## Measures to Facilitate Address Data Capture and Validation

*(This paper supersedes Address CSTF Paper 2004/03)*

**Introduction**

1.  Address information can be a useful attribute for correlating information[1] in electronic records. For example, by correlating dangerous goods licensing information through the goods' storage address, we can more easily analyse the fire risk of a building, or a region. Such correlation relies on good quality address information. One form of such information correlation is digital spatial analysis.

2.  At present, not many application systems capture address information in a consistent and structured way. This leads to inconsistent address information recorded in information systems, making it difficult to correlate information based on the address attribute.

3.  If an application system requires good quality address information, it usually validates a captured address against an address database before recording the address. We did a survey[2] in July 2004 to check government bureaux and departments (B/Ds)' demand for a common service for capturing and validating address information.

4.  Among the 64 B/Ds who responded to the survey, 7 B/Ds[3] reported that their

---

[1] *The correlation of information must be within legal bounds and must not infringe personal data privacy.*

[2] *Please refer to the ITG Infostation on the government Intranet ( http://itginfo.ccgo.hksarg/content/if/cstf/docs/consolidated_addr_db_survey.pdf) for the returns collected from B/Ds.*

[3] *The 7 B/Ds whose computer systems validate an address against an address database before recording the address are : Buildings Department, Electrical and Mechanical Services Department, Fire Services Department (Third Generation Mobilizing System), Housing Department, Land Registry, Rating and Valuation Department, Registration and Electoral Office. Apart from these 7 B/Ds, the Census and Statistics Department, Hong Kong Post, and the Lands Department also have a comprehensive address database which they use for purposes other than address validation.*

computer systems are already validating address information against an address database before the addresses are recorded.   Among those B/Ds that are not validating address information at present, 9 B/Ds considered it critical or highly desirable to do address validation in the future, and 9 B/Ds considered address validation a desirable (nice to have) service.   In other words, a total of 18 B/Ds indicated that they may implement address validation in the future.

5.   In view of the need for address validation, we are exploring the feasibility of putting in place some measures to facilitate address data capture and validation.   This paper[4] proposes some measures for consideration by the Members of the Address Common Schema Task Force (CSTF).   We start with an analysis of the criteria affecting the effectiveness of address validation, then we propose a standpoint for each criteria, and conclude with recommendations on the way forward.

6.   In the context of this paper, validating a captured address means checking the captured address against a comprehensive set of local addresses (i.e. a local address database) to better assure that the captured address refers to a location that physically exist.   The accuracy of such checking depends on the quality of the data in the address database.   (Another interpretation of "validating an address" is checking whether a person genuinely resides at the reported location, or whether an organization genuinely exists at the reported location. This interpretation is not applicable in the context of this paper, and the measures recommended in this paper cannot be used for such checking.)

7.   Some basic concepts regarding addresses are described in Address CSTF Paper 2004/01[5].   Readers are encouraged to read Paper 2004/01 in conjunction with this paper.


**Usage of Validated Addresses**

8.   Most of the B/Ds that may implement address validation in the future have suggested 2 or more areas where validated addresses may be used.   To

---

[4]  *http://www.xml.gov.hk/download/address_cstf_200404.pdf*

[5]  *http://www.xml.gov.hk/download/address_cstf_200401.pdf*

summarize : :
- 6 B/Ds are planning to use the address to enable digital spatial analysis;
- 9 B/Ds are planning to use it for other forms of information correlation; and
- 7 B/Ds want more accurate addresses to reduce mail delivery problem.
   Other usages mentioned include :
- To facilitate the conversion of addresses from English to Chinese and vice versa;
- To facilitate information retrieval via the address attribute;
- To better assure data integrity and accuracy in official records; and
- To ensure data consistency among information systems.

9. Please refer to Appendix A for an elaboration of some usages of validated address information.


**Criteria Affecting the Effectiveness of Address Validation**

10. The effectiveness of address validation is affected by 3 criteria : the coverage of addresses that can be validated, the level of details we validate, and the timeliness of the data in the address database.    These criteria directly affect the cost of maintaining the address database.   Therefore, we must strike the right balance between need and cost.


*Coverage of the Address Database*

11. There are 2 issues we need to consider here.    First, the address of **"WHAT" objects** are we talking about ?   Second, having agreed on the "WHAT", should we aim to enable the validation of the address of **each and every** object ?

12. The "WHAT" can cover persons, organizations, telecommunication facilities installed in a building, advertisement boards located at the top or external wall of buildings, etc.   Among the 18 B/Ds that may implement address validation in the future :
- 9 indicated that they need to validate the address of persons and/or organizations only;

- 6 mentioned other objects[6] in addition to persons and/or organizations;
- 1 B/D indicated that it just want to check the address of planning zones[7];
- 1 B/D indicated that it just want to check the address of complaint locations; and
- 1 replied 'to be defined later'.

13. Say for example, we decide to cover the address of persons and organizations only.   While persons and organizations are usually located in buildings, a person can reside on a fishing boat which parks in a boat parking lot, and an organization can be a cake shop in an MTR station (which is an underground property structure and not regarded as a building from the Data Alignment Measure (DAM)[8] project's point of view).

14. Ideally, our address database should cover all possible locations where a person or an organization may occupy, including buildings, boat parking lots, underground property structures, etc.   While "building" as a spatial object is clearly defined in the DAM project, at present, there may not be authoritative definition for other spatial objects (locations) like underground shopping center, open-space car park, etc.

15. Among the 18 B/Ds that may implement address validation in the future, all

---

[6] *These objects include properties (2 B/Ds mentioned this), construction sites, monuments and historical buildings, telecommunications equipment, water service points, etc.*

[7] *Planning zone refers to development / application site in town planning terminology.*

[8] *The DAM project lead by the Housing, Planning and Lands Bureau has successfully aligned a definition and an identification mechanism for 5 spatial units across 13 B/Ds.   One of the 5 spatial units covered in DAM is building, referered to as the DAM Building Common Spatial Unit (CSU).   The DAM Building CSU serves as a medium for exchanging various attributes of a building. Through the DAM Building CSU Identifier, B/Ds can associate attributes like geo-representation, population, floor plan, height, address, etc. to a building.*

*The scope of the DAM Building CSU 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.; and*
- *Miscellaneous structures including temporary and open structures that are surveyed as building polygons.*

B/Ds indicated that they need the address of buildings. While 9 B/Ds indicated that they **only** need to validate the address of buildings, other B/Ds have mentioned locations like cargo yards, embarking / disembarking points for passengers and cargoes, oil depots, sports ground, parks, gardens, flea markets, farms, construction sites, village house, tertiary planning unit / street blocks, etc.

16. Before deciding whether to cover locations like underground shopping centers and open-space car park or not, we should assess how we can define such locations (e.g. do we need an accurate survey of the boundary, or do we simply need a loose definition base on ownership or occupancy) and whether we can find data sources to provide the address of those locations.

17. In view that it may not be feasible to cover some locations, project teams should be prepared to handle locations that cannot be validated and devise exception handling procedures to deal with them.

*Level of Detail in Address Validation*

18. The question here is whether we need to validate an address up to floor and unit level, or can we simply validate an address up to building level (i.e. just check the address of the concerned building which usually comprises the district, street name, building number and building name).

19. Although 3-dimensional[9] (3-D) spatial analysis was useful in the SARS investigation in Amoy Garden in 2003, the majority of spatial analyses in use today are at 2-dimensional (2-D) level. If you validate addresses solely for enabling digital 2-D spatial analysis, must floor and unit information be validated ?

20. We should also consider the feasibility of validating unit information. Can we build an address database with accurate unit information in all cases ? Although the property units in residential buildings stay pretty stable, the units in commercial buildings can change from time to time. For example a

---

[9] *In the context of this paper, a 3-dimensional address refers to an address with floor and unit information, and a 2-dimensional address refers to the same address without floor and unit information.*

company that used to occupy the whole floor moved away and the landlord splits the floor space and lends them to 3 different companies.   As these changes happen every day, would it be cost effective to have our address database keep track of unit level information accurately in a timely manner ?

21. Among the 18 B/Ds that may implement address validation in the future, 2 B/Ds considered it critical to have address validated at floor and unit level, 2 considered it highly desirable, and 11 considered it desirable.   While most B/Ds considered it acceptable to validate up to floor level only, one B/D considered unit level validation a must.


*Timeliness of the Address Database*

22. We validate an address by matching it against the addresses in an address database.   If the street and building name combination of a captured address matches an entry in the address database, then the captured building address is considered valid.   However, there may be cases where a valid address cannot pass this test: this is often because the address database is out-dated.   For example, newly erected buildings and renamed buildings are not included yet.

23. Currently B/Ds rely on field officers and members of the public to report changes to an address database.   For example, when a postman notices that a building has been renamed, he reports to the Post Office.   It will take time for these changes to be discovered and recorded in the address database. Therefore, applications must be prepared to handle valid but fail-to-be-validated address.   For example, the application can record the address with an indicator to reflect that it is invalid, and trigger exception handling procedure such as report the case to the address database administrator.

24. Normally, an address database designed to support address validation only needs to cover the active addresses.   However, in some cases, it may need to cover the buildings that are still under construction.   For example, the applications used by estate agents may need to handle the address of buildings that are still under construction. And in the case of the conversion of historic records from a paper system to an electronic system, we may need to validate addresses that refer to demolished buildings or the old name of a building. Therefore, the addresses in the address database should have some means of

indicating whether it is an "active" address or an "under-construction" address, or the address of a demolished or renamed building.   Applications can then choose to accept all or only the active addresses.


**Proposed Measures to Facilitate Address Data Capture and Validation**

*What the Measures Could be ?*

25.  A local address database will be the core of the measures to be put in place. On top of this address database, we can build tools (such as open source program code) to facilitate application systems to make use of the address database to perform address data capture and validation.   Since these measures are regarded as a common service for use by B/Ds, we will call the address database a "**common address database**" in the rest of this paper.

26.  At present, the 7 B/Ds that are performing address validation are maintaining their own address databases, based on self-collected addresses, as well as addresses from other B/Ds such as the Lands Department (LandsD), the Rating and Valuation Department (RVD), and Hong Kong Post (HKPost). Since the addresses from LandsD, RVD and HKPost have a good coverage, we can request them to contribute address data on an on-going basis.

27.  In addition, there will be a database of all post boxes offered by HKPost. The address validation program will check against this when it encounters a post box address.


*Prime Usage of the Common Address Database*

28.  As the majority of B/Ds need to validate the address of persons and organizations, the prime usage of the common address database should be for validating the postal and physical address of persons and organizations, as well as the address of properties where persons or organizations may occupy. The validated addresses can serve as postal addresses or as an address where a person or an organization may be contacted physically.   These address may or may not correspond to those addresses that appear on the deed of the concerned properties.

29. The most effective way to do address validation is to incorporate the validation logic in e-government applications that capture address data directly from the data subject. We can also incorporate the validation logic in back-office data entry applications that process data collected via paper forms, but this is less effective as additional effort would be required to follow-up with the information provider if the address on the paper form is invalid.

30. In order to speed up the implementation of the common address database to serve e-government applications, other usages of the common address database should be considered at a later stage after the common address database is stable and in operation.

*Locations to be Covered by the Common Address Database*

31. To reasonably cater for the address of persons and organizations, the common address database should cover all buildings as a minimum. We propose to cover underground shopping centers and MTR stations also, because a fair number of shops (organizations) may be found there.

32. A working group comprising the major address contributors (LandsD, RVD, HKPost) and the major potential users of the address database should be formed to decide what locations (e.g. underground shopping centers) to be covered in the address database. The working group can work out a list of locations other than buildings, then set priority for assessing these locations one-by-one. A systematic way to work out the list of locations is to go through RVD's (or HKPost's) address database to identify those locations that are not covered by the DAM Building Common Spatial Units (CSUs).

33. During this exercise, the working group should lay down principles on what locations to cover, and set rules for defining those locations. These principles and rules should be followed by the address database administrator.

34. The working group should also design the mechanism and workflow for the creation and on-going maintenance of the addresses in those locations, in particular :
    - define the roles and responsibilities of the data contributor(s) and the address database administrator;

- ensure that appropriate supporting facilities are available to fulfill critical user requirements (e.g. ensure that there are matching digital spatial units for relevant 2-D addresses); and
- advise on resource implications on the data contributor(s).

35. Since many departments (under different bureaux) might be involved in the demand and supply chain, strong leadership with appropriate authority will be required to lead the working group. The governance framework adopted by the DAM project may serve as a reference in designing the composition of the working group.

36. Since this exercise is similar to the DAM exercise, we should learn from DAM's experience. Data alignment is rarely successful unless all participants share a common vision and are willing to compromise. Therefore, members of this working group must be open in considering trade-offs and be prepared to make their own sacrifices in order to reach consensus on what locations to cover and how to define those locations.

*Linkage Between Addresses and Spatial Objects on Digital Maps*

37. Since digital spatial analysis will be a major usage of the validated addresses, the building addresses in the common address database should be linked to the buildings on LandsD's digital maps (which will be aligned with the DAM Building CSUs after LandsD's implementation of relevant DAM recommendations). In other words, there should be a direct mapping between building addresses and DAM Building CSUs.

38. As for those locations other than buildings, in order to enable more thorough spatial analysis, we can request LandsD to create digital spatial units to cover those locations, and to link relevant 2-D addresses to those locations.

*Handling Floor and Unit Information*

39. Since the majority of B/Ds find it desirable to check floor and unit information (e.g. to reduce mail delivery problem), we should explore how their demands can be met in the long run.

40. As floor descriptions[10] are pretty stable, we should eventually cover these in the common address database. RVD (and HKPost[11]) can provide trustworthy floor descriptions for most buildings.

41. With regard to unit information, we can make use of the unit descriptions[12] from RVD (and HKPost) to provide a "confidence level" for reference by application systems. For example, we can incorporate the unit descriptions from RVD (and HKPost) into the common address database and let an application choose whether it wants to check against HKPost's unit description, or RVD's or both.

42. However, it will take time to generate floor and unit descriptions against each DAM Building CSU. Although both RVD and HKPost have floor and unit descriptions, those floor and unit descriptions cannot be used until we have done a mapping between RVD / HKPost's addresses against the DAM Building CSUs.

43. In order to take advantage of validated addresses sooner, B/Ds that plan to implement address validation should consider implementing 2-D address validation first, and enhance the validation to 3-D level when floor (and unit) information becomes available.

*Timeliness of the Address Database*

44. The common address database should be designed to cater for the fact that each address has an effective time period.

45. As not many applications need to use old addresses, we can populate the common address database with active addresses and "under-construction" addresses first.

---

[10] *Examples of floor descriptions are : Basement Level 2, Ground Floor, Podium, Floor 12A, Level 15, etc.*

[11] *Whether we can use HKPost's addresses or not will depend on whether there is a mapping between HKPost's building addresses and the DAM Building CSUs.*

[12] *Examples of unit description are : Room 2301, Flat C, Unit 151-153, Suite 307A, etc.*

## Address Databases on Hand

46. A number of B/Ds are currently maintaining address databases to support their operation.   Examples include LandsD, RVD and HKPost.   We should best utilize these valuable assets to build our common address database.

47. LandsD has the 2-D address of the buildings in Hong Kong, and each building address is cross-linked with a building spatial object on LandsD's maps. Over 40 government departments are now using LandsD's digital maps and associated building addresses.   A lot of organizations in the private sector are also using LandsD's digital maps and associated building addresses.   For example, the ESD[13]  Change of Address transaction makes use of LandsD's building addresses (which ESD acquires from an intermediary) to facilitate address data capture and validation at building level.   Because LandsD's building addresses will be linked with the DAM Building CSUs (subject to the completion of LandsD's implementation of this DAM recommendation around the 2nd half of 2005) and LandsD's addresses are compiled based on actual surveys, LandsD would be a good source for building addresses.

48. RVD has the address of all rateable properties up to property unit (i.e. 3-D) level, and as a data contributor for the DAM Building CSU, RVD shall conduct a data mapping[14] exercise to map its buildings against DAM's and will provide to DAM the buildings' recorded addresses.   Subject to the completion of this mapping around Q3 2006, RVD will also be a good source for building addresses, floor descriptions and unit descriptions.   At present, RVD keeps property unit records according to how rates are charged on those properties; in cases where the whole building is rated as one tenement (e.g. when the building is used to offer service apartments), then RVD does not keep floor and unit descriptions for that building, although RVD's address database design can record properties at unit level.   RVD would need additional resources if it has to record all properties at unit level (regardless of how rates are charged), because this is additional work on top of its current operation need.

---

[13]  *www.esd.gov.hk*

[14]  *Since RVD has adopted a slightly different definition on building (different from that used in DAM), a data mapping exercise is required before RVD can provide building addresses to the DAM Building CSU data agent.*

49. HKPost has a comprehensive address database comprising the 3-D address of all possible postal delivery locations. However, HKPost has no plan to map its building addresses against DAM's Building CSUs in the near future. If such mapping can be done, HKPost would be an ideal source for floor descriptions and unit descriptions.

**Business Model**

50. Considerable effort will be required to maintain and offer the common address database, in particular, the user support effort should not be under-estimated. With regard to the business model for maintaining and offering the common address database, there are at least 3 models :

    - government-operated : a government body acquires address data from the relevant B/Ds, reorganizes the data and offers it to B/Ds and other organizations;

    - public-private-partnership (PPP) model 1 : let the private sector run the business. A private company buys address data from LandsD, RVD, and HKPost, reorganizes the data and offers it to B/Ds and other organizations. In other words, this private company will act as the address database administrator, plus it will provide user support; and

    - PPP model 2 : a government body acquires address data from the relevant B/Ds, reorganizes the data and offers it to B/Ds. In other words, the government remains the address database administrator. When the address database is stable, the government can sell the address database to one or more wholesalers (intermediaries) who will resell the database to other organizations and provide user support for their customers.

51. Given the complexity of the address mapping involved, PPP model 2 is considered more feasible than PPP model 1.

52. We can consider riding on the next phase of the DAM project to attach floor and unit descriptions against each building, and to include additional spatial units like underground shopping centers and MTR stations, etc.

53. The major benefit of a PPP arrangement is to tap resources and creativity of the private sector, and to provide more business flexibility. The potential of the private market in providing value-added services could be explored, which would enhance the commercial viability and success of this initiative.

54. In designing a PPP arrangement, we must make it a win-win-win situation for the government, the intermediary, and the users of the address database. Since LandsD, RVD, and HKPost have to contribute considerable recurrent resources to maintain their address data, and perhaps even more resources would be required to handle queries from the intermediaries when the common address database is sold to the public, the PPP scheme should be designed to recover some of the recurrent expenditure that may be induced by these additional effort.

55. In supporting a PPP arrangement, we must also take into account the effort required to administer the PPP contract.

56. With regard to the program code for capturing and validating address, the IFCG Standing Office can make available sample program code for capturing and validating address on the EGIS-SPICA application hosting platform. These code can be made accessible to all B/Ds and the public.

**Justification for this Initiative**

57. Infrastructure projects like the common address database initiative are difficult to justify because the costs and the benefits from different parties do not match. While the costs are accrued to the data contributor departments, the benefits are mainly notional and external. It is difficult to justify such infrastructure project purely on a cost benefit analysis. It is a very difficult business case that merits top level commitment at the very first beginning and such commitment should be sustainable to the end.

**Conclusion and Recommended Way Forward**

58. Base on the prime objective of enabling applications to record address information in a consistent manner, so as to facilitate information correlation, digital spatial analysis and postal delivery, we recommend to put in place

some measures to facilitate address data capture and validation.

59. The core of these measures is a common address database which shall comprise the postal and physical address of persons and organizations, as well as the address of properties where person or organization may occupy.

60. The following steps are recommended to take this initiative forward :

- Form a working group to determine and define the locations to be covered by the address database, and to define the roles and responsibilities, mechanism and workflow for creating and maintaining the addresses in these locations;

- When LandsD finishes creating the DAM Building CSUs (with the CSU identifier, the geometry of the spatial unit, and the surveyed building address) around the 2nd half of 2005, promote DAM's building address database as the common address database (this will serve address validation at 2-D level);

- When RVD finishes mapping its buildings against DAM's around Q3 2006, review the success rate of the mapping. If the success rate is high (exact percentage to be determined by the working group), then adopt RVD's floor descriptions and unit descriptions. If the success rate is low or if the working group needs a more thorough coverage of floor and unit information, then seek management support and acquire funding to map HKPost's addresses onto DAM's Building CSUs;

- The Office of the Government Chief Information Officer (OGCIO) to initiate a pilot project to host the common address database on the EGIS-SPICA application hosting platform, and to build address data capture and validation tools for applications running on EGIS-SPICA, as well as for applications running on other application platforms, where feasible; and

- Explore the private sector's interest in acting as the address database administrator, in particular the private sector's interest in mapping HKPost's addresses against the DAM Building CSUs.

**Comment Sought**

61. Members are requested to comment on the recommendations proposed in this paper.

**Office of the Government Chief Information Officer**
**October 2004**

**Some Usages of Validated Address Information**

1.  Although currently not many applications validate address information before they record the information, the benefit of using validated addresses should not be overlooked, in particular in combination with Geographic Information Systems (GIS) for spatial analysis.   A suitably equipped GIS application will allow you to use address as a search key.   Therefore, if the address information in your electronic records has been validated against the address database used by the GIS application, then you can apply digital spatial analysis on your records.

2.  To illustrate this, let us imagine that we have a GIS application equipped with digital map to show all the buildings in Hong Kong, and this GIS application has the textual address of all buildings in Hong Kong.   If an address in your electronic record indicates the building "Cyberport 1", you can tell the GIS application to find and highlight this building on the digital map.   The GIS application can do this because it "recognizes" Cyberport 1 as a valid building (i.e. Cyberport 1 is defined in the GIS's building address database). However, if an address in your electronic record indicates the building "Cyberport 8", then the GIS application will return an error saying this building cannot be found.   Therefore, if you want to use the GIS application to highlight the buildings recorded in your electronic records, then you must validate the building address against the GIS's building address database[15].

3.  The Registration and Electoral Office (REO) is currently using this approach to facilitate the assignment of geographical constituencies and polling stations to the electors.   REO has acquired building related GIS data and building addresses from the Lands Department.   The REO validates and maps electors' address onto the defined buildings, and base on the distribution of electors, REO assigns geographical constituencies and polling stations to the electors.

---

[15]  *This is feasible only when a linkage has been created between the building objects on the digital map and the address database in the GIS application, and such linkage is being kept up-to-date continuously.*

4. Another usage scenario of validated addresses is when you want to record the same address in a consistent manner.   Say for example, you run a pizza delivery service and you accept orders through the Internet as well as over the phone.   If you capture and record the delivery address as free text without validation, you may record the same address as :
    - Floor 6, Guardian House, Oi Kwan Road, Happy Valley
    - 6/F, 32 Oi Kwan Rd, Wanchai

depending on how the customer reports his address.   Both forms of the address illustrated above are acceptable from the pizza delivery business' point of view.   However, it will be difficult if you want to co-relate the two orders with a view to analysing the order pattern of your customers.

5. Imagine that now you have a textual address database that contains all the valid building addresses in Hong Kong.   Such a database typically has :
    - a list of valid building names;
    - a list of valid street names; plus
    - a relationship between building name and street name + building number.

With this address database, you can design the user interface of your pizza order entry application to capture address information in a user-friendly manner.

6. For example, the application first prompts for the street name and building number.   If the customer enters "32 Oi Kwan Rd", then the application immediately knows that the customer is referring to "Guardian House" which is located at 32 Oi Kwan Road.   Then the application :
    - tells the customer to confirm the complete building address "Guardian House, 32 Oi Kwan Road, Wanchai"; and
    - captures the floor and unit information.

7. Say for example, the customer enters "Oi Kwan Rd" without a building number, then the order entry application can extract all building names associated with Oi Kwan Road and let the customer choose a building.   The customer chooses "Guardian House" and the application further prompts for floor and unit information.

8. With an address database and a well-designed address data entry user interface, applications can capture address in a user-friendly and controlled manner.   The building addresses captured under control can then be recorded either in a

consistent textual form or as a building code defined by the address database. With addresses recorded in a consistent manner, we can use address as an attribute to co-relate information such as the pizza order pattern.