Address CSTF Paper 2004/01

Common Schema for Address (Paper No. 2)

(revised to incorporate suggestions collected during the 2nd meeting of the Address CSTF)

Introduction

1. In December 2003, a survey[1] was conducted to explore whether it is feasible to use a common format to exchange address information among bureaux and departments (B/Ds). This discussion paper :

   - analyses the information collected from the survey;

   - recaps the basic concepts in relation to the processing of address information;

   - makes recommendations on how to facilitate address information exchange.

2. It is important that we agree on the concepts first, before we proceed to standardization. Readers are requested to read this paper with this attitude in mind.

3. An address can refer to a variety of locations (e.g. a unit in a building, a bus stop on the street, etc.). This paper only covers the physical and postal address of persons and organizations, and the address of property units where person and organization may occupy. Such addresses normally refer to a property unit in a building or perhaps the entire building.

4. This paper does not touch upon the problem of presenting an address in different languages (such as the different floor numbering systems used in Chinese and English addresses). Such problems will be dealt with at a later stage of this standardization exercise.

[1] The survey form, its appendix, and the findings are accessible from the Government Intranet at http://itginfo.ccgo.hksarg/content/if/cstf/index_cstf.htm

5. This paper only proposes the strategy for recording and exchanging address, the exact structure of the address has yet to be worked out by the Address Common Schema Task Force (CSTF). The structures used in this paper are for illustration only.

Observations from the Survey

6. Among the 21 B/Ds who indicated that they need to exchange address information, 16 indicated that they will have difficulty using the proposed format to exchange address information. The reasons being :
   - a different free text address format is now used in their applications, converting free text address information to another dimension may distort the presentation of the address;
   - the exchanged address information needs to be in a different structure which is designed to facilitate data analysis; and
   - a different pre-agreed address exchange format is now in use, changing it will invoke re-negotiation with other parties.

7. Among the 21 B/Ds who indicated that they need to analyse address information, 12 indicated that they will have difficulty using the proposed address structure to support data analysis, the reasons being :
   - they need to analyse land lot information;
   - they need a more detailed structure for data analysis; and
   - their address structure currently in use is optimal for their operation (e.g. address is encoded as estate code + building code + room number).

8. 10 B/Ds indicated that they need to keep land lot information as a discrete piece of information for analysis.

9. Many applications keep address information as one or more lines of free text. There is little commonality in the free text address format they use. Please refer to **Appendix A** for examples of the free text address formats in use by e-government applications.

10. Address always refers to a location, be it a farmhouse, an office, a residence or a post box. However, the object associated with the address varies from case to case. For some applications, the object associated with the address is a person or an organization. For others, the object associated with the

address is the location itself (e.g. a property, a piece of land). Although persons / organizations occupy locations and therefore a person / organization's address should correspond to a location's address, these addresses are structured differently due to different business requirements.

11. For cases where an application's concern is the address of a person or an organization, address information is usually kept in a less structured way because not many applications need to perform detail analysis on the address of a person or an organization. However, there are exceptions. The Registration and Electoral Office (REO) maintains the address of electors in a relatively structured way to facilitate the assignment of geographical constituencies and polling stations.

12. For cases where an application's concern is the address of a location, address information is usually kept in a more structured way to facilitate data analysis.

13. Some B/Ds keep record of all properties of a certain type in a highly structured way. For example, the Census and Statistics Department (C&SD) keeps record of quarters in buildings and structures where persons may reside for supporting population censuses / by-censuses and other household surveys. Such records cover each and every unit of quarters in a specific building and information including address is held in a database comprising multiple tables. Similarly, the Rating and Valuation Department (RVD) keeps record of all ratable properties, and the Post Office (PO) keeps record of all possible postal delivery locations.

14. Different business requirements drive applications to keep address information in different ways.

15. From the survey, we also observed that the RVD receives deed registration data from the Land Registry (LR), the RVD receives stamp duty data from the Inland Revenue Department (IRD), the Electrical and Mechanical Services Department (EMSD) receives buildings' electricity supply report from the electric companies, the Housing Department sends billing address to utility companies when a new public housing tenant moves in, the Lands Department receives government rent payment data from the Treasury, etc.

16. At present, when one application needs to send address information to

another application, the project teams negotiate among themselves on a case-by-case basis and design project-specific mapping tables to deal with the incompatibilities in address formats. In some cases, B/Ds have to match incompatible addresses contributed by different sources using both computer and manual approach each time they create or update their address data based on the contribution. Manual matching is resource consuming and sometimes the mis-matches cannot be resolved. What should we do to facilitate address information exchange ? The next section analyses this problem, starting with basic concepts and basic assumptions.

Analysis

What do we record address for ?

17. Applications record address for different purposes. From the data processing point of view, we can categorize the purposes into the following key areas :
    - for human interpretation (e.g. for correspondence or for on-site visits);
    - for computer-aided analysis; and
    - for more automated analysis.

18. In cases where an address is recorded in a legal-binding document, the application processing the document may need to capture the address as how it is recorded on the legal document (as free text in most cases). At the same time, the application may need to capture the address as structured components (e.g. flat, floor, block, building number, street, etc.) for automated analysis. Whether the addresses recorded in different formats truly correspond with each other is a business issue that the application owner needs to handle; it is an issue outside the context of this paper.

Storing and exchanging an address for human interpretation

19. If we record an address purely for human consumption (e.g. for correspondence or for on-site visit), a common practice is to capture and store the address as free-text without data validation[2].

[2] In the context of this paper, the purpose of validating an address is to check, as far as possible, whether the addressed location physically exist or not. Application systems usually implement

20. While we agree that in such cases it is not practical to mandatory impose address validation because this will incur additional overhead on the information collector to contact the information provider to rectify inaccurate information, we recommend new applications that plan to capture address as free text to try associating implicit meaning[3] onto the free text address. This will help generate good quality correspondence address which can help postal delivery.

21. Applications can either
    - guide the information provider what to write on which line; or
    - capture address as structured components without validation.

22. We should work out with the PO a guideline to instruct the information provider what to write on each line.

23. With regard to the exchange of free-text address, as there are many forms of free text format, we should build tools to facilitate the conversion from one dimension to another.

24. In a situation where the sender stores address in 3 lines, say 70+70+100, and the receiver stores address using a maximum of 4 lines with each line having a maximum of 35 characters (i.e. 4x35), when the sender sends out an address in XML, he sends out something like this :

```
<ClientAddress>
    <AddressLine>Flat B, Floor 15, Block 3, Cherry Mansions</AddressLine>
    <AddressLine>Whampao Garden</AddressLine>
    <AddressLine>Hung Hom, Kowloon</AddressLine>
</ClientAddress>
```

Upon receiving the address in XML, the receiver uses the conversion tool to

---

the address validation function by checking the captured address against an address database. The accuracy of such checking depends on the quality of the data in the address database.

In the context of this paper, validating an address does not refer to the checking of whether a person genuinely resides at the reported location, or whether an organization genuinely exists at the reported location, or whether the reported address location is actually the location the information provider wants to refer to.

[3] Implicit meaning can be associated by providing instructions on what to write on each line; e.g. write flat and floor on the 1st line.

convert the address to fit into the 4x35 dimension used in his database. The address may be converted to something like the following:

```
<ClientAddress>
    <AddressLine>Flat B, Floor 15, Block 3, </AddressLine>
    <AddressLine>Cherry Mansions, Whampao Garden</AddressLine>
    <AddressLine>Hung Hom, Kowloon</AddressLine>
</ClientAddress>
```

"Cherry Mansions" is moved from the 1st line to the 2nd line because the first line of the incoming address exceeds 35 characters.

25. In other words, the exchange format is simply one or more lines of text, and each line has one or more characters. It should be noted that in some cases, the address content may contain special characters like <carriage-return> and <line-feed>. The conversion tool should take note of this.

26. During the conversion from one dimension to another, when the incoming address is too long to fit into the destination dimension, the conversion tool will try to abbreviate the address based on abbreviations suggested by the PO (e.g. abbreviate "Building" as "Bldg") and do some specific truncation (e.g. drop "New Territories" from the string "Yuen Long, New Territories"). If the converted address still cannot fit into the destination dimension, the conversion tool will return an error and the receiver's application should handle such error.

27. The conversion tool should be intelligent enough to avoid breaking address components (e.g. a building name) between lines. However, there may be odd cases where the conversion tool makes a wrong guess and distorts the presentation of the address. To minimize such distortions, **NEW** applications are recommended to store un-validated address in a structured form, or to use a common format to store free text address. We suggest the common free text address format to be 5x40 base on the analysis presented in **Appendix B**.

28. Please note that we are NOT suggesting all existing applications to use the 5x40 format to exchange free text address information. If the sender keeps address as 70+70+100 and the receiver keeps address as 4x50, there is no point for them to exchange as 5x40. We are simply suggesting new applications that do not validate address to either use a structured address format or to adopt a 5x40 format with a view to progressive convergence of

the free text address format. We may never be able to achieve 100% convergence, this is a fact that we are prepared to accept. It is desirable but not critical for free text address formats to converge. After all, free text addresses serve their purpose as long as a human can understand the address.

Storing and exchanging an address for computer-aided analysis

29. In the context of this paper, computer-aided analysis refers to simple data manipulation like searching and sorting with the aim to support manual analysis. To facilitate analysis, applications usually break down an address into structured components like flat, floor, block, building number, street, etc.

30. In the next section, we will analyse the case where an application can afford to validate the information captured in each component.

31. In this section, we will cover the case where an application does not validate the captured information although it captures the information as structured components.

32. For new applications that capture an address as structured components but do not validate the captured information, we propose the following structure for storing and exchanging address (this structure is similar to that described in the address survey conducted in December 2003):
    - Local address / foreign address / post box address indicator
    - Local address (optional)
        - Room / Flat (optional, max 20 characters, e.g. Suite 1301-03)
        - Floor (optional, max 20 characters, e.g. Level 13)
        - Block (optional, max 20 characters, e.g. Tower 1)
        - Building Name / Phase (optional, max 30 characters, e.g. ABC Centre)
        - Estate / Village Name (optional, max 30 characters)
        - Building number (optional, max 20 characters)
        - Street name (optional, max 30 characters)
        - Lot number (optional, max 40 characters)
        - District (optional, max 40 characters)
        - Area (optional, max 20 characters)
        - Supplementary information (optional, max 5 lines of max 40 characters each)
    - Foreign address (optional)
        - Address details (max 5 lines of max 40 characters each)

    - Country (optional, base on the country alpha 3 code in ISO 3166 if exist)
- Post box address (optional)
    - Post office (max 20 characters)
    - Post box number (max 6 characters)

33. When information is captured without validation, we cannot ensure that the captured information is meaningful and we cannot ensure that information is captured in a consistent manner. For example, “Whampoa Garden” may be captured as “Wong Po Garden” or “Whampoa Gdn”. B/Ds that use such data for computer-aided analysis must take note of the poor quality of such data.

Storing and exchanging an address for more automated analysis

34. If we store an address for automated analysis, we must capture, validate and store the address as meaningful structured components. In most cases, the validation is done by checking the captured address component (e.g. a building name) against an address database; in such cases, a building name is considered meaningful if it is defined in the address database . After validation, the application may store the validated address component as text or as codified information. Since different applications use different conventions, information meaningful to one application may not be meaningful to another. For example, “Whampoa Garden” may be represented as estate identifier “1234” in one application which may be incomprehensible to other applications.

35. If we not only want to analyse the information ourselves, but also want to share the information with other applications, then we need to use a common reference that is meaningful to all. In other words, we need to find a way to represent a location and all applications that need accurate addresses should use the same approach to refer to that location.

36. Before we suggest what to use as the common reference, let us first examine the characteristics of address.

The characteristics of address

37. The address of a location can be written in many ways. For example, if a building is located at the junction of Nathan Road and Austin Road, we can either address it as number X, Nathan Road or number Y, Austin Road. In addition, very often, we can either use a building name or a street name plus a building number, or both, to address a building. It is difficult to restrict how one writes his address (unless addresses are **always** captured from the information provider electronically in a controlled manner, which is not always the case at present).

38. Because different B/Ds have different need on the analysis of address, they may use different address structures. For example, "Block 3, Cherry Mansions, Whampoa Garden" may be recorded in the following 2 ways :

Data element	Content
Block description	Block
Block value	3
Phase name	Cherry Mansions
Estate name	Whampoa Garden

Data element	Content
Block description	Block
Block value	3
Development name, phase name / number	Whampoa Garden, Cherry Mansions

39. There is no limit to the division of a property unit. For example, a property owner can subdivide a property unit for sale or lease. Since property units can split and merge, it is hard to determine whether a referenced property unit exist without on-site survey.

40. Different B/Ds have different definition for property units. For example, if a property owner divides a commercial property into several units and lets the units to different tenants, then the PO will see additional property units but RVD will see no change if the property owner remains the rates payer.

41. In the past, different B/Ds have different definition for building. For example, some treat the Landmark in the Central as one building; some treat it as 3 buildings comprising the Gloucester Tower, the Edinburgh Tower and the Podium. The Data Alignment Measure (DAM) project has successfully

aligned a definition and an identification mechanism for building across 13 B/Ds. Through the DAM building identifier (known as Building Common Spatial Unit[4] Identifier[5] in the DAM project), B/Ds can associate attributes like geo-representation, population, floor plan, height, address, etc. to a building.

42. An address refers to a spatial location. In most cases, the address of a person, an organization or a property refers to a property unit inside a building (or the entire building). Theoretically, these addresses can be represented using the DAM building identifier, plus floor and flat description. In some other cases, the address may refer to an open space (e.g. an open space car park, a construction site, etc.) or an underground property structure (e.g. a shop in a Mass Transit Railway Station), etc. For these cases, we either look for another identification mechanism, or we can treat them as exception and record details as free text.

43. For addresses that refer to a property unit inside a building, in most cases, the property unit can be differentiated by flat and floor information. However, some buildings do not use the flat attribute, they use the building number to differentiate the units in the building (or more precisely the units that share the same staircase). One example is the building located at 83-85 Sing Woo Road. For these cases, we cannot rely solely on the DAM building identifier + flat and floor information to identify a property unit. Additional attributes like building number and street name would be required in such cases.

---

[4] The scope of the DAM Building Common Spatial Unit covers :
- towers and podiums of legal private buildings, and Housing Authority (HA) / Housing Society buildings under jurisdiction of the Building Ordinance;
- New Territories small houses
- HA Buildings (including towers and podiums) – public housing and HA's Home Ownership Scheme estates;
- Towers and podiums of other government buildings such as government offices, public schools, hospitals, etc.
- Miscellaneous structures including temporary and open structures that are surveyed as building polygons.

[5] The DAM Building Common Spatial Unit Identifier is comprised of the geo-reference number of the building, concatenated with the record creation date and the building polygon type (the building polygon type can either be T for tower or P for podium).

Exchanging an accurate address in a way that is meaningful to all

44. Since 13 B/Ds have agreed on the definition of a building and have agreed on a building identifier through the DAM project, we can consider exchanging accurate address using the DAM building identifier, plus flat, floor, building number and street information.

45. The DAM building identifier is the most prominent one among the various building identifiers in use or to be adopted by B/Ds to identify a building because it has been defined through thoughtful alignment across 13 B/Ds. Like all data alignment processes, trade-offs were considered and the 13 B/Ds have made their own sacrifices before they finally reach consensus on the DAM building identifier. Data alignment is rarely successful unless all participants share a common vision and are willing to compromise. So, other B/Ds not previously involved in DAM should adopt a similar attitude. Having said that, we cannot rule out the possibility of ultimately using a different building identifier, but a separate alignment process will have to be conducted across all B/Ds.

46. A pre-requisite of using the DAM building identifier to refer to a building is that the data owners must be able to map their buildings to DAM's buildings.

47. For addresses that refer to other spatial objects (e.g. open space or underground property structure), we need to decide whether these fall under our scope of address, then we need to align across all B/Ds a common identification mechanism for the relevant spatial objects. One approach is to let the data owners map their buildings against DAM's first, and see what are the spatial objects that cannot be mapped, and the Address CSTF should work out an identification mechanism for the relevant spatial objects so that the data owners can do the other mappings.

48. We also need to align the structure for flat and floor information if we want to exchange such information accurately.

49. A structure for exchanging validated address is illustrated as follows :
    - building or other spatial object indicator (e.g. 'B' for building, 'U' for underground property structure)
    - identifier for building (follow DAM's)
    - identifier for other spatial objects (to be defined)

- building number
- street name / code
- flat / room / shop free format description (e.g. Shop B1-20)
- floor / level free format description (e.g. Basement Level 1)
- a list of structured description for flat and floor information
  - flat descriptor (e.g. Shop)
  - flat value (e.g. B1-20)
  - flat supplementary description (e.g. Rear)
  - floor descriptor (e.g. Basement Level 1)
  - floor value

50. When the receiver receives such an address, it should use the building identifier (or other spatial object's identifier) to locate the corresponding textual address in its address database.

51. For the purpose of identifying a building in a data exchange, we only need a common reference for buildings (i.e. the building identifier mentioned earlier), we do not need a commonly aligned textual address for that building. However, for address validation, we do need to validate the captured textual address against a set of "meaningful addresses". At present, some B/Ds are maintaining their own set of "meaningful addresses" (however, an address meaningful to one party may not be meaningful to other parties). Do we need a common reference for textual address ? We will analyse this in the next section.

Capturing an address and turning it into something that is meaningful to all

52. As discussed in paragraph 38, different B/Ds may have adopted different structure for the textual address in their address database, and the application logic in their systems may be coupled with the address structure. Therefore, it will be costly for them to modify their prevailing address structure. Provided that an application can map its buildings and other relevant spatial objects against a common reference, the application can continue to validate address in its own way, using its own address database.

53. Having said that, it should be noted that maintaining an up-to-date address database is also a costly exercise because B/Ds have to do on-site surveys regularly or rely on field officers to report changes promptly in order to take

into account new buildings, demolished buildings, and the change of a building's name. In order to lower the cost of address database maintenance, the address database owners should consider putting in place a simple mechanism (such as mailing list and notification templates) whereby they can share knowledge about a change to the address of some relevant spatial objects.

54. For new applications that need to capture address in a meaningful way, we recommend them not to maintain their own address database but to use a common reference for textual address. If there is a common reference for textual address, we can develop common modules for address data entry and validation, etc.

55. Whether existing address database owners should continue to maintain their own address database or switch to use the common textual address database (after modifying their application logic) is a project decision.

56. The coverage of the common textual address database (as well as the coverage of the common set of identifiers for buildings and other relevant spatial objects) should be in line with the scope of the addresses we want to cover, e.g. the scope described in paragraph 3. The Address CSTF should define the scope of addresses they intend to cover.

57. With regard to the development of the common textual address database, it can be built by extracting and converting accurate addresses from prevailing address databases, such as those maintained by the PO, the Lands Department, RVD and C&SD.

58. The PO has the most comprehensive address database in relation to person and organization addresses. It even contains unofficial aliases that represent how people usually write an address, and it covers addresses written in English and Chinese. Furthermore, it offers flat and floor information. However, the PO has not mapped its buildings to DAM's yet.

59. The Lands Department is the data agent for the building common spatial unit in DAM. As the data agent, the Lands Department will disseminate to the DAM participating departments those building attributes contributed by various B/Ds. The Lands Department itself will maintain the surveyed address (including alias) of the DAM building common spatial units. At

present, the Lands Department maintains the surveyed address of all buildings shown on its B1000 maps. The address associated with the DAM building common spatial units would be an ideal source of accurate building addresses because it provides direct mapping to the DAM building identifier. However, as described earlier, some spatial objects that might be relevant to addresses (e.g. underground property structures, open space, etc.) are currently not defined in DAM, we will have to deal with these separately.

60. The RVD maintains the address of all rateable properties up to property unit level in both English and Chinese, and RVD is going to provide to DAM the buildings' recorded addresses, including government properties and buildings in public housing estates.

61. The C&SD maintains a Register of Quarters System (RQS) which contains the address of all the quarters in buildings and structures where people may possibly reside in and hence does not cover all the quarters in Hong Kong. In addition, while quarters for residential purpose are individually listed in the RQS, non-residential units may only be recorded in RQS at compound, building or floor level (e.g. if a school comprises 3 buildings and some priests reside in the school, the school will be recorded at a compound level, i.e. as one single school). Moreover, the list of quarters in the RQS serves as a sampling frame for supporting the fieldwork operation of the population censuses/by-censuses and other household surveys. Hence, in the course of maintenance of the addresses, the address information is updated based on the information "observed" on the spot during field verifications and the address information as maintained in the RQS does not carry any official meaning.

62. The address structure and format to be used in this common textual address database will be designed later. This format is likely to affect how NEW applications capture, validate, and store accurate address.

Conclusion and Recommended Way Forward

63. We have introduced three types of addresses in this paper and they serve different roles in e-government applications. At present, Members of the Address CSTF generally believe that free text address and validated structured address will continue to be used in e-government applications but

not all Members agree that we should spend effort in defining a common schema for un-validated structured address. The Address CSTF has to discuss further before they can reach consensus on this point. Nevertheless, this paper forms a good basis for further discussion.

64. If the outcome is to recommend all 3 types of addresses, then the guideline to B/Ds on the choice of address model could be something like the following :

- B/Ds are highly recommended to use the "structured and validated" approach to capture and store address information. This approach allows accurate exchange and analysis of address information;

- B/Ds are recommended to use the "structured but not validated" approach to store address information if they cannot afford to validate addresses. This approach can facilitate computer aided analysis;

- In case B/Ds need to use free text address, they are recommended to provide guidelines to the information provider on what to write on which address line, in order to facilitate postal delivery. And new applications that need to use free text address should adopt the 5x40 format.

65. It is straightforward to convert structured address to free text address; the IFCG Standing Office can build tools for such conversion once the structures are defined. However, converting free text address captured from the public to structured ones **without error** is difficult, if not impossible. Similarly, converting structured but not validated address to structured and validated ones is equally difficult.

66. This paragraph summarizes the other recommendations drawn from the analysis :

Free text address

- The IFCG Standing Office should make available software to facilitate the conversion of free text address from one dimension to another and such software should try to make the converted address comprehensible (e.g. avoid breaking address components such as building names between lines);

Structured but not validated address

- If the Address CSTF decides that structured but not validated address should be standardized, they should work out the structure and format for recording and exchanging this type of address. A draft is proposed in paragraph 32 for consideration;

- The IFCG Standing Office should work with the PO to derive guidelines on how to generate well-formed free text address from structured but not validated address components, and should develop software for such conversion;

- When an application that stores address as structured components without validation wants to print an address for postal delivery, it should print the address in a well-formed manner to facilitate postal delivery;

Validated address

- The Address CSTF should define the scope of the addresses to be covered (e.g. the physical address of persons and organizations, and the address of property units where person and organization may occupy);

- The Address CSTF should identify those spatial objects that may be referenced in a person / organization / property address but are not spatially defined as a building in DAM, and should devise an identification mechanism for the spatial objects that fall within the scope of the intended address;

- Address database owners should map their buildings and other relevant spatial objects against the set of common identifiers;

- Validated addresses should be exchanged as a combination of an identifier for building or other relevant spatial objects, plus building number, street, flat and floor information. A draft structure is proposed in paragraph 49;

- Prevailing address database owners may continue to validate addresses using their prevailing address databases;

- Prevailing address database owners are recommended to put in place a

simple mechanism (e.g. mailing lists and notification templates) to facilitate the sharing of knowledge about changes to the address of buildings and other relevant spatial objects. The Address CSTF should work out the mechanism;

- A common reference for textual address (a common textual address database) should be maintained and new applications that need to validate addresses should validate address against this common textual address database.

Information Technology Services Department
June 2004

Appendix A

Examples of free text address formats in use by various B/Ds

B/D	Free text address format in use
Census & Statistics Department	Central Register of Establishments uses 4x30
Companies Registry	3 lines 60+60+90 bytes
Customs & Excise Department	ACCS uses 5x35 FRT uses 4x35 EDI CO/PN uses 4x35 RTEL uses 5x35 TDEC uses 4x35 TTRS uses 4x35 EMAN uses 140 EDI-DCP uses 5x50
Environmental Transport and Works Bureau	CCMIS uses 240
Fire Services Department	260 characters
Housing Department	CPMS uses 4x60
Intellectual Property Department	3 lines 100+60+60 characters
Judiciary	3 lines each having 30 to 100 characters (English) 3 lines each having 20 to 50 characters (Chinese)
Land Registry	IRIS uses 1x320
Lands Department	LACO CPMS uses 1x250 and 1x200
Official Receiver's Office	4x40
Rating &Valuation Department	Receive from ESD as 5x35 (English) and 5x28 (Chinese) May receive from IRD as 150 characters (English) & 60 characters (Chinese) in exceptional cases
Trade and Industry Department	BLICS uses 3x40 CPNS uses 4x50 EDI-CO uses 4x35 EDI-TTRS uses 5x45 SMECS uses 4x50 Web Portal uses 4x60 RICE uses 4x35 Rough Diamond uses 5x50
Treasury	Most systems use 3x40 (or 3x20 for Chinese)

Appendix B

Analysis of whether B/Ds should be encouraged to converge the free-format addresses to a common format

Option	Rationale	Pros	Cons
Do not need to converge to a common format, let it be "genuine free format" forever	Relying on a conversion tool to convert free-format address from 1 dimension to another may be good enough. Existing systems and their extensions to be developed in the future may prefer to stick to their prevailing format and hence the adoption rate of the proposed convergence format may be low anyway	Nil	Will have to rely on the address dimension conversion tool forever. However, this tool could make a "wrong guess" and make an address difficult to read
Converge to use the 5x40 format for address in English in the future	The 5x40 format provides for 200 characters which approximates to the mean size of the various free-format address windows currently in use. Based on the information in Appendix A, the mean size of the address window is 192.67 characters. Using the mean size of the address window **(not the mean length of the actual address content)** is reasonable because the actual content is usually smaller /shorter than the window. Looking at the address structures in use today, the	If more and more e-government applications use a common format for free-format address in the future (e.g. say 10 years later, 80% of applications that use free-format address converge to a common format), there will be less conversion from one dimension to another, thus better preserving the look of an address as provided by the information provider.	Application owners currently using another dimension are worried that as more and more applications converge to a different format, they will have a higher chance of hitting "conversion problem". For example, if the convergence format is longer than the format they use, they will have a higher chance of receiving long addresses that cannot fit into their existing address

	C&SD's Central Register of Establishments' address window is 120 (4x30) characters, the Companies Registry's address window is 210 (60+60+90) characters and the RVD's rates payer correspondence address window is 175 (5x35) characters. These records cover all establishments (organizations), registered companies and all rates payers. **Base on this comprehensive sample, 200 Characters should be adequate to cover the address of persons and organizations in the HKSAR.** In addition, a 40-character line, with abbreviation as necessary, should be adequate to cover most address lines if people follow the Post Office's guideline when they write an address. There are exceptions, but the percentage is tolerable. Please refer to **Note 1** for more information on the exceptions. However, please note that while 5x40 is acceptable to cover residential address or an organization's address, it may not be big enough to cover special "addresses" used in a land or property sale transaction if such transaction includes special property objects. A made-up example of such		format. [In reality, this may not be a real problem. The basic principle underlying the Post Office's guideline is to **use the minimum composition of address data fields necessary to identify a unique address for mail delivery**. Furthermore, the length of the address content would not change significantly simply because one is given more space to write his address]

	"special property objects" is illustrated below : "BIC (Brilliant International Corporation) Commercial Building, 123 – 129 Sha Tin Rural Committee Road, Town Centre, Sha Tin, and underground car park on land lot IL 1234 Section 5 Subsection 6 Section 7, and standalone wall between BIC Commercial Building and ABC Commercial (Sha Tin) Building, 131-135 Sha Tin Rural Committee Road, Town Centre, Sha Tin, and standalone wall between BIC Commercial Building and DEF Industrial Centre, 119-121 Sha Tin Rural Committee Road, Town Centre, Sha Tin" This type of "special property objects" (the BIC building + underground car park + 2 walls in the above fictitious example) would need special treatment.		
Converge to use the 5x60 format for address in English in the future	Most applications store address as AxB, with 5 being the maximum for A and 60 the maximum for B. If these applications ultimately migrate to adopt the 5x60 format, then they do not need to rearrange their address data (e.g. 5x35 can easily fit into 5x60). The rationale is simply to let B/Ds preserve	If more and more e-government applications use a common format for free-format address in the future (e.g. say 10 years later, 80% of applications that use free-format address converge to a common	Application owners currently using another dimension are worried that as more and more applications converge to a different format, they will have a higher chance of hitting "conversion problem".

	the look of addresses currently in their database as far as possible when they migrate their address format. As in the case of 5x40, even a 5x60 format cannot cover special “addresses” used in a land or property sale transaction if such transaction includes “special property objects”.	format), there will be less conversion from one dimension to another, thus better preserving the look of an address as provided by the information provider.	For example, if the convergence format is longer than the format they use, they will have a higher chance of receiving long addresses that cannot fit into their existing address format. [In reality, this may not be a real problem. The basic principle underlying the Post Office’s guideline is to **use the minimum composition of address data fields necessary to identify a unique address for mail delivery**. Furthermore, the length of the address content would not change significantly simply because one is given more space to write his address]

Note 1 :

At present, if people write their address in accordance with the Post Office’s guidelines, there are about 57,000 addresses with a line over 40 characters, this constitute 2.1% of the total. These long address lines, **when forced into a 5x40 format**, will affect the Post Office’s letter sorting process.

If we tell people to write these long address lines on a 40 character line, a normal person will split the line at a convenient point and put the overflown characters on the following line. E.g.
"PACKAGE WASTE RECEPTION BUILDING CHEMICAL WASTE TREATMENT FACILITIES"
will be written as
"PACKAGE WASTE RECEPTION BUILDING"
"CHEMICAL WASTE TREATMENT FACILITIES"

In addition, a normal person may use abbreviations, e.g. abbreviate
"HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY"
as
"HKUST"

If people make these splits and abbreviations consistently, then the Post Office can perhaps tune their address database (behind their letter sorting system) to cater for these exceptions. Nevertheless, these long lines will **increase the probability of inconsistencies** (e.g. abbreviations made in different ways), thus affecting the efficiency of the letter sorting process.

Worse still, people may split a line within an address component, this not only affects the letter sorting system, but will cause inconveniences to the postman also. e.g.
"PACKAGE WASTE RECEPTION BUILDING CHEMICAL WASTE TREATMENT FACILITIES"
may be written as
"PACKAGE WASTE RECEPTION BLDG CHEMICAL"
"WASTE TREATMENT FACILITIES"

Nevertheless, after discussion with the Post Office, we believe the percentage of 2.1% is still tolerable. As such, we believe 5x40 is acceptable as the proposed convergence format.