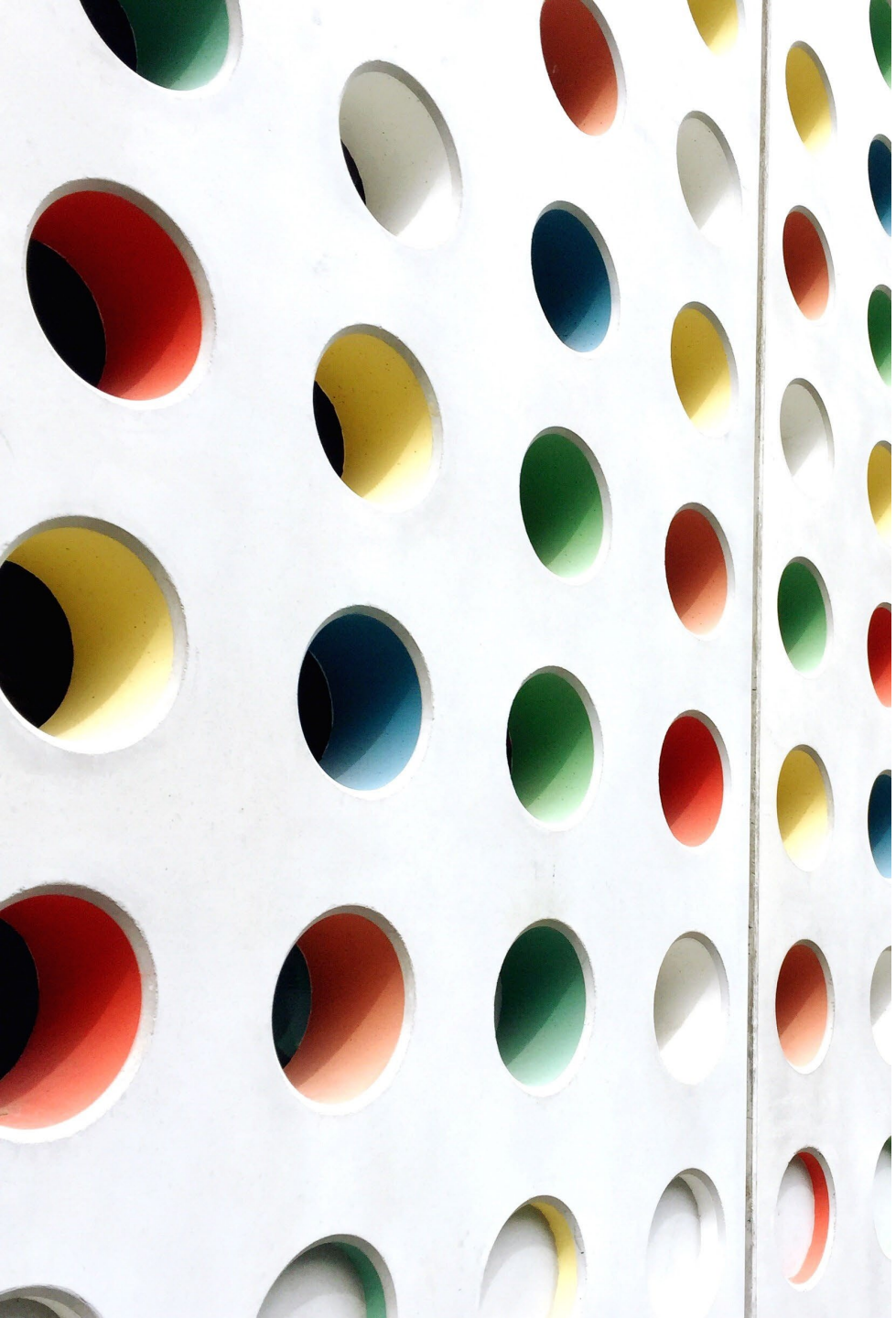


NCI-DOE Tools:

Natural Language Processing (NLP)

SEER*DMS WORKSHOP

FEBRUARY 13, 2023

SEER*DMS Workshop: Use Cases for the Natural Language Processing APIs

Meeting Insights

Organizer	Adams, Suzanne (IMS)	Sent Fri 1/6/2023 1:04 PM
Time	Monday, February 13, 2023 2:30 PM-4:30 PM	
Location	Microsoft Teams Meeting	
Response	Accepted Change Response	

Enterprise Vault + Get more add-ins

API stands for Application Programming Interface. It is a mechanism for two pieces of software to interface or communicate with each other. SEER*DMS uses an API to call the Natural Language Processing (NLP) algorithms developed as part of the NCI-DOE project. NLP is the application of linguistics and computer science to extract and interpret data from text-based documents (e.g., pathology reports, radiology reports, treatment summaries, clinical notes).

The path extraction API is currently used in SEER*DMS to auto-code a percentage of path reports. The percent auto-coded will increase in 2023 when the new case-level version of the API is deployed as part of the workflow. In addition to auto-coding, the path extraction API is now being used in support of rapid case ascertainment in two registries. It will also be used to auto-link path reports to CTCs even when the API could not fully auto-code the report. All current use cases will be described in more detail during the workshop.

The purpose of the SEER*DMS Workshop - NLP:

- Inform all registries of current uses cases for NLP in SEER*DMS
- Encourage registry managers and PIs to consider alternate ways to use the APIs.
    - The workshop will include a brainstorming session to consider novel ways of using the APIs.
    - Registries are welcome to ask questions and make suggestions.
    - If you have an idea that you'd like to explore in advance and need more information about the APIs, please contact Linda Coyle.
    - Also - feel free to submit suggestions for new use cases via email or the tech support issue (11687).

NCI-DOE Collaboration – NLP Algorithms

Aims

- Develop scalable NLP and machine learning tools
- Deep text comprehension of unstructured clinical text
- Accurate, automated capture of cancer surveillance data elements

Current Activities

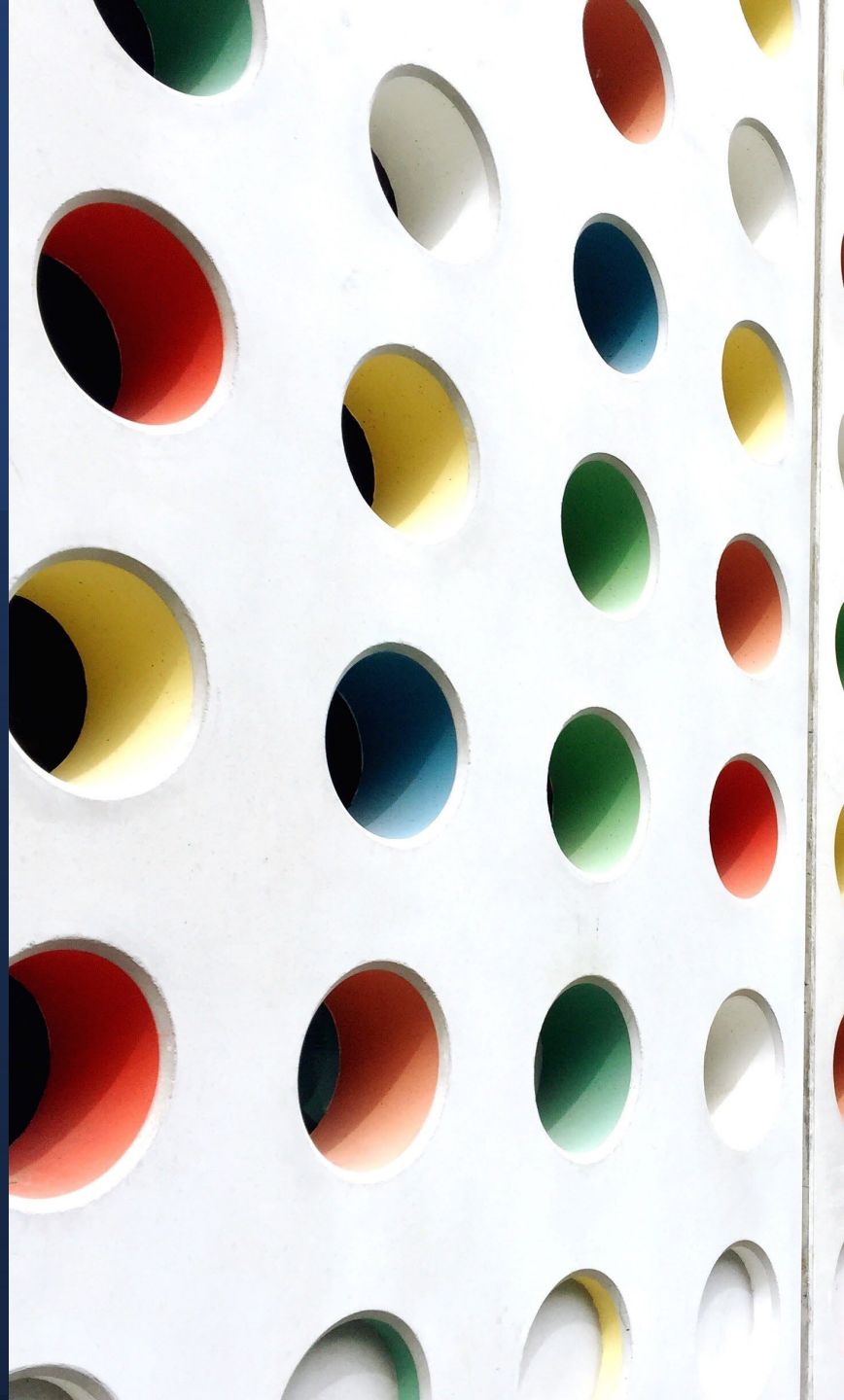
- Extraction of four key data elements from pathology reports
- Determination of whether a pathology or radiology report is related to cancer ("reportability")
- Extraction of relevant biomarker information
- Identification of recurrence

NLP Projects & SEER*DMS

PathExtraction	• Using report level endpoint (v11rc5) in SEER*DMS for path coding • Working to integrate case-level version – will increase % autocoded • Adding support for other use cases to reduce manual tasks (to be discussed today)
Reportability	• API (v2rc1) deployed in SEER*DMS for testing. • Evaluating use of development version in combination with Path Extraction
Bucketing	• Available as an API (v10rc2) in SEER*DMS. • Evaluating possible workflow use cases.
ICCC	• Available as an API (v11rc2) in SEER*DMS for evaluation and testing
Recurrence	• Algorithms under development at DOE • SEER*DMS registries annotated path reports to create “gold standard” data

Current Workflow:

Report Level
Path Extraction API



Using the Report Level Path Coding API in SEER*DMS

Deployed in production instances of SEER*DMS

- May thru August 2021 – Georgia, Utah, Louisiana
- It is now the default workflow for processing path in SEER*DMS

Workflow goals

- To accurately auto-code **four** fields **on individual path reports** and **reduce the level of effort** related to manual path coding
  - Site
  - Laterality
  - Histology
  - Behavior
- To identify reports that cannot be auto-coded and forward those to a **manual coding task**
- To **increase efficiency** in registry process where the API results provide value (Rapid Case Ascertainment, auto-linking, etc)

Project goals

- Increase auto-processing of path reports in multiple workflows
- Collect data to train **future versions of the API**.

SEER*DMS Admin PRODUCTION

Overview Queries Searches L

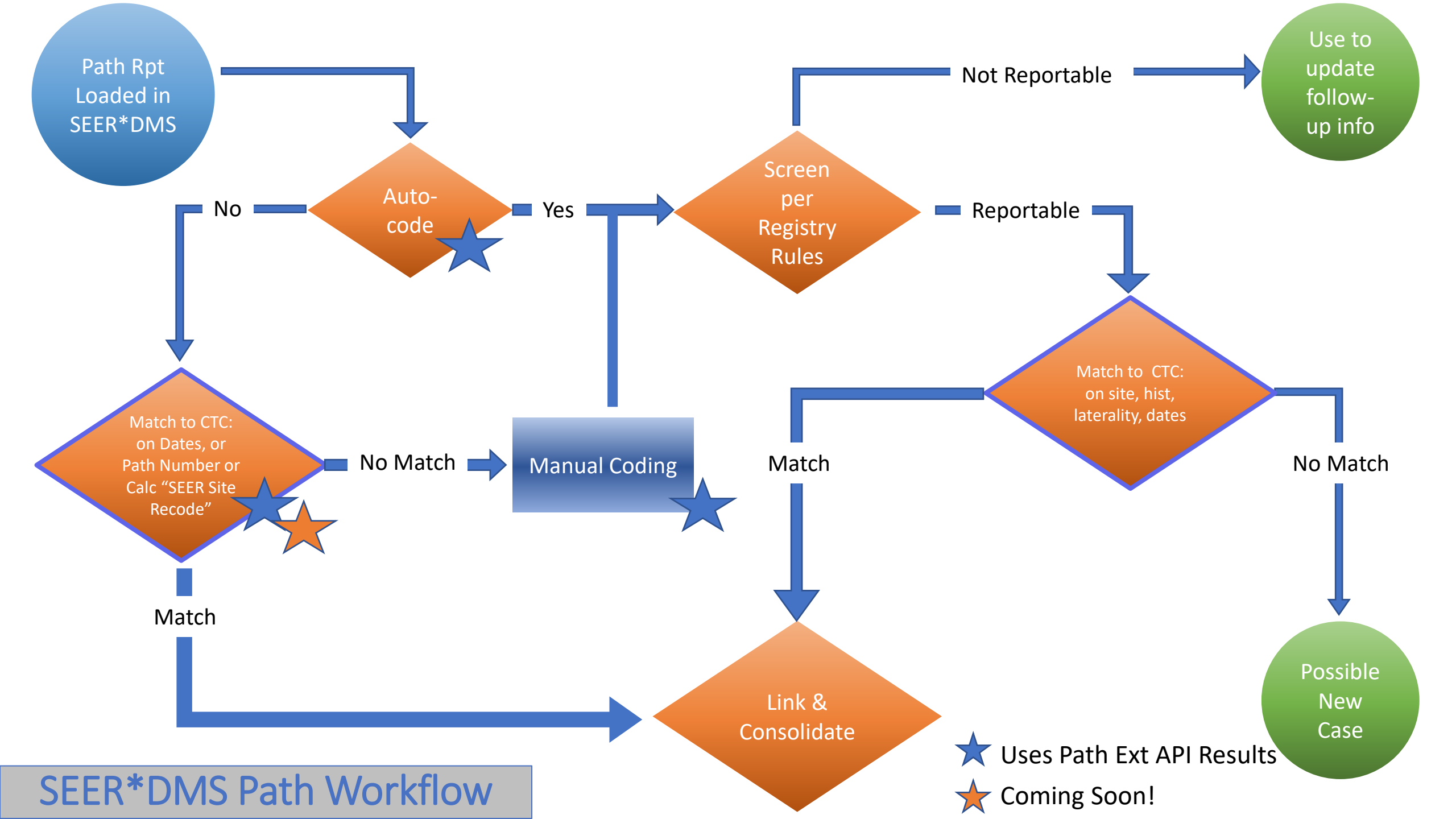
#	URL	Registry	Island
1	AK	Alaska	Sterling
2	AR	Arkansas	Sterling
3	CA	California	Baltimore
4	CN	Cherokee Nation	Sterling
5	CT	Connecticut	Baltimore
6	DT	Detroit	Baltimore
7	FL	Florida (LITE)	Sterling
8	GA	Georgia	Baltimore
9	HI	Hawaii	Baltimore
10	IA	Iowa	Sterling
11	ID	Idaho	Sterling
12	IL	Illinois	Sterling
13	KY	Kentucky	Sterling
14	LA	Louisiana	Baltimore
15	MA	Massachusetts	Sterling
16	MN	Minnesota	Baltimore
17	NCCR	NCCR	Sterling
18	NJ	New Jersey	Baltimore
19	NM	New Mexico	Sterling
20	NY	New York	Baltimore
21	OH	Ohio (LITE)	Baltimore
22	SE	Seattle	Baltimore
23	TN	Tennessee (LITE)	Sterling
24	TX	Texas	Baltimore
25	UT	Utah	Baltimore

16 registries process path data in SEER*DMS and use the Path Extraction API to auto-code site, histology, behavior, laterality.

California is in the process of migrating to SEER*DMS. Deployment is scheduled for June 2024. APIs can be run on LAX data for testing; and could be tested on other CA path data in SEER*DMS this year.

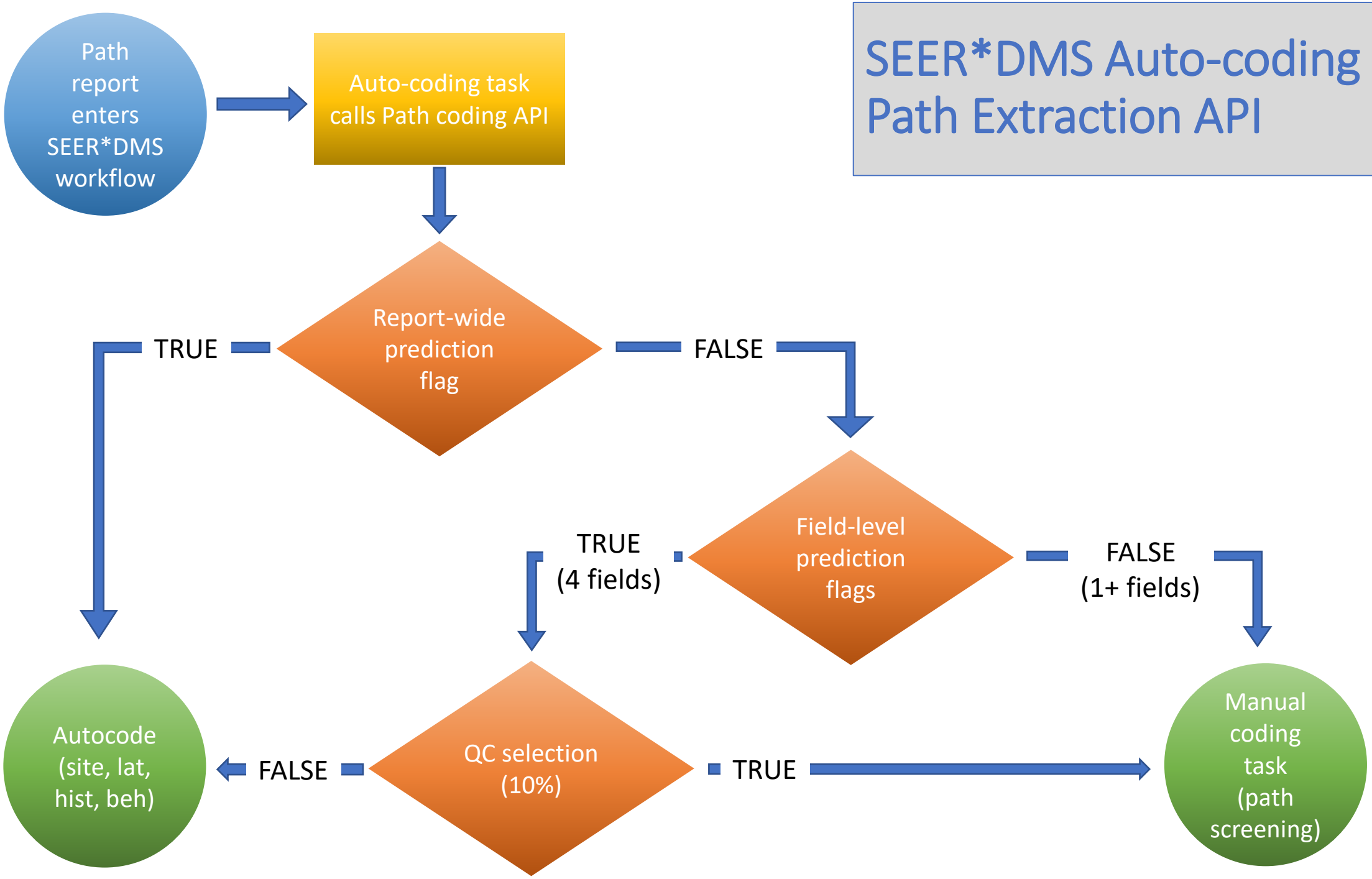
Path reports are not processed in 7 instances of SEER*DMS. AR and IL will process path in SEER*DMS in 2023 or 2024. The "lite" instances support NCCR activities, the data may include path sometime soon.

Path reports are loaded into Seattle's instance of SEER*DMS Seattle but are processed in an external system prior to import. APIs can be run on SE data for testing.



SEER*DMS Path Workflow

SEER*DMS Auto-coding Rules Path Extraction API



NY BETA SEER*DMS NYSCR 23.17+ Matching

View Manage Tools System

Keywords

ctc

Filter Clear Show All

CTC Matching: HL7 (E-path)

An HL7 record will be linked to a CTC in the linked patient set if the following are true:

- Record matches one CTC in the patient set using the multiple primary rules.
- Or the record matches one CTC in the patient set based on SEER site recode:
  - Record SEER Site Recode or NLP predicted SEER site recode matches the CTC.
  - And there is no conflict in laterality (see definition below).
  - And the record's event date is within 30 days of the date of diagnosis or is later than the date of diagnosis and within one year.
- Or the record matches based on dates:
  - Patient set has one non-deleted CTC.
  - No conflict in site or laterality (see definitions below).
  - Record's event date is compared to:
    - DX Date
    - Rx Summ Date Surgery
    - Most Def Surgery
  - It is a match if the Record's event date is within 30 days or is later than the date of diagnosis and within one year. For melanoma (C44) and breast (C50) primaries, it is a match if Record's event date is within 30 days.
- Or there is a single CTC with a linked HL7 with the same Pathology Number, and there is no conflict in site or laterality (see definitions below).

Definitions for conflicts:

- No conflict in site is defined as:
  - Site on the HL7 record is (C809 or blank) and the SEER site recode calculated by the NLP algorithm is null, abstains, or matches the CTC site recode.
  - Or the HL7 record is a match to the CTC using the SEER multiple primary rules.
- No conflict in laterality is defined as:
  - CTC site is not paired or the CTC laterality matches the record laterality. If the record laterality is blank then the NLP predicted laterality is used. Blank and unknown values do not create a conflict.

New logic to be deployed in Feb

Logic varies by registry. Registries will be notified in Squish to confirm their registry's logic.

Registry Variations:

1. Num Days – most use 10 or 30
2. Whether the X days timeline is applied to all sites or only to melanoma and breast. If only melanoma and breast, then 1 year is used for other sites.

SEER*DMS Path Screening Task
Using API results to facilitate manual coding

HL7 E-Path

Save | Cancel

Reportability *
Based on coded values (recommended)

Primary Diagnosis

Site *

C778 : Lymph nodes of multiple regions	*13%*
C509 : Breast, NOS	*12%*
C421 : Bone marrow	*12%*

Laterality *

0 : Not Paired	*73%*
2 : Left (origin of primary)	*24%*
1 : Right (origin of primary)	*2%*

Histology *

9823/3 : Chronic lymphocytic leukemia/small lymphocytic lymphoma (ICD-O-3 update)	*65%*
9670/3 : Malignant lymphoma, small B lymphocytes, NOS [OBS], see 9823/3 (ICD-O-3 update)	*12%*
8500/3 : Invasive carcinoma of no special type (C50._)	*7%*

Behavior *
3 : Malignant Primary

3 : Malignant Primary

Add Primary

REC-1000000535 edit full record

OLMSTED, CEDRIC D	810-46-8176	Male	MRN: MR401960	Collected 07-12-2018; Received 07-14-2018
Unlinked	b. 01-25-1934	State: UT	Path#: 189109	IMP-1001 on 08-16-2018
	Rpt Hosp#: FAC-9999		FAC-9999	

Clinical Hist

Evaluate for non-Hodgkin's lymphoma: ALL: myelodysplastic syndromes: chronic Lymphoproliferative disorders, CLL. Prior therapy: chemotherapy, Fludarabine more than one month ago. CBC report received.

Nature of Spec

Bone marrow.

Gross Path

Part #1 is labeled "left breast biopsy" and is received fresh after frozen section preparation. It consists of a single firm nodule measuring 3cm in circular diameter and 1.5cm in thickness surrounded by adherent firbrofatty tissue. On section a pale gray, slightly mottled appearance is revealed. Numerous sections are submitted for permanent processing.\.br\Part #2 is labeled "apical left axillary tissue" and is received fresh. It consists of two amorphous fibrofatty tissue masses without grossly discernible lymph nodes therein. Both pieces are rendered into numerous sections and submitted in their entirety for history.\.br\Part #3 is labeled "contents of left radical mastectomy" and is received flesh. It consists of a large ellipse of skin overlying breast tissue, the ellipse measuring 20cm in length and 14 cm in height. A freshly sutured incision extends 3cm directly lateral from the areola, corresponding to the closure for removal of part #1. Abundant amounts of fibrofatty connective tissue surround the entire beast and the deep aspect includes and 8cm length of pectoralis minor and a generous mass of overlying pectoralis major muscle. Incision from the deepest aspect of the specimen beneath the tumor mass reveals tumor extension gross to within 0.5cm of muscle. Sections are submitted according to the following code: DE- deep surgical resection margins; SU, LA, INF, ME -- full thickness radila samplings from the center of the tumor superiorly, laterally, inferiorly and medially, respectively: NI- nipple and subjacent tissue. Lymph nodes dissected free from axillary fibrofatty tissue from levels I, II, and III will be labeled accordingly.

Micro Desc

Final DX

A small population of monoclonal B-cells (Kappa) is present in the bone marrow. The antigenic profile is consistent with chronic lymphocytic leukemia/small lymphocytic lymphoma (CLL/SLL).

Supp Rpt Add

Staging

Comments

Correlation with a comprehensive bone marrow morphology examination, CBC data/blood smear, and other relevant clincial and laboratory data is recommended.

Full Text

Snomed

Path Ref Range

CAP Synoptic

Registry Comment

Discussion Notes: Questions or comments on the SEER*DMS workflow and the Report Level Path Extraction API?

Name & Registry	Comment
Jenifer Hafterson (SE)	Is there a way to use the SE beta instance. Freeze it. Then load path files loaded into prod vs beta.
Linda Coyle	We may be able to do the analysis by evaluating the NLP prediction data; may not need to freeze beta.
Serban	Site recode is an official classification for reporting cancer statistics in the US. It is used by SEER, NPCR, and we believe NAACCR.
Gary	Is there a mechanism where the manual reviews feedback/train information back to the NLP algorithm?
Linda	Training packages are sent 2x per year; not an automated feedback loop
Betsy	The only data sent to DOE are from registries with agreements with Oakridge National Lab
April	I am wondering if clinical history is used for the coding
Linda	Yes, all path report text fields are used by the API
Amy (MN)	Why is clinical history used?
Serban	Does the question come from the perspective that the manuals indicate to excl clinical history. Algorithm is trained to look at everything and return the same answer to a certain level of accuracy.

Using the Current Path Extraction API in SEER*DMS

- API prediction values are available to SEER*DMS processes immediately after the record is loaded.
- If a report is not auto-coded, the top 2 predictions can still be put to good use.
- Other use cases to be discussed today:
  - **Kevin Ward:** Prioritize manual Path Screening tasks for reports related to Rapid Case Ascertainment studies (Corpus and Ovarian study)
  - **Colleen Sherman:** Identify the best manual Path Screening tasks for training new path coders. Registry manager selects reports based on the probable site so that the path coder can concentrate on coding rules for that site.

NY Live SEER*DMS NYSCR 23.24 Worklist

Actions

Apply Reset

Search

Task ID

Task Type
Is Path Screen

User
Is *(Missing)*

Task Date

Flag
Any of PR ML TH BR CL

Status

In Progress

ID

Data Type

Event Date
Is __-__-2022

Facility

Import

Import Date

CTC Reportability

Results | Saved Searches

15,783 items

F	Task	Type	User	Age	ID	Data	Event
BR	TSK-40587038	Path Screen		26d	REC-25709520	HL7	2022
BR	TSK-40951042	Path Screen		11d	REC-25743730	HL7	2022
CL	TSK-31298852	Path Screen		4mon	REC-23971481	HL7	2022
ML	TSK-37016856	Path Screen		4mon	REC-25371615	HL7	2022
TH	TSK-39647990	Path Screen		1mon	REC-25594643	HL7	2022
BR	TSK-31306650	Path Screen		3mon	REC-23978980	HL7	2022
BR	TSK-31306668	Path Screen		3mon	REC-23979031	HL7	2022
BR	TSK-32174249	Path Screen		7mon	REC-24074871	HL7	2022
BR	TSK-39409200	Path Screen		2mon	REC-25584969	HL7	2022
BR	TSK-32176016	Path Screen		7mon	REC-24076586	HL7	2022
BR	TSK-39686072	Path Screen		1mon	REC-25624092	HL7	2022
BR	TSK-39686087	Path Screen		1mon	REC-25624089	HL7	2022
BR	TSK-40160399	Path Screen		25d	REC-25641544	HL7	2022
ML	TSK-40203728	Path Screen		1mon	REC-25662701	HL7	2022
ML	TSK-40207925	Path Screen		1mon	REC-25665736	HL7	2022
ML	TSK-40940746	Path Screen		11d	REC-25735575	HL7	2022
ML	TSK-40950594	Path Screen		11d	REC-25743387	HL7	2022
BR	TSK-40982624	Path Screen		11d	REC-25766905	HL7	2022
BR	TSK-36173707	Path Screen		3mon	REC-25253629	HL7	2022
ML	TSK-39675876	Path Screen		1mon	REC-25616088	HL7	2022
ML	TSK-39678884	Path Screen		1mon	REC-25618613	HL7	2022

Discussion:

API Workflows