HSA Study PHI Corpus - Annotation guidelines

Version	Comments	Date	Drafted by
1.0	Initial guidelines	15 November 2016	Jitendra Jonnagaddala (TCRN/PoWCS/Med/UNSW)
4.0	Updated guidelines to accommodate records not meeting criteria	20 February 2017	Jitendra Jonnagaddala (TCRN/PoWCS/Med/UNSW)
5.0	Updated guidelines to include temporal annotations	11 January 2023	Hong-Jie Dai (NHRI/KMU/KMUH/NKUST) Jitendra Jonnagaddala (TCRN/PoWCS/Med/UNSW)

IRB Approval: HSA346a/2016

IRB Amendment: HSA346b/2016

IRB Amendment: 2023

** These guidelines are adapted based on the i2b2 PHI shared task 2014/2016 guidelines

1. Overview

The process of de-identification includes removing or masking all the personal health information in a document that may lead to identification of the person. This information includes things like person name, locations, date, age, contact numbers etc. It is extremely important to annotate this information in clinical texts before using them for any research purpose.

The process of de-identifying information includes the annotation or marking all such information in the text and then substituting this information with surrogates. Thus, the process of de-identification involves 2 steps- annotation of PHI information and then surrogate generation. Annotation of PHI information is a step where an annotator who is generally a subject matter expert would read through a document and mark all the PHI within the text. In general there are more than 1 annotator and a reviewer, all of them repeat the same process in order to capture all the PHI. Once the PHI are marked then they can be substituted with similar information by surrogate generation process. To properly generate shifted date surrogate, it is necessary to normalize the temporal narrative description to the ISO 8601 standard form.

The document here outlines the various categories of PHI and the guidelines that would help an annotator to identify the PHI and create the TIMEX3 annotations for temporal descriptions.

2. Corpus Pre-processing - Surrogate Generation

Surrogate generation is the process of masking the PHI data in order to protect the patient information. Any medical report like the clinical notes, discharge summary, pathological report etc. would contain sensitive patient information like name, address, age, profession etc. This information should not be shared publicly and needs to be removed from the records before sharing the data. Thus, the entire process of generating pseudo names, dates etc. is considered as surrogate generation. One way to generate surrogates is to use automated systems and then manually curate...sort of semi-automatic

The process of surrogate generation would include the following steps:

1. Identify the PHI within a record.
2. Replace it with a surrogate using the following rules:
    a. Surrogate DATEs are generated through date-shifting. For each document, shift all of the DATEs forward by the same random number of years, months, and days.
    b. For NAMEs, for each new document first randomly map each letter in the alphabet to another letter, and pre-decide that all NAMEs starting with, for example, letter A would be replaced by a surrogate that started with G as a way of simplifying initial generation for NAMEs. Then, pay attention to maintain gender information and to replace NAMEs with NAMEs of the appropriate gender by selecting from lists generated from census data. For example, assuming a mapping from A to G, and from F to D, “Angie Ferrerro” became “Grace Dollard”. In order to preserve coreference information, all occurrences of an authentic NAME should be replaced by the same surrogate. The surrogates mimic the surface form of the authentic PHI, mapping “A. Ferrerro”, “Mrs. Ferrerro”, “Angie” to “G. Dollard”, “Mrs. Dollard” and “Grace”, respectively.
    c. LOCATIONs and PROFESSIONs are replaced with a pre-compiled list of surrogates, while preserving coreference information.
    d. Hospital DEPARTMENTs are handled differently from other LOCATIONs: instead of replacing them with random surrogate department names, specific department names are matched to more generic ones. For example, the “Department for radiology, imaging, and oncology” became “Radiology”. If such a generic name did not exist, we changed the DEPARTMENT to “Internal Medicine”.
    e. AGE is left unchanged, unless they are 90 or over, in which case they are changed to “90”.

f. All numbers, including PHONEs, FAXes, and all sub-categories of IDs are replaced by randomly selecting new strings of digits/letters of the same length and format.

3. Corpus Pre-processing - Enrichment

- Manually "enrich" to include PHI that is especially difficult to identify (such as "foley catheter" and "Parkinson's disease")
- include more instances of infrequently appearing types of PHI.
- As these reports are pathological reports for tissue specimens they didn't contain much of patients' clinical details therefore no manual enrichment of dataset was done.

1 HIPAA and PHI

HIPAA (Health Insurance Portability and Accountability Act of 1996) is United States legislation that provides data privacy and security provisions for safeguarding medical information. HIPAA requires that patient medical records have all identifying information removed in order to protect patient privacy. This information belonging to the patients that help to identify them is known as Protected Health Information or PHI. There are 18 categories of Protected Health Information (PHI) identifiers of the patient or of relatives, employers, or household members of the patient that must be removed in order for a file to be considered de-identified.

In order to de-identify the records, each file must have the PHI marked up so that it can be removed/replaced later. For the purposes of this annotation project, the 18 HIPAA categories have been grouped into 8 main categories and 30 subcategories[1]:

PHI category	**Sub-category**	**Example**		
NAME	PATIENT, DOCTOR, USERNAME			
PROFESSION	(none)	Lawyer, teacher		
LOCATION	ROOM, DEPARTMENT, HOSPITAL, ORGANIZATION, STREET, CITY, STATE, COUNTRY, ZIP, OTHER	PERI-OPERATIVE UNIT-POW, MACQUARIE WARD - RHW		
AGE	(none)	23, 98		
DATE	DATE, TIME, DURATION, SET	Text	Normalization	Type
		24/12/1987	1987-12-24	DATE
		September 26th	09-26	DATE
CONTACT	PHONE, FAX, EMAIL, URL, IPADDRESS	abc@gmail.com, 194.223.1.1		
IDs	SOCIAL SECURITY NUMBER, MEDICAL RECORD NUMBER, HEALTH PLAN NUMBER, ACCOUNT NUMBER, LICENSE NUMBER, VEHICLE ID, DEVICE ID, BIOMETRIC ID, ID NUMBER	MRN: 9174338 ID NUMBER: 12R1500257 Block Id: B10		
OTHER	None	Finger Print, Photograph, Company Logo		

Thus, all the data mentioned above needs to be identified and annotated. The annotators need to identify the relevant text from the pathological reports and mark them according to the categories mentioned above. Furthermore, the time expression is required to determine the type and to be normalized to the ISO 8601 standard to encode the temporal information. The following sections define in details for the various categories along with the examples that would help the annotators to mark the PHI data and create the TIMEX3 annotations.

2 Annotation Guidelines

The annotator needs to annotate the PHI data in the pathological reports. This would be done using the annotation software MAE, and all PHI will be given an XML tag indicating its category, type and value, where applicable. This section defines each category along with examples to identify and tag the information correctly. There are a total of 8 categories as mentioned above. Whenever the annotator identifies text as PHI he/she needs to mark that text and annotate it under one of the categories. The ways to identify and annotate the relevant text are mentioned below.

[1] The subcategories TIME and DURATION, which are not covered by the HIPPA regulations, have been added in order to capture more detailed temporal information.

2.1 Name

This is the PHI category that mentions the name of a patient or a doctor or nurse etc. Basically any mention of a person's name should be annotated under this category. There are 3 sub-categories-Patient, Doctor and Username. The following rules should be followed while annotating a name:

1. Annotate the entire name or the initials in the text.
2. Titles (Dr., Mr., Ms., etc.) do not have to be annotated.
3. Information such as "M.D.", "R.N." do not have to be annotated.
4. If a name is possessive (e.g., Sam's) do not annotate the 's
5. USERNAME category be used for PHI that are initials followed by numbers (i.e., as4)
6. All the names of any professor, registrar, nurse etc. should be annotated as Doctor.
7. The initials of the doctors should be annotated.

The examples for name are as follows:

Sub-category	Examples(text to be annotated is marked in bold)
Patient	**JOHN,JONES** I had the pleasure to see Mrs. **Smith** today. SL Pathology (labelled 1111 **WILLIAMS**);
Doctor	Dr. **Alex Lamb** (**YG**/ta 17/1/14) MICROSCOPIC (reported by Dr **H Noah**): Prof **David Brown**
Username	**Smi123**

2.2 Location

This PHI is the mention of any physical location. It may be the hospital name, the department in a hospital, the name of an organization, address of the patient, city, country etc. It has multiple sub-categories including the parts of an address or an organization, hospital, department etc. The rules for annotating a location are as follows:

1. Annotate state/country names as well as addresses and cities. Each part of an address will get its own tag.
2. Generic locations (i.e., "hair salon", "golf course", "retail store") need to be annotated along with named organizations (i.e., "UNSW").
3. If there's a name of a medical facility and you're not sure if it should be a HOSPITAL or a DEPARTMENT, you should probably go with HOSPITAL.
4. "POW C" should be tagged as HOSPITAL(POW) ROOM(C)
5. Hospital room numbers should be annotated as ROOM
6. Floors and suites can also be annotated as ROOM
7. "Floor 2, room 2" can all be one ROOM tag
8. Annotate state/country names as well as addresses and cities. Each part of an address will get its own tag.
9. The departments inside of hospitals should be annotated, but only if they are unique. There is a list of generic hospital units at the end of this file; if a department is not on that list, it should be annotated.
10. Medical facilities should not be marked as organizations.
11. All the components of an address should be marked including the street, city, state, zip. For example if the address is mentioned as :
    1 TODMAN AVENUE
    FARMBOROUGH HEIGHTS NSW 2526
    Annotate the following:
    '1 TODMAN AVENUE'-Street

‘FARMBOROUGH HEIGHTS’-City
‘NSW’- State
‘2526’-Zip

Sub-category	Examples(text to be annotated is marked in bold)
Hospital	**St. George** Hospital Received from the **Prince of Wales** Hospital Randwick 8.MEDIC/SURGERY WARD-**POWP**
Department	**8.MEDIC/SURGERY WARD**-POWP the previous excision from **Southern IML** Pathology Slides sent to **SEALS Central**
Room	**Room 12** **OBG floor, Room3**
Street	**12 Kings Avenue** **345 Beach Street**
City	**Randwick** **Bondi** **CONISTON**
State	**NSW** **ACT**
Zip	**2344** **2567**
Country	**Australia** **USA** **South Africa**
Organization	**Oracle, Microsoft**
Other	**P.O. BOX-31112**

2.3 Age

Any mention of age should be annotated. The annotator should annotate the number for age and not the associated text. Annotators should annotate all ages, not just those over 90, including those for patient's families if they are mentioned.

Example:

1. **72**yr old.
2. TIA age **42**
3. **60** yr old, squamous cervix carcinoma.

2.4 Date

In this project, any kind of temporal expression within the text should be annotated. Temporal expressions are phrases that contain time information. The types of temporal expressions that we need to mark include date (represented by the subcategory DATE), time (represented by the subcategory TIME), duration (represented by the subcategory DURATION) and frequency (represented by the subcategory SET). Dates are an important PHI because the dates include date of births which can be used to get the age of the patient. Dates can take a varieties of formats, like, ‘/’ or ‘-‘ as the separator, just the date and month, only the year, the month as a number or in words etc. Thus, the annotators need to identify any format of temporal expressions and annotate by using one of the subcategory tags. The detailed guidelines to annotate the temporal expressions are as follows:

1. Any calendar date, including years, seasons, months, holidays and time of day, should be annotated.
2. Days of the week should also be annotated
3. If the phrase has 's (i.e., "in the '90's), annotate "'90's"
4. Include annotations of seasons ("Fall '02")
5. Include quote marks that stand in for years ("'92")

We use the DATE tag to annotate dates (absolute and relative dates), the TIME tag to annotate time and date with the time expression, the DURATION tag to annotate duration and the SET tag to exclusively annotate expressions which give both a quantifier and an interval (e.g. "two times weekly", "1/day"). Each tag has the "val" attribute conveying the normalized value of the various forms of temporal phrases. The purpose of the attribute field is to store the temporal information in a standard, normalized format in order for machines to understand and process them. More specifically, we use the ISO 8601 standard to encode the temporal information in this guideline. Therefore, for each temporal expression, you need to determine its value representation and fill out the "val" field in MAE.

We suggest the annotators to follow the rules described in the "i2b2 2012 Temporal Relations Challenge Annotation Guidelines"[2] to determine the value representation. In the following paragraphs, we excerpt the rules for the reference.

- **DATE value representation**: The format of calendar date in ISO 8601 is: **YYYY-MM-DD**. If the temporal expression doesn't include the exact date information, the format **YYYY-MM** is allowed. If the temporal expression only includes the year information, you can simply put **YYYY** in the "val" field.
- **TIME value representation**: The extended time format of ISO 8601 is used for annotation, specified as: **hh:mm:ss.hh** ranges from 00 to 24 (where 24 is only used to represent the end point of a calendar day, and the hours are zero-padded, e.g. 9 am -> 09); **mm** and **ss** both range from 00 to 59, zero-padded, and represent minutes and seconds respectively. For reduced accuracy, **ss** and/or **mm** can be omitted, *i.e.* **hh** and **hh**:**mm** are allowed. A combined date and time representation follows the format YYYY-MM-DD**T**hh:mm:ss. In your annotation, if a time expression can be anchored to a specific date, as is usually the case (see the examples below), make sure to include the date information in the "val" field as well.
- **DURATION value representation**: Durations are temporal expressions that describe a period of time, *e.g.* for eleven days. The syntax of duration representation is **Pn**[**YMWD**]. So, "for eleven days" will be represented as "P11D", meaning a period of 11 days. The **n** field doesn't have to be an integer; "P0.5Y" is a valid notation for a period of half a year. The designators can be combined to express a compound period, so "P2M3D" means a period of 2 months and 3 days. As tabulated below, the designator "M" can mean both "month" and "minute". In order to differentiate, we use the "T" designator to signal time periods, e.g. "P20M" = 20 month, "PT20M" = 20 minutes.

Designator	Meaning	Designator	Meaning
P	Period	T	Time

[2] Refer to the supplementary data 1 of the paper published by Journal of Biomedical Informatics entitled "Annotating temporal information in clinical narratives" written by Dr. Weiyi Sun *et al.* The guideline should be available at the following URL: https://www.sciencedirect.com/science/article/pii/S1532046413001032

Y	Year	H	Hour
M	Month	M	Minute
W	Week	S	Second
D	Day		

- **SET value representation**: The frequency of events can be represented by using the ISO 8601 repeated intervals. The ISO repeated interval syntax is: **R[*n*][*duration*]**, where ***n*** denotes the number of repeats and ***duration*** denotes the DURATION value representation. When the ***n*** is omitted, the expression denotes an unspecified amount of repeats. For example, "once a day for 3 days" is "**R3P1D**" (repeat the time interval of 1 day (**P1D**) for 3 times (**R3**)), twice every day is "**RP12H**" (repeat every 12 hours). In clinical text, we may see frequencies such as p.r.n (take as needed), which can be represented by "R" (repeat for unspecified times of unspecified duration).

Sub-category	Examples (text to be annotated is marked in bold)	Normalized Value
DATE	D.O.B: **16/01/1941** (JL/ta **20/11/13**) Additional history (… HI-12-111111 **3.Apr.12**) osteosarcoma identified in the previous resection from **2009** lesion removed from the left arm in **October 2011** … Collected: **16/07/2013** at :… **Now** having palliative TAH ...	**1941-01-16** **2013-11-20** **2012-04-03** **2009** **2011-10** **2013-07-16 2013-07-16**
TIME	Collected: **02/09/2014 at 11:42** .. A/Prof E Salisbury **at 9:30am on 18/3/14**. Dr N Lambie **at 9:16am on the 18th of September 2013**	**2014-09-02T11:42** **2014-03-18T09:30** **2013-09-18T09:16**
DURATION	⋯ whilst **20weeks** pregnant in 2013 Lung cancer resected **two weeks** ago.	**P20W** **P2W**
SET	… tested **twice** with positive controls;	**R2**

2.5 Identifiers (ID)

Identifiers are numeric or alphanumeric identifiers in the text that would help in the identification of the patient. These identifiers can be hospital numbers, record numbers, etc. The guidelines to annotate the ids and the examples of each category are as follows:

1. Any kind of numeric or alphanumeric identifier should be annotated.
2. Only the ID with the hospital name should be annotated as MEDICAL RECORD.
3. Block numbers on which procedures are carried out should be annotated as IDNUM.
4. Lab IDs and Episode Numbers should be annotated as IDNUM.
5. When in doubt, call something a IDNUM
6. Doctor or nurse IDs should be annotated as "other id"
7. No need to label names of devices (for example: "25 mm Carpentier-Edwards magna valve", "3.5 mm by 32 mm Taxus drug-eluting stent", Angioseal")

Sub-category	Examples(text to be annotated is marked in bold)
Medical Record Number	**022213.PWP** **55555774.RAN**
Social Security Number	
Account Number	
Health plan number	
License number	
Vehicle id	**CB 33 GO**

	LCS 23 **CC 12 MS**
Device id	.035 x 180cm Fixed Core 3mm J wire - Qty: 1Each Part #: **17722GNP** Boston Scientific 6F JL 4.0 - Qty: 1 Each Part#: **30058**{Catheters}
Biometric id	Aventis Fluzone Lot **L5317YR**
Id number	Episode No: **12R423044B** Lab No: **12H08861** Block **B1** Block **1**

2.6 Contact

This PHI is the data that specifies the details that may help to contact the patient. The examples of this include the phone number, email, fax etc. Thus, the annotator should annotate any data in the text record that specifies a contact number for the patient. The following guidelines should be followed:

1. Annotate all the phone numbers, fax numbers, e-mail, URL, IP address, etc.
2. Pager numbers should be annotated as phone numbers.
3. The phone numbers are usually 4 or 8 digit and have the format "02-xxxx-xxxx". It often starts with "029385" because that's the telephone code for the hospitals to which the records belong.
4. Do not annotate numbers starting with #. They need to follow the format defined below.

Sub-category	Examples(text to be annotated is marked in bold)
Phone	**02-2222-3455** **8282 7154**
fax	**02-2222-3455** **8123 9876**
email	**abc@gmail.com**
url	**www.xyz.com**
ipaddress	**192.1.1.1**

2.7 Profession

The job or the profession of any patient should be annotated under this category. The annotators should annotate any profession mentioned in the text like carpenter, teacher, lawyer etc. The only exception is the medical profession. The medical profession related jobs should not be annotated.

1. Any job mentioned for the patient should be annotated.
2. any job that is mentioned that is not held by someone on the medical staff should be tagged
3. College majors are considered professions

Examples, the following should be annotated:

1. The patient is a **carpenter**.
2. Lauri has a full-time **teaching** job in the local school.

The following should not be annotated:

1. The results were conveyed to the anaesthetic.
2. The nurse recorded the results of the procedure.

2.8 Other

Any data that would potentially help in identifying a patient but is not included in the categories above should be marked as other. The examples of this category include the finger prints, or a company logo etc. The PHI may be an image rather than just the text.

The section above explains 8 categories and the associated sub-categories along with the examples for each. The next section mentions some of the general guidelines and tips that would help an annotator in the task:

1. Multiple occurrences of the same PHI should be annotated. If the same PHI exists multiple times within a record then all the instances should be annotated. There are several PHIs that may exists multiple times like,name, city, hospitals, identifiers etc. For example,
   Receptors (Block **B1**): ….
   Her-2 in-situ hybridisation: (Her-2 CISH Invitrogen SPOT-Light assay, block **B1**)
2. Only annotate the information that would need to be replaced when the file is re-identified.
3. When in doubt, annotate!
4. When tagging something that is PHI but it's not obvious what to tag it as, think about what it should be replaced by, and whether that will make sense in the document

The annotator should follow the guidelines as defined above and try to annotate as much as information as possible from the text records in order to de-identify the records. If any of the information is not annotated that information would not be included in the de-identification process and hence, that would put the patient at the risk of re-identification.

Appendix - GENERIC HOSPITAL DEPARTMENTS

Acute assessment unit
Cardiology
Coronary care unit/CCU
Critical care
Ear nose and throat (ENT)
Emergency department/ED
Emergency room/ER
Emergency ward/EW
Gastroenterology
General surgery
Geriatric intensive-care unit
Gynaecology
Haematology
Intensive care unit (ICU)
Internal medicine
Maternity
Medical records department
Neonatal unit
Neonatal intensive care unit (NICU)
Nephrology
Neurology
Obstetrics
Occupational therapy
Oncology
Operating room/OR
Ophthalmology
Orthopaedics
Pediatric intensive care unit (PICU)
Pharmacy
Physical therapy
Post-anesthesia care unit
Psychiatric Unit/Psychiatry
Radiology
Rheumatology
Surgery
Urgent care
Urology

Appendix - 18 HIPAA PHI categories (45 CFR 164.514).

1. Names;
2. All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly-available data from the Bureau of the Census:
    (a)The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and
    (b)The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
4. Telephone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. Web Universal Resource Locators (URLs);
15. Internet Protocol (IP) address numbers;
16. Biometric identifiers, including finger and voice prints;
17. Full face photographic images and any comparable images; and
18. Any other unique identifying number, characteristic, or code.

Appendix – Original 2016 DTD

```
<!ENTITY name "deId">

<!ELEMENT NAME ( #PCDATA ) >
<!ATTLIST NAME TYPE ( PATIENT | DOCTOR | USERNAME ) >
<!ATTLIST NAME comment CDATA >

<!ELEMENT PROFESSION ( #PCDATA ) >
<!ATTLIST PROFESSION TYPE ( PROFESSION ) >
<!ATTLIST PROFESSION comment CDATA >

<!ELEMENT LOCATION ( #PCDATA ) >
<!ATTLIST LOCATION TYPE ( ROOM | DEPARTMENT | HOSPITAL | ORGANIZATION |
STREET | CITY | STATE | COUNTRY | ZIP | LOCATION-OTHER ) >
<!ATTLIST LOCATION comment CDATA >

<!ELEMENT AGE ( #PCDATA ) >
<!ATTLIST AGE TYPE ( AGE ) >
<!ATTLIST AGE comment CDATA >

<!ELEMENT DATE ( #PCDATA ) >
<!ATTLIST DATE TYPE ( DATE ) >
<!ATTLIST DATE comment CDATA >

<!ELEMENT CONTACT ( #PCDATA ) >
<!ATTLIST CONTACT TYPE ( PHONE | FAX | EMAIL | URL | IPADDR ) >
<!ATTLIST CONTACT comment CDATA >

<!ELEMENT ID ( #PCDATA ) >
<!ATTLIST ID TYPE ( SSN | MEDICALRECORD | HEALTHPLAN | ACCOUNT | LICENSE |
VEHICLE | DEVICE | BIOID | IDNUM ) >
<!ATTLIST ID comment CDATA >

<!ELEMENT OTHER ( #PCDATA ) >
<!ATTLIST OTHER comment CDATA >
```

Appendix – UNSW/HSA DTD

```
<!ENTITY name "deid">

<!ELEMENT NAME ( #PCDATA ) >
<!ATTLIST NAME TYPE ( PATIENT | DOCTOR | USERNAME ) >
<!ATTLIST NAME comment CDATA >

<!ELEMENT PROFESSION ( #PCDATA ) >
<!ATTLIST PROFESSION TYPE ( PROFESSION | DONT USE ) >
<!ATTLIST PROFESSION comment CDATA >

<!ELEMENT LOCATION ( #PCDATA ) >
<!ATTLIST LOCATION TYPE ( ROOM | DEPARTMENT | HOSPITAL | ORGANIZATION |
STREET | CITY | STATE | COUNTRY | ZIP | LOCATION-OTHER ) >
<!ATTLIST LOCATION comment CDATA >

<!ELEMENT AGE ( #PCDATA ) >
<!ATTLIST AGE TYPE ( AGE | DONT USE ) >
<!ATTLIST AGE comment CDATA >

<!ELEMENT DATE ( #PCDATA ) >
<!ATTLIST DATE TYPE ( DATE | DONT USE ) >
<!ATTLIST DATE comment CDATA >

<!ELEMENT CONTACT ( #PCDATA ) >
<!ATTLIST CONTACT TYPE ( PHONE | FAX | EMAIL | URL | IPADDR ) >
<!ATTLIST CONTACT comment CDATA >

<!ELEMENT ID ( #PCDATA ) >
<!ATTLIST ID TYPE ( SSN | MEDICALRECORD | HEALTHPLAN | ACCOUNT | LICENSE |
VEHICLE | DEVICE | BIOID | IDNUM ) >
<!ATTLIST ID comment CDATA >

<!ELEMENT OTHER ( #PCDATA ) >
<!ATTLIST OTHER comment CDATA >
```

Appendix – UNSW/NKUST AICUP DTD

```
<!ENTITY name "deid">

<!ELEMENT NAME ( #PCDATA ) >
<!ATTLIST NAME TYPE ( PATIENT | DOCTOR | USERNAME ) >
<!ATTLIST NAME comment CDATA >

<!ELEMENT PROFESSION ( #PCDATA ) >
<!ATTLIST PROFESSION TYPE ( PROFESSION | DONT USE ) >
<!ATTLIST PROFESSION comment CDATA >

<!ELEMENT LOCATION ( #PCDATA ) >
<!ATTLIST LOCATION TYPE ( ROOM | DEPARTMENT | HOSPITAL | ORGANIZATION |
STREET | CITY | STATE | COUNTRY | ZIP | LOCATION-OTHER ) >
<!ATTLIST LOCATION comment CDATA >

<!ELEMENT AGE ( #PCDATA ) >
<!ATTLIST AGE TYPE ( AGE | DONT USE ) >
<!ATTLIST AGE comment CDATA >

<!ELEMENT DATE ( #PCDATA ) >
<!ATTLIST DATE TYPE ( DATE | TIME|DURATION|SET ) >
<!ATTLIST DATE val CDATA >
<!ATTLIST DATE comment CDATA >

<!ELEMENT CONTACT ( #PCDATA ) >
<!ATTLIST CONTACT TYPE ( PHONE | FAX | EMAIL | URL | IPADDR ) >
<!ATTLIST CONTACT comment CDATA >

<!ELEMENT ID ( #PCDATA ) >
<!ATTLIST ID TYPE ( SSN | MEDICALRECORD | HEALTHPLAN | ACCOUNT | LICENSE |
VEHICLE | DEVICE | BIOID | IDNUM ) >
<!ATTLIST ID comment CDATA >

<!ELEMENT OTHER ( #PCDATA ) >
<!ATTLIST OTHER comment CDATA >
```