caBIG Data Extract Specifications

Two University Office Park
51 Sawyer Road
Waltham, MA 02453

Telephone: 781-434-2200
Fax: 781-642-6222

Contents

This Page Was Intentionally Left Blank.

Introduction

CoPathPlus offers an option, starting with V2.5, to extract data for the purpose of processing it into modules of caBIG. The initial effort is focused on caTIES, but it is expected that the same data format could be used by Clinical Annotation Engine and eventually perhaps by caTISSUE CORE. This document provides a specification for the data format. We intend to draft a second document describing how to configure and manage the interface.

The data extraction functionality is implemented as a "batch-oriented interface." In a production system it will be scheduled to run at a regular interval, and its output will be to a file in a designated directory. For example, it could be scheduled to run weekly or monthly, and generate one file with the data from all cases that had a report signed out within the specified date range. The same functionality can also be scheduled to pull historical cases. For example, a job could be defined to run one night to generate a batch of data from 1/1/2000 through 3/31/2000, another job could be scheduled for the next night to generate a file for 4/1/2000 through 6/30/2000, and so on, to extract all historical cases up to the present. Then the routine weekly or monthly job can continue to pull new data thereafter.

A case may be pulled in multiple date ranges, if a report (Final, Addendum or Procedure) is signed out within the time frames of multiple batches. Within a single batch the case will only appear once (even if multiple reports were signed out during that timeframe). When a case is included in the data extract, ALL data for that case are pulled. If a case is received a second time in a later batch, the new data should replace the old data.

Note that it is expected that this functionality will run as a set of scheduled jobs that will normally be run off peak hours on an interface agent. For most sites an existing agent will be able to handle this additional load, since it will running off peak times. But for some sites there may not be enough capacity available, and a new agent PC (plus CoPath license and appropriate 3rd party licenses) would be required.

Data Format Specifications

The data are formatted into messages, segments and fields based upon HL7 version 2.3.1. The file consists of a file header message, followed by any number of messages describing the individual cases, followed by a file trailer message.

FHS
{ORU^R01 messages} Case data {repeating}
FTS

The following message segments may appear for each case. An OBR is sent for the complete specimen result and an additional OBR is sent for each procedure or addendum that has been signed out. Each OBR is followed by corresponding OBX segments.

MSH	Message header
PID	Patient information
PV1	Patient visit (if available)
ORC	Common order data
{OBR	Observation request {multiple}
ZLR	Extensions to HL7 – one segment per OBR
{OBX}}	Observation result {multiple per OBR}

In the message descriptions below, there are references to an “ObjectID”. This is a field that communicates the relationships between various specimen-related entities in CoPathPlus. It is a composite ID where the number of component pieces is determined by the level of the object in the hierarchy where “specimen” in at the top, and “part” is an example of an object under specimen, and “block” is an example of an object under each part. Each component piece of the ID is an internal code that is not intended to have any external meaning – it defines the hierarchy of objects without specifying any sequence/order for the “sibling objects.”

The full set of entities in the hierarchy is shown below. In parentheses is an example of a possible ObjectID for each entity.

Specimen (abc123)
Part (abc123.1)
Block (abc123.1.1)
Stain/Process (abc123.1.1.1 or abc123.1.0.1 if no block specified)
Procedure (abc123.P1)
Addendum (abc123.A1)
Synoptic Worksheet (abc123.W1)

Sample File with One Specimen Record
(detailed specification follows)

FHS|^~\&|CoPathPlus||||20051208073742

MSH|^~\&|CoPathPlus|X|caBIG||20051208073742||ORU^R01|110008700000032949|P|2.3

PID|1||123321^^^GH|cop14021|holland^ramona||19410401|F|||||||||||||||||||N

PV1|1|||||||||||||||||S2005-345

ORC|RE||^CoPathPlus||CM||||200511162230|^Holland^Zeke||^Jones^Mike

OBR|1||^CoPathPlus^cop45423|SURG^Surgical Pathology|||200511162228|||||||
200511162228|COLON|||||^64^Y|-1|200512062219||SP|C|||||||^Adrian^Lee |^Holland^Zeke

OBX|1|CE|ANT|cop45423.1|COLON^Colon resection (tumor)^L|||||C||1^A

OBX|2|CE|ANT|cop45423.2|APPEND^Appendix^L|||||C||2^B

OBX|3|ST|BLOCK|cop45423.1.1|1|||||||1^1|200511170042

OBX|4|ST|BLOCK|cop45423.1.2|2|||||||2^2|200512062216

OBX|5|ST|BLOCK|cop45423.2.3|1|||||||1^1|200512062217

OBX|6|CE|PROCESS|cop45423.1.1.1|$hinit^Initial H&E^L|||||||1^H&E|200511170042

OBX|7|CE|PROCESS|cop45423.1.0.2|co041^CELL BLOCK^L|||||||1^CB|200511170051

OBX|8|CE|PROCESS|cop45423.1.2.3|$hinit^Initial H&E^L|||||||1^H&E|200512062216

OBX|9|CE|PROCESS|cop45423.1.2.4|co28^ALCIAN BLUE W/HYALURONIDASE^L|||||||2^ALC W/HY|200512062216

OBX|10|CE|PROCESS|cop45423.2.3.5|$hinit^Initial H&E^L|||||||1^H&E|200512062217

OBX|11|CE|PROCESS|cop45423.2.3.6|14^Additional H&E, Level I^L|||||||2^Level I|200512062217

OBX|12|TX|GDT|cop45423|This is the Gross Description. Received in formalin... \.br\||||||C

OBX|13|TX|MDT|cop45423|This is the microscopic description.\.br\||||||C

OBX|14|TX|FIN|cop45423|See synoptic text. (This is the Final Diagnosis.)\.br\ ||||||C

OBX|15|CE|IMP|cop45423|E0004^^SNO|||||||1||M

OBX|16|CE|IMP|cop45423|D0586^^SNO|||||||1||M

OBX|17|CE|IMP|cop45423|D4684^^SNO|||||||1||M

OBX|18|CE|SYNOP|cop45423.W2|72592005^^CT|||||||406031003^2||A

OBX|19|CE|SYNOP|cop45423.W2|373197004^8^CT|cm|||||406031003^2||A

OBX|20|CE|SYNOP|cop45423.W2|103679000^^CT|||||||406031003^2||A

OBX|21|CE|SYNOP|cop45423.W2|32913002^^CT|||||||406031003^2||A

OBX|22|CE|SYNOP|cop45423.W2|60815008^^CT|||||||406031003^2||A

OBX|23|CE|SYNOP|cop45423.W2|395534006^^CT|||||||406031003^2||A

OBX|24|CE|SYNOP|cop45423.W2|399652005^^CT|||||||406031003^2||A

OBX|25|CE|SYNOP|cop45423.W2|405980004^^CT|||||||406031003^2||A

OBX|26|CE|SYNOP|cop45423.W3|122659008^^CT|||||||406020006^3||B

OBX|27|CE|SYNOP|cop45423.W3|67109009^^CT|||||||406020006^3||B

OBX|28|CE|SYNOP|cop45423.W3|4797003^^CT|||||||406020006^3||B

OBX|29|CE|SYNOP|cop45423.W3|54102005^^CT|||||||406020006^3||B

OBX|30|CE|SYNOP|cop45423.W3|84921008^^CT|||||||406020006^3||B

OBX|31|CE|SYNOP|cop45423.W3|54452005^^CT|||||||406020006^3||B

OBX|32|CE|SYNOP|cop45423.W3|372309006^5^CT|||||||406020006^3||B

OBX|33|CE|SYNOP|cop45423.W3|372308003^3^CT|||||||406020006^3||B

OBX|34|CE|SYNOP|cop45423.W3|17076002^^CT|||||||406020006^3||B

OBX|35|CE|SYNOP|cop45423.W3|384960007^This is a fill-in^CT|||||||406020006^3||B

OBX|36|CE|SYNOP|cop45423.W3|395553001^^CT|||||||406020006^3||B

OBX|37|CE|SYNOP|cop45423.W3|370051000^^CT|||||||406020006^3||B

OBX|38|CE|SYNOP|cop45423.W3|399525009^^CT|||||||406020006^3||B

OBX|39|CE|IMP|cop45423|794.31^^ICD|||||||1

OBX|40|CE|IMP|cop45423|376.43^^ICD|||||||A

OBR|2||cop45423^CoPathPlus^cop45423.A2|ADD^Addendum|||||||||200512062201|||||||2|200512062201||SP|F|||||||^Robinson^David|^Holland^Zeke

OBX|1|TX|ADC|cop45423|This is the comment for the addendum.\.br\||||||C

OBX|2|TX|ADX|cop45423|This is the addendum Dx text.\.br\||||||C

OBR|3||cop45423^CoPathPlus^cop45423.P1|EM^Electron Microscopy|||||||||200512062156|||||||1|200512062156||SP|F|||||||UXD^Cerner^User|^Holland^Zeke

OBX|1|TX|PIN|cop45423|This is the interpretation text for the EM procedure.\.br\||||||C

OBX|2|TX|PRC|cop45423|This is the "result" or is it the "comment" for the EM procedure.\.br\||||||C

FTS|1

NOTE: Most Cerner DHT specifications start counting the fields in a segment beginning with the segment ID, while this field is not counted in the HL7 specification. Thus (with the exception of the MSH segment) the Cerner DHT specs would show field numbers 1 greater than on the HL7 spec. However, in this document the numbering has been started at 0, so that the field (SEQ) numbers match between this spec and the HL7 spec.

NOTE: The HL7 standard is to separate message segments with the <CR> character. However, the caBIG Extract files have segments separated by <CR><LF>. This makes it easier to read the contents of the files, and in some cases makes it easier for other systems to parse the data.

MSH Segment

The MSH (Message Header) segment defines the intent, source, destination, and some specifics of the syntax of a message.

SEQ	DT	Element Name	Description
0	ST	Segment ID	"MSH"
1	ST	Field Separator	"|"
2	ST	Encoding Characters	"^~\&"
3	ST	Sending Application	"COPATHPLUS"
4	ST	Sending Facility	Will be added with caBIG changes later, so that files from different facilities can be distinguished. Has an X as a placeholder for now.
5	ST	Receiving Application	"caBIG"
6	ST	Receiving Facility	not used
7	TS	Date/Time of Message	YYYYMMDDHHMMSS
8	ST	Security	not used
9	ID	Message Type	"ORU^R01"
10	ST	Message Control ID	DHTI counter (seqn_key from i_request_q table)
11	ID	Processing ID	"P"
12	NM	Version ID	"2.3"
13	NM	Sequence Number	not used
14	ST	Continuation Pointer	not used

PID Segment

The PID (Patient Identification) segment is used as the primary means of communicating patient identification information. This segment contains permanent patient identifying and demographic information that is not likely to change frequently.

SEQ	DT	Element Name	Description
0	ST	Segment ID	"PID"
1	SI	Set ID - Patient ID	"1"
2	CK	Patient ID - External	not used
3	CK	Patient ID - Internal	Patient Medical Record Number^^^Client dictionary Interface Link for the client associated with the MRN. **Note**: if no MRN is available, the internal patient ID is substituted as a means of providing a unique identifier.
4	ST	Alternate Patient ID	Internal CoPathPlus key of patient record
5	PN	Patient Name	Lastname^Firstname^Middlename^Suffix^ Prefix
6	ST	Mother's Maiden Name	not used
7	DT	Date of Birth	YYYYMMDD
8	ID	Sex	"M", "F", or "U"
9	PN	Patient Alias	not used
10	ID	Ethnic Group	Race dictionary Interface Link
11	AD	Patient Address	street^other^city^state^zip^country
12	ID	County Code	not used
13	TN	Phone Number - Home	
14	TN	Phone Number - Bus.	
15	ST	Language - Patient	not used
16	ID	Marital Status	(dictionary interface link)
17	ID	Religion	not used
18	CK	Patient Account Number	not used
19	ST	SSN Number - Patient	NNN-NN-NNNN (Govt. Number)
20 – 28			not used
29	TS	Date/time of Death, if expired	YYYYMMDDHHMM
30	ID	Patient Death Indicator	"Y" or "N"

PV1 Segment

The PV1 (Patient Visit) segment is used to communicate information about a specific visit.

SEQ	DT	Element Name	Description
0	ST	Segment ID	"PV1"
1	SI	Set ID - Patient Visit	"1"
2	ID	Patient Class	not used
3	ID	Assigned Patient Location	Location dictionary Interface Link^Room
4	ID	Admission Type	not used
5	ST	Pre-admit Number	not used
6	ID	Prior Patient Location	not used
7	CN	Attending Doctor	Person dictionary Interface Link^Lastname^Firstname (From Attending MD)
8	CN	Referring Doctor	Person dictionary Interface Link^Lastname^Firstname (From Additional Attending MD)
9	CN	Consulting Doctor	Person dictionary Interface Link^Lastname^Firstname (From Additional Attending MD)
10	ID	Hospital Service	not used
11	ID	Temporary Location	not used
12	ID	Pre-admit Test Indictor	not used
13	ID	Re-admission Indicator	not used
14	ID	Admit Source	not used
15	ID	Ambulatory Status	not used
16	ID	VIP Indicator	not used
17	CN	Admitting Doctor	Person dictionary Interface Link^Lastname^Firstname
18	ID	Patient Type	Patient Type dictionary Interface Link
19	NM	Visit Number	Encounter Number
20	ID	Financial Class	Financial Class dictionary Interface Link
21 – 35			not used
36	ID	Discharge Disposition	Discharge Disposition dictionary Interface Link
37 – 43			not used
44	TS	Admit Date/Time	YYYYMMDDHHMM
45	TS	Discharge Date/Time	YYYYMMDDHHMM

ORC Segment

The ORC (Common Order Data) segment is used to communicate information common to orders.

SEQ	DT	Element Name	Description
0	TX	Segment ID	"ORC"
1	ID	Order Control	“RE” = Results (could be NW for new order or OC for order canceled, but these don’t apply to caBIG)
2	CM	Placer Order Number	Requisition number
3	CM	Filler Order Number	Accession Number^"CoPathPlus"
4	CM	Placer Group Number	not used
5	ID	Order Status	“CM” = Specimen completed (could be IP = specimen in progress or CA = specimen canceled, but these don’t apply to caBIG)
6	ID	Response Flag	not used
7	TQ	Quantity/Timing	not used
8	CM	Parent	not used
9	TS	Date/Time of Transaction	YYYYMMDDHHMM (Accession Date/Time)
10	CN	Entered By	Person dictionary Interface Link^Lastname^Firstname (From Accessioner)
11	CN	Verified By	not used
12	CN	Ordering Provider	Person dictionary Interface Link^Lastname^Firstname (From Ordering/Submitting MD)
13	CM	Enterer’s Location	not used
14	TN	Call Back Phone Number	not used
15	TS	Order Effective Date/Time	not used
16	CE	Order Control Code Reason	not used
17	CE	Entering Organization	not used
18	CE	Entering Device	not used
19	CN	Action By	not used

OBR Segment

The OBR (Observation Request) segment serves as the report header in the reporting of clinical data. There will be one OBR for the reporting for the specimen, and additional OBRs for each procedure and addendum. Each OBR will be followed by the corresponding OBXs (Observation/Results).

SEQ	DT	Element Name	Description
0	TX	Segment Name	"OBR"
1	SI	Set ID - Observation Request	Sequential OBR counter
2	CM	Placer’s Order #	Same as ORC-2. If there is an OE Interface, this will be the order entry number assigned by the HIS system. Otherwise, it will be blank.

SEQ	DT	Element Name	Description
3	CM	Filler's Order #	Accession Number^"CoPathPlus"^ObjectID
4	CE	Universal Service Ident.	Department dictionary Interface Link^Depoartment Name (e.g., "S^Surgical", if the ObjectID has just 1 piece (implying that this is the case-level report). *Or* Procedure dictionary Interface Link^Procedure Name (e.g., "EM^Electron Microscopy", if the ObjectID has a 2nd piece beginning with "P". *Or* "ADD^Addendum" if the ObjectID has a 2nd piece beginning with "A".
5	ID	Priority	not used
6	TS	Requested Date/Time	not used
7	TS	Observation Date/Time	YYYYMMDDHHMM (Date/Time taken of first part). This is the Autopsy Date/Time for autopsy cases. Not applicable for procedures/addenda.
8	TS	Observation End Date/Time	not used
9	CQ	Collection Volume	not used
10	CN	Collector Identifier	not used
11	ID	Specimen Action Code	not used
12	CM	Danger Code	not used
13	ST	Relevant Clinical Info.	not used
14	TS	Specimen Received Date/Time	YYYYMMDDHHMM Date received in lab (of 1st part), or procedure/addendum order date/time
15	CM	Specimen Source	Part Type dictionary Interface Link for Part #1&Part #1's Description. **NOTE:** The Description is from the specimen and not the Part Type Dictionary Not applicable for procedures/addenda
16	CN	Ordering Provider	Person dictionary Interface Link^Lastname^Firstname (From Ordering/Submitting Physician) Not applicable for procedures/addenda.
17	TN	Order Call-Back Phone #	not used
18	ST	Placer's Field #1	not used
19	ST	Placer's Field #2	Normalcy code, used only for cytology cases. First determine if the specimen was inadequate: If any of the specimen adequacies have Unsatisfactory = "Y", send an "I' to indicate "inadequate". If the specimen is adequate, determine normal, abnormal, or atypical based on the value in the Interpretation dictionary for the specimen's Primary Interpretation. Values will be Y = abnormal, N = normal, M = atypical (MD Review). Not applicable for procedures/addenda.
20	CM	Filler's Field #1	Age at accession: ^64^Y Not applicable for procedures/addenda.
21	ST	Filler's Field #2	Procedure/addendum internal report instance number, or -1

SEQ	DT	Element Name	Description
			if this OBR corresponds to the case-level report.
22	TS	D/T Rslts Rpt/Status Chg	YYYYMMDDHHMM Date and time of sign out (of specimen or procedure/addendum, depending on ObjectID).
23	CM	Charge to Practice	not used
24	ID	Diagnosis Serv Sect ID	"SP" (for Surgical Pathology)
25	ID	Result Status	“F” (final report) normally “C” if report has been amended (corrected) "I" Specimen in lab; no result yet (not expected for caBIG)
26	CM	Linked Results	not used
27	TQ	Quantity/Timing	not used
28	CN	Result Copies To	not used
29	CM	Parent Accession #	not used
30	ID	Transportation Mode	not used
31	CE	Reason for Study	not used
32	CM	Principal Result Interpreter	Person dictionary Interface Link^Lastname^Firstname (From Primary Pathologist or procedure/addendum pathologist, depending on ObjectID).
33	CM	Assist Result Interpreter	Person dictionary Interface Link^Lastname^Firstname (From Signout Person; Neuro Pathologist for Autopsy Cases)
34	CM	Technician	Person dictionary Interface Link^Lastname^Firstname (From Cytotechnologist) Not applicable for procedures/addenda.
35	CM	Transcriptionist	not used
36	TS	Scheduled Date/Time	not used

OBX Segment

Each OBR (Observation Request) segment will be followed by the corresponding OBX (Observation/Result) segments. OBX's are used for several different types of result data – text fields (such as the gross description and final diagnosis), information about the "parts" (tissue sources), synoptic values, SNOMED codes, and blocks and stain/processes.

SEQ	DT	Element Name	Description
0	TX	Segment ID	"OBX"
1	SI	Set ID	Sequential Number (starts at 1 for each OBR)
2	ID	Value Type	"TX" for text fields "CE" for ANT/PROCESS/SYNOP/IMP "ST" for BLOCKs
3	CE	Observation ID	Observation type. The types can be: ANT – part (tissue source) information BLOCK – block information PROCESS – stain/process information SYNOP – synoptic value IMP – coded "impression" (SNOMED II, SNOMED CT and/or ICD9 codes), but not SNOMED CT codes from synoptic worksheets, which have the SYNOP type. {texttype} – textual information, see **NOTE 1** for text types
4	ST	Observation Sub-ID	ObjectID for ANT/BLOCK/PROCESS/SYNOP For SYNOPs this will be the ID of the worksheet. For IMPs this will be the case-level ID. For text types this will be the case-level ID
5	ST	Observation Value	The meaning of this value depends upon the type in SEQ 3. For part/ANT OBX's this is the Part Type dictionary Interface Link^^L (local code). For BLOCKs this is the block label designator^optional block description. For PROCESSes this is the Stain/Process dictionary Foreign Identifier^ dictionary description^L (local code). For SYNOPs this is the SNOMED CT code^possible fill-in value^CT, where an example of a fill-in value is the maximum tumor dimension. For IMPs this is the coded value^^code scheme. For SNOMED II codes (from the CoPath automatic coder, or entered manually) the code scheme is SNO, and ICD9 codes (which are likely to be very incomplete from CoPathPlus!) the coding scheme is ICD. For text-type OBX's this is the entire "blob" of text (can be multiple lines). The end of each line will be represented by the HL7 escape sequence \.br\. Other HL7 escape sequences (for bolding or underlining) can be embedded in the text.
6	ST	Units	Used only for SYNOP types that have fill-in values that are numeric (tumor size, for example, in which case the Units would most likely be "cm").

SEQ	DT	Element Name	Description
7	ST	Reference Range	Not used
8	ST	Abnormal Flag	Not used
9	NM	Probability	Not used
10	ID	Nature of Abnormal Test	Not used
11	ID	Observation Result Status	"I" Specimen in Lab, no result yet "F" Final Report "C" Corrected Report
12	TS	Date Last Obr Normal Values	Not used
13	ST	User Defined Access Checks	Miscellaneous add'l info that varies by type: For ANT=part sequence#^part designator (may equal the sequence# or be the alpha equivalent) For BLOCK=block sequence#^block label designator For PROCESS=stain/process sequence#^slide label text For SYNOP=worksheet interface link^worksheet instance For IMPS this is the "group designator" for SNOMED II's ("1" if not specified) or the "source" for ICDs (also "1" if none). In both cases this is a "hint" as to where the code came from. E.g., it is commonly the part designator, but not always. Empty for text fields.
14	TS	Date/Time of the Observation	Status Date/Time for BLOCKs and PROCESSes.
15	CE	Producer's ID	For SNOMED II type IMPs only: A=autocoded, L=linked (e.g., from part type dictionary), M=manually entered For SYNOPs the Part Designator for the associated part.
16	CN	Responsible Observer	not used

NOTE 1: The following text codes are used for "TX" text types:

At the case level:
"FIN" – Final Diagnosis
"GDT" – Gross Description
"MDT" – Microscopic Description

For addenda:
"ADX" – Addendum Diagnosis
"ADC" – Addendum Comment

For procedures:
"PIN" – Procedure Interpretation
"PRC" – Procedure Result/Comment