

Common Schema for Person Name

1. This discussion paper analyses how we may proceed to develop a Common Schema for “person name”.
2. A person’s name is just an attribute of a person. It cannot be used to uniquely identify a person because two individuals can have the same name.
3. A person’s name may appear differently in different computer systems. If no clear instruction were specified to guide how one inputs his name, the captured content would be un-predictable. For example, one could input his name as “Chan Tai Man” or “Tai Man Chan”, and it would be hard to tell whether he is Mr Chan or Mr Tai.
4. A person’s name can be used as one of the searching criteria to help retrieve a person’s record. However, there is no guarantee that a person’s record can be retrieved from a given name because, as explained in the paragraph above, a person can write his name in many ways, for example, “Mrs Chan Mei Yee” and “Ms May Fung” can be the same person.
5. Person name can also be used for sorting persons’ records. A convention for writing the name is necessary to make the sorting meaningful. It can also be used for printing mailing labels.
6. First, we describe how person names are stored and used in some government systems. Then we discuss the considerations for developing the Person Name Common Schema.

Types of “Person Names” Used in Government Services

7. Person Names are stored in departmental systems either as :
 - a string of unstructured text;
 - a string of semi-structured text (e.g. use a comma to delimit the

- surname and the given names); or
- a structured data element comprising sub-components like surname, maiden name, and given name, etc.
8. A person's title (e.g. Mr, Mrs, Ms, Dr, etc.) is sometimes captured and maintained as a discrete piece of information. If a person's title is captured for the purpose of addressing a person politely (e.g. Mrs Chiang, Ms Sung), then the input form should explicitly spell out this purpose; otherwise, the person providing information may treat "title" as a means of capturing his/her marital status and the captured title may not match with the surname.
9. A simple survey of the downloadable forms (640 in total) from the HKSARG Website shows that the majority of forms capture person name in an unstructured format. Only 9% capture surname and given name separately, and 25% capture person titles separately.
10. The Immigration Department (IMMD) stores a registered person's name in IMMD's internal database using the following format:
- English name as a string of text with surname followed by a comma and the given names; and
 - Chinese name as a string of digits representing the Chinese Commercial Code (CCC) of a maximum of 6 Chinese characters (no comma is used to separate the surname from the given name).
11. In the early 90s, the Information Technology Services Department (ITSD) suggested bureaux and departments (B/Ds) to store person name (in English) using a format similar to that of the IMMD's.
12. There may be reasons driving a B/D to store person name as a semi-structured (like IMMD's format) or structured data element, e.g. :
- the B/D needs to sort persons by their surname and given name (e.g. producing the sorted register of electors)
 - the B/D lets user input person name in a structured way to facilitate data entry, and adopts the same structure in its database;
 - the B/D wants to address the person in a more polite way (e.g. printing letter head as Dear Mr. Ho or Dear Dr. Lee).

Considerations for Developing a Common Schema for “Person Name”

(Note : The issues (in bold italics) below are open issues with no conclusive solution yet. The analysis we have put down are just one of the many possible viewpoints, they serve to stimulate discussion only.)

Issue #1 : What data format should be used to exchange person name ? Should we have different person name structures for different languages ?

13.As most B/Ds are keeping person name as unstructured text, they cannot convert unstructured names to semi-structured or structured names. As such, the exchange format must cater for unstructured names. However, we should also facilitate the exchange of structured information when they are available. The format for exchanging English names is proposed as follows :

Person Name in English

1. Person Full Name
2. Structured Components
 - 2.1. Person Surname
 - 2.2. Person Given Names
3. Person Title (optional)

(Note : either the Person Full Name or the Structured Components must exist. B/Ds should provide the structured component where possible. If the person title and person surname are provided, they are expected to match for addressing the person accurately)

14.It is arguable what to be included in the structured components. To one extreme, we can include surname and given names for male, but husband’s surname (if any), maiden name and given names for female. To the other extreme, we can just include surname and given names for all persons and leave it to individual female whether to input her

husband's surname, her maiden name, or both. Our decision should be based on business need and practical constraints such as what information is maintained in existing systems. Since most departmental systems do not maintain maiden name separately, it is suggested not to include it as a separate component.

15. Both the person name in Chinese and in English have similar components. The difference is in the maximum length of each component. Full person name in English commonly used in HK has at most 40 characters while that in Chinese has at most 6 characters only. In addition, the CCC is stored by some B/Ds. Therefore, the Chinese Person Name is proposed to adopt a slightly different structure as follows :

Person Name in Chinese

1. Person Full Name (6 characters)
2. Structured Components
 - 2.1. Person Surname
 - 2.2. Person Given Name
3. Person Title (optional)
4. Chinese name in CCC code (optional)

(Note : either the Person Full Name or the Structured Components must exist. B/Ds should provide the structured component where possible. If the person title and person surname are provided, they are expected to match for addressing the person accurately)

Issue #2 : What principles should we observe when we develop a Common Schema for person name ?

16. The principles may include :

- No matter B/Ds store person name as a structured or unstructured data element, they should be able to convert their data to the common data exchange format. (It should be noted that it may not be possible to convert unstructured person name received from

another party to a structured format)

- Using a common data exchange format will have impact on B/Ds that are using a different format in their backend systems. We should ensure that the overall impact to all B/Ds is minimal

Issue #3 : Should Chinese Commercial Code (CCC) be used to exchange the Chinese person name ?

17.The IMMD stores a person's Chinese name using CCC in its internal database. Some other B/Ds also keep the CCC of a person's Chinese name. CCC consists of 4 numeric digits and an optional field to represent the glyph of a Chinese character that can be written in different ways. The 4 numeric digits are printed on the Hong Kong Identity Card without the optional field.

18.Some Chinese characters can be written in different ways (i.e. represented using different glyphs). Some of these glyphs are treated as the same character in BIG5/ISO 10646/HKSCS but not in the CCC. Using the CCC (including the optional digit) can preserve a specific written representation of the Chinese character. However CCC is not widely adopted in the public and private sector.

19.The Hong Kong Smart ID Card has the CCC printed and it also stores the person's Chinese name in ISO 10646 with an extra byte indicating which glyph is used for printing.

20.Since CCC is not commonly used in HK, it should not be used as the primary means to encode a Chinese person name. However, CCC can be included in the exchange format as an optional field.

Issue #4 : How to convert between structured components and unstructured name ?

21.A structured name can be converted to an unstructured name easily. For person name in English, it is recommended to adopt IMMD's

convention, i.e., surname followed by a comma and the given names. For person name in Chinese, we can simply concatenate the surname and the given names.

22. However, without “tools”, it is not possible to convert an unstructured name into its structured components. If we capture a person’s name without specific instruction on how to write the name, we may get : “Chan Mei Yee”, “Mei Yee, Chan”, “May Chan”, “Chan Mei Yee, May”, etc.. Even when there are guidelines to instruct people to write their surname first then followed by given names, there may still be uncertain cases. For example, if a lady writes “Au Yeung Mei Yee” as her name, we would not know if “Au Yeung” is her husband’s surname, or her maiden name, or “Au Yeung” reflects her husband surname “Au” followed by her maiden name “Yeung”. In this case, we have to ask the person to confirm.

23. Knowing the limitation of unstructured person name, data owners should use such information carefully. To improve the situation slightly, the data sender should provide structured components where possible. However, as explained earlier in the “May Fung” (see paragraph 4) example, even structured names cannot always help you locate a person’s record.

24. Based on the exchange format suggested under issue #1, below is an analysis of what the sender and receiver needs to do in some possible scenarios:

		receiver	
		Store structured components in backend	Store unstructured name in backend
sender	Store structured components in backend	Sender and receiver exchange person name as structured components	Sender sends person name as structured components, receiver converts name to unstructured text

	Store unstructured name in backend	Sender sends person name as unstructured text. Receiver figures out how to get structured names, if necessary	Sender and receiver exchange person name as unstructured text
--	------------------------------------	--	---

Issue #5 : How to handle maiden name ?

25. Not many B/Ds keep maiden name as a discrete piece of information; most B/Ds ask the public for surname and given name and leave it to the public whether and how they put down their maiden name.

26. The IMMD allows a married female to include her husband's surname as part of her full name, or change her surname to her husband's surname. It is up to individual to make this request or not. If both the husband's surname and the maiden name were to be included, IMMD would concatenate the two surnames and store it as the female's surname. For Chinese people, the common practice is to store the husband's surname and the maiden name together as surname. For Westerners, the common practice is to drop the maiden name.

27. We should not expect the person name exchange format to reflect "maiden name" accurately because B/Ds rarely need this piece of information.

28. We also have no way to tell from a female's surname alone whether that is her husband's surname or her maiden name.

Issue #6 : What are the industry standards we can make reference to ?

29. Please refer to Appendix A

Issue #7 : What are the person name formats currently in use by B/Ds that we can make reference to ?

30. Please refer to Appendix B

Issue #8 : What are the person name formats adopted by other economies that we can make reference to ?

31. Please refer to Appendix C

Way Forward

32. The Task Force is recommended to:

- align the data structure for exchanging person name;
- define each data element in the structure (e.g.
 1. the full name (or part of it) can help locate a person's record. Depending on how the full name was captured and whether the input has been verified, the full name can be written in an un-predictable way;
 2. the surname is a person's family name. Depending on how it was captured and whether the input has been verified, it may or may not match with that on the HKID card
 3. the title is how a person expects others to address him/her, it should correspond to the person's surname if both exist)
- align the field length of each data element
- align any necessary controlled vocabularies such as person title.

**Information Technology Services Department
November 2003**

Appendix A - Industry Standards Related to Person name

Industry Standard/ Project Team	Organization	URL
xNAL (includes xNL: eXtensible Name Language and xAL- eXtensible Address Language)	OASIS Customer Information Quality (CIQ) Technical Committee	http://www.oasis-open.org/committees/ciq/download.shtml select Download the HTML formatted documents for CIQ SCHEMAS only
HR-XML	HR-XML Consortium	The HR-XML Consortium will develop a common HR vocabulary and model, and will develop schemas for common HR objects used in Recruiting and Staffing, Benefits Enrollment, Payroll, etc. It is working on the many attributes of the Person object which must be handled consistently by different HR processes. The first two attributes defined are PersonName and PostalAddress http://ns.hr-xml.org/2_1/HR-XML-2_1/CPO/PersonName.pdf
vCard	Internet Mail Consortium	vCard is used in applications such as Internet mail, voice mail, Web browsers, telephony applications, call centers, video conferencing, PIMs (Personal Information Managers), PDAs (Personal Data Assistants), pagers, fax, office equipment, and smart cards. vCard is commonly used in exchanging personal information among those systems. http://www.imc.org/pdi/ http://www.ietf.org/rfc/rfc2426.txt There are 2 “name” types defined in vCard: FN and N. FN (formatted name) is based on X.520 Common name attribute and N is based on X.520 individual name attributes. N is a

		structured type value that corresponds, in sequence, to the Family Name, Given Name, Additional Names, Honorific Prefixes, and Honorific Suffixes. The text components are separated by the SEMI-COLON character.
--	--	---

Appendix B – Person Name Formats In use by B/Ds

B.1 Person name format used to capture user input in ESD Change of Address transaction

Field Description	length
Surname in English	20
Other name in English	20
Chinese full name	30
Person title	Controlled list

B.2 Person name format used for searching in Government Directory

Field Description	length
Surname in English	64
Other name in English	32
Surname in Chinese	64
Other name in Chinese	32

B.3 Immigration Department

Field Description	length
Person full name in English (Surname first followed by a comma and a space to separate surname and given names)	40
Chinese Commercial Code (No Chinese character is stored in the system)	6 x (4+1 numeric digits)

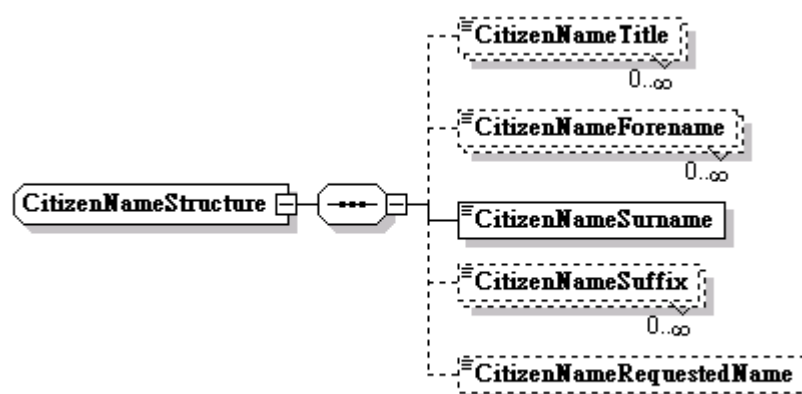
Appendix C – Person Names Used in Other Economies

- C.1 Person name structure in UK as stipulated in “UK Online – Information Architecture – Address and Personal Details Fragment”

http://www.govtalk.gov.uk/documents/IA_APD_1_1.doc

The following examples on person name is compiled according to data specifications published on the UK Govtalk homepage (Schemas & Standards - Agreed Schemas - Address & Personal Details)

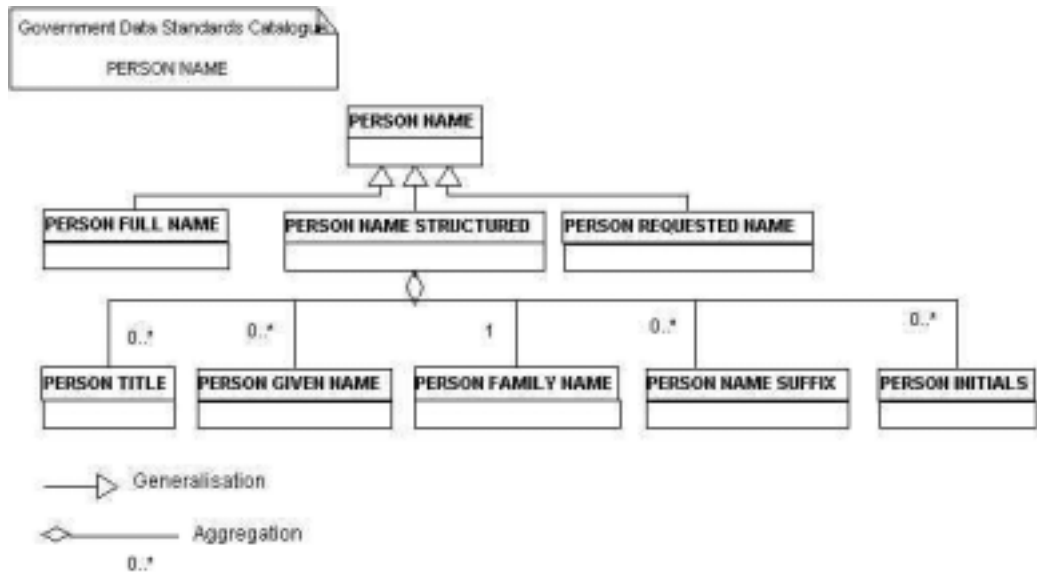
http://www.govtalk.gov.uk/schemasstandards/agreedschema_schema.asp?schemaid=182



Generated with XMLSpy Schema Editor www.xmlspy.com

The following UML is captured from UK Government Data Standard Catalogue on Person name. It represents possible variance that person name may compose of.

<http://www.govtalk.gov.uk/schemasstandards/eservices.asp>



C.2 Person name structure in US Department of Justice – Global Justice XML Data Dictionary Version 3.0 Prerelease
http://it.ojp.gov/topic.jsp?topic_id=43

