12 Incidence [3,339 char]	233,039

Add... Remove View...

Edits (610 available edits, 610 selected)

Check or uncheck an edit to turn it ON or OFF. Double-click an Edit Id to view the edit's documentation.

- ☑ SEER [edits version SE12-004-02, released on 03/18/2011]
  - ⊞ ☑ Submission
  - ⊟ ☑ Field
    - ☑ Age_at_Diagnosis {Age at diagnosis is not valid.}
    - ☑ Ambiguous_Terminology_DX {Ambiguous Terminology DX is not valid.}
    - ☑ Behavior_73-91_ICD-O-1 {Behavior (73-91) ICD-O-1 is not valid.}
    - ☑ Behavior_92-00_ICD-O-2 {Behavior (92-00) ICD-O-2 is not valid.}
    - ☑ Behavior_Code_ICD-O-3 {Behavior Code ICD-O-3 is not valid.}
    - ☑ Birthplace {Birthplace is not valid.}
    - ☑ Casefinding_Source {Casefinding Source is not valid.}
    - ☑ Cause_of_Death {Cause of death is not valid.}

Edit Sets

Select... Reset

Search

● Free Text

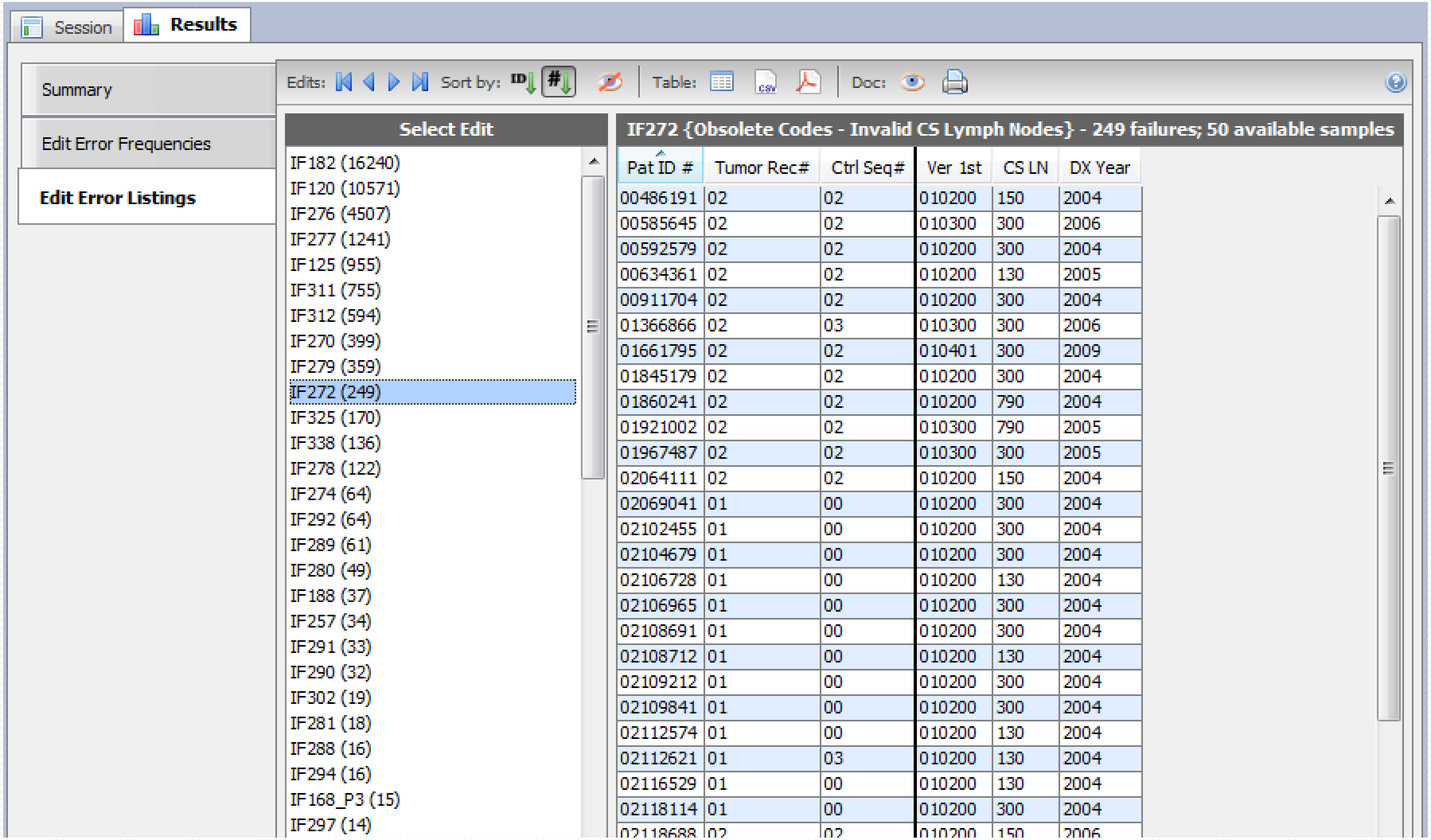
○ NAACCR Item #

Apply Reset

Reports

☑ **Edit Reports** □ Create a copy of the file with edit flags added to each line File Name: nm.nov11.edits .zip □ Include records with errors only

Use case – SEER*Edits (Results)

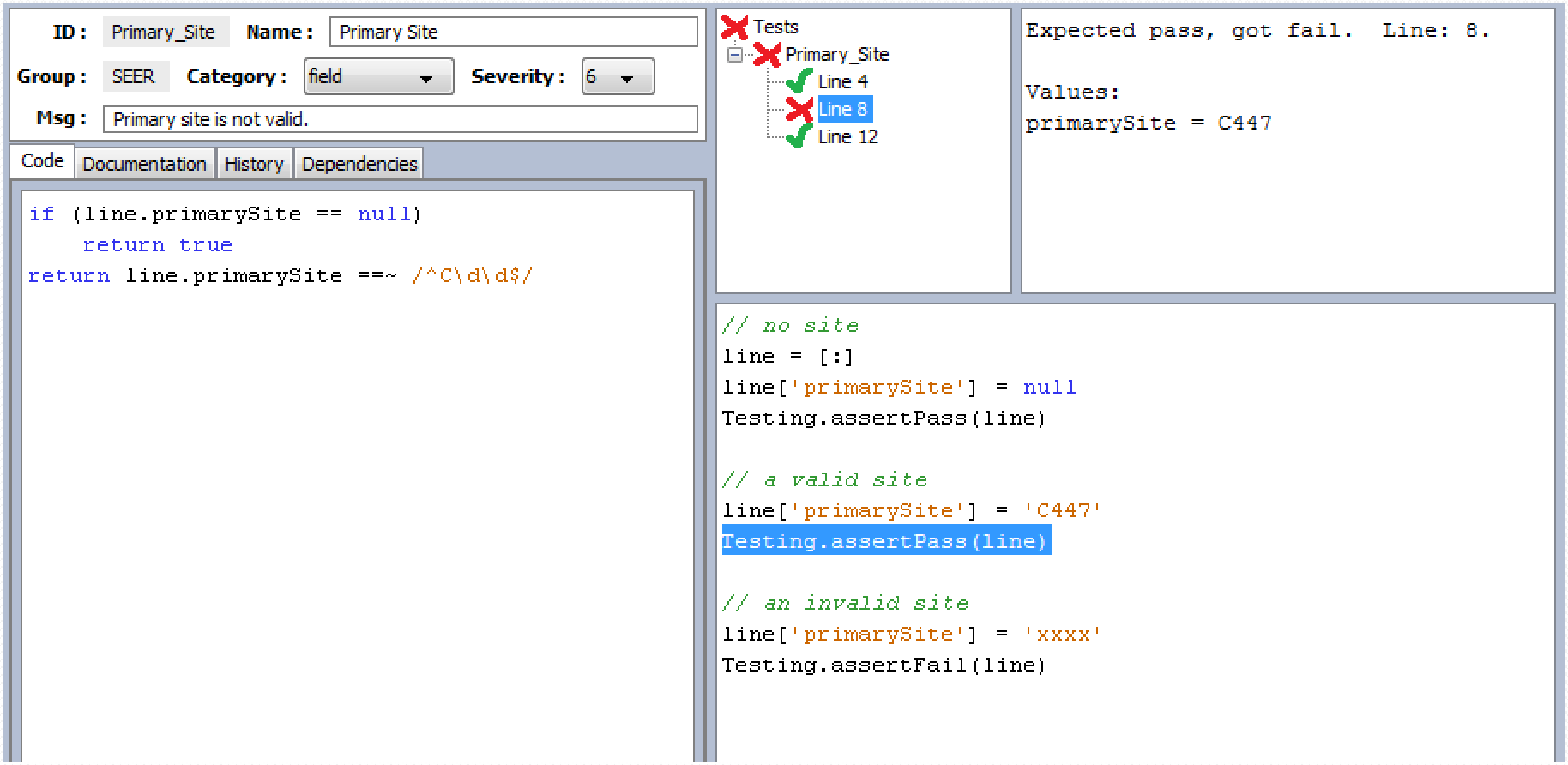


IF182 (16240)
IF120 (10571)
IF276 (4507)
IF277 (1241)
IF125 (955)
IF311 (755)
IF312 (594)
IF270 (399)
IF279 (359)
IF272 (249)
IF325 (170)
IF338 (136)
IF278 (122)
IF274 (64)
IF292 (64)
IF289 (61)
IF280 (49)
IF188 (37)
IF257 (34)
IF291 (33)
IF290 (32)
IF302 (19)
IF281 (18)
IF288 (16)
IF294 (16)
IF168_P3 (15)
IF297 (14)

IF272 {Obsolete Codes - Invalid CS Lymph Nodes} - 249 failures; 50 available samples

Pat ID #	Tumor Rec#	Ctrl Seq#	Ver 1st	CS LN	DX Year
00486191	02	02	010200	150	2004
00585645	02	02	010300	300	2006
00592579	02	02	010200	300	2004
00634361	02	02	010200	130	2005
00911704	02	02	010200	300	2004
01366866	02	03	010300	300	2006
01661795	02	02	010401	300	2009
01845179	02	02	010200	300	2004
01860241	02	02	010200	790	2004
01921002	02	02	010300	790	2005
01967487	02	02	010300	300	2005
02064111	02	02	010200	150	2004
02069041	01	00	010200	300	2004
02102455	01	00	010200	300	2004
02104679	01	00	010200	300	2004
02106728	01	00	010200	130	2004
02106965	01	00	010200	300	2004
02108691	01	00	010200	300	2004
02108712	01	00	010200	130	2004
02109212	01	00	010200	300	2004
02109841	01	00	010200	300	2004
02112574	01	00	010200	130	2004
02112621	01	03	010200	130	2004
02116529	01	00	010200	130	2004
02118114	01	00	010200	300	2004
02118688	02	02	010200	150	2006

Use case – Edit Writer



ID: Primary_Site Name: Primary Site

Group: SEER Category: field Severity: 6

Msg: Primary site is not valid.

```
if (line.primarySite == null)
    return true
return line.primarySite ==~ /^C\d\d$/
```

- ✗ Tests
  - ✗ Primary_Site
    - ✓ Line 4
    - ✗ Line 8
    - ✓ Line 12

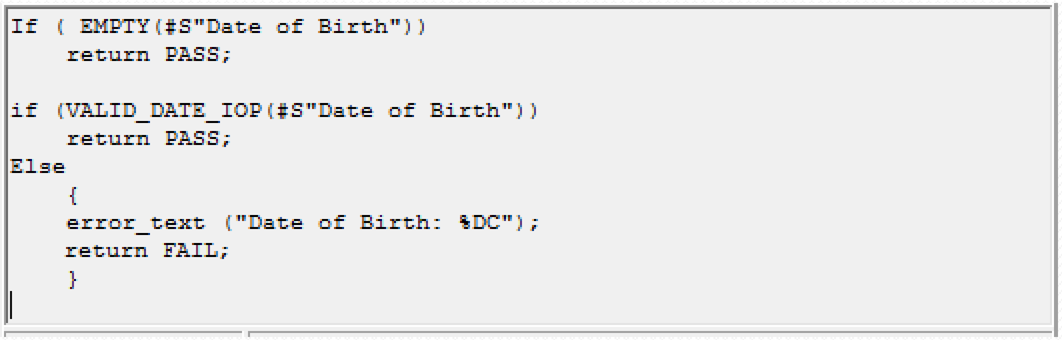
```
Expected pass, got fail.  Line: 8.

Values:
primarySite = C447
```

```
// no site
line = [:]
line['primarySite'] = null
Testing.assertPass(line)

// a valid site
line['primarySite'] = 'C447'
Testing.assertPass(line)

// an invalid site
line['primarySite'] = 'xxxx'
Testing.assertFail(line)
```

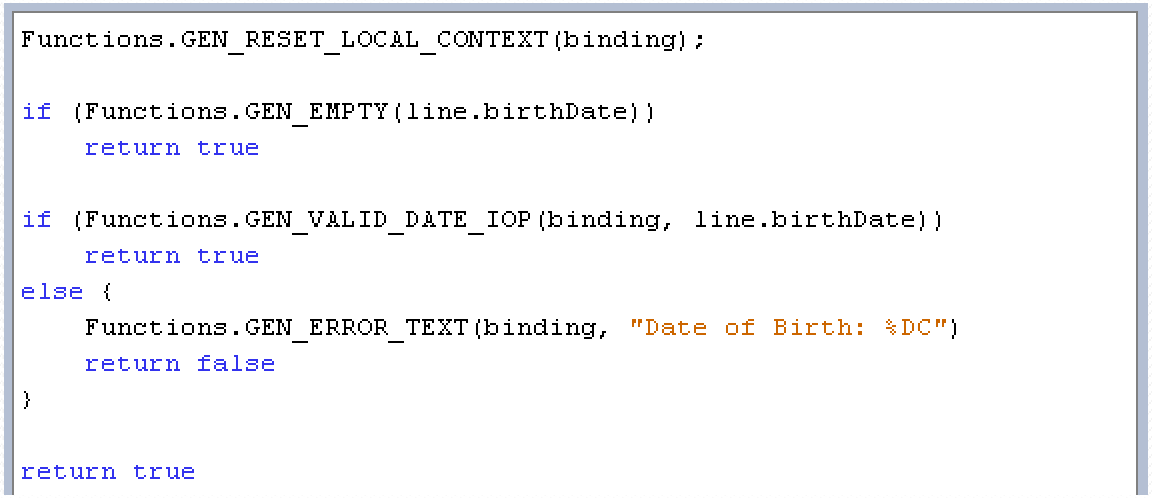
Using the GenEdits Metafile

- Many SEER Registries also use NPCR or NCDB edit sets
  - Goal: use these edit sets in SEER*DMS, SEER*Abs, SEER*Edits
- GenEdits
  - Uses a simple and well-defined language
  - In theory, it is possible to write a compiler to translate GenEdits source code into Groovy code that could be made available to the Java programs via the SEER Edits Engine
- Edits Compiler
  - Work in progress
  - Nearly all edits have been translated, but not all
  - Updates made only in the GenEdits version; the compiler would be used to create the updated edit sets for the Java programs

Translating a Simple Edit: Date of Birth

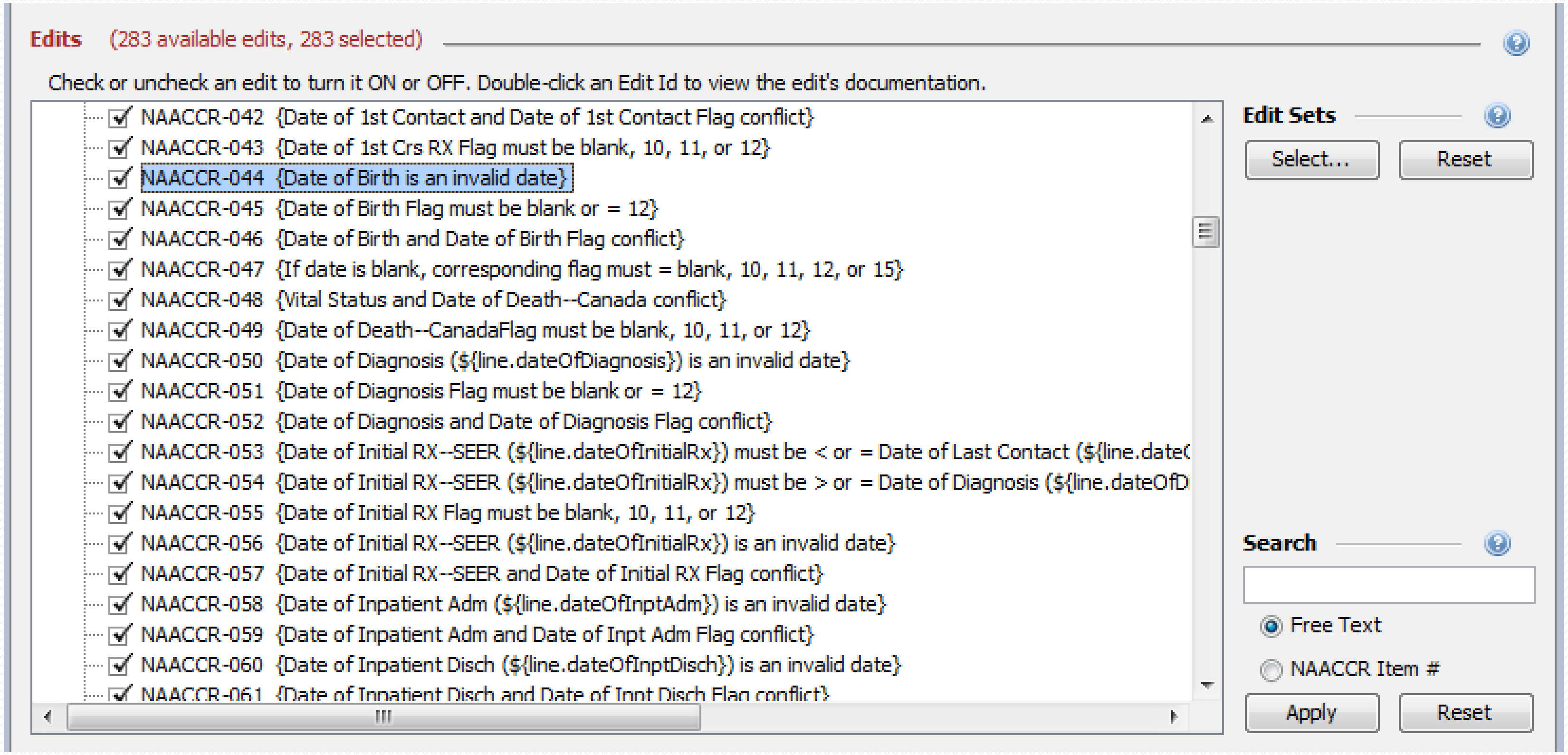


```
If ( EMPTY(#S"Date of Birth"))
    return PASS;

if (VALID_DATE_IOP(#S"Date of Birth"))
    return PASS;
Else
    {
    error_text ("Date of Birth: %DC");
    return FAIL;
    }
```



```
Functions.GEN_RESET_LOCAL_CONTEXT(binding);

if (Functions.GEN_EMPTY(line.birthDate))
    return true

if (Functions.GEN_VALID_DATE_IOP(binding, line.birthDate))
    return true
else {
    Functions.GEN_ERROR_TEXT(binding, "Date of Birth: %DC")
    return false
}

return true
```

Loading Translated Edits in SEER*Edits



Edits (283 available edits, 283 selected)

Check or uncheck an edit to turn it ON or OFF. Double-click an Edit Id to view the edit's documentation.

- ☑ NAACCR-042 {Date of 1st Contact and Date of 1st Contact Flag conflict}
- ☑ NAACCR-043 {Date of 1st Crs RX Flag must be blank, 10, 11, or 12}
- ☑ NAACCR-044 {Date of Birth is an invalid date}
- ☑ NAACCR-045 {Date of Birth Flag must be blank or = 12}
- ☑ NAACCR-046 {Date of Birth and Date of Birth Flag conflict}
- ☑ NAACCR-047 {If date is blank, corresponding flag must = blank, 10, 11, 12, or 15}
- ☑ NAACCR-048 {Vital Status and Date of Death--Canada conflict}
- ☑ NAACCR-049 {Date of Death--CanadaFlag must be blank, 10, 11, or 12}
- ☑ NAACCR-050 {Date of Diagnosis (${line.dateOfDiagnosis}) is an invalid date}
- ☑ NAACCR-051 {Date of Diagnosis Flag must be blank or = 12}
- ☑ NAACCR-052 {Date of Diagnosis and Date of Diagnosis Flag conflict}
- ☑ NAACCR-053 {Date of Initial RX--SEER (${line.dateOfInitialRx}) must be < or = Date of Last Contact (${line.date(
- ☑ NAACCR-054 {Date of Initial RX--SEER (${line.dateOfInitialRx}) must be > or = Date of Diagnosis (${line.dateOfD
- ☑ NAACCR-055 {Date of Initial RX Flag must be blank, 10, 11, or 12}
- ☑ NAACCR-056 {Date of Initial RX--SEER (${line.dateOfInitialRx}) is an invalid date}
- ☑ NAACCR-057 {Date of Initial RX--SEER and Date of Initial RX Flag conflict}
- ☑ NAACCR-058 {Date of Inpatient Adm (${line.dateOfInptAdm}) is an invalid date}
- ☑ NAACCR-059 {Date of Inpatient Adm and Date of Inpt Adm Flag conflict}
- ☑ NAACCR-060 {Date of Inpatient Disch (${line.dateOfInptDisch}) is an invalid date}
- ☑ NAACCR-061 {Date of Inpatient Disch and Date of Inpt Disch Flag conflict}

Edit Sets

Search

● Free Text

○ NAACCR Item #

Next Steps: SEER*Utils

- Utility programs available in a single Java library
- SEER*Utils currently contains
    - SEER Edits Engine
    - Java bridge to the Collaborative Stage DLL
    - SEER Site Recode mappings
    - SEER*Rx drugs and regimens data
    - Hematopoietic and Lymphoid Database
    - Multiple Primaries Calculator
- NAACCR mappings with an API to facilitate reading and writing of NAACCR data files

SEER*Utils

- SEER*Utils can be very easily integrated with a Groovy script:

```
import com.imsweb.seerutils.*
import com.imsweb.seerutils.cstage.*

SeerUtils.initializeAll()
println ‘CStage DLL version is ‘ + CollaborativeStage.getVersion()
SeerUtils.uninitializeAll()
```